UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

COMBINING CLASSIFICATION AND BAYESIAN METHODS TO BETTER MODEL DRUG ABUSE

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE
in Data Science and Analytics

By

MATTHEW J. BEATTIE

Norman, Oklahoma

2018

COMBINING CLASSIFICATION AND BAYESIAN METHODS TO BETTER MODEL DRUG ABUSE

A THESIS APPROVED FOR THE

GALLOGLY COLLEGE OF ENGINEERING

BY

Dr. Randa Shehab, Chair

Dr. Sridhar Radhakrishnan

Dr. Wayne Stewart

Table of Contents

Abstract

Illicit drug use in the United States has shown no signs of abating, and the morbidity from drug abuse has risen sharply over the last several years. This is primarily due to a rise in the abuse of opioids, including prescription opioids, heroin, and most recently, fentanyl. Finding potential predictors of heroin use could help to reduce fatalities from drug overdose.

There have been many studies to identify correlates of heroin use, and most follow the same methodological pattern. A literature search leads to a pre-selection of a set of predictors, which are then analyzed using traditional statistics – frequentist summaries and logistic regression. This approach limits the potential for finding unexpected combinations of predictors, predictors that correlate with small subject classes, and previously undiscovered predictors. The regression component of the approach is limited by the inability to accept the null hypothesis and it makes no use of information gathered during the literature search.

We propose an improvement to this approach. We believe that principal component analysis and regression tree classification provide methods to objectively identify potential heroin use correlates from the data itself. The information gained by these methods, combined with a review of existing research, allows us to create prior distributions for regressors. We can then use Bayesian Markov Chain-Monte Carlo (MCMC) methods to build predictive models that are more robust than those from traditional approaches.

We demonstrated this approach by modelling the probability of heroin use based upon self-reported factors from a multiyear national dataset. The ease of implementing classification and MCMC modelling allowed us to examine multiple years of data, which in turn enabled us to see how predictors of heroin use have evolved over time. We found that our methods reinforced some beliefs, such as that OxyContin is correlated to heroin use. We also found that they refuted other beliefs, such as that early age of use of drugs is strongly related to potential heroin use. We found that certain usage patterns, such as polyabuse, easy access to heroin, and low perceived risk of heroin use are correlated to heroin use.

Introduction

Evolution of the Opioid Epidemic

Illicit drug use has long been a problem in the United States. Heroin, which was initially created as a medicine in 1898, has a particularly long history of abuse. New York's Bellevue hospital had its first admission for heroin addiction in 1910. Federal government regulation of heroin began with the Harrison Narcotic Act in 1914, and by 1924, the US Congress banned all domestic manufacture of heroin (Scott, 1998).

A century later, opioid use, including heroin, has seen a marked increase, and with that increase has come a dramatic rise in overdose deaths. As shown in Fig. 1 below, in 1999, there were a total of 8,050 opioid-related deaths in the United States. By 2016, that figure had risen to 42,249. While opioid overdoses have risen throughout this period, the steepest rise has occurred since 2013.

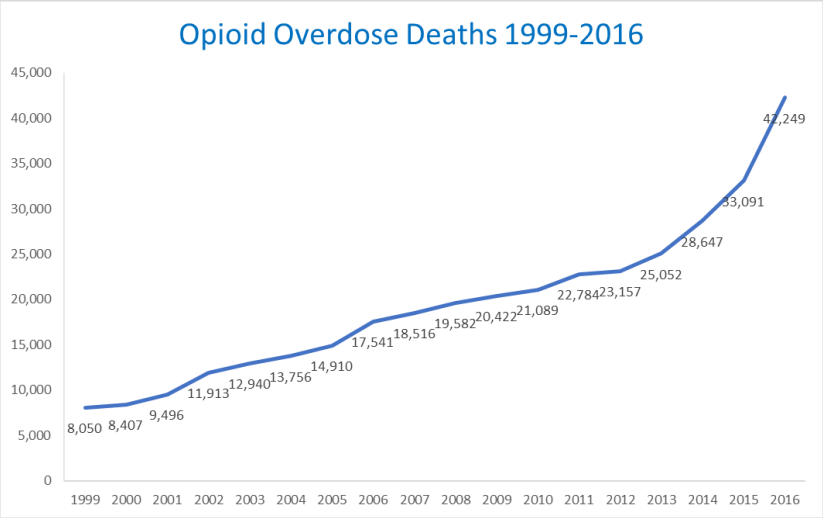

Figure 1: Opioid Overdose Deaths (Henry J. Kaiser Family Foundation, 2018)

Since the 1960s, the demographics of opioid users have changed. Over time, the proportion of female heroin users has risen from less than 20% to around half of opioid users. The ethnicity of users has changed as well. In the 1960s, whites made up half of heroin users seeking treatment, but by the 2010s, whites accounted for over 90% of treatment seeking users (Cicero, Ellis, Surratt, & Kurtz, 2014). Although whites make up the largest share of heroin users, this epidemic affects all races, and opioid related death rates have increased for all ethnicities. Fig. 2 below shows that while non-Hispanic whites retain the highest death rates, non-Hispanic blacks have seen a recent sharp increase.

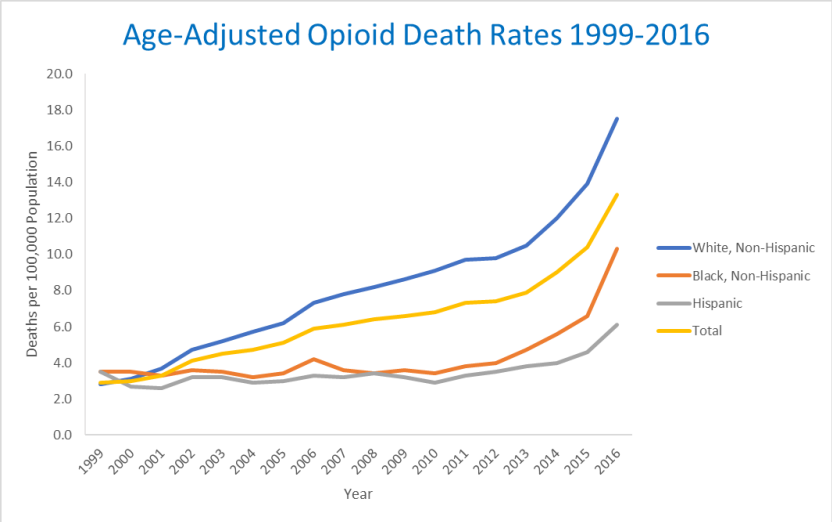

Figure 2: Opioid Death Rates by Race (Henry J. Kaiser Family Foundation, 2018)

One potential driver of this increase has been a change in usage patterns. Since the middle of the last century, the way opioid abusers initiate their habits has changed dramatically. In the 1960s, over 80% of users' first opioid abuse was with heroin. In the 2000s, over 70% of users initiated abuse with prescription opioids. This trend reversed somewhat in the 2010s as the heroin became the first opioid of abuse for 35% of users (Cicero et al., 2014).

The rise in opioid deaths has progressed in three overlapping phases as shown in Fig. 3. The first wave began with increased prescribing of opioids such as oxycodone and hydrocodone in the 1990s. This increase occurred as attitudes shifted to help patients avoid pain despite a lack of objective studies to quantify the risks of an increase in opioid prescriptions (Wilkerson, et al., 2016). As a result, deaths due to this activity have increased since 1999. Abuse deterrent formulas of opioids helped to reduce their misuse (Wilkerson et al., 2016), but other abuse patterns arose. The second wave began in 2010 as heroin usage increased and caused a steep increase in overdose fatalities. The third wave began in 2013 and has seen a dramatic rise in deaths due to synthetic opioids such as fentanyl. Fentanyl, which is approved for cases of extreme pain, is 50 to 100 times more potent than morphine. Fentanyl presents an extremely challenging situation because it is often mixed with heroin or cocaine without the knowledge of the user.

For this study, we will focus on heroin usage. This is because the cohort of acknowledged heroin users is sufficiently large to study across multiple years. Also, because fentanyl use is often unintended, users may acknowledge heroin use in a survey while denying fentanyl use.

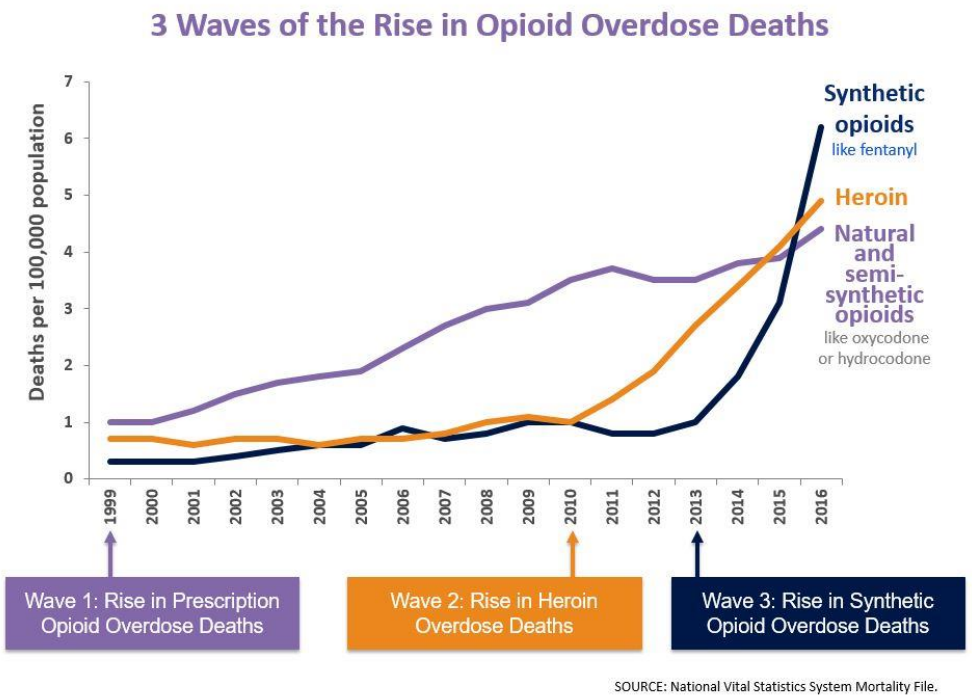

Figure 3: Waves of the Opioid Epidemic (Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, 2018)

We believe opioid deaths are avoidable. Our goal is to classify individuals and predict their likelihood to use opioids, particularly the most dangerous ones such as heroin. There are many possible correlates to heroin usage, including demographic variables such as race and sex, as well as behavioral variables such as the abuse of other illicit drugs or alcohol. Other factors, such as income and level of education could play a role as well. There are existing studies that identify correlates to heroin usage. Our study will make use of the findings of these studies and improve upon the methods and modelling techniques that they used. Ultimately, our goal is to identify predictors that could be used by medical practitioners to identify individuals who are at risk for heroin use. Those practitioners could then develop intervention strategies to reduce heroin use in America.

Data Concerning Drug Use

Data about drug use and addiction in America, like all health data, is subject to limitations due to privacy concerns. There are several types of datasets that can be used in research. We can track the distribution of controlled substances in the United States through ARCOS (Automation of Reports and Consolidated Orders System), which is published by the Drug Enforcement Agency. This system the monitors the flow of controlled substances from their point of manufacture through commercial distribution channels to point of sale or distribution at the dispensing/retail level - hospitals, retail pharmacies, practitioners, mid-level practitioners, and teaching institutions (US Department of Justice Drug Enforcement Agency, n.d.).

Another category of data to explore is medical claims information. Some states have mandated collection of claims and their publication in public datasets for researchers. The APCD (All-Payer Claims Database) Council has collected links to states that have taken this step, and researchers can gather volumes of data directly from the states. Medical claims data can also be obtained by working with managed care providers or insurance companies. While these datasets are only specific to one set of patients, they can contain more detailed information than is found in public data.

The US Federal government collects vast amounts of data on subjects related to healthcare and makes these datasets available to researchers. Studies include annual reports of vital statistics, surveys of health among specific categories of citizens, epidemiologic research, and many other valuable sources.[1] For those who seek to investigate drug use and addiction, there has never been more information available to researchers, and it has never been easier to obtain.

The National Survey on Drug Use and Health[2]

The Substance Abuse and Mental Health Services Administration (SAMHSA) is an agency within the U.S. Department of Health and Human Services whose mission is to reduce the impact of substance abuse and mental illness on America's communities (Substance Abuse and Mental Health Administration, n.d.). SAMHSA was established by Congress in 1992 to make substance use and mental disorder information, services, and research more available. Since then, SAMHSA has administered and collected data and disseminated it to the public through written reports, online query tools, and full datasets.

One of SAMHSA's publications is the National Survey on Drug Use and Health (NSDUH). NSDUH is an annually published dataset that is the result of a massive survey regarding substance abuse and mental health in the United States. The NSDUH data has tracked the prevalence and correlates of drug use in the United States since 1971. The goals of the study are to:

- Provide accurate data on the level and patterns of alcohol, tobacco and substance use and misuse
- Track trends in the use of alcohol, tobacco, and various types of drugs
- Assess the consequences of substance use and misuse
- Identify those groups at high risk for substance use and misuse

NSDUH is used by many government agencies, private organizations, individual researchers, and the public at large. Because the survey contains so many variables, it is an excellent dataset for exploration, and we have chosen to focus our study on NSDUH.

[1] The National Institutes of Health publishes a list of databases at https://www.nlm.nih.gov/hsrinfo/datasites.html
[2] The following discussion refers to information in the NSDUH Codebook.

NSDUH provides information about the use of illicit drugs, alcohol, and tobacco among members of the U.S. civilian, noninstitutionalized population aged 12 or older. The survey also includes several modules of questions that focus on mental health issues. Surveys have been conducted periodically since 1971, with the most recent ones in 1979, 1982, 1985, 1988, and annually from 1990 through 2014. Currently, public use files are available for surveys from 1979 onward. Due to improvements in the methodology of the survey, data from 2002 onward should not be compared to years prior to 2002. Also, some changes to the survey in 2015 and 2016 regarding prescription opioid use limit some comparisons between those years and the surveys from 2002-2014. We restricted our analysis to 2003-2014 to work within these limitations and to allow us to construct four datasets with three years of information in each.

NSDUH's sampling methodology is designed to capture as many geographic and demographic sections of the United States as possible. Each observation in the resulting dataset contains a weight which researchers can use to extrapolate the observation to a section of the population. The sum of the weights of the observations is equal to the population of the United States as measured by the most recent census. Necessarily, the survey is huge -- in 2014, NSDUH contained 55,271 observations gathered from 67,901 interviews conducted by 700 field investigators. There are some limitations to NSDUH:

- The data are comprised of self-reports of drug use, and their value depends on respondents' truthfulness and memory
- The survey is cross-sectional rather than longitudinal. That is, individuals were interviewed only once and were not followed for additional interviews in subsequent years.
- Because the target population of the survey is defined as the civilian, noninstitutionalized population of the United States, a small proportion (approximately 3 percent) of the population is excluded.

Nevertheless, the size, consistency over time, and thoroughness of NSDUH make it an exceptional resource for large dataset exploration.

NSDUH contains questions pertaining to drug usage history, mental health history, and demographic factors. The dataset also contains variables that have been imputed by SAMHSA to reduce missingness and to improve the accuracy of results. For example, if a respondent skips a question regarding whether she has ever used heroin but subsequently answers a question regarding the last time she used heroin, an imputed heroin use flag will be positive. SAMHSA recommends that researchers use imputed variables rather than direct responses. There are thousands of variables in the dataset -- in 2014, there were 3,148 variables for each observation.

To facilitate research across multiple years, SAMHSA has released a single dataset containing the NSDUH survey results from 2002-2014. We used this file for our study. There are 722,653 observations in the dataset, and each observation has 3,625 variables. Because the

survey changes slightly each year, the variables selected for this dataset are consistent across multiple, but not always all, years in the range.

The richness of the data in NSDUH allows researchers to explore a great number of topics. For example, data from 2005-2014 show that there has been an increase in binge drinking and alcohol use disorder among subjects aged 50 and over (Han et al., 2017). Another study correlated drug use to employment, finding that subjects who were unemployed following the 2008 recession were more likely to have been marijuana users prior to losing their jobs (Compton, Gfroerer, Conway, & Finger, 2014). Interestingly, this study also showed that NSDUH, despite being a self-reported survey, matches other objective data such as the National Bureau of Labor Statistics. In summary, NSDUH is ideally suited to help us employ data science methods on a large dataset to seek new information regarding heroin use.

Literature Review

The purpose of our literature review was twofold. First, we sought to learn what correlates to heroin use have been identified in other studies. This information helped to guide the assumptions that we used in exploring the NSDUH dataset. Second, we desired to see what methods other researchers have used and whether we could improve upon those. Due to the seriousness of the opioid epidemic, there is no shortage of studies that investigate correlates between heroin abuse and other factors. Because victims of the opioid epidemic use prescription opioids (POs), heroin, and synthetic opioids, we chose to review literature that was not limited to heroin alone.

Earlier Findings Regarding Opioid Abuse

Wilkerson et al. provide a summary of existing research and point out that there is a set of demographic characteristics that is tightly linked to opioid abuse. Opioid users are more likely to be Caucasian (non-Hispanic white), men aged 18-25, Medicaid eligible, have low household incomes, and initiate abuse with non-medical use of POs (Wilkerson et al., 2016).

Several studies use NSDUH data. These studies show that heroin users tend to be non-Hispanic whites, have used cocaine or POs within the last year, live in larger cities, and have either no health insurance or rely on Medicaid (Jones, 2013). Similarly, an evaluation of NSDUH data across several years shows that heroin initiation is strongly related to prior abuse of POs, and while there is no racial difference in the likelihood of heroin use among PO abusers, heroin users who were not PO abusers were more likely to be non-Hispanic whites (Cerd, An Santaella, Marshall, Kim, & Martins, 2015). Some researchers have used NSDUH to explore very specific topics. For example, one study shows that heroin users were more likely to have had prior use of inhalants (Wu & Howard, 2007).

Other datasets regarding opioid abuse include electronic medical records, prescription data, and smaller scale surveys. These datasets tend to be proprietary and are obtained from health care providers such as HMOs or state government records of hospital admissions. Despite

their different origin, these datasets confirm the findings from studies of NSDUH. PO abuse is linked to early first non-medical use of POs (McCabe, West, Morales, Cranford, & Boyd, 2007). Subjects with mental health disorders or other substance use disorders had a higher incidence of opioid abuse (Edlund et al., 2010).

Some studies have found patterns that can be exploited in the treatment or prevention of abuse. For example, opioid abusers tend to be male, receive prescriptions for opioids with more days of supply, receive more medical and psychiatric treatments, and are prescribed more concordant drugs with their opioid medications (Cochran et al., 2014). Stumbo attempts to generalize combinations of factors by defining five pathways to opioid abuse (Stumbo, et al., 2017):

- Inadequately controlled chronic pain leads to misuse
- Individual vulnerability to opioid dependence even after brief opioid exposure
- Individuals with prior substance use problems who are prescribed opioids
- Relief from emotional distress which reinforces misuse or abuse; and
- Abuse begins with recreational or non-medically supervised use of opioids

One way that knowledge of these factors has been used to reduce opioid abuse has been to restrict access to opioids via changes in prescriptions. Subjects who have been prescribed opioid medications with lower days' supply, lower average doses, and limitations to Type III and IV opioids have a lower likelihood of opioid abuse disorder (Edlund et al., 2010).

Methodologies of Earlier Studies

Except for Cerd et al. (2015), all the aforementioned studies used a similar methodology. Each started with an assumed set of likely predictors of opioid or heroin abuse. These assumed predictors were sometimes chosen from reviews of other studies, and sometimes the predictors were specifically chosen to test a hypothesis. For example, the question "Do lower prescription amounts reduce the likelihood of opioid abuse?" calls for a very specific analysis. The predictors were tested using frequentist techniques – researchers looked for differences in levels of the predictor between categories of outcome variables. The significance of these differences was then tested with null hypothesis significance tests (NHST). With a set of predictors in hand, researchers then built traditional regression models to determine how the factors changed the odds of belonging to an outcome category. For example, regression results have shown that the odds of having a non-fatal overdose are much higher (adjusted odds-ratio = 3.68) for a user who injects opioids than one who doesn't.

There are two areas for improvement in this approach. First, datasets such as NSDUH contain a great number of potential predictors, which also means that there is an enormous set of interactions among these factors, any of which could be a very strong predictor. If we can consider more factors, we may uncover findings which are less obvious. We can also improve the accuracy of estimating the effect of any predictor. We already know that many studies have

shown that prior non-medical use of POs is associated with heroin use. We should take such knowledge into account when conducting analyses that compare the odds of heroin or opioid abuse.

Classification and regression tree analysis (C&RT) provides one way of identifying factors and combinations of factors that can be predictors. C&RT analysis is a way to explore the relationships between an outcome variable and many potential predictors. The method starts by finding the most significant predictor variable that splits observations (branching) into two mutually exclusive subgroups (nodes). Each element of a node has the same value for the outcome variable as every other member. The method proceeds to then divide each one of the nodes into two other nodes and continues in this fashion until each lowest node (leaf) is completely homogenous, or as is more likely, some stopping criterion is satisfied.

Tree-based and rule-based methods generate models that are easy to interpret. They can also handle many types of predictors, including nominal, ordinal, and interval variables, and those predictors don't need to be pre-processed. Tree-based models also do not require to user to assume the underlying distribution of the predictor (Kuhn & Johnson, 2013). When examining datasets with many variables, and which have many potential interactions between those variables, C&RT is an effective way to reduce the number of interactions for further consideration (Piper, Loh, Smith, Japuntich, & Baker, 2011). For these reasons, C&RT methods are an ideal way to for us to identify potential predictors from NSDUH, a dataset containing thousands of variables. Indeed, other authors have advocated that classification trees should be used for data mining medical data (Koh & Tan, 2005). However, C&RT has some drawbacks – it is subject to model instability as data change and has sub-optimal predictive performance. Therefore, it should be seen as complimentary to other methods, such as regression, rather than as a replacement for them (Fernández, Mediano, García, Rodríguez, & Marín, 2016; Piper et al., 2011).

C&RT has occasionally been used in medical studies. For example, one study of HMO data found that subjects of 24-25 years old with at least three prescriptions from at least three different pharmacies were more likely to have abuse-related diagnoses (Chitwood-Dagner, Carlson, Friedman, & Skatter, 1995). C&RT's flexibility can be seen in the range of medical studies in which it has been applied. Some of these include predicting mastitis (Fernández et al., 2016), tobacco smoking relapse (Piper et al., 2011), HIV risk analysis (Frisman, Prendergast, Lin, Rodis, & Greenwell, 2008), and the impact of religiosity in avoiding suicide in Iran (Baneshi et al., 2017).

