Substance Abuse Data Analysis Project – Who Needs Aid?

Niki Z. Petrakos

Department of Biostatistics, University of Washington

BIOST 579: Data Analysis and Reporting

Dr. Brian Leroux

June 8, 2021

I. Background Information

Substance abuse is a significant public health issue that affects various communities in the United States. It not only contributes to other leading causes of death, including cancer, heart disease, and HIV/AIDS, but it also plays a leading role in the current opioid epidemic, where overdose deaths have increased by alarming rates in the past two decades[1]. It is evident that aid is needed, but who needs aid the most?

TEDS-A is a data set that comprises substance abuse treatment data from 48 states in the United States (in the 2018 data set, data from Georgia and Oregon were excluded due to insufficient data being collected). Records are for each admission to a substance abuse facility, meaning that some admissions may represent the same person as some people may be admitted to a facility more than once during a calendar year. TEDS-A contains data on patients aged 12 years and older and their demographics as well as their substance use characteristics. The following are some of the variables recorded (categorized by demographic data or substance use data):

- **Demographic data:**
    - Age at admission
    - Gender
    - Race
    - Ethnicity
    - Marital status
    - Education
    - Employment status
    - Detailed not in labor force
    - Pregnant at admission
    - Veteran status
    - Living arrangements
    - Source of income/support
    - Arrests in past 30 days
    - Census region
- **Substance use data:**
    - Type of treatment service/setting
    - Medication-assisted opioid therapy
    - Days waiting to enter substance use treatment
    - Referral Source
    - Previous substance use treatment episodes
    - Substance use (primary, secondary, and/or tertiary)
    - Route of administration (primary, secondary, and/or tertiary)
    - Frequency of use (primary, secondary, and/or tertiary)
    - Age at first use (primary, secondary, and/or tertiary)
    - Current IV drug use
    - Heroin reported at admission
    - Non-rx methadone reported at admission
    - Other opiates/synthetics reported at admission
    - Other drug reported at admission
    - Substance use type
    - Co-occurring mental and substance use disorders
    - Health insurance

[1] NIDA. (2005, June 1). Drug Abuse and Addiction: One of America's Most Challenging Public Health Problems. https://archives.drugabuse.gov/publications/drug-abuse-addiction-one-americas-most-challenging-public-health-problems.

II. Scientific Questions and Aims

Since the data set is for a population in which everyone was admitted to a publicly-funded substance abuse facility, the general aim of this project will be to investigate certain types of admittance among those admitted. First, admittance for use of more than one drug will be analyzed, along with assessing whether or not living arrangements is an effect modifier. Then, admittance for opioids specifically will be explored, in order to better understand the impact that medication-assisted opioid therapy may (or many not) have on patients who experience opioid misuse.

1. Among people at least 12 years of age who were admitted to a publicly-funded substance abuse facility in 2018, is there an association between co-occurring mental and substance use disorders and admittance to a substance abuse facility for use of more than one drug, controlling for age and sex?

2. After having analyzed the potential association between co-occurring mental and substance use disorders and admittance to a substance abuse facility for use of more than one drug, is living arrangements an effect modifier in the relationship between these two variables, controlling for age and sex, among people at least 12 years of age who were admitted to a publicly substance abuse facility in 2018?

    - <u>Aim:</u> to investigate if a patient's living situation (e.g., whether or not they are houseless) has an effect on the potential association between co-occurring mental and substance use disorders and admittance to a facility for abuse of more than one drug.

        - Note: understanding which populations may be more vulnerable for substance misuse and which factors may affect vulnerability to substance misuse (and then leading to admittance to a facility) is important in order to reduce the burden of health in these vulnerable populations. In order to deduce which new therapies or solutions may be presented as aid, we must first know which populations need to be targeted or need the most assistance.

3. Controlling for age, sex, and living arrangements, is medication-assisted opioid therapy associated with a lower odds of admittance to a substance abuse facility for opioid abuse, compared to no therapy, among people at least 12 years of age who were admitted to a publicly-funded facility for substance abuse?

    - <u>Aim:</u> to investigate if there exists an association between a patient being on medication-assisted opioid therapy and being admitted to a substance abuse facility for opioid abuse, while controlling for age, sex, and living arrangements.

- Note: age, sex, and living arrangements were chosen as covariates to be controlled for because these variables likely have an impact on both access to medication-assisted opioid therapy as well as being admitted to a treatment center for opioid use. Moreover, the addition of living arrangements to this model will allow for inference from this model to be related to the inference from the previous model, where living arrangements is investigated as a potential effect modifier.

III. Statistical Methods – General Overview with Descriptive Statistics

Question 1: Multiple Logistic Regression Model

To answer the first research question (which will henceforth be referenced as Question 1), a multiple logistic regression model will be fit following the guidelines presented below, where the percentages in grey text represent the descriptive statistics of each variable.