There are assertions that Bayesian methods are superior to traditional methods such as NHST and regression for medical studies. Bayesian methods are better for sparse data, and the use of prior distributions allows for the inclusion of regressors that might otherwise be excluded by p-value selection (Greenland, 2007). Traditional methods also don't have the ability to prove null hypotheses because p-values don't change with the addition of additional results in favor of

the null. In contrast, Bayesian techniques, such as the use of the Bayes Factor to compare models, allow the proof of nulls (Dienes, 2008).

Perhaps the best reason to use Bayesian techniques is the ability to use prior knowledge in the construction of predictive models. In our case, we already know that abuse of POs is a strong predictor for heroin use. If our model includes PO abuse as a predictive variable, we can use a prior distribution that is positive for the effect of PO abuse when we calculate odds ratios via Markov Chain-Monte Carlo (MCMC) analysis. Taking advantage of prior knowledge requires a careful approach. Priors should be fair, well reviewed, and should consider previous studies (Dienes, 2008; Greenland & Poole, 2013). Even though they are difficult to construct, informed priors, even weak priors, are better than equal odds priors or frequentist methods, which assume no prior knowledge at all. The assumption of no prior knowledge is inappropriate for most medical studies (Greenland & Poole, 2013).

While setting priors could be complex, we can adopt some simplifying methods to estimate a weakly informative distribution. One suggestion for the construction of priors is to have three categories: "uncertain", "probably positive", and "probably strong". The prior for each category would be of the form $P(parameter) \sim f(Median = m, variance = v)$. The variance for "uncertain" would be the highest, and the one for "probably strong" would be the lowest (Greenland, 2007). Another even simpler method for setting priors would be to take an estimate of the mean of a predictor, round that value up to the next "large" value, and set a standard deviation equal to half of that value (Dienes, 2008).

We propose a better way to conduct analyses of drug abuse data than the traditional ones we saw in the literature search. We shall combine classification and regression tree (C&RT) methods with Bayesian MCMC methods. C&RT will allow us to uncover likely heroin use correlates without overly limiting the vast set of variables in NSDUH to consider. C&RT will also allow us to find combinations of variables that are potential correlates. Some of these correlates will match those found in earlier studies. For those, we will develop prior distributions that recognize their impact on heroin use. We will then use MCMC methods to determine posterior distributions and odds ratios for the correlates that make use of those informed prior distributions. This approach should be more comprehensive and accurate than traditional methods.

Methods

The goal of our study is to construct a model that can define the probability of heroin use based upon the values of a set of variables from the NSDUH dataset. We used this model on multiple datasets that span different blocks of years to see if the predictors of heroin use have evolved over time. There are four steps in this process: dataset preparation, variable selection, model construction, and model validation. In dataset preparation, we reduced the NSDUH 2003-2014 dataset, which consists of 722,653 observations and 3,625 variables, into four smaller datasets. Each of these datasets contained three years of data and a useful and manageable

collection of variables. In variable selection, we considered each three-year dataset and extracted a set of variables that is relatively small yet best correlates with heroin usage. In model construction, we determined how changes in each of those variables are related to changes in the probability of heroin usage over a three-year period. In model validation, we applied our model to dataset samples that we held out from model construction and evaluated the accuracy of its predictions.

Dataset Preparation

Sections of the NSDUH Dataset

Each observation in the NSDUH dataset represents information about a single respondent. The variables in the dataset are data about the respondent. The variables for each observation are organized into sections. The first is the group of responses by the participant to questions related to substance use. SAMHSA calls these "self-administered substance use" questions, and they cover topics such as age of first use, frequency of use, and other usage characteristics for each of a great number of substances.

The second section consists of variables that are imputed by SAMHSA from the first set. SAMHSA recommends the use of the resulting imputed variables for multivariate analyses (SAMHSA, 2016). We used imputed variables and ignored self-administered variables whenever possible.

There are several other self-administered sections that provide information beyond simple substance use statistics. These include treatment history, social environment factors, youth experiences, mental health and depression history, income, insurance coverage, and many other special topics. As with the drug-use section, each special topic self-administered section is followed by a recoded section containing imputed variables.

The demographic section of self-administered and imputed variables contains data about marital status, education level, and employment type, history, and environment. Next, the dataset contains geographic variables regarding population density and whether the respondent is in an American Indian area. The last section in the dataset consists of a set of weighting factors. These factors allow researchers to extrapolate information in NSDUH to the entire population of the United States. For example, the first observation in the dataset has a one-year weighting factor of 10773.49 – the values for the variables of that observation represent those of approximately 10,773 Americans.

Manual Reduction of the Dataset

The first step in dataset preparation was to manually review all 3,625 variables, section-by-section, to eliminate unnecessary or redundant predictors. Whenever possible we used imputed values rather than self-administered ones to improve the accuracy of the data. We also eliminated obvious duplicates and variables that recategorized values from other variables. For example, the variable "ALCAFU" has values that represent categories of age-of-first-use of

alcohol. One such value is "15-17 Years Old". This value occurs in observations for which the value of imputed alcohol age of first use ("IRALCAGE") is 15, 16, or 17. In cases like these, we kept the ordinal variable rather than the categorical variable.

We then eliminated substance-use variables for uncommon drugs that were also grouped together under another single variable. One grouped variable is "BENZOS", a binary variable that is true if the respondent has used any of the following drugs: klonopin, clonazepam, Xanax, alprazolam, Ativan, lorazepam, valium, diazepam, Librium, limbitrol, rohypnl, serax, or tranxene. In this case we were able to use one variable that represented a true answer for any of thirteen other variables.[3]

We did not include any of the weighting factor variables. Our model seeks to find relationships between heroin and other predictors. It does not seek to provide aggregate statistics regarding substance use, so there is no need to include the extrapolation factors. Finally, we eliminated any variables that were obviously completely correlated with heroin use. For example, we did not include frequency-of-use of heroin or age-of-first-use of heroin variables because they hold no meaning for respondents who have never used heroin.

We removed nine additional variables that did not have enough data to be considered in our study. The impact of the manual reduction of the dataset was to drop the number of variables for consideration from 3,625 to 371.

Manual Imputation and Modification

Even though SAMHSA conducted imputation that reduced errors and missingness, we were still left with missing values in our dataset. Knowing that random forest classification trees must not have missing data, we chose to eliminate missing values directly. Using the MICE and VIM package in the R programming language, we identified all variables that retained missing values. There were 127 of these variables, most of which fell into one of three categories. The first category was the set of variables that was specific to youths (under 18). An adult respondent would have missing values for any variable that was only asked of youths. The second category was specific to adults. In this case, a youth respondent would have missing values for any variable that was asked only of adults. In these two categories, we replaced missing values with "false", 0, or whatever value would appropriately convey a negative response.

The third category consisted of sets of variables that required specific intervention. One such set indicates a respondent's history with physical or mental ailments. For example, the variable "HEPATLIF" is true if a respondent indicated that he had ever had hepatitis. When we

[3] We did not create any other groupings of substances beyond those contained in the NSDUH.

encountered missing values in this category, we again replaced them with "false", 0, or an appropriately negative value.[4]

Another block contained nominal variables with specific values for skips or refusals to answer. The variable "TOTDRINK", which represents the number of days a respondent used alcohol in the past twelve months, has the following values as shown in Fig. 4:

TOTDRINK Variable		
Value	Interpretation	Frequency
Range from 1-372	Number of days used alcohol	433,507
994	Don't know	2,977
997	Refused	756
998	Blank (No answer)	285,413

Figure 4: Variable with Missing Values

There were 285,413 respondents who did not indicate how many days they had consumed alcohol in the past year. From other variables, we know that 282,124 respondents said that they had not used alcohol in the past year. Since these two values are very close, we conclude that blanks can be assumed to be 0. Consequently, we simplified the "TOTDRINK" variable by replacing the values "994", "997", and "998" with 0. We treated all such nominal variables in this manner.

Questions regarding treatment history had three potential answers – 0 if the respondent felt the need for treatment but didn't receive it, 1 if the respondent received treatment, and missing if the respondent didn't answer or didn't feel the need for treatment. We replaced the missing value with -9, a value that is used throughout the NSDUH dataset to indicate questions that were not asked in a given year. We also used this method with other variables that contained missing as a valid answer – we replaced missing with -9.

Questions regarding the ability to obtain substances had three possible values: 1 = "Fairly or Very Easy", 0 = "Other", or missing, which represented "Unknown". We replaced missing with 0 for these questions. By doing this, we are allowing the most interesting response, "Fairly or Very Easy", to maintain its impact. We used this method with other blocks of similarly structured questions, such as the group regarding a respondent's perception of the risk associated with use of specific substances.

The last category of variables we imputed manually included questions regarding needle use and crime. These questions were worded in a way that should have generated a binary response. The question regarding cocaine needle use was, "Have you ever, even once, used a needle to inject cocaine?" The answers included: "YES", "NO", "YES (Logically assigned)",

[4] Because the distribution of positive responses for these variables was so low, we assumed that that correct answer for a skip was most likely negative.

"BAD DATA", "NEVER USED COCAINE", "DON'T KNOW", "REFUSED", and "BLANK (No answer)". We simplified this set of answers to simply "YES" and "NO". The number of "BAD DATA", "DON'T KNOW", "REFUSED", and "BLANK" values represented less than 0.02% of all answers, so assigning them to "NO" maintained the integrity of the data. Questions regarding whether a respondent had committed a particular crime were similarly structured. We manually imputed the values for these variables into binary values as well.

The combination of these interventions and imputations left us with a dataset containing no missing values and a cleaner set of answers to straightforward questions. The final manual change we made was to eliminate observations from 2002. We did this to prepare the dataset for separation into three-year blocks: 2003-2005, 2006-2008, 2009-2011, and 2012-2014.

Data Preparation: Exploration

Before conducting variable selection with classification tree methods, we wanted to see if there were any patterns that we could identify in the data that would influence our analysis. Specifically, we wanted to determine the following:

- Are there any variables that explain so much of the variance of the data that they warrant special handling?
- Which variables are so highly correlated to each other that they can be combined?

In this step, we needed an efficient method that could provide us answers to these two questions in an unsupervised fashion. We elected to use Principal Component Analysis, a method that satisfies these criteria.

Principal Component Analysis Review

Principal Component Analysis (PCA) is a method that is used to simplify high-dimension multivariate data analysis. PCA transforms possibly correlated variables into a smaller number of uncorrelated features called principal components. The first principal component accounts for as much variation in the original dataset as possible, and each subsequent component accounts for the next most amount of variation.

PCA is an efficient technique because it is ultimately a solution to a set of linear equations. If we let $\boldsymbol{X}$ be an $m\ x\ n$ matrix of our data, where m is the number of observations and n is the number of variables, then the covariance matrix for our data is given by:

$$\mathbf{S} = \text{cov}(\mathbf{X}) = \text{E}[(\mathbf{X} - \text{E}[\mathbf{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^T$$

and

$$\boldsymbol{W}^{-1}\,\boldsymbol{SW} = \boldsymbol{D}$$

where $\boldsymbol{W}$ is the set of eigenvectors of $\boldsymbol{S}$, and $\boldsymbol{D}$ is a diagonal matrix of the eigenvalues of $\boldsymbol{S}$. The ith eigenvector w_i is the ith principal component of our original dataset. The ith eigenvalue λ_i,

where $\sum_i \lambda_i = 1$, describes the proportion of variance of the original data that is explained by w_i. For example, if $\lambda_i = 0.20$, then the principal component given by w_i explains 20% of the variance of the data.

There are some drawbacks to using PCA to explain data. First, the method will favor components that have more variation. In our case, this means that variables with a wider range of values will tend to fall on the first components. We can compensate for this effect by z-score standardizing the variables in the dataset so that each variable will have a mean of zero and standard deviation of 1:

$$z - score = \frac{X - \bar{X}}{Standard\ Deviation}$$

Another drawback is that the features generated by PCA are complex linear combinations of all the variables in the dataset. In our case, each feature would be a function of 371 variables. These features are hard to interpret and do not allow us to make explanatory conclusions regarding heroin use. That said, the features do satisfy the goals of our exploration.

Principal Component Analysis Results

Results from the First Pass

We conducted three passes of PCA on the NSDUH dataset. In each case, we randomly selected 25% of the observations from the 2003-2014 dataset for consideration. We then used the PCA() function from the R package FactoMineR (Husson, Josse, Le, Mazet, & Husson, 2018) to determine the principal components and their corresponding eigenvalues. We conducted a preliminary run that showed the effect of the components decreases significantly after the first three. To ensure we captured the most important components for each run, we set the ncp (number of principal components) parameter for PCA() to ten. We analyzed the results for each pass to determine the twenty most positively and twenty most negatively correlated variables for each component. From this information we determined whether to remove any variables from the dataset or to restructure it entirely. In this section we will describe in detail our analysis of the first pass to illustrate our methodology. We will then summarize our findings for the second and third passes.

In the first pass, we started with the manually reduced 2003-2014 dataset: 668,574 observations with 371 variables. We then z-scored the dataset and extracted a random sample of 167,143 observations (25% of the total). We ran PCA() on this dataset and found that 18 principal components explained at least 50% of the variance in the dataset. The first component explained 11.8% of the variance, the second accounted for 9.2%, and the third 5.9% (see Fig. 5).

NSDUH PCA Analysis	First Pass	Second Pass		Third Pass	
	Total	Youth	Adult	Youth	Adult
Components to Explain 50% Variance	18	30	29	33	37
First Component Variance	11.8%	10.2%	8.9%	10.2%	8.8%
Second Component Variance	9.2%	4.6%	4.5%	3.7%	3.5%
Third Component Variance	5.9%	3.7%	4.2%	3.6%	3.1%

Figure 5: Variance Explained by Principal Components

The variables most positively correlated with the first component in the first pass were all related to respondents' histories of physical illness. The top twenty all had virtually identical correlation scores of 0.997. The most highly correlated variable was whether the respondent had lung cancer in the past year (luncayr). A raw value of 1 for luncayr indicated that she had, 0 indicated that she hadn't, and -9 indicated that data wasn't collected for the year of that survey. Thus, respondents who had lung cancer were positively correlated with the first component. Similarly, the remainder of the top twenty positively correlated variables were responses to questions about illnesses such as cirrhosis, tuberculosis, pancreatitis, and HIV.

The variable most negatively correlated with the first component indicated whether the respondent had ever used oxycontin (OXYCONT2[5]). This variable was highly negatively correlated and had a score of -0.997. Because the next most negative variable had a score of -0.069, we conclude that oxycontin use dominated the negative correlation of the component. A raw value of 1 for OXYCONT2 indicated that the respondent had used oxycontin, 0 indicated that he hadn't, and -9 indicated that the data wasn't collected in that year.

Interpreting the principal components in this analysis is challenging due to the number of variables involved, but we surmise that each represents a set of features of respondents that collectively can describe a sub-population. In the case of the first component of the first pass, we see that respondents who have had physical illness but have not ever used oxycontin account for the most variability in the dataset.

The variables that most positively correlated with the second component included ismother and isfather, which had values of 1 if the mother or father was in the household, 2 if

[5] We will see later that this variable was problematic.

not, 3 for 'don't know', and 4 if the respondent was 18 or older. ismother and isfather were most correlated with the component when a parent was not in the house. The next set of positively correlated variables were flags that indicated the use of marijuana, cigarettes, hallucinogens, and alcohol.

The variable that was most negatively correlated with the second component included a youth's participation in youth activities, where a high value indicated participation. The next two were a flag indicating that the respondent was between 12 and 17 years of age and another flag indicating that the respondent was a youth. The third most negatively correlated flag was one that indicated whether a youth had seen a drug prevention message outside school.

The third principal component was like the second. It was dominated by variables that were specific to youths. Its primary difference from the second component was that it was reversed in direction.

Interpreting the First Pass

The first pass of PCA provided us with information that enables us to improve the preparation of the data for classification and regression trees. The first component was dominated by physical illness flags, all of which were almost identically correlated with the component. When we examined the number of positive respondents for any of these flags, we saw that the numbers were very low. For example, the number of respondents who had lung cancer in the past year was only 207, and only 9 of those had ever used heroin. We concluded that we could simplify the dataset by replacing all the individual physical illness flags with one flag, PHYSICKEVR. This flag would equal 1 if the respondent had indicated that she had at least one of the individual physical illnesses, and 0 otherwise. We also replaced the individual mental health flags with one summary flag, MHSICKEVR. This flag would equal 1 if the respondent had ever had anxiety or depression, and 0 otherwise.

The second and third components were both dominated by variables specific to youth. Here the PCA presented us with information that we should have caught in the manual reduction of the dataset. The NSDUH survey contains large sections that are age-specific. There are blocks of questions that are reserved for youths (respondents under the age of 18) that are not asked of adults. There are also blocks of questions reserved only for adults. To retain these variables, we must therefore split the dataset into two portions – one containing youth respondents and one containing adult respondents.

After the first pass of PCA, we were able to eliminate a large group of variables through consolidation and we split the dataset into two subsets. We then had one dataset for youths that contained 212,558 observations with 331 variables, and one dataset for adults with 456,016 observations with 351 variables.

Results from the Youth Second Pass

We ran both the youth and adult datasets through the second pass of PCA and again found variables we could remove from the dataset. The first thirty components explained 50% of

the variance in the youth dataset. The first component accounted for 10.2% of the variance, the second 4.6%, and the third 3.7%. The variables most positively correlated with the first component were led by the flag indicating hallucinogen use, whether the respondent needed treatment for drug or alcohol use during the past year, whether the respondent had ever used psychotherapeutics (a category including sedatives, tranquilizers, stimulants, and analgesics), and two variables that both indicated the number of days the respondent had used marijuana in the past year. The variables most negatively correlated with the first component were all drug and alcohol specific age-of-first use indicators. The first component therefore represented drug and frequent marijuana users who began their substance use at an early age.

The most positive correlates to the second component were variables that indicated why a respondent did not receive treatment for drug and alcohol use despite feeling the need. The negative correlates were led by flags indicating the respondent did not feel the need for treatment for drug and alcohol use. These were followed by flags indicating alcohol and tobacco use in the past year. The second component represented users who felt the need for substance abuse treatment, didn't receive that treatment, and had low levels of alcohol use.

The third component represented respondents who had suffered mental illness or physical illness, indicated sources of sedatives, stimulants, or tranquilizers, had used oxycontin, had no cigarette use, and had not received treatment for mental health illness in the prior year.

Results from the Adult Second Pass

The first 29 components explained 50% of the variance in the adult dataset. The first component accounted for 8.9%, the second 4.5%, and the third 4.2%. The positive correlates to the first component were flags indicating the use of specific drugs, the most positive of which was the flag for psychotherapeutics. The negative correlates were the specific drug age-of-first-use variables. The first component represented drug users who began their use at an early age.

The second component of the adult dataset represented users who felt the need for mental health treatment, did not receive that treatment, and supplied reasons for not doing so. These respondents also did not smoke or use hallucinogens. The third component represented respondents who felt the need for drug and or alcohol abuse treatment yet didn't receive it for various reasons. These respondents also described reasons for not receiving treatment for mental health illness.

Interpreting the Second Pass

Both the youth and adult datasets had similar principal components in the second pass. They both contained components that had nearly identically correlated variables indicating reasons for not receiving mental health or substance abuse treatment. The number of positive responses for any one of these variables was very low. For example, the total number of adults across all years who did not receive substance treatment because they didn't feel the need for it (despite abuse) was only 287, and only 26 of those had ever used heroin. We concluded that we

could drop the variables for specific reasons for not receiving mental health or substance treatment without impacting our analysis of heroin use.

We also uncovered some anomalies with specific variables. We discovered that the flag for psychotherapeutics was redundant. It duplicated responses to the individual drug use questions for sedatives, tranquilizers, stimulants, and analgesics. We decided to simplify the dataset by removing he psychotherapeutic flag. Also, we discovered that there were two different imputed variables for oxycontin use: oxyflag and OXYCONT2. The second of these, OXYCONT2, involved additional data manipulation by SAMHSA in the preparation of NSDUH. Because the imputation methodology for oxyflag was the same as that for other drugs, we dropped OXYCONT2 in favor of oxyflag. Also, we found a set of substance frequency use variables that were duplicated by imputed frequency variables. We eliminated the non-imputed variables from the dataset. With these changes, the adult dataset was reduced to 301 variables and the youth dataset was reduced to 296.

Results from the Youth Third Pass

The third and final PCA pass on the youth dataset produced 34 components that explained 50% of the variance of the dataset. The first component accounted for 10.1% of the variance, the second accounted for 3.8%, and the third 3.6%. The variables most positively correlated with the first component included whether the respondent had used hallucinogens, whether he had needed substance abuse treatment in the past year, the frequency of his marijuana use in the past year, whether he had hallucinogen and marijuana use in the past year, and whether he had ever used hydrocodone products. The most negatively correlated variables were age-of-first use variables and the level of alcohol use in the past month (higher values indicated lower use). The first component represented youths who were recent users of hallucinogens and/or marijuana, began drug use early, were binge drinkers, and needed substance abuse treatment in the past year.

The variables most positively correlated to the second component included whether a respondent had ever been mentally ill, physically sick, indicated sources for nonmedical use of drugs, and had used oxycontin in the past year. The most negatively correlated variables were whether the respondent had ever used cigarettes, had received in-patient mental health services in the past year, had recent alcohol and/or cigarette use, and had a perceived great risk of LSD use. The second component represents youths who had been ill but not received treatment, were substance abusers, used oxycontin, and had a low perceived risk of drug use. The third component represented tobacco users who began their use early but did not have recent use of smokeless tobacco or snuff.

Results from the Adult Third Pass

The third PCA pass on the adult dataset produced 36 components that explained 50% of the variance. The first component accounted for 8.8%, the second 3.6%, and the third 3.1%. The variables most positively correlated to the first component were variables that indicated that

the respondent had ever used hallucinogens, cocaine, pain relievers, tranquilizers, and hydrocodone products. The most negatively correlated variables were age-of-first use variables for various substances. The first component represents respondents who were drug users and who had started their use early.

The variables most positively correlated to the second component included whether the respondent had ever been mentally or physically ill, indications of sources of stimulants and tranquilizers, and indications of recent oxycontin use. The negatively correlated variables included whether the respondent perceived great risk of frequent marijuana use, participated in government assistance programs (low values indicate participation), had easy access to LSD, and had been prescribed medicine for mental health illness in the past year (low values indicate prescriptions). The second component represents respondents who had been sick, used oxycontin, had received prescriptions for mental health illness, were on government assistance, and did not perceive risk from frequent marijuana use. The third component represents respondents who began tobacco use early, do not drink heavily, and had no recent smokeless tobacco or snuff use.

Interpreting the Third Pass

The completion of the third pass of PCA demonstrated that we now had two datasets without unnecessary and highly correlated variables. We did not identify any further needs for variable elimination, and we had well defined components. Consequently, the adult and youth datasets were ready for the next stage of our study, classification and regression tree analysis. The youth dataset now consisted of 212,558 observations and 296 variables. The adult dataset contained 416,016 observations and 301 variables. Fig. 6 summarizes the top three components in the final datasets.

Dataset	Dataset Size		Principal Component Description		
	Observations	Variables	First	Second	Third
Youth	212,558	296	Recent drug users who began use early, were binge drinkers, and needed substance abuse treatment	Substance abusers with untreated illness, had used oxycontin, and had a low perceived risk of drug use	Tobacco users who began use early but who had not had recent use of smokeless tobacco or snuff.
Adult	416,016	301	Drug users who began their use early	Oxycontin users who had been sick, received prescriptions for mental illness, did not perceive risk from frequent marijuana use, and who were on government assistance.	Tobacco users who began use early, did not drink heavily, and who had not had recent use of smokeless tobacco or snuff.

Figure 6: Dataset Split and Principal Components

Separation into Three-Year Blocks

To evaluate changes in heroin usage over time, we separated the post-PCA 2003-2014 datasets into several smaller datasets. Heroin usage is relatively rare, and we wanted to have as many users as possible in each dataset while still allowing us to analyze changes over time. Based upon a recommendation from SAMHSA, we split the file into three-year blocks: 2003-2005, 2006-2008, 2009-2011, and 2012-2014.

	Years in Dataset	Number of Observations	Number of Variables
Youth	2003-2005	55,176	296
	2006-2008	53,462	296
	2009-2011	55,185	296
	2012-2014	48,735	296
Adult	2003-2005	111,561	301
	2006-2008	111,732	301
	2009-2011	111,759	301
	2012-2014	119,964	301

Figure 7: Data Subsets by Time Period

Variable Selection by Random Forests of Conditional Inference Trees

Review of Tree-Based Methods

The next step in our study was to select a set of variables that we could carry forward into a Bayesian regression model. In the previous step, data exploration through PCA, we found variables that explained much of the variance in the total datasets. In this step, we sought to identify variables that most explain the difference between heroin users and non-users. Throughout the rest of this study, we refer to these variables as *important*.

Tree-based models are well suited for this task for several reasons. They are easy to implement and generate results that are easy to interpret. They can handle many types of predictors without the need to pre-process them. This is important in our case because our datasets contain binary, sequential, and categorical variables. Tree-based models also do not require the investigator to specify the relationship between the predictor and response variables (Kuhn & Johnson, 2013).

The oldest and most common tree method is the Classification and Regression Tree (CART) method (Breiman, et al. 1984). In this method, the model starts with the entire dataset, $\boldsymbol{S}$, and searches every value of every predictor to find the combination of predictor and value that minimizes the sum of squares within two partitions of the dataset, S_1 and S_2:

$$\text{SSE} = \sum_{i \epsilon S_1} (y_i - \overline{y_1})^2 + \sum_{i \epsilon S_2} (y_i - \overline{y_2})^2$$

$\overline{y_1}$ and $\overline{y_2}$ are the average outcomes of the response variables within the two sets S_1 and S_2. The method then partitions the subsets and continues recursively partitioning until a stopping criterion, such as minimum number of set members, is reached.

CART has some limitations. One that affects our study is that the method tends to favor predictors with large numbers of distinct values. This is because the method maximizes a splitting criterion across all possible splits simultaneously (Hothorn, Hornik, & Zeileis, 2006). In our case, we have predictors, particularly the age-of-first-use variables, that have many different values. These predictors could be inappropriately assigned high importance by CART. To avoid this, we used conditional inference trees, which eliminate the bias towards many-values correlates by splitting variable selection and variable splitting into two separate steps (Hothorn et al., 2006). A simplified version of the conditional inference tree algorithm is:

1. Test the global null hypothesis that the variables $X_j \epsilon \boldsymbol{X}$ in the dataset $\boldsymbol{S}$ are independent from the values of the response variable $\boldsymbol{Y}$. Use hypothesis testing and a threshold parameter, such as a p-value, to test for independence. If the global null hypothesis cannot be rejected, stop. Otherwise select the predictor X_j with the strongest association to the response variable.
2. Find the optimal binary split of the values for X_j into X_{j1} and X_{j2} so that $\boldsymbol{S}$ can be partitioned into two subsets such that all values of X_j in $S_1 \epsilon X_{j1}$, and all values of X_j in $S_2 \epsilon X_{j2}$. The values for the response variable $\boldsymbol{Y}$ are homogenous within each of the two subsets S_1 and S_2.
3. Recursively repeat steps 1 and 2.

Another limitation of tree-based models is that single-tree models are unstable. If the data on which they are designed alters slightly, the model can produce very different splits (Breiman, 1996b). In our case, this means that single-tree models can be highly dependent on the selection of training data from the dataset. Breiman (1996a) proposed bagging, a way to reduce the variance of predictions of individual trees. He did this by building trees from bootstrap samples of the modeling dataset and then averaging the prediction across the trees. This method still has a drawback – because at each step the algorithm considers all possible predictors, the individual trees built from the bootstrap data can still be structurally similar. We can overcome this drawback by randomly selecting a set of possible predictors at each step of tree construction. Breiman (2001) developed an algorithm that unified bootstrapping and random predictor selection – random forests. The remaining drawback of the random forest algorithm is that the ensemble of trees it uses for prediction is no longer easily interpretable as a simple set of binary splits. However, the method provides stable variable importance scores that

we can use to select the most important variables from the model for our Bayesian regression analysis.

Results from Random Forest Modelling: Adults

We used the random forest algorithm with the conditional inference tree method of selecting predictors and split values. To run this algorithm, we used the cforest() method from the 'party' package (Hothorn, 2018). This method allows the user to select the number of trees used to build the forest ('ntree'), the number of predictors to sample at each step ('mtry'), and runs in R. We ran cforest() in a powerful compute environment – a Microsoft Data Science Azure virtual machine with 8 CPUs and 16 GB of RAM. This VM includes R libraries that are designed to run multiple threads and take advantage of multiple processors.

In this compute environment, the cforest() routine ran quickly, but the Predict() method in the 'party' package consumed a great deal of memory and wouldn't complete when more than 20,000 observations were involved.[6] We were able to build models using a variety of sizes for our training set, but the evaluation of those models was limited to a random sample of 20,000 observations for both the training and test sets of data.

We ran several iterations of random forest on the 2003 to 2005 adult dataset to determine the best tuning parameters for the algorithm. To evaluate the models, we examined three statistics: the training set κ_{train}, the test set κ_{test}, and the Area Under the Receiving Operator Characteristic curve (AUROC). κ describes the relationship between the observed accuracy of the model and the expected accuracy, which is the accuracy of a random classifier given the actual and predicted data.

$$\kappa = \frac{(observed\ accuracy - expected\ accuracy)}{(1 - expected\ accuracy)}$$

The maximum value for κ is 1, and higher values of κ indicate better performance of the model.

AUROC is a function based upon the specificity and sensitivity of the model, which are defined as follows:

$$sensitivity = \frac{true\ positives}{(true\ positives + false\ negatives)}$$

$$specificity = \frac{true\ negatives}{(true\ negatives + false\ positives)}$$

[6] In contrast, we found that the CART based random forest routines, rforest() and predict(), were able to run completely. We nevertheless chose to work with cforest() to leverage the advantages of conditional inference tree methods.

Sensitivity is also known as the "True Positive Rate". (1 – specificity) is known as the "False Positive Rate". When we plot the True Positive Rate versus the False Positive Rate, we generate the Receiver Operating Characteristic (ROC) curve. The area under this curve (AUROC)

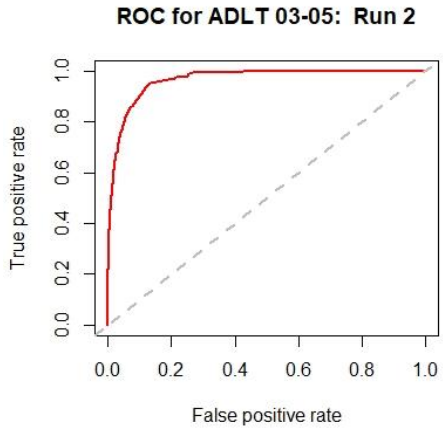

Figure 8: Example ROC Curve

approaches 1 as the performance of a model improves. As an example, the ROC curve for the second run we conducted on the adult 2003-2005 dataset is shown in Fig. 8. The area under this curve is 0.9657.

In summary, the steps we use to evaluate the random forest model are:

1) Work with the 2003-2014 adult and youth datasets for parameter optimization
2) Select a size for the training set, such as 20% or 35%
3) Run the cforest() method on the training set to build a model
4) Calculate the variable importance scores for the model
5) Apply the model to a randomly chosen 20,000 observations from the training set and evaluate its performance by computing a confusion matrix
6) Apply the model to a randomly chosen 20,000 observations from the test set and evaluate its performance by computing a confusion matrix
7) Calculate κ_{train} and κ_{test}
8) Calculate AUROC for the test set
9) Select the model with the best κ_{test} and AUROC combination

The results of the parameter optimization runs are shown in Fig. 9 below.

			Tuning Parameters								
Trial	Dataset	Observations	Trainset Size (%)	mtry	ntree	cforest() Run Time (mins)	Prediction Threshold	Training Set Kappa	Test Set Kappa	AUROC	TS Kappa * AUROC
1	Adult 2003-2005	111,561	25%	5	200	1.87	0.2	0.5766	N/A	N/A	N/A
2	Adult 2003-2005	111,561	25%	5	300	2.78	0.2	0.5821	0.4430	0.9657	0.4278
3	Adult 2003-2005	111,561	25%	10	200	2.19	0.2	0.6257	0.4673	0.9690	0.4528
4	Adult 2003-2005	111,561	25%	10	300	3.34	0.2	0.6203	0.4723	0.9687	0.4575
5	Adult 2003-2005	111,561	35%	5	200	3.52	0.2	N/A	N/A	N/A	N/A
6	Adult 2003-2005	111,561	35%	5	300	5.23	0.2	0.5658	0.4408	0.9691	0.4272
7	Adult 2003-2005	111,561	35%	10	200	4.09	0.2	0.6063	N/A	N/A	N/A
8	Adult 2003-2005	111,561	35%	10	300	6.04	0.2	0.6063	0.4884	0.9707	0.4741

Figure 9: Tuning Parameter Analysis

The tuning parameters for the adult model that resulted in the highest combination of κ and AUROC were: training set equal to 35%, mtry = 10, and ntrees = 300.

Using Set Intersection to Validate Variable Choice

Each of the random forest trials provides a list of variable importance scores. The lists vary somewhat from trial to trial. Ideally, our final list of important variables would be common across the top variables from each of the trials of our model. We can see if this is the case by comparing the top variables from the best performing trial (the one with the best κ and AUROC combination) to those from the other trials.

We took the top 10% of variables from each of the adult dataset runs. We then computed the Jaccard distance between the sets to determine the agreement among them:

$$Jaccard\ distance = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

where S_i is the set of top 10% of the variables from run i. The distances from run to run and between all the runs are shown in Fig. 10.

Trial Comparison	Jaccard
2 vs 1	0.8182
3 vs 2	0.7143
4 vs 3	0.7647
5 vs 4	0.7143
6 vs 5	0.7647
7 vs 6	0.7647
8 vs 7	0.8750
Intersection of All vs Union of All	0.4390

Figure 10: Trial Comparisons

For each iteration of random forest, between 70% and 80% of the top variables were common with the prior run. 44% of the top variables from the combination of all runs were common to all the runs. This intersection of all the top variables was a set of 18 predictors. To sort this list, we used the variable importance scores from the best performing random forest model, Trail 8. The scores are shown in Fig 11.

Variable	Trial 8 Importance	Description
crkflag	3.99E-04	Crack - ever used
rdifher	3.99E-04	Heroin fairly or very easy to obtain
ircrkage	3.66E-04	Crack age of first use
cocneedl	2.47E-04	Ever used needle to inject cocaine
cocflag	1.89E-04	Cocaine - ever used
ircocage	1.61E-04	Cocaine age of first use
pcpflag	1.41E-04	PCP - ever used
irpcpage	1.25E-04	PCP age of first use
otdgnedl	9.41E-05	Ever used needle to inject any other drug
halflag	8.03E-05	Hallucinogens - ever used
txilalev	7.83E-05	Received treatment for drug or alcohol use in lifetime
lsdflag	7.60E-05	LSD - ever used
METHDON2	7.30E-05	Methadone - ever used
MORPHIN2	7.11E-05	Morphine - ever used
benzos	7.02E-05	Benzodiazepine products - ever used
irlsdage	6.17E-05	LSD age of first use
irtrnage	5.56E-05	Tranquilizer age of first use
MESC2	5.25E-05	Mescaline - ever used

Figure 11: Common Important Variables

Of these 18 variables, all but MESC2 matched the top 18 variables in Trial 8, the run with the best κ and AUROC combination. Due to this high level of agreement, we can conclude that selecting the top variables from the best performing model does indeed provide us with a valid way to select the variables for regression modelling. We didn't need to run multiple random forest models for remaining datasets – we just needed to run one model per dataset with the parameters: training set equal to 35%, mtry = 10, and ntrees = 300. Using these parameters, we identified the top twenty variables[7], ranked by importance, for each of the multiyear datasets. The Adult Important Variables are shown in Fig. 12.

[7] See Appendix A for a list of variable descriptions from NSDUH

Adult Variables and Importance Scores

2003-2005	
Variable	Importance
crkflag	3.99E-04
rdifher	3.99E-04
ircrkage	3.66E-04
cocneedl	2.47E-04
cocflag	1.89E-04
ircocage	1.61E-04
pcpflag	1.41E-04
irpcpage	1.25E-04
otdgnedl	9.41E-05
halflag	8.03E-05
txilalev	7.83E-05
lsdflag	7.60E-05
METHDON2	7.30E-05
MORPHIN2	7.11E-05
benzos	7.02E-05
irlsdage	6.17E-05
irtrnage	5.56E-05
trqflag	5.41E-05
MESC2	5.25E-05
PSILCY2	5.21E-05

2006-2008	
Variable	Importance
ircrkage	4.23E-04
cocneedl	3.46E-04
crkflag	3.23E-04
rdifher	1.74E-04
ircocage	1.67E-04
cocflag	1.54E-04
MORPHIN2	1.40E-04
METHDON2	1.40E-04
pcpflag	1.22E-04
lsdflag	1.17E-04
irpcpage	1.16E-04
txilalev	1.04E-04
halflag	1.02E-04
irlsdage	9.57E-05
otdgnedl	9.37E-05
OXYCODP2	6.94E-05
MESC2	6.67E-05
irhalage	5.92E-05
iroxyage	5.47E-05
oxyflag	5.38E-05

2009-2011	
Variable	Importance
ircrkage	3.75E-04
crkflag	3.54E-04
cocneedl	2.35E-04
cocflag	2.00E-04
METHDON2	1.97E-04
ircocage	1.89E-04
otdgnedl	1.85E-04
MORPHIN2	1.41E-04
irlsdage	1.36E-04
oxyflag	1.35E-04
lsdflag	1.31E-04
rdifher	1.24E-04
iroxyage	1.15E-04
irpcpage	1.11E-04
irhalage	1.11E-04
txilalev	9.57E-05
halflag	8.28E-05
pcpflag	7.54E-05
irecsage	7.49E-05
PSILCY2	6.94E-05

2012-2014	
Variable	Importance
ircrkage	4.46E-04
crkflag	4.32E-04
cocflag	2.92E-04
cocneedl	2.87E-04
otdgnedl	2.74E-04
rdifher	2.65E-04
ircocage	2.61E-04
METHDON2	1.81E-04
OXYCODP2	1.59E-04
iroxyage	1.49E-04
othanl	1.40E-04
MORPHIN2	1.35E-04
txilalev	1.28E-04
oxyflag	1.24E-04
DILAUD2	1.08E-04
lsdflag	9.08E-05
irhalage	8.72E-05
irlsdage	8.56E-05
pcpflag	8.53E-05
irtrnage	8.49E-05

Figure 12: Adult Important Variables by Time Period

The model fit results for the multiple datasets is given in Fig. 13.

			Tuning Parameters								
Trial	Dataset	Observations	Trainset Size (%)	mtry	ntree	cforest() Run Time (mins)	Prediction Threshold	Training Set Kappa	Test Set Kappa	AUROC	TS Kappa * AUROC
8	Adult 2003-2005	111,561	35%	10	300	6.04	0.2	0.6063	0.4884	0.9707	0.4741
9	Adult 2006-2008	111,561	35%	10	300	5.89	0.2	0.6020	0.4810	0.9719	0.4675
10	Adult 2009-2011	111,561	35%	10	300	6.10	0.2	0.6231	0.4570	0.9685	0.4426
11	Adult 2012-2014	111,561	35%	10	300	6.22	0.2	0.6522	0.5201	0.9740	0.5066

Figure 13: Tuning Parameters for Trials 8-11

Results from Random Forest Modelling: Youths

Selecting variables from the youth datasets was more difficult than from the adult datasets. This is because the number of heroin users among youth NSDUH survey respondents was very low. In 2003-2005, there were only 186 youth heroin users (out of 55,176 respondents) compared to 2,008 adult heroin users (out of 111,561 respondents). Due to the low number of youth users, we could not get meaningful results when we split the data between training and test sets, even when we increased the training set size to 50%. As seen in Fig. 14, the combination of κ and AUROC never went above 0.13, and the κ for the training set never surpassed 0.242. We had to forego splitting the youth dataset into training and test components to have enough users to build a meaningful model. When we did so, we obtained better values of between 0.4454 and 0.6301 for the model. As with the adult runs of PCA(), the youth runs would not complete prediction on test sets larger than 20,000 observations.

			Tuning Parameters								
Trial	Dataset	Observations	Trainset Size (%)	mtry	ntree	cforest() Run Time (mins)	Prediction Threshold	Training Set Kappa	Test Set Kappa	AUROC	TS Kappa * AUROC
12	Youth 2003-2005	55,176	35%	10	300	1.14	0.2	0.0407	0.0000	0.9773	0.0000
13	Youth 2006-2008	55,176	35%	10	300	59.81	0.2	0.2409	0.0725	0.9831	0.0713
14	Youth 2003-2005	55,176	50%	10	300	1.18	0.2	0.2420	0.1328	0.9744	0.1294
15	Youth 2003-2005	55,176	100%	20	300	1.40	0.2	N/A	0.6079	0.9917	0.6029
16	Youth 2006-2008	55,176	100%	20	300	1.17	0.2	N/A	0.6341	0.9938	0.6301
17	Youth 2009-2011	55,176	100%	20	300	1.20	0.2	N/A	0.5997	0.9920	0.5949
18	Youth 2011-2014	55,176	100%	20	300	1.02	0.2	N/A	0.4490	0.9920	0.4454

Figure 14: Tuning Parameters for Trials 12-18

Using training data of 100% of the dataset and PCA() tuning parameters of mtry = 20 and ntree = 300, we identified the top twenty variables, ranked by importance, for the youth datasets (See Fig 15).

Youth Variables and Importance Scores

2003-2005	
Variable	Importance
ircocage	2.18E-04
cocflag	2.03E-04
ircrkage	1.27E-04
crkflag	1.19E-04
irlsdage	1.12E-04
irecsage	1.06E-04
cocyr	8.52E-05
irhalage	8.30E-05
halflag	7.17E-05
rdifher	7.14E-05
ecsflag	5.28E-05
lsdflag	5.06E-05
irmthage	5.03E-05
METHDES2	4.46E-05
crkyr	3.96E-05
halyr	3.90E-05
grskhreg	3.73E-05
irpcpage	2.87E-05
cpnmthfg	2.84E-05
grskhtry	2.81E-05

2006-2008	
Variable	Importance
cocflag	1.76E-04
ircocage	1.36E-04
ircocfy	1.07E-04
pcpflag	9.26E-05
crkflag	8.86E-05
irpcpage	8.45E-05
ecsflag	7.79E-05
irecsage	6.74E-05
ircrkage	6.45E-05
cocyr	5.85E-05
irlsdage	5.79E-05
lsdflag	5.44E-05
halflag	5.15E-05
rdifher	4.41E-05
PSILCY2	3.70E-05
irhalage	3.63E-05
ecsyr	3.63E-05
mthneedl	3.43E-05
crkyr	3.25E-05
pcpyr	3.16E-05

2009-2011	
Variable	Importance
cocflag	1.93E-04
ircocage	1.47E-04
rdifher	1.14E-04
irecsage	1.02E-04
ecsflag	9.64E-05
halflag	7.22E-05
lsdflag	6.54E-05
irhalage	6.51E-05
irmthage	6.31E-05
cpnstmfg	6.25E-05
irlsdage	5.46E-05
ircrkage	5.45E-05
irstmage	5.45E-05
irtrnage	4.97E-05
PSILCY2	4.89E-05
benzos	4.34E-05
irpcpage	4.23E-05
MORPHIN2	4.16E-05
trqflag	3.67E-05
METHDES2	3.32E-05

2012-2014	
Variable	Importance
ircocage	1.19E-04
cocflag	8.45E-05
ircrkage	5.14E-05
irhalage	3.62E-05
crkflag	3.07E-05
mthneedl	2.97E-05
halflag	2.79E-05
rdifher	2.27E-05
ecsflag	2.19E-05
OXYCODP2	2.08E-05
halyr	1.88E-05
irpcpage	1.87E-05
irmthfy	1.76E-05
cpnmthyr	1.62E-05
ircocfy	1.41E-05
irhalfy	1.37E-05
grskhreg	1.37E-05
cocyr	1.33E-05
CODEINE2	1.30E-05
pcpflag	1.24E-05

Figure 145: Youth Important Variables by Time Period

The complete set of adult and youth variables, with their descriptions, is listed in Appendix A.

One Last Imputation

There were several composite variables that arose in the important variable lists. For example, the variable "halflag" is a flag that represents the use of LSD, PCP, ecstasy, psilocybin, and other hallucinogens. Leaving "halflag" in the list without modification makes drugs such as ecstasy, whose flag variable "ecsflag" is also an important variable, get double counted. To eliminate double counting, we created new imputed flags that separated out common flags from

the composite variables. In the case of "halflag", we created "otherhal", which represented the use of hallucinogens other than LSD, ecstasy, PCP, psilocybin, and mescaline. Similarly, we created the flags "otherstim" for stimulants other than methamphetamine, "otherpain" for pain killers other than morphine, and "onlyoxycod" for oxycodone derivatives other than OxyContin. If we had perfect knowledge of the dataset, we could have performed these splits prior to PCA and random forest modelling. However, one of the strengths of our approach is that it identifies factors worthy of close inspection from an otherwise unwieldy dataset.

Interpretation of Variable Selection

Although the NSDUH dataset contains a great many variables across many different aspects of a respondent's experiences and environment, the variables that are most important in identifying correlation to heroin use fall into six categories for adults, and seven for youths (See Fig. 16).

Variable Category
Drug Use Flags
Access to Heroin
Perceived Risk of Heroin Use
Age of First Use of Specific Drugs
Needle Use
Prior Substance Abuse Treatment
Frequency of Use of Specific Drugs (Youth Only)

Figure 16: Variable Categories

The first and largest category of important variables is the set of flags indicating whether a respondent has ever used a specific drug. Tobacco, alcohol, and marijuana did not appear in this list, even for youths, whose experience with a wide variety of drugs would be expected to be narrower than for adults. As shown in Fig. 17 below, we found that heroin users also used many of the drugs found in our most important variable list. However, only users of certain drugs had a high rate of heroin use. For example, over 45% of OxyContin and methadone users also used heroin, while less than 10% of hallucinogen and oxycodone (excluding OxyContin) users did.

The second category is access to heroin. Did the respondent find heroin easy to acquire? The third is perceived risk of heroin use. Did the user find heroin risky to use regularly? If not, did she find it risky to try?

The fourth set is age-of-first-use (AFU) variables for specific drugs. While the NSDUH does not include AFU variables for all drugs, it covers some commonly abused ones, including cocaine, crack, and methamphetamine.[8] Inclusion of AFU drugs in our study does not explain which drugs were used before heroin, but other studies have compared AFU values between

[8] It is interesting to note that AFU variables for tobacco, alcohol, and marijuana, which are often described as "gateway" drugs, were not selected by our methods.

heroin and other drugs (Jones, 2013). Our focus is instead limited to the relation between AFU for various drugs and any heroin use, whether it occurred before or after those drugs.

Drug	Total Users of Drug	Heroin Users Ever Used Drug	Heroin Users Never Used Drug	% of Heroin Users That Used Drug	% of Drug Users That Used Heroin
Methadone	4,927	2,566	6,710	27.66%	52.08%
Oxycontin	8,503	3,875	5,401	41.77%	45.57%
Crack	13,897	5,739	3,537	61.87%	41.30%
Morphine	7,778	3,168	6,108	34.15%	40.73%
PCP	9,304	3,080	6,196	33.20%	33.10%
Mescaline	10,252	2,740	6,536	29.54%	26.73%
Meth, Desoxy,	19,366	3,681	5,595	39.68%	19.01%
Codeine	18,940	3,559	5,717	38.37%	18.79%
Methamphetamine	25,150	4,397	4,879	47.40%	17.48%
Oxycodone Products	36,374	5,383	3,120	63.31%	14.80%
LSD	43,702	6,406	2,870	69.06%	14.66%
Ecstasy	40,876	5,359	3,917	57.77%	13.11%
Benzodiazeprine	46,881	6,006	3,270	64.75%	12.81%
Psilocybin	47,162	6,008	3,268	64.77%	12.74%
Cocaine	69,124	8,521	755	91.86%	12.33%
Tranquilizers	50,362	6,125	3,151	66.03%	12.16%
Stimulants	48,706	5,739	3,537	61.87%	11.78%
Hallucinogens	84,361	8,082	1,194	87.13%	9.58%
Oxycodone Only	21,080	1,508	7,768	16.26%	7.15%

Figure 17: Other Drug Use by Heroin Users

The fifth category was a set of variables that indicated that the respondent had used needles to inject cocaine, methamphetamine, or other drugs. The sixth category was a flag indicating if a respondent had prior treatment for substance abuse.

The seventh category of variables applies only to youths. Frequency of use variables were important across each block of years, but especially in the 2012-2014 dataset. These variables fell into two sub-categories. The first was how often a respondent used a given drug in the past year. The second was whether a respondent had used a given drug even once in the past year. These variables overlap – if the flag for use at least once is positive, the frequency of use flag must be at least one. This fact allowed us to consider only the variable for how often a respondent used the drug.

Model Construction

Given that the largest category of variables available to the model are those that are flags for specific drugs, we decided to insulate their effects by creating a two-phase model. The first phase of the model was to predict heroin usage as a function of individual drug use flags. This phase allows us to do two things. First, it enables us to identify the drugs that are most associated with heroin for a block of time. In this analysis we refer to these drugs as *critical drugs*. Second, it shows us how the strength of correlation between other drugs and heroin has evolved over time. The general form for this phase is:

$$heroin\ use_i = f\big(lin(use\ of\ drug\ j)\big)$$

$$i \in \{all\ respondents\}, \qquad j \in \{drugs\ identified\ by\ random\ forest\ model\}$$

The critical drugs identified in Phase One dictate the structure of Phase Two, which focuses on how a respondent used those critical drugs. Phase Two focuses only on use of the critical drugs and ignores use of non-critical drugs. It focuses on critical drugs by considering *how* a respondent used those drugs. In Phase Two, the behaviors we considered were: how many critical drugs did a respondent use, did the respondent use needles to administer drugs, what was the minimum age of first use for the critical drugs, had the respondent received treatment for substance abuse, what was the respondent's perceived risk of heroin use, how easy was it for the respondent to obtain heroin, and how frequently did the respondent use critical drugs (youths only). The general form of this phase is:

$$heroin\ use_k = f\big(lin(polyabuse, needle, min\ AFU, treatment, risk, access, frequency)\big)$$

$$k \in \{users\ of\ critical\ drugs\}$$

Our model will not be able to absolutely define heroin use as a linear combination of predictors. Because the NSDUH survey describes human behavior, it is inherently indeterminate. No combination of factors will absolutely dictate the actions of all respondents. Instead, the predictors combine to define the *probability* of heroin use. We consider the following definition:

$$\gamma_i = \text{heroin use by respondent i}$$

$$\gamma_i \in \{0,1\}$$

We assume the distribution of γ, a dichotomous value that indicates the respondent has used heroin, can be described by the Bernoulli function with mean μ:

$$\gamma_i \sim \text{Bernoulli}(\mu)$$

$$0 \leq \mu \leq 1$$

In this case, where γ_i is either 0 or 1, $\mu = P(\gamma_i = 1)$. The parameter μ is defined by a linear combination of the predictor variables we identified in prior sections. The function that maps the predictors to μ is known as the *inverse link function*. It must have an asymptote at 0 when the combination of predictors is increasingly negative, and it must have an asymptote at 1 when the combination of predictors is increasingly large. A commonly used inverse link function for this situation is the logistic function (Kruschke, 2015c):

$$y = \text{logistic}(x) = \frac{1}{1 + e^{-x}}$$

The inverse of the logistic function is the logit function:

$$logit(y) = log\left(\frac{x}{1-x}\right)$$

To link the linear combination of predictors, $lin(\boldsymbol{X})$, back to μ, we use the logit function:

$$logit(\mu) = log\left(\frac{\mu}{1-\mu}\right) = log\left(\frac{P(\gamma_i = 1)}{P(\gamma_i = 0)}\right) = \text{lin}(\boldsymbol{X})$$

The quantity $\frac{P(\gamma_i=1)}{P(\gamma_i=0)}$ is known as the *odds* that $\gamma_i = 1$. Note that unlike a probability, which ranges from 0 to 1, odds ranges from 0 to ∞. It is common to describe the change in likely outcome between two scenarios of factors by comparing their odds via an *odds ratio*:

$$OR_{x_j=1} = \frac{odds\left(\gamma_i = 1 \middle| x_j = 1\right)}{odds\left(\gamma_i = 1 \middle| x_j = 0\right)}$$

If $OR_{x_j=1} > OR_{x_k=1}$, we conclude that variable x_j more closely predicts a positive outcome for γ_i than does variable x_k.

We define our two-phase model by substituting its predictors for $\boldsymbol{X}$. Phase 1 thus becomes:

$$\text{PHASE 1:} \qquad \gamma_i \sim \text{Bernoulli}(\mu_1), \qquad \text{logit}(\mu_1) = \beta_{1_0} + \sum_{j \in D_1} \omega_j d_j \quad (\text{Eq. } 1)$$

and Phase 2 becomes:

$$\text{PHASE 2:} \qquad \gamma_k \sim Bernoulli(\mu_2), \qquad logit(\mu_2) = \beta_{2_0} + \sum_{m=1}^{7} \alpha_m g_m \, (\text{Eq. } 2)$$

where:

- $\beta_{1_0}, \beta_{2_0}, \boldsymbol{\omega}, \boldsymbol{\alpha}$, are constants
- $i \in \{all\ respondents\},\ k \in \{users\ of\ critical\ drugs\}$
- $d_j =$ any use of jth drug
- $D_1 = \{all\ drugs\}$
- g_1= number of drugs used among the critical drug set (polyabuse)
- g_2= needle use flag
- g_3 = minimum age of first use of a critical drug
- g_4 = prior treatment use flag
- g_5 = perceived risk of heroin use (2=no risk of regular use, 1=no risk of occasional use, 0=otherwise)
- g_6 = access to heroin (1=easy access, 0=otherwise)
- g_7 = frequency of use of cocaine, hallucinogens, and methamphetamine.[9]

[9] g_7 is used in the youth model only

Reducing Autocorrelation via Normalization

g_2 is an example of a dichotomous parameter – it can only have the values 1 or 0. g_1 is a range parameter – it can be equal to any integer between 1 and the number of critical drugs. We must normalize the range parameters so that their values are distributed such that the mean of a normalized range parameter becomes 0. This is done to reduce autocorrelation in MCMC analysis. Normalization is done via z-scoring:

$$z_x = \frac{x - \bar{x}}{s_x}$$

Where z_x is the normalized variable corresponding to x, and s_x is the standard deviation of x. When we normalize the range variables in our model, the linear function in the second phase becomes:

$$logit(\mu_2) = \zeta_0 + \zeta_1 z_{g_1} + \alpha_2 g_2 + \zeta_3 z_{g_3} + \alpha_4 g_4 + \zeta_5 z_{g_5} + \alpha_6 g_6 + \zeta_7 z_{g_7} \quad \text{(Eq. 3)}$$

We will run the second phase of the model using Eq. 3. However, we wish to be able to use the model on the original values of the range parameters, not on their normalized values. We can convert the normalized parameters $\zeta_0, \zeta_1, \zeta_3, \zeta_5, \zeta_7$ back to the non-normalized parameters as follows:

$$\beta_{2_0} = \zeta_0 - \frac{\zeta_1 \overline{g_1}}{s_{g_1}} - \frac{\zeta_3 \overline{g_3}}{s_{g_3}} - \frac{\zeta_5 \overline{g_5}}{s_{g_5}} - \frac{\zeta_7 \overline{g_7}}{s_{g_7}}$$

$$\alpha_1 = \frac{\zeta_1}{s_{g_1}}, \alpha_3 = \frac{\zeta_3}{s_{g_3}}, \alpha_5 = \frac{\zeta_5}{s_{g_5}}, \alpha_7 = \frac{\zeta_7}{s_{g_7}}$$

The set of parameters $\beta_{1_0}, \beta_{2_0}, \boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\varsigma}$ are not deterministic – our model cannot calculate with 100% certainty their values. Instead, each parameter is subject to a distribution of its own. The remainder of our analysis focused on the calculation of these distributions, which in turn determine the distribution of μ_1, μ_2, the parameters that determine the distribution of heroin users when given a set of values for the predictive variables from the dataset. In the following sections, we will describe our methods using Bayesian methods, solve the model, and interpret its outcome.

Bayes Rule

Each of the parameters of our model, $\beta_{1_0}, \beta_{2_0}, \boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\varsigma}$, has a distribution. If the model is accurate, then it will accurately predict the number of heroin users in the NSDUH dataset. Said another way, the correct distributions of our parameters will maximize the likelihood that we observe model predictions that match the NSDUH dataset.[10] Determining these distributions can be done via Bayesian analysis. We will briefly review Bayes Rule and describe its relationship to our model.

[10] The following discussion on Bayes Rule is drawn from (Kruschke, 2015a)

Bayes Rule is easily derived from the definition of conditional probabilities. Let θ represent a parameter, and let D represent a set of observed data. That definition of conditional probability is:

$$p(\theta|D) = \frac{p(\theta, D)}{p(D)}$$

which can be rewritten as:

$$p(\theta, D) = p(\theta|D)p(D)$$

It is also true that $p(\theta, D) = p(D|\theta)p(\theta)$. Combining these two equations, we see that:

$$p(\theta|D)p(D) = p(D|\theta)p(\theta)$$

Solving for $p(\theta|D)$ yields Bayes Rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

In Bayes Rule, $p(\theta|D)$ is known as the *posterior distribution* for the parameter θ. $p(D|\theta)$ is the likelihood of the data given a value for the parameter θ. $p(\theta)$ is the *prior distribution* for the parameter θ. $p(D)$ is known as the *evidence*. Bayes Rule says that the posterior distribution of a parameter when given a set of data is equal to our prior belief about the distribution of that parameter, multiplied by the likelihood of the data we see for that parameter, divided by the evidence. If we define the probability of the data as a conditional probability given the set of all possible values of the parameter, Bayes Rule becomes:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int d\theta^* p(D|\theta^*)p(\theta^*)}$$

where θ^* indicates the entire range of parameter values, distinct from a single parameter value under consideration.[11]

The Importance of the Prior Distribution

In Bayesian statistics, determining the prior distribution is critically important. The data that we see in an experiment will modify, not replace, our belief in the prior distribution. If our assumption of a prior distribution is inaccurate, the posterior distribution will most likely be inaccurate as well. Without any prior knowledge, we can set *uninformed* priors. An example of an uninformed prior is the normal distribution with a mean of zero. Most studies of medical data, because they are not Bayesian, are similar to a Bayesian analysis with uninformed priors. However, an uninformed prior can be invalid. In our case, we know from many studies that

[11] For the remainder of this discussion, we will consider Phase One of our model. The concepts described below apply equally to Phase Two.

OxyContin abuse is correlated with heroin use. We argue that we *must* take that knowledge into account with an informed prior. Our approach, because it allows for recognition of previous studies via informed priors, is a better way to study medical data than by simple null hypothesis testing.

Bayes Rule and the Model Parameters

Recall that in Phase 1, the probability that a respondent has used heroin is distributed as a Bernoulli trial with parameter μ_1, which in turn is a linear function of parameters $\boldsymbol{\beta_{1_0}}, \boldsymbol{\omega}$, each of which has their own distribution that we assume to be independent of the other parameters. We define the data for our model as the set of respondents' use of heroin: $\mathrm{D} = \{\gamma_i\},\ i \in \{\text{respondents}\}$, or $\mathrm{D} = \boldsymbol{\gamma}$. Using these definitions, we can use Bayes Rule to determine the parameters' distributions:

$$p(\mu_1|\boldsymbol{\gamma}) = \frac{p(\boldsymbol{\gamma}|\mu_1)p(\mu_1)}{p(\boldsymbol{\gamma})}$$

$$p\left(\boldsymbol{\beta_{1_0}}, \boldsymbol{\omega}|\boldsymbol{\gamma}\right) = \frac{p\left(\boldsymbol{\gamma}|\boldsymbol{\beta_{1_0}}, \boldsymbol{\omega}\right)p\left(\boldsymbol{\beta_{1_0}}, \boldsymbol{\omega}\right)}{p(\boldsymbol{\gamma})} \qquad \text{(Eq. 4)}$$

$$\mathrm{p}(\boldsymbol{\gamma}) = \int_{\beta_{1_0}} \int_{\boldsymbol{\omega}} d\boldsymbol{\beta_{1_0}}\, d\boldsymbol{\omega}\, p\left(\boldsymbol{\gamma}|\boldsymbol{\beta_{1_0}}, \boldsymbol{\omega}\right)p\left(\boldsymbol{\beta_{1_0}}, \boldsymbol{\omega}\right) \qquad \text{(Eq. 5)}$$

For our model, $p\left(\boldsymbol{\beta_{1_0}}, \boldsymbol{\omega}\right)$ is the set of prior distributions for our parameters. $p\left(\boldsymbol{\gamma}|\boldsymbol{\beta_{1_0}}, \boldsymbol{\omega}\right)$ is the likelihood of heroin use given a set of our parameters. $p(\gamma)$ is the evidence for our data, and $p\left(\boldsymbol{\beta_{1_0}}, \boldsymbol{\omega}|\boldsymbol{\gamma}\right)$ is the posterior distribution of our parameters. Finding the set of posterior distributions is the goal of our model.

Monte Carlo-Markov Chain Methods to Estimate Posterior Distributions

Evaluation of the compound integral for $p(\boldsymbol{\gamma})$ is extremely difficult. To determine the posterior distributions for the model parameters, we must use approximations using Monte Carlo-Markov Chain (MCMC) methods[12]. MCMC methods approximate the posterior distribution of a parameter (or set of parameters) by generating a sequence of values for the parameter. Values are added to the sequence in a random-walk manner dictated by a combination of the prior probability of the value and the likelihood of the data as a function of that prior value. Said another way, the MCMC methods select a set of values from the prior distributions that are likely to generate the data seen by the researcher.

[12] The following discussion is based upon (Kruschke, 2015b)

Gibbs Sampling

For each iteration of the MCMC random walk, we select one of the following parameters to consider for change: $\beta_{1_0}; \omega_i, i \in D_1$. We are therefore selecting one of $1 + |D_1|$ total variables to consider for change. To generate enough values of each parameter to estimate all posterior distributions, we iterate through each parameter in turn. Within an iteration we then generate a proposed value for the considered parameter based upon its distribution conditioned against the data and the fixed values of the other parameters. We then proceed to the next iteration by considering the next parameter. We stop the random walk when we reach *N*, the number of iterations dictated by the researcher. Using the parameters from our model, we describe Gibbs sampling as follows:

1) Define $\boldsymbol{\theta}$ as an ordered sequence of the parameters for our model, where $\theta_0 = \beta_0$, $\theta_1 = \omega_1$, $\theta_2 = \omega_2, \ldots, \theta_{1+|D_1|} = \omega_{|D_1|}$. Set $i = 0$.
2) Select a value for θ_i according to the distribution $p\left(\theta_i \middle| \{\theta_{j \neq i}\}, \boldsymbol{\gamma}\right)$ and add to the set $\theta_i = \{\theta_{i_k}\}, k \in \{0, \ldots, \frac{N}{1+|D_1|}\}$
3) Set $i = i + 1$
4) Repeat steps 2-3 until $i = N + 1$

Gibbs Sampling works because at each step we are generating a parameter value directly from the posterior conditional distribution for that parameter. The posterior conditional distribution is set by the combination of the likelihood of the data given the parameter value and the prior probability of that parameter value. The set of all values we generate for a particular parameter is thus a representation of the posterior distribution of that particular parameter.

In our analysis, we implemented Gibbs Sampling by using the JAGS software package, which stands for "Just Another Gibbs Sampler" (Plummer, 2003). JAGS automatically builds MCMC samplers for hierarchical Bayesian models. JAGS reduces the complexity of implementing Gibbs Sampling and allows the researcher to focus on more interesting areas of concern, such as selecting the prior distributions for model parameters. To speed up the performance of our model, we used the runjags package (Denwood, 2016), which optimizes JAGS by making use of multiple CPUs. We also used previously written routines to provide graphical output for the sampler and script templates for our models (Kruschke, 2015d).

Setting the Priors

Setting the priors was a critical step in our analysis. Unlike many studies of medical data, our model would build upon previous knowledge by incorporating the findings of other studies as prior probabilities for our model parameters. To set the prior distributions for our model parameters, we performed a literature search in which we looked for studies that found significant correlation between any member of our factor set and heroin use. Our factor set was the group of variables indicated by our random forest analysis as important. We sought to

maintain the concept of "prior" knowledge by considering only studies that did not use the NSDUH datasets.

We first considered a concept that has received a great deal of attention -- the link between nonmedical prescription opioid (NPO) abuse and heroin use. Progression from NPO abuse, particularly OxyContin, to heroin involves crossing thresholds of stigma associated with heroin as the cost of NPOs increases (Mars, Bourgois, Karandinos, Montero, & Ciccarone, 2014). One quantitative study found that 70% to 80% of heroin-dependent respondents used an NPO as their first abused opioid, and that heroin abuse rose as a new, abuse resistant formula of OxyContin was released (Cicero et al., 2014). Another study found that young intravenous drug users frequently initiated NPO use and transitioned to heroin. The same study found that heroin users initiated or continued NPO abuse, particularly that of OxyContin, for a variety of reasons, including easier access or lower price than heroin, to boost the effects of heroin use, and to avoid heroin withdrawal symptoms (Lankenau et al., 2012). From these prior studies, we can conclude that there is a *probably strong correlation* between OxyContin use and heroin. Furthermore, respondents in the studies frequently mention OxyContin, but don't cite other oxycodone formulations as regularly. For this reason, we *don't* assume a link between non-OxyContin oxycodone and heroin.

Unsurprisingly, easy access to heroin was found to be correlated to heroin use (Maher et al., 2007). Some studies found a link between low age-of-first-use (AFU) of various drugs and eventual heroin use but did not quantify a linear relationship between them (Kandel, et al., 1992; Pugatch et al., 2001). The prevalence of cocaine abuse among heroin users was high, and 50% of intravenous cocaine users were found to also use heroin (Leri, Bruneau, & Stewart, 2003). We also found studies that identified links between heroin and crack (Beswick et al., 2001; Mc Bride et al., 1992) and between heroin and dilaudid (McBride, 1980). Multiple studies indicated a strong correlation between intravenous use of other drugs and heroin use (Leri et al., 2003; Rhodes, Briggs, Kimber, Jones, & Holloway, 2007). For the remainder of the factors, we either found no evidence in the literature for a link to heroin use or inconclusive evidence. The table in Appendix B summarizes the findings of our search.

Translating the results of a literature search into probability distributions for priors is at the discretion of the researcher and is therefore subject to scrutiny. Fortunately, Greenland has studied this problem extensively and has provided guidelines for researchers who wish to apply Bayesian techniques to medical data (Greenland, 2000, 2001, 2006, 2007).[13]

The priors for a Bayesian analysis should reflect results from previous studies in order to seem reasonable or credible. If true frequency distributions exist and are known for parameters, those should be used for the prior distributions. And while they aren't exact, the use of estimated priors is still better than relying on the datasets in the current study alone. Greenland

[13] The following discussion on how to set priors summarizes Greenland's approach.

recommends thinking of priors as making bets, which are commonly expressed as odds ratios. An odds ratio of 1 represents even odds – we don't know whether a factor will increase or decrease the chance of a positive outcome. An odds ratio of less than 1 means that the presence of a factor will reduce the chance of a positive outcome, and an OR of greater than 1 means the factor will increase the chance.

Bets are uncertain things. Not only do we guess at an odds ratio, we have varying degrees of uncertainty regarding that guess. We can express this uncertainty like this: "We are 95% certain that the odds ratio for the correlation between LSD and heroin use is between ¼ and 4." As our certainty of an OR increases, the range between the high and low estimates of the OR shrinks. The more certain that we are that there is a positive effect of a factor, the greater the bottom and top of the range become.

Determining a shape for the probability distribution function (pdf) of a parameter is particularly difficult. Sometimes the attributes of a factor can help determine the pdf. For example, if a factor can only have a positive impact on an outcome, we can choose the beta distribution, which begins at 0. In the absence of such information, we should stick with a less informative prior, such as the normal distribution, which can be positive or negative. In our analysis, we had no guiding information or factor attributes that would point to a specific pdf for any prior. We assumed that all our priors were normally distributed.

Many of the studies we encountered in our literature search consisted of surveys of a small number of respondents. When compared to the scope of the NSDUH dataset, they were *extremely* small. The findings from these studies therefore provide us with what Greenland describes as *subjective priors*. In such cases, we can use three categories of factors commonly used by clinicians:

- Uncertain direction: mean OR of 1, 95% certainty of a range between ¼ and 4
- Probably positive: mean OR of 2, 95% certainty of a range between ½ and 8
- Probably strong: mean OR of 4, 95% certainty of a range between 1 and 16

As we mentioned earlier, a regression parameter is equal to the logarithm of an odds ratio. The mean value for a parameter is simply $\overline{\beta_i} = ln(OR_i)$. The 95% confidence interval for a parameter is equal to $\overline{\beta_i} \pm 1.96\sigma^{1/2}$, where σ is the variance of the parameter. Using these definitions, we can convert Greenland's subjective bets into probability distributions of the form Normal(mean, variance).

- Uncertain direction: Normal (0, 0.5)
- Probably positive: Normal (0.6931, 0.5)
- Probably negative: Normal (-0.6931, 0.5)
- Probably strongly positive: Normal (1.396, 0.5)
- Probably strongly negative: Normal (-1.396, 0.5)

We can now set the prior distributions for the parameters associated with the factors in our model (See Fig. 18). All other parameters were set with uninformative priors of Normal (0, 0.5).

Variable	Description	Model Parameter	z-score Normalized?	Subjective Prior	Prior Distribution: Normal (mean,
rdifher	Heroin fairly or very easy to obtain	alpha6	N	Probably Positive	Normal (0.6931, 0.5)
minafu	Minimum AFU for high risk drugs	zeta3	Y	Probably Negative	Normal (-0.6931, 0.5)
cocflag	Cocaine - ever used	omega2	N	Probably Positive	Normal (0.6931, 0.5)
crkflag	Crack - ever used	omega6	N	Probably Positive	Normal (0.6931, 0.5)
DILAUD2	Dilaudid - ever used	omega7	N	Probably Positive	Normal (0.6931, 0.5)
oxyflag	OxyContin - ever used	omega16	N	Probably Strongly Positive	Normal (1.386, 0.5)
onlyoxycod	Oxycodone products (excl OxyContin) - ever used	omega15	N	Probably Positive (YOUTH ONLY)	Normal (0.6931, 0.5)
needleuse	Ever used needle to inject any drug	alpha2	N	Probably Strongly Positive	Normal (1.386, 0.5)
polyabuse	Flag for multiple high risk drug use (equal to number of drugs)	zeta1	Y	Probably Positive	Normal (0.6931, 0.5)

Figure 18: Prior Distributions

Running the Gibbs Sampling Model

Phase One of Gibbs sampling involved considering nineteen parameters from datasets of up to 120,000 observations. Despite using runjags on a powerful virtual machine,[14] we had to limit ourselves to random samples of the model so that it would complete. Because heroin users represent a small portion of the respondents of the NSDUH dataset, we tried to use as large a sample as possible. We ran Phase One on the 2003-2005 adult dataset with three different sizes of sample, 25% of the total, 35%, and 50%,[15] to determine an appropriate sample size. We found that the 50% sample ran in between eight to twelve hours. We were unable to calculate DIC samples for the model, which limited our ability to compare different versions of Phase One. However, in Phase One we were focusing on identifying the *relative* strength of correlations between factors and heroin use, and not a definitive prediction of heroin use. We therefore opted to focus on the largest sample, the 50% sample, and concede the calculation of DIC samples in our analysis. For consistency, we chose to also use 50% of the adult datasets for Phase Two of the model. Because the number of youth heroin users is small, and because the youth dataset is smaller, we used 70% of the dataset for Phases One and Two. This allowed us to maximize the number of observations in the model while still preserving a holdout sample for model validation.

When conducting a Bayesian analysis, we can reject the null hypothesis when a value for a parameter that corresponds to the null is outside of the highest density interval (HDI). The HDI commonly consists of 95% of the distribution of a parameter. In our model, we have assumed that all the parameters are normally distributed. If the 95% interval of a parameter's distribution includes 0, the null is within the HDI. In Phase One, we considered nineteen drugs

[14] An 8 CPU Linux machine with 16GB of RAM running on Microsoft Azure
[15] We were unable to get the model to complete on 100% of the dataset. Nor were we able to compute DIC samples for the 50% sample – the program ran over 24 hours before we stopped it.

for inclusion in the critical drug set. We dropped any drug d_i whose distribution of parameter ω_i was found to include 0 in the HDI from the list. In Phase Two, any factor g_i whose distribution of parameter α_i included 0 was found to not correlate with heroin use. In both Phases, we sought to find non-null parameters, the relative strength of the parameters, and changes in these outcomes over time.

A Note on Odds Ratios

The parameter mean values and HDIs from Gibbs sampling directly indicate the strength of the factors in our model, but we have calculated odds ratios (OR) for consistency with other studies. This is because medical and epidemiological studies often express the strength of predictive factors this way. It is important to point out that an odds ratio is exactly that – a *ratio*. We must have something in the numerator *and* the denominator. In Phase One, the OR is simple: the numerator indicates the odds of heroin use when the respondent has also used drug d_i. The denominator indicates the odds of heroin use when the respondent has not used any other drug $d_i, i \in D_1$:

$$OR_i = \frac{odds\left(\gamma = 1 \middle| d_i = 1, d_j = 0, i \neq j\right)}{odds(\gamma = 1 | d_i = 0)}$$

$$OR_i = \frac{e^{\beta_0 + \beta_i}}{e^{\beta_0}} = e^{\beta_i}$$

In Phase Two, things get more complicated. The factors are not just dichotomous flags, they include ranges: polyabuse of drugs ($g_i \geq 1$), minimum age of first use of a critical drug ($g_3 \leq 99$), and perceived risk of heroin abuse ($g_5 \in \{0,1,2\}$). Fortunately, proper consideration of the factors in the numerator and denominator of the OR will cancel out all but the factor of interest. To illustrate this, suppose we want to calculate the OR for age of first use. We begin by creating a base case for the denominator. We define this case as a construction of likely values for each of the factors: no needle use, heroin is difficult to get, the respondent perceives high risk with heroin use, he has never received prior treatment for substance abuse, he has only used one critical drug, and his first use of a critical drug occurred at age 19. In the numerator we place a case identical to the base case, except the age of first use occurred at value *a*, which is not equal to 19:

$$OR_2 = \frac{e^{\beta_0 + \beta_1 + \beta_5(a)}}{e^{\beta_0 + \beta_1 + \beta_5(19)}} = e^{\beta_5(a-19)}$$

In the case where $a = 20$, $OR_{2,a=20} = e^{\beta_5}$. When $a = 21$, $OR_{2,a=21} = e^{2\beta_5} = OR_{2,a=20}^2$. While we can calculate odds ratios for range values, we must apply them differently when considering values more than one unit from the base case for the factor.

In general, the method we used to illustrate odds ratios for a range variable can be extended to comparing any two combinations of factors. For example, suppose we wish to estimate how much more likely heroin use by respondent A is than heroin use by respondent B (See Fig. 19).

	Respondent A	
Description	A	B
Polyabuse (number of critical drugs used)	2	4
Ever used needle to inject any drug	N	Y
Minimum AFU for any critical drug	18	16
Ever received treatment for substance abuse	N	Y
Perceived risk of heroin use	High	High
Easy to access heroin	Y	Y

Figure 19: Comparing Two Potential Heroin Users

The odds ratio that compares respondent A to respondent B becomes:

$$OR_{A:B} = \frac{OR_A}{OR_B} = \frac{e^{\alpha_0+2\alpha_1-\alpha_3+\alpha_6}}{e^{\alpha_0+4\alpha_1+\alpha_2-3\alpha_3+\alpha_4+\alpha_6}} = \frac{1}{e^{2\alpha_1+\alpha_2-2\alpha_3+\alpha_4}}$$

Knowing the values for the parameters of the model thus gives us a very powerful means of estimating how different behaviors can increase the likelihood of corresponding heroin use.

Results

In this section, we discuss the results from the MCMC analysis for the Adult and Youth models. For each model, we will show which drugs were determined critical – those whose parameters had null values outside of the high-density interval (HDI). We will also show changes in critical drugs over time and in their relative impacts by comparing their odds ratios. We will then discuss how usage patterns and respondent attributes affect the likelihood of heroin use by users of critical drugs.

Adult Critical Drug Analysis (Model Phase One)

The parameter values, HDIs, and odds ratios for the adult model are shown in Appendix C. We will discuss the results from 2003-2005 to illustrate results interpretation for all the iterations of the analysis.

As we see in Fig. 20, 11 of the 19 drugs under consideration were critical. Interestingly, OxyContin, which our literature search indicated was closely linked to heroin use, was *not* critical. Despite that fact that we set a prior for OxyContin to Probably Strongly Positive, the data from this period shifted the distribution towards the null. The drug most closely correlated with heroin usage was cocaine. Cocaine users are almost 16 times more likely than non-cocaine

Adult Years 2003-2005: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-7.0472	0.1744	-7.4053	-6.7184	1.0000
Cocaine	2.7466	0.2048	2.3550	3.1583	15.5895
Crack	1.5521	0.1004	1.3564	1.7495	4.7213
Dilaudid	0.8575	0.2124	0.4398	1.2725	2.3573
Methadone	0.8470	0.1639	0.5251	1.1670	2.3326
PCP	0.7317	0.1109	0.5148	0.9493	2.0785
Tranquilizers	0.7285	0.2966	0.1408	1.3035	2.0721
Mescaline	0.6997	0.1126	0.4775	0.9201	2.0131
LSD	0.6831	0.1313	0.4246	0.9377	1.9800
Morphine	0.6023	0.2064	0.1959	1.0045	1.8262
Oxycodone (not OxyContin)	*0.3053*	*0.1375*	*-0.0360*	*0.5756*	*1.3571*
Ecstasy	0.2997	0.1073	0.0891	0.5090	1.3495
OxyContin	*0.2856*	*0.1640*	*-0.0349*	*0.6058*	*1.3306*
Other Pain Killers	*0.2570*	*0.1468*	*-0.0349*	*0.5425*	*1.2931*
Codeine	*0.0120*	*0.1356*	*-0.2543*	*0.2784*	*1.0121*
Psilocybin	*-0.0413*	*0.1189*	*-0.2738*	*0.1917*	*0.9595*
Other Stimulants	*-0.0705*	*0.1540*	*-0.3755*	*0.2264*	*0.9319*
Other Hallucinogens	*-0.1996*	*0.6916*	*-1.6655*	*1.0316*	*0.8190*
Benzos	*-0.2659*	*0.2930*	*-0.8311*	*0.3170*	*0.7665*
Methamphetamine	-0.3371	0.1153	-0.5638	-0.1120	0.7139

Figure 20: Adult Critical Drugs 2003-2005 (nulls in red italic)

users to use heroin. Crack was the second most strongly correlated drug. It is possible that these two drugs are most strongly linked to heroin by the action of "speedballing". In speedballing, users seek to reinforce the effects of both cocaine/crack and heroin by using them together (Leri et al., 2003; McBride et al., 1992; Rhodes et al., 2007). Dilaudid, the third most strongly linked drug, is less common in current literature, but was linked to heroin use in the past (McBride, 1980). It is possible that dual dilaudid/heroin users are older, and this could be determined through further analysis of the NSDUH data. Methadone is often a treatment for recovering heroin addicts, but the NSDUH study asks specifically about methadone abuse. When we see it here as a correlate to heroin use, we are seeing the link between illicit methadone use and heroin use.

One of the most powerful attributes of MCMC is the ability to examine the distributions of the model parameters. We will illustrate this attribute for this run of the model here. We compare the distributions for the parameters associated with cocaine (ω_2) and OxyContin (ω_{16}) in Fig. 21.

In the fourth quadrant charts, we see that the parameters follow the normal distribution, as they should since we set the priors that way. The HDI for ω_2 is well above 0, the null value.

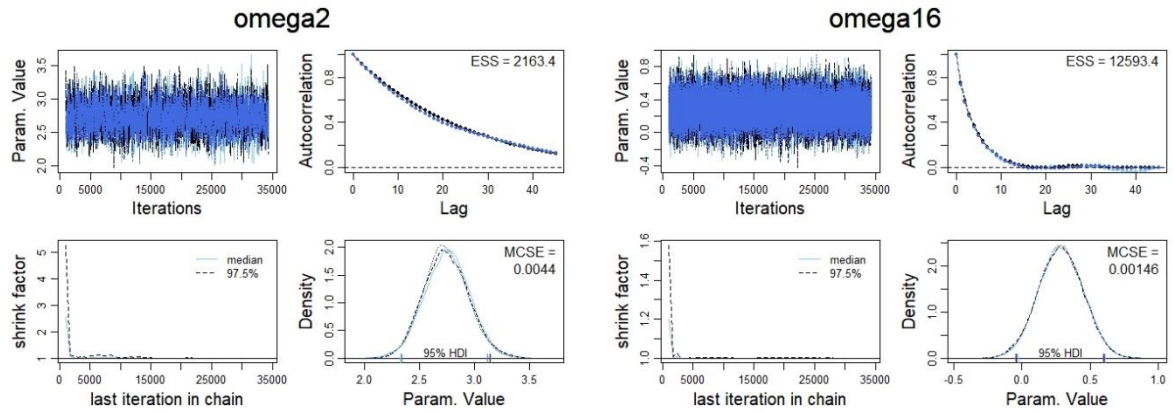

Figure 21: Sample Adult Phase One Parameter Distributions

In contrast the HDI for ω_{16} includes 0. The model execution wasn't perfect – autocorrelation of ω_2 approaches 0 yet doesn't do so quickly. Since the factors in the model are all dichotomous flags, normalizing them won't reduce autocorrelation. When we examine the cross-correlations for the model (Fig. 22), we see negative correlation between ω_2 and the intercept β_0. It is possible that normalizing the intercept could reduce autocorrelation, but due to the extraordinary runtimes for the model (over nine hours), we elected to accept the outcomes as they are.[16]

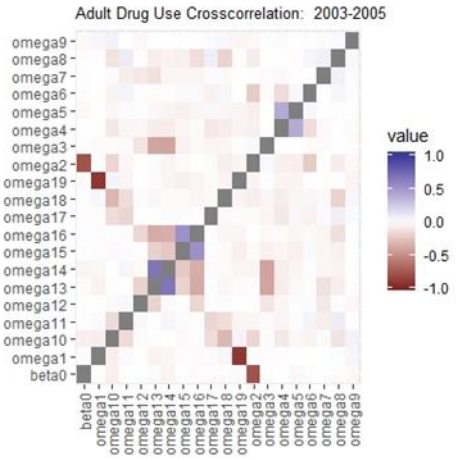

Figure 22: Adult Drug Use Crosscorrelation

To easily compare the changes of critical drugs across time, we ranked the drug flag parameters by odds ratio, with 1 being the most impactful parameter and 19 being the least (See Fig. 23). Again, drugs in red are those whose HDI included 0, making them null. Cocaine remains the drug with the highest odds ratios across all four time periods, and crack remains the

[16] We also see relationships between oxycodone (no OxyContin) and OxyContin. This could be due to use of multiple drugs in this category by users.

second highest. We also see consistency among most of the low odds-ratio drugs – psilocybin and other hallucinogens are not positively associated with heroin use in any time period.

Adult Ranking and Odds Ratios of Drug Use Flags (Nulls in *Red Italic*)

	2003-2005		2006-2008		2009-2011		2012-2014	
Drug	Rank	OR	Rank	OR	Rank	OR	Rank	OR
Cocaine	1	15.5895	1	11.8626	1	14.3677	1	11.3874
Crack	2	4.7213	2	4.3120	2	4.3479	2	3.7005
Dilaudid	3	2.3573	4	2.4495	5	1.9429	5	2.1821
Methadone	4	2.3326	8	1.6352	4	2.0846	7	1.7290
PCP	5	2.0785	3	2.6355	3	2.1977	4	2.5568
Tranquilizers	6	2.0721	*18*	*0.9656*	*11*	*1.3386*	*15*	*0.9161*
Mescaline	7	2.0131	9	1.5572	7	1.8197	11	1.5315
LSD	8	1.9800	5	1.8669	10	1.4814	12	1.4325
Morphine	9	1.8262	6	1.8551	9	1.7128	6	2.1604
Oxycodone (not OxyContin)	*10*	*1.3571*	12	1.3383	*14*	*1.0866*	8	1.7146
Ecstasy	11	1.3495	11	1.4146	8	1.7288	10	1.5827
OxyContin	*12*	*1.3306*	7	1.7323	6	1.8538	3	2.6672
Other Pain Killers	*13*	*1.2931*	*14*	*1.0971*	*15*	*1.0820*	*14*	*1.0822*
Codeine	*14*	*1.0121*	*19*	*0.9549*	*17*	*1.0427*	18	0.7766
Psilocybin	*15*	*0.9595*	*16*	*1.0246*	*19*	*0.9363*	*16*	*0.8961*
Other Stimulants	*16*	*0.9319*	*17*	*0.9949*	12	1.2570	*17*	*0.8649*
Other Hallucinogens	*17*	*0.8190*	*10*	*1.5175*	*16*	*1.0686*	*19*	*0.4298*
Benzos	*18*	*0.7665*	*13*	*1.3366*	*18*	*1.0130*	*9*	*1.6866*
Methamphetamine	19	0.7139	*15*	*1.0367*	*13*	*1.1048*	13	1.3021

Figure 23: Adult Drug Use Ranks and Odds Ratios (nulls in red italic)

Note that we see a marked increase in the odds ratio for OxyContin. In 2006-2008 it ceases to be null, and by 2012-2014 OxyContin has the third highest OR. Because cocaine and crack can both be associated with simultaneous heroin use, by 2012-2014 OxyContin has become the independent drug with the highest association with heroin. This finding reinforces what we found in the literature – non-medical prescription opioid use can lead to heroin abuse. OxyContin abusers are over 2.6 times as likely to use heroin as non-abusers. Abuse of non-OxyContin oxycodone climb to 1.7 times as likely to abuse heroin as non-abuse. PCP had consistently high odds ratios across all years despite a lack of prior evidence from the literature.

Adult Usage Pattern Analysis (Model Phase Two)

The parameter values, HDIs, and odds ratios for the usage pattern model are shown in Appendix D. Again, we will discuss the results from 2003-2005 (Fig. 24) to illustrate results interpretation for all the iterations of the analysis. Recall that for Phase Two, the respondent set is restricted to users of the critical (non-null) drugs identified in Phase One.

Adult Years 2003-2005: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-5.5831	0.2498	-6.0731	-5.0935	1.0000
Needle Use	1.4793	0.1031	1.2776	1.6819	4.3898
Access to Heroin	1.3083	0.0816	1.1487	1.4676	3.7000
Prior Treatment	0.6782	0.0827	0.5167	0.8399	1.9703
Perceived Risk	0.6122	0.0576	0.4990	0.7252	1.8444
Polyabuse	0.5105	0.0207	0.4701	0.5514	1.6661
Age of First Use	*-0.0037*	*0.0113*	*-0.0261*	*0.0180*	*0.9963*

Figure 24: Adult Usage Pattern Factors (nulls in red italic)

Despite having a Probably Negative prior, the Age of First Use (AFU), which is the minimum age that a respondent used any critical drug, is null. In fact, the parameter for AFU, α_3, is nearly centered on zero. In contrast, the correlation between needle use and heroin use is strongly positive. The odds ratio for the needle use parameter, α_2, is 4.390. The diagnostic plots in Fig. 25 show that the HDI for α_2 does not include zero, while the HDI for α_3 does.

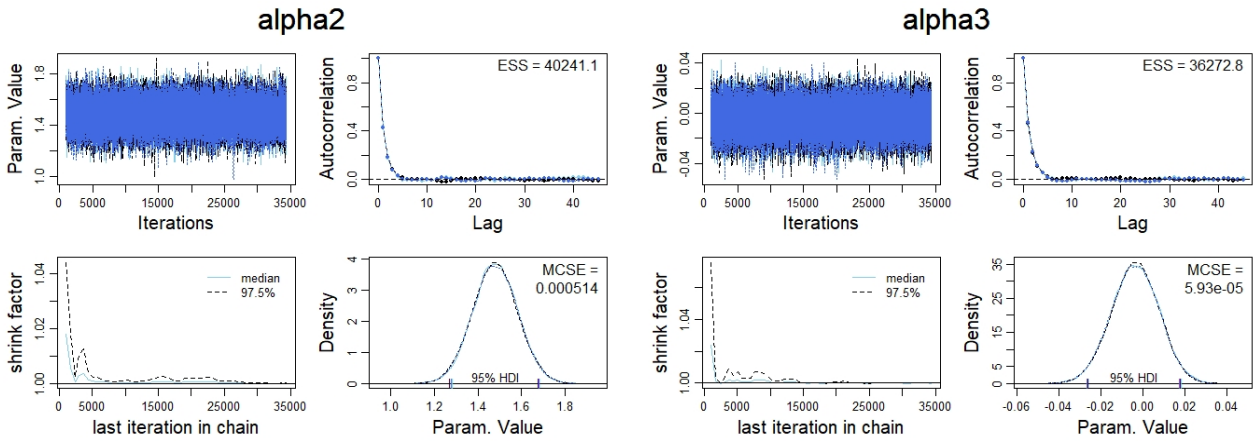

Figure 25: Sample Adult Phase Two Parameter Distributions

As shown in Fig. 26, the rankings for the usage pattern parameters are virtually unchanged over time. Needle use, access to heroin, and prior substance abuse treatment strongly affect the probability of heroin use. The parameter for AFU remains null for two time periods and then becomes correlated, but only slightly so, with heroin use.

Adult Ranking and Odds Ratios of Usage Pattern Flags (Nulls in *Red Italic*)

	2003-2005		2006-2008		2009-2011		2012-2014	
Drug	Rank	OR	Rank	OR	Rank	OR	Rank	OR
Needle Use	1	4.3898	1	5.9625	1	5.8058	1	4.8345
Access to Heroin	2	3.7000	2	2.9607	2	2.8875	2	3.3878
Prior Treatment	3	1.9703	3	2.0161	4	1.8925	3	2.2835
Perceived Risk	4	1.8444	4	1.9365	3	1.7977	4	1.9691
Polyabuse	5	1.6661	5	1.7934	5	1.7704	5	1.6172
Age of First Use	*6*	*0.9963*	*6*	*1.0001*	6	0.9728	6	0.9699

Figure 26: Adult Usage Pattern Ranks and Odds Ratios

The value of the AFU parameter for 2012-2014 is $\alpha_3 = -0.0306$. Consider two respondents from that time period. Respondent B began critical drug use at the age of 19, the average for all respondents. Respondent A began critical drug use at the age of 16. As we described above, we can see how more likely Respondent A is than B to use heroin by comparing their odds ratios:

$$OR_{A:B} = \frac{OR_A}{OR_B} = \frac{e^{\alpha_3(16)}}{e^{\alpha_3(19)}} = \frac{1}{e^{\alpha_3(3)}} = \frac{1}{e^{(-0.0306)(3)}} = 1.096$$

The odds that Respondent A uses heroin are less than 10% higher than that of Respondent B. We can thus drop the AFU variable from our model. This greatly simplifies the Phase Two model and makes it easier to apply in practice. In order to include AFU in the Phase Two model, we had to restrict the dataset to only those respondents who had used a critical drug. By removing AFU, we can once again apply the model to the entire dataset. For example, if we encounter a respondent who has used three critical drugs (polyabuse = 3), a model without AFU allows us to compare their odds of heroin use to a non-user or to a user of a non-critical drug such as marijuana. When we remove AFU, the Adult Revised Phase Two Model becomes:

$$heroin\ use_j = f\big(lin(polyabuse, needle, treatment, risk, access)\big)$$

$$j \in \{all\ respondents\}$$

$$\gamma_j \sim Bernoulli(\mu_2), \qquad logit(\mu_2) = \zeta_0 + \zeta_1 z_{g_1} + \alpha_2 g_2 + \alpha_4 g_4 + \zeta_5 z_{g_5} + \alpha_6 g_6 \quad (Eq.4)$$

$$\beta_{2_0} = \zeta_0 - \frac{\zeta_1 \overline{g_1}}{s_{g_1}} - \frac{\zeta_5 \overline{g_5}}{s_{g_5}}$$

$$\alpha_1 = \frac{\zeta_1}{s_{g_1}}, \alpha_5 = \frac{\zeta_5}{s_{g_5}}$$

With the removal of AFU, none of the remaining predictors are null. When we examine the parameter values for the 2003-2005 dataset in Fig. 27, we see that the values for the predictors increase slightly as well.

Revised Adult Years 2003-2005: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.6272	0.9081	-6.8090	-6.4524	1.0000
Needle Use	1.5487	0.1104	1.3301	1.7633	4.7053
Access to Heroin	1.2778	0.0815	1.1181	1.4374	3.5887
Prior Treatment	0.8380	0.0869	0.6670	1.0077	2.3117
Polyabuse	0.6830	0.0166	0.6508	0.7160	1.9798
Perceived Risk	0.6261	0.0586	0.5104	0.7405	1.8703

Figure 27: Revised Adult Usage Pattern Factors

In Fig. 28, we again see stability in the rankings of the usage pattern factors. Needle Use and Access to Heroin remain the strongest correlates with heroin usage. In the 2012-2014 dataset, needle users were 5.61 times as likely as non-needle users to also use heroin. Access to heroin also greatly increases the odds of heroin use – a respondent with easy heroin access is 3.93 times more likely to use than somebody without access.

Revised Adult Ranking and Odds Ratios of Usage Pattern Flags (Nulls in *Red Italic*)

	2003-2005		2006-2008		2009-2011		2012-2014	
Drug	Rank	OR	Rank	OR	Rank	OR	Rank	OR
Needle Use	1	4.7053	1	6.5666	1	5.5367	1	5.6062
Access to Heroin	2	3.5887	2	2.9910	2	3.0600	2	3.9318
Prior Treatment	3	2.3117	3	2.2129	4	2.1417	3	2.6977
Polyabuse	4	1.9798	4	2.1285	3	2.1572	5	1.8699
Perceived Risk	5	1.8703	5	1.7918	5	1.7522	4	2.0555

Figure 28: Revised Adult Usage Pattern Ranks and Odds Ratios

Youth Critical Drug Use Analysis (Model Phase One)

The parameters for the youth (respondents under the age of 18) critical drug analysis are shown in Appendix F. As we interpret this data, it is important to keep in mind that there were far less youth heroin users than adults. As a result, we should expect more variability in our results. We can see that there are significantly less critical drugs for youths than for adults. While we aren't sure of the reason for this, we can guess that one explanation would be that youths simply haven't had as much time as adults to use many drugs. In 2003-2005, only five of the 19 candidate drugs were critical (Fig. 29).

Youth Years 2003-2005: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-7.1958	0.2175	-7.6401	-6.7917	1.0000
Cocaine	2.9517	0.3707	2.2217	3.6809	19.1381
PCP	1.3126	0.3548	0.6150	2.0048	3.7158
Methamphetamine	1.1352	0.3489	0.4528	1.8229	3.1118
OxyContin	1.1166	0.4224	0.3055	1.9593	3.0543
Morphine	*0.8977*	*0.5028*	*-0.1112*	*1.8628*	*2.4539*
Ecstasy	0.7971	0.3248	0.1603	1.4343	2.2191
Other Stimulants	*0.3301*	*0.4183*	*-0.5112*	*1.1344*	*1.3911*
Mescaline	*0.2928*	*0.4665*	*-0.6312*	*1.1979*	*1.3401*
Oxycodone (not OxyContin)	*0.2716*	*0.4662*	*-0.6610*	*1.1681*	*1.3121*
Psilocybin	*0.2359*	*0.3480*	*-0.4526*	*0.9176*	*1.2660*
Other Pain Killers	*0.2256*	*0.4171*	*-0.6097*	*1.0290*	*1.2531*
Crack	*0.1750*	*0.3622*	*-0.5429*	*0.8757*	*1.1912*
Other Hallucinogens	*0.1084*	*0.9102*	*-1.8566*	*1.7172*	*1.1145*
LSD	*0.0649*	*0.3666*	*-0.6606*	*0.7794*	*1.0671*
Methadone	*0.0205*	*0.4225*	*-0.8240*	*0.8317*	*1.0207*
Tranquilizers	*-0.2946*	*0.7654*	*-1.8808*	*1.1179*	*0.7448*
Benzos	*-0.3027*	*0.7717*	*-1.7337*	*1.2857*	*0.7388*
Codeine	*-0.3049*	*0.3687*	*-1.0286*	*0.4126*	*0.7372*
Dilaudid	*-0.4328*	*0.8450*	*-2.1814*	*1.1406*	*0.6487*

Figure 29: Youth Critical Drugs 2003-2005 (nulls in red italic)

When we look at the rankings of critical drugs over time (Fig. 30), we again see a dramatic rise in the correlation strength of OxyContin. In 2003-2005, it was a critical drug, but in the next two periods, its correlation with heroin use became negligible, only to rise again in 2012-2014. It is possible that the relatively high ranking of OxyContin in 2003-2005 is a function of the small youth heroin user cohort. The subsequent time periods mirror the rise in importance of OxyContin we saw in the adult datasets. We also see a rise in the importance of oxycodone and crack across the time periods.

Youth Ranking and Odds Ratios of Drug Use Flags (Nulls in *Red Italic*)

	2003-2005		2006-2008		2009-2011		2012-2014	
Drug	Rank	OR	Rank	OR	Rank	OR	Rank	OR
Cocaine	1	19.1381	1	14.8626	1	5.2059	1	11.4210
PCP	2	3.7158	8	2.1782	*10*	*1.8406*	*12*	*1.2916*
Methamphetamine	3	3.1118	*11*	*1.6281*	6	2.5710	*10*	*1.3963*
OxyContin	4	3.0543	*13*	*1.3688*	*17*	*0.7471*	5	2.8702
Morphine	*5*	*2.4539*	*9*	*2.0831*	2	4.5150	*8*	*1.7896*
Ecstasy	6	2.2191	3	2.6789	4	3.1607	4	3.1125
Other Stimulants	*7*	*1.3911*	*16*	*0.8555*	*15*	*0.9731*	*14*	*1.0491*
Mescaline	*8*	*1.3401*	*7*	*2.2746*	*13*	*1.2050*	*6*	*2.8601*
Oxycodone (not OxyContin)	*9*	*1.3121*	*12*	*1.3721*	*18*	*0.6410*	3	3.3133
Psilocybin	*10*	*1.2660*	*15*	*0.9894*	7	2.3011	*15*	*0.9525*
Other Pain Killers	*11*	*1.2531*	5	2.3475	5	2.7638	*16*	*0.8602*
Crack	*12*	*1.1912*	2	3.3380	9	1.9897	2	3.7705
Other Hallucinogens	*13*	*1.1145*	*6*	*2.3057*	*11*	*1.7198*	*7*	*2.1039*
LSD	*14*	*1.0671*	4	2.6735	*12*	*1.6167*	*13*	*1.2229*
Methadone	*15*	*1.0207*	*17*	*0.6438*	8	2.1128	*18*	*0.8051*
Tranquilizers	*16*	*0.7448*	*18*	*0.6301*	3	3.7040	*9*	*1.5651*
Benzos	*17*	*0.7388*	*10*	*1.9758*	*19*	*0.4688*	*17*	*0.8215*
Codeine	*18*	*0.7372*	*14*	*1.0087*	*16*	*0.8160*	*11*	*1.3521*
Dilaudid	*19*	*0.6487*	*19*	*0.4518*	*14*	*1.0259*	*19*	*0.6183*

Figure 30: Youth Drug Use Ranks and Odds Ratios

Youth Usage Pattern Analysis (Model Phase Two)

The parameter values, HDIs, and odds ratios for the youth usage pattern model are shown in Appendix G. As with the adult analysis we restricted our Phase Two analysis to youths who had used any of the critical drugs identified in Phase One. We then modelled the probability of heroin use based upon polyabuse of critical drugs, needle use, minimum age of first use of any critical drug, prior substance abuse treatment, perceived risk of heroin use, access to heroin, and maximum frequency of use of any critical drug. Fig. 31 shows the results of this analysis for the time period 2003-2005, and Fig. 32 shows the relative ranks of the factors across all time periods.

Youth Years 2003-2005: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-4.3374	1.0489	-6.4709	-2.3591	1.0000
Access to Heroin	1.0172	0.2775	0.4780	1.5611	2.7655
Needle Use	0.9897	0.4158	0.1661	1.7947	2.6904
Polyabuse	0.9063	0.1235	0.6657	1.1504	2.4753
Perceived Risk	0.6158	0.1622	0.2977	0.9376	1.8511
Prior Treatment	*0.0910*	*0.3049*	*-0.5165*	*0.6763*	*1.0953*
Frequency of Use	*0.0018*	*0.0017*	*-0.0016*	*0.0051*	*1.0018*
Age of First Use	*-0.1046*	*0.0678*	*-0.2339*	*0.0310*	*0.9007*

Figure 31: Youth Usage Pattern Factors (nulls in red italic)

Youth Ranking and Odds Ratios of Usage Pattern Flags (Nulls in *Red Italic*)

	2003-2005		2006-2008		2009-2011		2012-2014	
Drug	Rank	OR	Rank	OR	Rank	OR	Rank	OR
Access to Heroin	1	2.7655	2	3.5766	2	4.7070	2	3.9701
Needle Use	2	2.6904	1	12.2177	1	5.4456	1	7.2955
Polyabuse	3	2.4753	3	2.3531	3	1.8898	3	2.1975
Perceived Risk	4	1.8511	5	1.5766	4	1.5604	5	1.8242
Prior Treatment	*5*	*1.0953*	4	1.7719	*5*	*1.4004*	4	2.0739
Frequency of Use	*6*	*1.0018*	*7*	*1.0003*	*6*	*1.0007*	7	1.0071
Age of First Use	*7*	*0.9007*	*6*	*1.0350*	*7*	*0.9369*	*6*	*1.0606*

Figure 32: Youth Usage Pattern Ranks and Odds Ratios

For youths, Access to Heroin and Needle Use are the two most important variables across all time periods. Frequency of Use and Age of First Use are the two least important variables, and both are null for three of the four time periods. For this reason, we decided to revise the Phase Two youth model by removing these two variables from consideration. This left us with the same model that we saw for the Adult Revised Phase Two Model (Eq. 4). The youth revised model results are shown in Appendix H. Fig. 33 shows a sample table from 2003-2005, and Fig. 34 shows the revised ranks and odds ratios across all time periods.

Revised Youth Years 2003-2005: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.7899	0.1755	-7.1408	-6.4543	1.0000
Needle Use	3.6752	0.3042	3.0694	4.2628	39.4577
Prior Treatment	2.3401	0.2181	1.9036	2.7585	10.3827
Access to Heroin	1.8479	0.1783	1.4984	2.1978	6.3467
Perceived Risk	0.4660	0.1060	0.2577	0.6733	1.5935
Polyabuse	0.0723	0.0234	0.0279	0.1193	1.0750

Figure 33: Revised Youth Usage Pattern Factors

Revised Youth Ranking and Odds Ratios of Usage Pattern Flags (Nulls in *Red Italic*)

	2003-2005		2006-2008		2009-2011		2012-2014	
Drug	Rank	OR	Rank	OR	Rank	OR	Rank	OR
Needle Use	1	39.4577	1	13.1143	2	4.5272	4	2.8826
Prior Treatment	2	10.3827	3	3.7468	4	2.0740	3	3.3505
Access to Heroin	3	6.3467	2	4.1153	1	5.2925	2	4.4611
Perceived Risk	4	1.5935	5	1.7179	5	1.4592	5	1.9521
Polyabuse	5	1.0750	4	3.6136	3	2.4963	1	4.9744

Figure 34: Revised Youth Usage Pattern Ranks and Odds Ratios

Interestingly, we see a significant shift in the ranks of the odds ratios of the youth usage pattern factors. We see Needle Use drop steadily from period to period in importance and we see polyabuse grow in importance. Again, the small datasets could explain some of this variability, but the fact that the changes in these two variables occur consistently across all periods lead us to believe that there is indeed a consistent usage pattern change for the youth respondents. We also see that the top two factors don't dominate the results for 2012-2014, where they did in 2003-2005. The number of youth heroin users in the 2012-2014 dataset was 101, which is much less than 186, the number in the 2003-2005 dataset. It is possible that as the number of heroin users increases in the youth dataset, the usage pattern factor ranks approach those of the adult datasets, where needle use is a dominant variable.

Model Validation

We focused our validation efforts on the latest dataset, 2012-2014. We elected to use this dataset because our goal is to predict future heroin use, and future use is more likely to follow the patterns we see in our most recent dataset. For both the adult and youth datasets, we applied our models to observations that had not been used in model construction. We began by creating diagnostic plots (Figs. 35 and 36). In these charts, we plot the value for the heroin use flag (1 = use, 0 = no use) versus the probability of use calculated by our final revised usage pattern models. We forced a slight jittering of the data points so that we could better see the volume of points at any position on the plot.

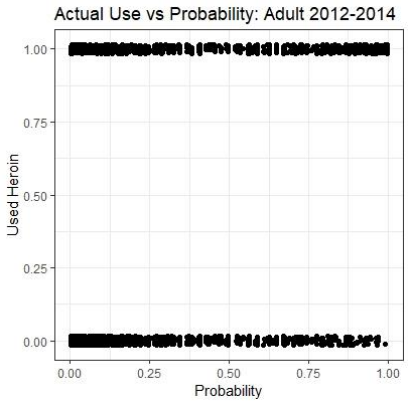

Figure 35: Adult Use vs. Probability

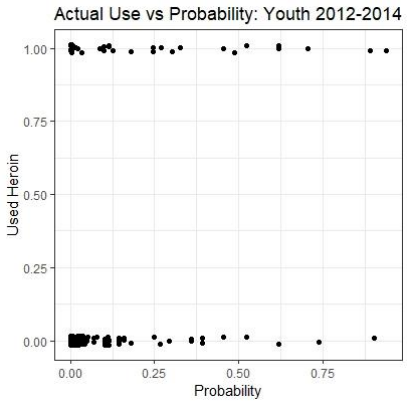

Figure 36: Youth Use vs. Probability

In the adult plot, we see that the density of the plots of non-users decreases with an increase in the probability of use. The density of the users remains consistent across the probability of use, despite the reduction in respondent counts at higher probability levels. We see a similar pattern in the plot for the youth respondents. These plots show that our models are successfully predicting heroin use, but the plots don't quantify the degree to which the models are accurate.

We validated our models using area under the receiver operating curve (AUROC) as a diagnostic metric. For both the adult and youth revised models, we plotted the ROC (Figs. 37 and 38). The AUROC for the adult revised model was 0.9703, which indicates a very good model. The AUROC for the revised youth model was 0.9233 – less than that of the adult model, but still a very good score. From the AUROC scores we can conclude that our revised Phase Two models fit both the adult and youth datasets well.

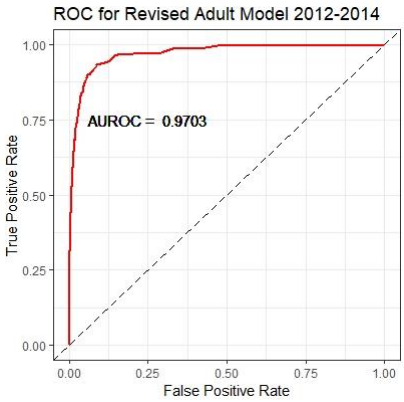

Figure 37: ROC for Revised Adult Model

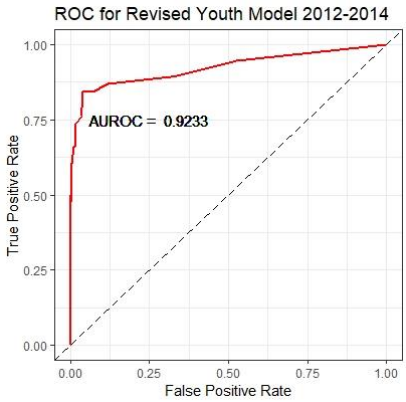

Figure 38: ROC for Revised Youth Model

Finally, we plotted the distribution of μ for both the Adult and Youth Revised Models (Figs. 39-42). There are two plots for each model. The first shows the distribution of μ for non-users. For both the adult and youth datasets, we see that the mean and median values for μ are very low. The second plot portrays the distribution of μ for users. For both datasets, the mean and median values for μ are well above those for the non-users, which we would expect. However, those values are lower than 0.5, which means that if we are to use these models for classification or prediction of heroin users, the threshold value, the value over which we would identify a potential heroin user, is only 0.35 for adults and 0.11 for youths.

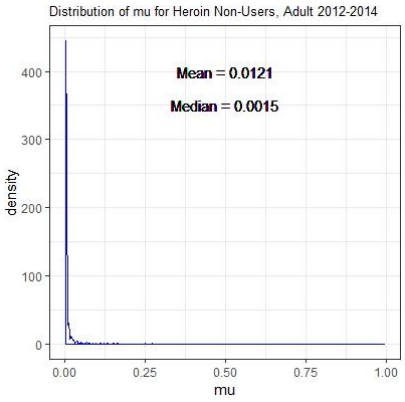

Figure 39: Adult Non-User Distribution of μ

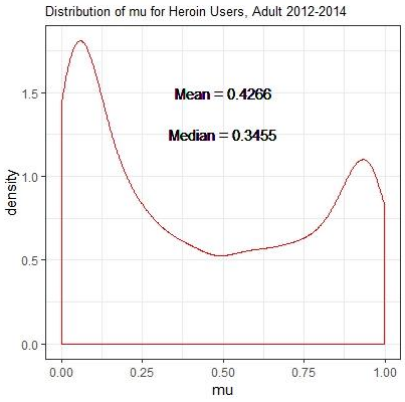

Figure 40: Adult User Distribution of μ

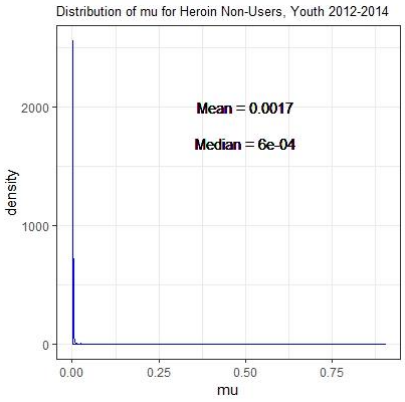

Figure 41: Youth Non-User Distribution of μ

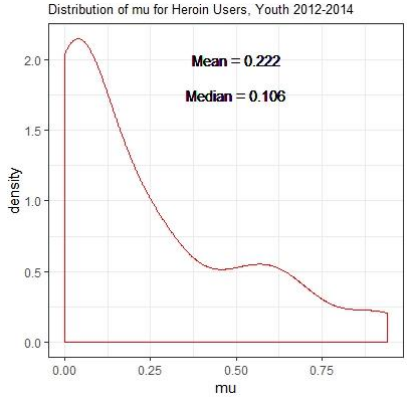

Figure 42: Youth User Distribution of μ

Conclusions

We can draw conclusions from our analysis regarding the methodology of our model. These conclusions are relevant to researchers wishing to explore large, complex datasets. We can also draw conclusions about correlates to heroin use. These conclusions are relevant to clinicians and practitioners wishing to impact the opioid crisis. Before presenting these summary findings, we will review some limitations of our study.

Limitations

The most important limitation to our study is that it is based upon data that was voluntarily provided by users through surveys. This means that our findings rely on the accuracy and honesty of those surveys. One way to validate our conclusions would be to use different data, such as insurance records, and redo the model using our findings as a set of prior distributions. Another limitation is due to our imputation approach. In the initial reduction of the dataset, we manually imputed data for missing values based upon reasonable assumptions. While most of the variables for which we did this were ultimately not important, we could use

more robust imputation techniques to remove this limitation entirely. Finally, our method for setting prior distributions for model parameters is open to debate. We chose to set subjective priors based upon a review of prior literature. A competing approach would be to set more specific priors using the values for parameters found in that literature. Given that the studies in the literature involved many fewer observations than our dataset, we believe that such an approach was unwarranted.

Methodological Findings

Traditional medical and epidemiological research begins with a set hypothesis and then uses a dataset, sometimes from a clinical study, sometimes from an existing dataset, and conducts null hypothesis tests to determine an outcome. When we are presented with a massive dataset from which many hypotheses can be tested, we believe that the data itself should guide researchers to important investigations.

The NSDUH dataset presents a good example of this situation – it has many observations, covers multiple time periods, and has thousands of variables. When we confront such a dataset, we should be able to ask vague questions, such as "what factors are closely correlated with heroin use" and let the data guide us to the answers. Unsupervised methods such as principal component analysis and supervised methods such as random forest classification trees can uncover variables that should be used by the researcher in model construction. We applied these techniques without restricting the potential variables in the NSDUH dataset that could be relevant to our question. When we did so, we uncovered some variables that we would have expected from other studies, such as OxyContin abuse and needle use, and others that had not been well studied, such as respondents' perceived risk of heroin and their ability to acquire heroin easily. Our methods rejected other variables which we may have thought were relevant such as sex, race, and other socio-demographic factors.

Our methods also incorporated Bayesian statistics, which allowed us to consider previous studies when we constructed our models. When doing research, we should not ignore previous studies, especially those based upon other data sources. Instead, research should either reinforce conclusions found elsewhere or refute them by finding them invalid according to a higher standard than we would have in the absence of prior knowledge. Bayesian statistics, through the incorporation of prior distributions, accomplishes this goal and in doing so lets a single study represent the accumulated knowledge of many. We drew some conclusions that reinforced prior knowledge, such as the correlation between needle use and heroin use. We also refuted others. Despite prior knowledge that suggested a correlation between age of first use and heroin use, our study found this relationship to be rejected at the 95% high density interval.

We advocate the use of our methodology, the application of PCA and classification trees to large datasets to determine variables for inclusion in a Bayesian model, on other sources. We are seeing an increase in availability of large medical datasets from Federal, state, and local government sources, as well as from insurance providers. We have reached a point where the

volume of data available to researchers requires advanced data analytics in order to uncover new and potentially surprising findings that would otherwise be undiscovered by traditional medical research techniques

Heroin Use Findings

We were able to quantify correlates to heroin use and how those have changed over time. We demonstrated that the impact of OxyContin use has increased since 2003. The correlation between OxyContin and heroin has been well documented, but we have shown that it arose after 2005 and grew through 2014. We also showed that while a great deal of attention has been paid to the relationship between non-prescription opioids and heroin, cocaine and crack have consistently been the drugs most associated with heroin use among adults from 2003 through 2014. We found that for both adults and youths, hallucinogen use, especially of PCP and ecstasy, was highly correlated to heroin use.

We showed that not all drugs are relevant. Instead, there is a list of critical drugs, which differs for adults and youths, that is definitively correlated to heroin use. For example, we can say that there is no evidence in the NSDUH dataset that marijuana use is correlated to heroin use for either adults or youths. This is a significant finding considering the national debate on marijuana legalization. Additionally, we found that not only is the use of critical drugs relevant, but other factors related to their use is also important. The likelihood of heroin use increases with the number of critical drugs that a potential user has abused.

We found two other factors of interest to practitioners wishing to suppress heroin usage. The first of these pertains to access to heroin. We found that respondents who had easy access to heroin were more likely to use it. While this seems obvious, we had not found any studies that considered this relationship. Advocates of policies regarding decriminalization of heroin use or the creation of "safe space" areas for use should take note that they may create unintended consequences of increased usage among the population by easing access to the drug. Similarly, respondents with a low perceived risk of heroin use are more likely to use it. Again, this may seem obvious, but by proving the relationship, we are providing evidence that education programs for likely users may reduce the chance of heroin use.

Our study has created opportunities for further research. Our parameter values can be used to create a likely-heroin-use scoring model. This model could be used by health care providers or counsellors to determine whether more aggressive intervention strategies are warranted for potential users. We recommend a longitudinal clinical study to see if a scoring model based upon our findings could reduce heroin use in a population.

.

Appendix A: Important Variables from Random Forest Analysis

Variable	Category	Description	Adult Imp	Youth Imp
rdifher	Access	Heroin fairly or very easy to obtain	Y	Y
ircocage	Age of First Use	Cocaine age of first use	Y	Y
ircrkage	Age of First Use	Crack age of first use	Y	Y
irecsage	Age of First Use	Ecstasy age of first use	Y	Y
irhalage	Age of First Use	Hallucinogens - age of first use	Y	Y
irlsdage	Age of First Use	LSD age of first use	Y	Y
irmthage	Age of First Use	Methamphetamine - age of first use		Y
iroxyage	Age of First Use	Oxycontin age of first use	Y	
irpcpage	Age of First Use	PCP age of first use	Y	Y
irstmage	Age of First Use	Stimulants age of first use		Y
irtrnage	Age of First Use	Tranquilizer age of first use	Y	Y
benzos	Ever Used	Benzodiazepine products - ever used	Y	Y
cocflag	Ever Used	Cocaine - ever used	Y	Y
CODEINE2	Ever Used	Codeine - ever used		Y
cpnmthfg	Ever Used	Methamphetamine - ever used		Y
cpnstmfg	Ever Used	Stimulants - ever used		Y
crkflag	Ever Used	Crack - ever used	Y	Y
DILAUD2	Ever Used	Dilaudid - ever used	Y	
ecsflag	Ever Used	Ecstasy - ever used		Y
halflag	Ever Used	Hallucinogens - ever used	Y	Y
lsdflag	Ever Used	LSD - ever used	Y	Y
MESC2	Ever Used	Mescaline - ever used		Y
METHDON2	Ever Used	Methadone - ever used	Y	Y
MORPHIN2	Ever Used	Morphine - ever used	Y	Y
othanl	Ever Used	Other pain relievers - ever used		Y
OXYCODP2	Ever Used	Oxycodone products (excl Oxycontin) - ever used	Y	Y
oxyflag	Ever Used	Oxycontin - ever used	Y	Y
pcpflag	Ever Used	PCP - ever used	Y	Y
PSILCY2	Ever Used	Psilocybin - ever used	Y	Y
trqflag	Ever Used	Tranquilizers - ever used	Y	Y
ircocfy	Frequency	Cocaine frequency past year		Y
irhalfy	Frequency	Hallucinogen frequency past year		Y
irmthfy	Frequency	Methamphetamine frequency past year		Y
cocneedl	Needle Use	Ever used needle to inject cocaine	Y	
mthneedl	Needle Use	Methamphetamine - ever used needle to inject		Y
otdgnedl	Needle Use	Ever used needle to inject any other drug	Y	
grskhreg	Perceived Risk	Great risk - use heroin 1-2 times per week		Y
grskhtry	Perceived Risk	Great risk - trying heroin once or twice		Y
txilalev	Treatment	Received treatment for drug or alcohol use in lifetime	Y	

Appendix B: Evidence in Literature for Priors

Variable	Category	Description	Evidence in Literature
rdifher	Access	Heroin fairly or very easy to obtain	Yes
ircocage	Age of First Use	Cocaine age of first use	Yes
ircrkage	Age of First Use	Crack age of first use	Yes
irecsage	Age of First Use	Ecstasy age of first use	Yes
irhalage	Age of First Use	Hallucinogens - age of first use	Yes
irlsdage	Age of First Use	LSD age of first use	Yes
irmthage	Age of First Use	Methamphetamine - age of first use	Yes
iroxyage	Age of First Use	Oxycontin age of first use	Yes
irpcpage	Age of First Use	PCP age of first use	Yes
irstmage	Age of First Use	Stimulants age of first use	Yes
irtrnage	Age of First Use	Tranquilizer age of first use	Yes
benzos	Ever Used	Benzodiazepine products - ever used	Inconclusive
cocflag	Ever Used	Cocaine - ever used	Yes
CODEINE2	Ever Used	Codeine - ever used	No
cpnmthfg	Ever Used	Methamphetamine - ever used	Inconclusive
cpnstmfg	Ever Used	Stimulants - ever used	No
crkflag	Ever Used	Crack - ever used	Yes
DILAUD2	Ever Used	Dilaudid - ever used	Yes
ecsflag	Ever Used	Ecstasy - ever used	No
halflag	Ever Used	Hallucinogens - ever used	No
lsdflag	Ever Used	LSD - ever used	No
MESC2	Ever Used	Mescaline - ever used	No
METHDON2	Ever Used	Methadone - ever used	Inconclusive
MORPHIN2	Ever Used	Morphine - ever used	No
othanl	Ever Used	Other pain relievers - ever used	No
onlyoxycod	Ever Used	Oxycodone products (excl Oxycontin) - ever used	Inconclusive
oxyflag	Ever Used	Oxycontin - ever used	Strong
pcpflag	Ever Used	PCP - ever used	No
PSILCY2	Ever Used	Psilocybin - ever used	No
trqflag	Ever Used	Tranquilizers - ever used	Inconclusive
ircocfy	Frequency	Cocaine frequency past year	No
irhalfy	Frequency	Hallucinogen frequency past year	No
irmthfy	Frequency	Methamphetamine frequency past year	No
cocneedl	Needle Use	Ever used needle to inject cocaine	Strong
mthneedl	Needle Use	Methamphetamine - ever used needle to inject	No
otdgnedl	Needle Use	Ever used needle to inject any other drug	Strong
grskhreg	Perceived Risk	Great risk - use heroin 1-2 times per week	No
grskhtry	Perceived Risk	Great risk - trying heroin once or twice	No
txilalev	Treatment	Received treatment for drug or alcohol use in lifetime	No
polyabuse	Ever Used	Constructed flag for multiple high risk drug use	Yes

Appendix C: Adult Phase One Gibbs Sampling Results (Null in *Red Italic*)

Adult Years 2003-2005: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-7.0472	0.1744	-7.4053	-6.7184	1.0000
Cocaine	2.7466	0.2048	2.3550	3.1583	15.5895
Crack	1.5521	0.1004	1.3564	1.7495	4.7213
Dilaudid	0.8575	0.2124	0.4398	1.2725	2.3573
Methadone	0.8470	0.1639	0.5251	1.1670	2.3326
PCP	0.7317	0.1109	0.5148	0.9493	2.0785
Tranquilizers	0.7285	0.2966	0.1408	1.3035	2.0721
Mescaline	0.6997	0.1126	0.4775	0.9201	2.0131
LSD	0.6831	0.1313	0.4246	0.9377	1.9800
Morphine	0.6023	0.2064	0.1959	1.0045	1.8262
Oxycodone (not OxyContin)	*0.3053*	*0.1375*	*-0.0360*	*0.5756*	*1.3571*
Ecstasy	0.2997	0.1073	0.0891	0.5090	1.3495
OxyContin	*0.2856*	*0.1640*	*-0.0349*	*0.6058*	*1.3306*
Other Pain Killers	*0.2570*	*0.1468*	*-0.0349*	*0.5425*	*1.2931*
Codeine	*0.0120*	*0.1356*	*-0.2543*	*0.2784*	*1.0121*
Psilocybin	*-0.0413*	*0.1189*	*-0.2738*	*0.1917*	*0.9595*
Other Stimulants	*-0.0705*	*0.1540*	*-0.3755*	*0.2264*	*0.9319*
Other Hallucinogens	*-0.1996*	*0.6916*	*-1.6655*	*1.0316*	*0.8190*
Benzos	*-0.2659*	*0.2930*	*-0.8311*	*0.3170*	*0.7665*
Methamphetamine	-0.3371	0.1153	-0.5638	-0.1120	0.7139

Adult Years 2006-2008: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.7543	0.1288	-7.0146	-6.5113	1.0000
Cocaine	2.4734	0.1560	2.1761	2.7862	11.8626
Crack	1.4614	0.0829	1.2997	1.6246	4.3120
PCP	0.9691	0.0936	0.7853	1.1521	2.6355
Dilaudid	0.8959	0.1572	0.5880	1.2042	2.4495
LSD	0.6243	0.1003	0.4274	0.8191	1.8669
Morphine	0.6179	0.1680	0.2833	0.9433	1.8551
OxyContin	0.5495	0.1391	0.2813	0.8240	1.7323
Methadone	0.4918	0.1306	0.2354	0.7461	1.6352
Mescaline	0.4429	0.1002	0.2465	0.6397	1.5572
Other Hallucinogens	*0.4171*	*0.4788*	*-0.5993*	*1.2805*	*1.5175*
Ecstasy	0.3469	0.0897	0.1705	0.5210	1.4146
Oxycodone (not OxyContin)	0.2914	0.1209	0.0542	0.5283	1.3383
Benzos	*0.2901*	*0.3101*	*-0.2850*	*0.9428*	*1.3366*
Other Pain Killers	*0.0927*	*0.1270*	*-0.1587*	*0.3382*	*1.0971*
Methamphetamine	*0.0360*	*0.0930*	*-0.1467*	*0.2178*	*1.0367*
Psilocybin	*0.0243*	*0.0961*	*-0.1640*	*0.2119*	*1.0246*
Other Stimulants	*-0.0051*	*0.1304*	*-0.2628*	*0.2486*	*0.9949*
Tranquilizers	*-0.0350*	*0.3138*	*-0.6894*	*0.5485*	*0.9656*
Codeine	*-0.0462*	*0.1111*	*-0.2637*	*0.1715*	*0.9549*

Adult Years 2009-2011: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.7367	0.1229	-6.9820	-6.5021	1.0000
Cocaine	2.6650	0.1468	2.3820	2.9538	14.3677
Crack	1.4697	0.0791	1.3152	1.6246	4.3479
PCP	0.7874	0.0968	0.5981	0.9769	2.1977
Methadone	0.7346	0.1162	0.5066	0.9635	2.0846
Dilaudid	0.6642	0.1440	0.0382	0.9471	1.9429
OxyContin	0.6172	0.1311	0.3636	0.8770	1.8538
Mescaline	0.5987	0.0977	0.4057	0.7901	1.8197
Ecstasy	0.5475	0.0835	0.3852	0.7126	1.7288
Morphine	0.5382	0.1637	0.2123	0.8542	1.7128
LSD	0.3930	0.0898	0.2171	0.5699	1.4814
Tranquilizers	*0.2916*	*0.2939*	*-0.3098*	*0.8536*	*1.3386*
Other Stimulants	0.2288	0.1094	0.0135	0.4419	1.2570
Methamphetamine	*0.0997*	*0.0893*	*-0.0764*	*0.2741*	*1.1048*
Oxycodone (not OxyContin)	*0.0831*	*0.1191*	*-0.1497*	*0.3185*	*1.0866*
Other Pain Killers	*0.0788*	*0.1235*	*-0.1686*	*0.3182*	*1.0820*
Other Hallucinogens	*0.0663*	*0.4701*	*-0.9195*	*0.9182*	*1.0686*
Codeine	*0.0419*	*0.1018*	*-0.2406*	*0.1572*	*1.0427*
Benzos	*0.0129*	*0.2905*	*-0.5374*	*0.6075*	*1.0130*
Psilocybin	*-0.0659*	*0.0912*	*-0.2454*	*0.1121*	*0.9363*

Adult Years 2012-2014: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.4759	0.1054	-6.6857	-6.2740	1.0000
Cocaine	2.4325	0.1304	2.1803	2.6902	11.3874
Crack	1.3085	0.0802	1.1526	1.4660	3.7005
OxyContin	0.9810	0.1357	0.7167	1.2488	2.6672
PCP	0.9388	0.0970	0.7489	1.1287	2.5568
Dilaudid	0.7803	0.1276	0.5294	1.0302	2.1821
Morphine	0.7703	0.1607	0.4539	1.0825	2.1604
Methadone	0.5475	0.1158	0.3200	0.7754	1.7290
Oxycodone (not OxyContin)	0.5392	0.1182	0.3056	0.7702	1.7146
Benzos	*0.5227*	*0.3666*	*-0.0171*	*1.2723*	*1.6866*
Ecstasy	0.4592	0.0848	0.2923	0.0626	1.5827
Mescaline	0.4263	0.1029	0.2237	0.6264	1.5315
LSD	0.3594	0.0901	0.1834	0.5371	1.4325
Methamphetamine	0.2640	0.0886	0.0910	0.4376	1.3021
Other Pain Killers	*0.0790*	*0.1247*	*-0.1665*	*0.3200*	*1.0822*
Tranquilizers	*-0.0877*	*0.3692*	*-0.8422*	*0.6148*	*0.9161*
Psilocybin	*-0.1097*	*0.0911*	*-0.2890*	*0.0677*	*0.8961*
Other Stimulants	*-0.1451*	*0.1159*	*-0.3754*	*0.0801*	*0.8649*
Codeine	-0.2529	0.1029	-0.4543	-0.0516	0.7766
Other Hallucinogens	*-0.8445*	*0.6550*	*-2.2389*	*0.3249*	*0.4298*

Appendix D: Adult Phase Two Gibbs Sampling Results (Null in *Red Italic*)

Adult Years 2003-2005: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-5.5831	0.2498	-6.0731	-5.0935	1.0000
Needle Use	1.4793	0.1031	1.2776	1.6819	4.3898
Access to Heroin	1.3083	0.0816	1.1487	1.4676	3.7000
Prior Treatment	0.6782	0.0827	0.5167	0.8399	1.9703
Perceived Risk	0.6122	0.0576	0.4990	0.7252	1.8444
Polyabuse	0.5105	0.0207	0.4701	0.5514	1.6661
Age of First Use	*-0.0037*	*0.0113*	*-0.0261*	*0.0180*	*0.9963*

Adult Years 2006-2008: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-5.6274	0.2524	-6.1223	-5.1329	1.0000
Needle Use	1.7855	0.1071	1.5745	1.9933	5.9625
Access to Heroin	1.0854	0.0868	0.9148	1.2554	2.9607
Prior Treatment	0.7011	0.0870	0.5295	0.8695	2.0161
Perceived Risk	0.6609	0.0606	0.5421	0.7796	1.9365
Polyabuse	0.5841	0.0231	0.5393	0.6298	1.7934
Age of First Use	*0.0001*	*0.0114*	*-0.0228*	*0.0223*	*1.0001*

Adult Years 2009-2011: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-4.8827	0.2489	-5.3717	-4.3987	1.0000
Needle Use	1.7589	0.1049	1.5539	1.9646	5.8058
Access to Heroin	1.0604	0.0814	0.9013	1.2196	2.8875
Perceived Risk	0.6379	0.0555	0.5286	0.7462	1.8925
Prior Treatment	0.5865	0.0812	0.4268	0.7455	1.7977
Polyabuse	0.5712	0.0213	0.5298	0.6130	1.7704
Age of First Use	-0.0275	0.0120	-0.0512	-0.0043	0.9728

Adult Years 2012-2014: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-4.8569	0.2332	-5.3136	-4.3963	1.0000
Needle Use	1.5758	0.1023	1.3764	1.7761	4.8345
Access to Heroin	1.2202	0.0786	1.0668	1.3746	3.3878
Prior Treatment	0.8257	0.0798	0.6691	0.9808	2.2835
Perceived Risk	0.6776	0.0556	0.5690	0.7866	1.9691
Polyabuse	0.4807	0.0178	0.4463	0.5158	1.6172
Age of First Use	-0.0306	0.0109	-0.0522	-0.0095	0.9699

Appendix E: Revised Adult Phase Two Gibbs Sampling Results

Revised Adult Years 2003-2005: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.6272	0.9081	-6.8090	-6.4524	1.0000
Needle Use	1.5487	0.1104	1.3301	1.7633	4.7053
Access to Heroin	1.2778	0.0815	1.1181	1.4374	3.5887
Prior Treatment	0.8380	0.0869	0.6670	1.0077	2.3117
Polyabuse	0.6830	0.0166	0.6508	0.7160	1.9798
Perceived Risk	0.6261	0.0586	0.5104	0.7405	1.8703

Revised Adult Years 2006-2008: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.5278	0.0894	-6.7059	-6.3542	1.0000
Needle Use	1.8820	0.1103	1.6652	2.0964	6.5666
Access to Heroin	1.0956	0.0861	0.9253	1.2634	2.9910
Prior Treatment	0.7943	0.0913	0.6151	0.9720	2.2129
Polyabuse	0.7554	0.0185	0.7193	0.7917	2.1285
Perceived Risk	0.5832	0.0600	0.4655	0.7001	1.7918

Revised Adult Years 2009-2011: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.3796	0.0834	-6.5452	-6.2177	1.0000
Needle Use	1.7114	0.1123	1.4924	1.9311	5.5367
Access to Heroin	1.1184	0.0805	0.9599	1.2768	3.0600
Polyabuse	0.7688	0.0174	0.7350	0.8032	2.1572
Prior Treatment	0.7616	0.0848	0.5957	0.9273	2.1417
Perceived Risk	0.5609	0.0553	0.4529	0.6686	1.7522

Revised Adult Years 2012-2014: 50% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.4847	0.0852	-6.6551	-6.3200	1.0000
Needle Use	1.7237	0.1097	1.5089	1.9377	5.6052
Access to Heroin	1.3691	0.0791	1.2140	1.5242	3.9318
Prior Treatment	0.9924	0.0856	0.8230	1.1580	2.6977
Perceived Risk	0.7205	0.0540	0.6145	0.8265	2.0555
Polyabuse	0.6259	0.0151	0.5967	0.6558	1.8699

Appendix F: Youth Phase One Gibbs Sampling Results (Null in *Red Italic*)

Youth Years 2003-2005: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-7.1958	0.2175	-7.6401	-6.7917	1.0000
Cocaine	2.9517	0.3707	2.2217	3.6809	19.1381
PCP	1.3126	0.3548	0.6150	2.0048	3.7158
Methamphetamine	1.1352	0.3489	0.4528	1.8229	3.1118
OxyContin	1.1166	0.4224	0.3055	1.9593	3.0543
Morphine	*0.8977*	*0.5028*	*-0.1112*	*1.8628*	*2.4539*
Ecstasy	0.7971	0.3248	0.1603	1.4343	2.2191
Other Stimulants	*0.3301*	*0.4183*	*-0.5112*	*1.1344*	*1.3911*
Mescaline	*0.2928*	*0.4665*	*-0.6312*	*1.1979*	*1.3401*
Oxycodone (not OxyContin)	*0.2716*	*0.4662*	*-0.6610*	*1.1681*	*1.3121*
Psilocybin	*0.2359*	*0.3480*	*-0.4526*	*0.9176*	*1.2660*
Other Pain Killers	*0.2256*	*0.4171*	*-0.6097*	*1.0290*	*1.2531*
Crack	*0.1750*	*0.3622*	*-0.5429*	*0.8757*	*1.1912*
Other Hallucinogens	*0.1084*	*0.9102*	*-1.8566*	*1.7172*	*1.1145*
LSD	*0.0649*	*0.3666*	*-0.6606*	*0.7794*	*1.0671*
Methadone	*0.0205*	*0.4225*	*-0.8240*	*0.8317*	*1.0207*
Tranquilizers	*-0.2946*	*0.7654*	*-1.8808*	*1.1179*	*0.7448*
Benzos	*-0.3027*	*0.7717*	*-1.7337*	*1.2857*	*0.7388*
Codeine	*-0.3049*	*0.3687*	*-1.0286*	*0.4126*	*0.7372*
Dilaudid	*-0.4328*	*0.8450*	*-2.1814*	*1.1406*	*0.6487*

Youth Years 2006-2008: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-7.6004	0.2181	-8.0453	-7.1922	1.0000
Cocaine	2.6988	0.3660	1.9800	3.4134	14.8626
Crack	1.2054	0.3131	0.5930	1.8250	3.3380
Ecstasy	0.9854	0.3249	0.3552	1.6228	2.6789
LSD	0.9834	0.3284	0.3387	1.6269	2.6735
Other Pain Killers	0.8533	0.3733	0.1098	1.5728	2.3475
Other Hallucinogens	*0.8354*	*0.6567*	*-0.5439*	*2.0223*	*2.3057*
Mescaline	*0.8218*	*0.4174*	*-0.0005*	*1.6364*	*2.2746*
PCP	0.7785	0.3481	0.0920	1.4578	2.1782
Morphine	*0.7339*	*0.4866*	*-0.2344*	*1.6796*	*2.0831*
Benzos	*0.6810*	*0.7154*	*-0.6366*	*2.1694*	*1.9758*
Methamphetamine	*0.4874*	*0.3286*	*-0.1632*	*1.1272*	*1.6281*
Oxycodone (not OxyContin)	*0.3163*	*0.4353*	*-0.5636*	*1.1484*	*1.3721*
OxyContin	*0.3139*	*0.3859*	*-0.4338*	*1.0743*	*1.3688*
Codeine	*0.0087*	*0.3303*	*-0.6413*	*0.6528*	*1.0087*
Psilocybin	*-0.0107*	*0.3414*	*-0.6820*	*0.6625*	*0.9894*
Other Stimulants	*-0.1561*	*0.3822*	*-0.9301*	*0.5701*	*0.8555*
Methadone	*-0.4403*	*0.3874*	*-1.2118*	*0.3080*	*0.6438*
Tranquilizers	*-0.4619*	*0.7241*	*-1.9614*	*0.8795*	*0.6301*
Dilaudid	*-0.7945*	*0.7222*	*-2.2809*	*0.5518*	*0.4518*

Youth Years 2009-2011: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-7.0499	0.1653	-7.3837	-6.7361	1.0000
Cocaine	1.6498	0.3196	1.0245	2.2769	5.2059
Morphine	1.5074	0.4387	0.6371	2.3565	4.5150
Tranquilizers	1.3094	0.6232	0.0148	2.4472	3.7040
Ecstasy	1.1508	0.3182	0.5325	1.7747	3.1607
Other Pain Killers	1.0166	0.3448	0.3267	1.6766	2.7638
Methamphetamine	0.9443	0.3280	0.2972	1.5804	2.5710
Psilocybin	0.8334	0.3073	0.2277	1.4384	2.3011
Methadone	0.7480	0.3204	0.1143	1.3714	2.1128
Crack	0.6880	0.3392	0.0201	1.3473	1.9897
PCP	*0.6101*	*0.3370*	*-0.0575*	*1.2609*	*1.8406*
Other Hallucinogens	*0.5422*	*0.7491*	*-1.0590*	*1.8749*	*1.7198*
LSD	*0.4804*	*0.3066*	*-0.1237*	*1.0777*	*1.6167*
Mescaline	*0.1865*	*0.4210*	*-0.6427*	*1.0029*	*1.2050*
Dilaudid	*0.0256*	*0.6692*	*-1.3288*	*1.3153*	*1.0259*
Other Stimulants	*-0.0273*	*0.3623*	*-0.7549*	*0.6671*	*0.9731*
Codeine	*-0.2034*	*0.3089*	*-0.8119*	*0.3984*	*0.8160*
OxyContin	*-0.2915*	*0.3667*	*-1.0055*	*0.4311*	*0.7471*
Oxycodone (not OxyContin)	*-0.4447*	*0.4477*	*-1.3513*	*0.4037*	*0.6410*
Benzos	*-0.7576*	*0.6227*	*-1.9017*	*0.5244*	*0.4688*

Youth Years 2012-2014: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-7.1976	0.1913	-7.5878	-6.8377	1.0000
Cocaine	2.4355	0.4429	1.5718	3.3065	11.4210
Crack	1.3272	0.4587	0.4241	2.2192	3.7705
Oxycodone (not OxyContin)	1.1979	0.5009	0.1778	2.1506	3.3133
Ecstasy	1.1354	0.4583	0.2356	2.0289	3.1125
OxyContin	1.0544	0.5040	0.0679	2.0474	2.8702
Mescaline	*1.0508*	*0.5773*	*-0.1055*	*2.1583*	*2.8601*
Other Hallucinogens	*0.7438*	*0.8090*	*-0.9503*	*2.2085*	*2.1039*
Morphine	*0.5820*	*0.5561*	*-0.5284*	*1.6490*	*1.7896*
Tranquilizers	*0.4480*	*0.8529*	*-1.2883*	*2.0428*	*1.5651*
Methamphetamine	*0.3338*	*0.4409*	*-0.5448*	*1.1858*	*1.3963*
Codeine	*0.3017*	*0.4413*	*-0.5528*	*1.1709*	*1.3521*
PCP	*0.2559*	*0.4577*	*-0.6550*	*1.1479*	*1.2916*
LSD	*0.2012*	*0.4549*	*-0.6789*	*1.0976*	*1.2229*
Other Stimulants	*0.0479*	*0.4837*	*-0.9229*	*0.9775*	*1.0491*
Psilocybin	*-0.0487*	*0.4167*	*-0.8756*	*0.7620*	*0.9525*
Other Pain Killers	*-0.1506*	*0.5012*	*-1.1641*	*0.8030*	*0.8602*
Benzos	*-0.1967*	*0.8546*	*-1.8013*	*1.5441*	*0.8215*
Methadone	*-0.2168*	*0.4928*	*-1.1922*	*0.7402*	*0.8051*
Dilaudid	*-0.4807*	*0.8632*	*-2.2634*	*1.1257*	*0.6183*

Appendix G: Youth Phase Two Gibbs Sampling Results (Null in *Red Italic*)

Youth Years 2003-2005: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-4.3374	1.0489	-6.4709	-2.3591	1.0000
Access to Heroin	1.0172	0.2775	0.4780	1.5611	2.7655
Needle Use	0.9897	0.4158	0.1661	1.7947	2.6904
Polyabuse	0.9063	0.1235	0.6657	1.1504	2.4753
Perceived Risk	0.6158	0.1622	0.2977	0.9376	1.8511
Prior Treatment	*0.0910*	*0.3049*	*-0.5165*	*0.6763*	*1.0953*
Frequency of Use	*0.0018*	*0.0017*	*-0.0016*	*0.0051*	*1.0018*
Age of First Use	*-0.1046*	*0.0678*	*-0.2339*	*0.0310*	*0.9007*

Youth Years 2006-2008: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.8528	0.9338	-8.7801	-5.1098	1.0000
Needle Use	2.5029	0.3544	1.8074	3.1953	12.2177
Access to Heroin	1.2744	0.2605	0.7641	1.7867	3.5766
Polyabuse	0.8557	0.0935	0.6750	1.0413	2.3531
Prior Treatment	0.5721	0.2822	0.0096	1.1128	1.7719
Perceived Risk	0.4552	0.1543	0.1501	0.7554	1.5766
Age of First Use	*0.0344*	*0.0609*	*-0.0812*	*0.1577*	*1.0350*
Frequency of Use	*0.0003*	*0.0017*	*-0.0031*	*0.0035*	*1.0003*

Youth Years 2009-2011: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-5.5754	0.7243	-7.0464	-4.2107	1.0000
Needle Use	1.6948	0.3764	0.9564	2.4303	5.4456
Access to Heroin	1.5490	0.2415	1.0785	2.0260	4.7070
Polyabuse	0.6365	0.0698	0.5008	0.7752	1.8898
Perceived Risk	0.4450	0.1464	0.1549	0.7295	1.5604
Prior Treatment	*0.3368*	*0.2865*	*-0.2315*	*0.8859*	*1.4004*
Frequency of Use	*0.0007*	*0.0014*	*-0.0022*	*0.0035*	*1.0007*
Age of First Use	*-0.0652*	*0.0474*	*-0.1547*	*0.0304*	*0.9369*

Youth Years 2012-2014: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-7.0220	1.5457	-10.1357	-4.1107	1.0000
Needle Use	1.9873	0.4365	1.1301	2.8448	7.2955
Access to Heroin	1.3788	0.3268	0.7430	2.0219	3.9701
Polyabuse	0.7873	0.1711	0.4539	1.1258	2.1975
Prior Treatment	0.7294	0.3376	0.0563	1.3812	2.0739
Perceived Risk	0.6011	0.2017	0.2056	0.9978	1.8242
Age of First Use	*0.0588*	*0.0973*	*-0.1273*	*0.2545*	*1.0606*
Frequency of Use	0.0070	0.0022	0.0027	0.0113	1.0071

Appendix H: Revised Youth Phase Two Gibbs Sampling Results

Revised Youth Years 2003-2005: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-6.7899	0.1755	-7.1408	-6.4543	1.0000
Needle Use	3.6752	0.3042	3.0694	4.2628	39.4577
Prior Treatment	2.3401	0.2181	1.9036	2.7585	10.3827
Access to Heroin	1.8479	0.1783	1.4984	2.1978	6.3467
Perceived Risk	0.4660	0.1060	0.2577	0.6733	1.5935
Polyabuse	0.0723	0.0234	0.0279	0.1193	1.0750

Revised Youth Years 2006-2008: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-8.2731	0.2762	-8.8355	-7.7560	1.0000
Needle Use	2.5737	0.3916	1.7932	3.3350	13.1143
Access to Heroin	1.4147	0.2575	0.9095	1.9170	4.1153
Prior Treatment	1.3209	0.2972	0.7297	1.8970	3.7468
Polyabuse	1.2847	0.0777	1.1327	1.4380	3.6136
Perceived Risk	0.5411	0.1582	0.2304	0.8490	1.7179

Revised Youth Years 2009-2011: 70% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-7.7509	0.2219	-8.2006	-7.3292	1.0000
Access to Heroin	1.6663	0.2195	1.2354	2.0962	5.2925
Needle Use	1.5101	0.3975	0.7319	2.2862	4.5272
Polyabuse	0.9148	0.0491	0.8183	1.0111	2.4963
Prior Treatment	0.7295	0.2802	0.1675	1.2713	2.0740
Perceived Risk	0.3779	0.1309	0.1215	0.6359	1.4592

Revised Youth Years 2012-2014: 730% Sample (Informed Prior)					
Drug	Mean	SD	2.50%	97.50%	OR
Intercept	-8.0889	0.2856	-8.6721	-7.5526	1.0000
Polyabuse	1.6043	0.1180	1.3726	1.8347	4.9744
Access to Heroin	1.4954	0.2838	0.9346	2.0479	4.4611
Prior Treatment	1.2091	0.3581	0.4966	1.9041	3.3505
Needle Use	1.0587	0.5081	0.0489	2.0413	2.8826
Perceived Risk	0.6689	0.1652	0.3455	0.9941	1.9521

References

Baneshi, M. R., Haghdoost, A. A., Zolala, F., Nakhaee, N., Jalali, M., Tabrizi, R., & Akbari, M. (2017). Can Religious Beliefs be a Protective Factor for Suicidal Behavior? A Decision Tree Analysis in a Mid-Sized City in Iran, 2013. *Journal of Religion and Health*, *56*(2), 428–436. https://doi.org/10.1007/s10943-016-0215-x

Beswick, T., Best, D., Rees, S., Coomber, R., Gossop, M., & Strang, J. (2001). Multiple drug use: Patterns and practices of heroin and crack use in a population of opiate addicts in treatment. *Drug and Alcohol Review*, *20*(2), 201–204. https://doi.org/10.1080/09595230120058588

Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). *Classification and Regression Trees*. New York, NY: Chapman and Hall.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (1996b). Heuristics of instability in model selection. *The Annals of Statistics*, *24*(6), 2350–2383. https://doi.org/10.1214/aos/1032181158

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cerd, M., An Santaella, J., Marshall, B. D. L., Kim, J. H., & Martins, S. S. (2015). Nonmedical Prescription Opioid Use in Childhood and Early Adolescence Predicts Transitions to Heroin Use in Young Adulthood: A National Study. *The Journal of Pediatrics*, *167*, 605–612.e2. https://doi.org/10.1016/j.jpeds.2015.04.071

Chitwood-Dagner, K. K., Carlson, A. M., Friedman, F., & Skatter, B. A. (1995). An alternative method for the identification of potential habitual narcotic users from a managed care claims database. *Medical Interface*, *8*(11), 102–103, 105–109. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10153507

Cicero, T. J., Ellis, M. S., Surratt, H. L., & Kurtz, S. P. (2014). The changing face of heroin use in the United States a retrospective analysis of the past 50 years. *JAMA Psychiatry*, *71*(7), 821–826. https://doi.org/10.1001/jamapsychiatry.2014.366

Cochran, B. N., Flentje, A., Heck, N. C., Van Den Bos, J., Perlman, D., Torres, J., … Carter, J. (2014). Factors predicting development of opioid use disorders among individuals who receive an initial opioid prescription: Mathematical modeling using a database of commercially-insured individuals. *Drug and Alcohol Dependence*, *138*(1), 202–208. https://doi.org/10.1016/j.drugalcdep.2014.02.701

Compton, W. M., Gfroerer, J., Conway, K. P., & Finger, M. S. (2014). Unemployment and substance outcomes in the United States 2002–2010. *Drug and Alcohol Dependence*, *142*, 350–353. https://doi.org/10.1016/j.drugalcdep.2014.06.012

Denwood, M. J. (2016). runjags : An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS.

Journal of Statistical Software, *71*(9). https://doi.org/10.18637/jss.v071.i09

Dienes, Z. (2008). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Taper & Lele*. https://doi.org/10.1177/1745691611406920

Edlund, M. J., Martin, B. C., Fan, M. Y., Devries, A., Braden, J. B., & Sullivan, M. D. (2010). Risks for opioid abuse and dependence among recipients of chronic opioid therapy: Results from the TROUP Study. *Drug and Alcohol Dependence*, *112*(1–2), 90–98. https://doi.org/10.1016/j.drugalcdep.2010.05.017

Fernández, L., Mediano, P., García, R., Rodríguez, J. M., & Marín, M. (2016). Risk Factors Predicting Infectious Lactational Mastitis: Decision Tree Approach versus Logistic Regression Analysis. *Maternal and Child Health Journal*, *20*(9), 1895–1903. https://doi.org/10.1007/s10995-016-2000-6

Frisman, L., Prendergast, M., Lin, H.-J., Rodis, E., & Greenwell, L. (2008). Applying Classification and Regression Tree Analysis to Identify Prisoners with High HIV Risk Behaviors. *Journal of Psychoactive Drugs*, *40*(4), 447–458. https://doi.org/10.1080/02791072.2008.10400651

Greenland, S. (2000). When Should Epidemiologic Regressions Use Random Coefficients? *Biometrics*, *56*(3), 915–921. https://doi.org/10.1111/j.0006-341X.2000.00915.x

Greenland, S. (2001). Putting Background Information About Relative Risks into Conjugate Prior Distributions. *Biometrics*, *57*(3), 663–670. https://doi.org/10.1111/j.0006-341X.2001.00663.x

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *International Journal of Epidemiology*, *35*(3), 765–775. https://doi.org/10.1093/ije/dyi312

Greenland, S. (2007). Bayesian perspectives for epidemiological research. II. Regression analysis. *International Journal of Epidemiology*, *36*(1), 195–202. https://doi.org/10.1093/ije/dyl289

Greenland, S., & Poole, C. (2013). Living with P values: Resurrecting a bayesian perspective on frequentist statistics. *Epidemiology*, *24*(1), 62–68. https://doi.org/10.1097/EDE.0b013e3182785741

Han, B. H., Moore, A. A., Sherman, S., Keyes, K. M., & Palamar, J. J. (2017). Demographic trends of binge alcohol use and alcohol use disorders among older adults in the United States, 2005–2014. *Drug and Alcohol Dependence*, *170*, 198–207. https://doi.org/10.1016/j.drugalcdep.2016.11.003

Hothorn, T. (2018). A Laboratory for Recursive Partytioning. Retrieved from http://party.r-forge.r-project.org

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Husson, A. F., Josse, J., Le, S., Mazet, J., & Husson, M. F. (2018). Package ‘ FactoMineR .’ https://doi.org/10.1201/b10345-2>.License

Jones, C. M. (2013). Heroin use and heroin use risk behaviors among nonmedical users of prescription opioid pain relievers – United States, 2002–2004 and 2008–2010. *Drug and Alcohol Dependence*, *132*(1–2), 95–100. https://doi.org/10.1016/j.drugalcdep.2013.01.007

Kandel, D. B., Ph, D., & Chen, K. (1992). Stages of Progression in Drug Involvement from Adolescence to Adulthood: Further Evidence for the Gateway Theory*. *Journal of Studies on Alcohol*, 447–457.

Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *Journal of Healthcare Information Management : JHIM*, *19*(2), 64–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15869215

Kruschke, J. K. (2015a). Bayes’ Rule. In *Doing Bayesian Data Analysis* (pp. 99–120). Elsevier. https://doi.org/10.1016/B978-0-12-405888-0.00005-2

Kruschke, J. K. (2015b). Markov Chain Monte Carlo. In *Doing Bayesian Data Analysis* (pp. 143–191). Elsevier. https://doi.org/10.1016/B978-0-12-405888-0.00007-6

Kruschke, J. K. (2015c). Overview of the Generalized Linear Model. In *Doing Bayesian Data Analysis* (pp. 419–447). Elsevier. https://doi.org/10.1016/B978-0-12-405888-0.00015-5

Kruschke, J. K. (2015d). The R Programming Language. In *Doing Bayesian Data Analysis* (pp. 33–70). Elsevier. https://doi.org/10.1016/B978-0-12-405888-0.00003-9

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Lankenau, S. E., Teti, M., Silva, K., Bloom, J. J., Harocopos, A., & Treese, M. (2012). Patterns of prescription drug misuse among young injection drug users. *Journal of Urban Health*, *89*(6), 1004–1016. https://doi.org/10.1007/s11524-012-9691-9

Leri, F., Bruneau, J., & Stewart, J. (2003). Understanding polydrug use: review of heroin and cocaine co-use. *Addiction*, *98*(1), 7–22. https://doi.org/10.1046/j.1360-0443.2003.00236.x

Maher, L., Li, J., Jalaludin, B., Wand, H., Jayasuriya, R., Dixon, D., & Kaldor, J. M. (2007). Impact of a reduction in heroin availability on patterns of drug use, risk behaviour and incidence of hepatitis C virus infection in injecting drug users in New South Wales, Australia. *Drug and Alcohol Dependence*, *89*(2–3), 244–250. https://doi.org/10.1016/j.drugalcdep.2007.01.001

Mars, S. G., Bourgois, P., Karandinos, G., Montero, F., & Ciccarone, D. (2014). Every “Never” I Ever Said Came True: Transitions from opioid pills to heroin injecting. *International Journal of Drug Policy*, *25*, 257–266. https://doi.org/10.1016/j.drugpo.2013.10.004

McBride, D. C., Inciardi, J. A., Chitwood, D. D., McCoy, C. B., & The National AIDS Research Consorti. (1992). Crack Use and Correlates of Use in a National Population of Street Heroin Users. *Journal of Psychoactive Drugs*, *24*(4), 411–416.

https://doi.org/10.1080/02791072.1992.10471665

McBride, D. M. R. L. (1980). Dilaudid use: trends and characteristics of users. *Chemical Dependencies: Behavioral and Biomedical Issues*, *4*(1–2), 85–100.

McCabe, S. E., West, B. T., Morales, M., Cranford, J. A., & Boyd, C. J. (2007). Does early onset of non-medical use of prescription drugs predict subsequent prescription drug abuse and dependence? Results from a national study. *Addiction*, *102*(12), 1920–1930. https://doi.org/10.1111/j.1360-0443.2007.02015.x

Piper, M. E., Loh, W.-Y., Smith, S. S., Japuntich, S. J., & Baker, T. B. (2011). Using Decision Tree Analysis to Identify Risk Factors for Relapse to Smoking. *Substance Use & Misuse*, *46*(4), 492–510. https://doi.org/10.3109/10826081003682222

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *3d International Workshop on Distributed Statistical Computing*. Retrieved from http://www.r-project.org/conferences/DSC-2003/

Pugatch, D., Strong, L. L., Has, P., Patterson, D., Combs, C., Reinert, S., … Brown, L. (2001). Heroin use in adolescents and young adults admitted for drug detoxification. *Journal of Substance Abuse*, *13*(3), 337–346. https://doi.org/10.1016/S0899-3289(01)00081-5

Rhodes, T., Briggs, D., Kimber, J., Jones, S., & Holloway, G. (2007). Crack-heroin speedball injection and its implications for vein care: Qualitative study. *Addiction*, *102*(11), 1782–1790. https://doi.org/10.1111/j.1360-0443.2007.01969.x

SAMHSA. (2016). National Survey on Drug Use and Health, 2014: Codebook. https://doi.org/10.3886/ICPSR32722.v3

Scott, I. (1998). A hundred-year habit. *History Today*, *48*(6), 6–8. Retrieved from http://web.b.ebscohost.com.ezproxy.lib.ou.edu/ehost/pdfviewer/pdfviewer?vid=3&sid=31afa7d5-ef8c-4e8f-83f6-a03ef5e8ed25%40sessionmgr102

Stumbo, S. P., Yarborough, B. J. H., McCarty, D., Weisner, C., & Green, C. A. (2017). Patient-reported pathways to opioid use disorders and pain-related barriers to treatment engagement. *Journal of Substance Abuse Treatment*, *73*, 47–54. https://doi.org/10.1016/j.jsat.2016.11.003

Substance Abuse and Mental Health Administration. (n.d.). Who We Are | SAMHSA - Substance Abuse and Mental Health Services Administration. Retrieved August 13, 2018, from https://www.samhsa.gov/about-us/who-we-are

US Department of Justice Drug Enforcement Agency. (n.d.). Automation of Reports and Consolidated Orders System (ARCOS). Retrieved August 13, 2018, from https://www.deadiversion.usdoj.gov/arcos/index.html

Wilkerson, R. G., Kim, H. K., Windsor, T. A., & Mareiniss, D. P. (2016). The Opioid Epidemic in the United States. *Emergency Medicine Clinics of North America*, *34*(2), e1–e23. https://doi.org/10.1016/j.emc.2015.11.002

Wu, L.-T., & Howard, M. O. (2007). Is inhalant use a risk factor for heroin and injection drug use among adolescents in the United States? *Addictive Behaviors*, *32*(2), 265–281. https://doi.org/10.1016/j.addbeh.2006.03.043

Dedicated to John (Jack) Beattie

May 14, 1996 – October 14, 2016