- Response: admittance for misuse of more than one substance – Yes (57.9%) /No (42.1%) (binary outcome)
    - Missing data: 0.0%
- Predictor of interest: co-occurring mental and substance use disorders – Yes (34.4%) /No (49.3%) (binary predictor)
    - Missing data: 16.3%
- Other covariates:
    - Age at admission (categorized into the following bins):
        - 12-20 (6.1%), 21-29 years (25.6%), 30-39 years (30.9%), 40-49 years (18.2%), 50 years and older years (19.2%)
        - Missing data: 0.0%
    - Sex – Female (35.6%) /Male (64.3%) (binary covariate)
        - Missing data: 0.1%
- Interpretation: adjusted odds ratio, 95% C.I., p-value

Question 2: Multiple Logistic Regression Model

To answer the second research question (which will henceforth be referenced as Question 2), a second multiple logistic regression model will be fit with the same response, predictor of interest, and other covariates as in the previous model, with the addition of an interaction term between the predictor of interest (co-occurring mental and substance use disorders) and living arrangements (the hypothesized effect modifier).

- Response: admittance of misuse of more than one substance – Yes/No (binary outcome)
    - Missing data: 0.0%
- Predictor of interest: interaction term between co-occurring mental and substance use disorders and living arrangements

    - Co-occurring mental and substance use disorders – Yes (34.4%) /No (49.3%) (binary predictor)
        - Missing data: 16.3%
    - Living arrangements (categorized into the following bins):
        - Houseless (14.0%), Dependent living (16.3%), Independent living (57.6%)
        - Missing data: 12.1%
- Other covariates:
    - Age at admission (categorized into the following bins):
        - 12-20 (6.1%), 21-29 years (25.6%), 30-39 years (30.9%), 40-49 years (18.2%), 50 years and older years (19.2%)
        - Missing data: 0.0%
    - Sex – Female (35.6%) /Male (64.3%) (binary covariate)
        - Missing data: 0.1%
- Interpretation: odds ratios, 95% C.I, p-values

Question 3: Multiple Logistic Regression Model

To answer the third research question (which will henceforth be referenced as Question 3), a third multiple logistic regression model will be fit following the guidelines presented below.

- Response: admittance for opioid abuse – Yes (42.9%) /No (57.1%) (binary outcome)
    - 30.4% for heroin, 0.3% for non-prescription methadone, 12.2% for other opiates
- Predictor of interest: medication-assisted opioid therapy – Yes (15.3%) /No (77.3%) (binary predictor)
    - Missing data: 7.4%
- Other covariates:
    - Age at admission (categorized into the following bins):
        - 12-20 (6.1%), 21-29 years (25.6%), 30-39 years (30.9%), 40-49 years (18.2%), 50 years and older years (19.2%)
        - Missing data: 0.0%
    - Sex – Female (35.6%) /Male (64.3%) (binary covariate)
        - Missing data: 0.1%
    - Living arrangements (categorized into the following bins):
        - Houseless (14.0%), Dependent living (16.3%), Independent living (57.6%)
        - Missing data: 12.1%
- Interpretation: odds ratio, 95% C.I., p-value

IV. Statistical Analysis Plan

Version 3 – June 8, 2021

A. Background

Substance abuse is an issue in the United States that has led to a dramatic increase in the number of deaths due to overdoses in the past decade, and it also plays a role in the spread of certain diseases such as HIV. Understanding how best to treat people for substance abuse and targeting the populations that need the most help are two pertinent steps to lessen the burden of health.

B. Description of Data Set

The data set from which analyses will be performed is the Treatment Episode Data Set: Admissions 2018 (TEDS-A-2018). This data set is from the parent series entitled Treatment Episode Data Set – Admissions, which is a national data system of annual admissions to substance abuse treatment facilities. TEDS-A does not include all admissions to substance abuse treatment; instead, it includes the subset that constitutes the public burden for substance abuse treatment, as states are only required by state laws to report their publicly-funded admissions. TEDS-A contains records on admissions aged 12 or older, and includes information on admission demographics as well as substance abuse characteristics. It is important to note that this data set represents admissions, not individuals, as a person may be admitted to treatment more than once. TEDS-A includes data reported from as early as 1992 and as recent as 2018. For the purposes of this analysis, the 2018 data set was selected.

C. Research Questions

There are three research questions – the first two involve the association between the misuse of multiple drugs and co-occurring mental health issues and determining whether or not living arrangements is an effect modifier, and the third involves the association between medication-assisted opioid therapy and admittance to a substance abuse facility for opioid misuse. The first hypothesis is that houseless people have higher odds of being admitted for misuse of multiple substances, given a co-occurring mental health issue. The second hypothesis is that medication-assisted opioid therapy is associated with lower odds of being admitted for opioid misuse (meaning medication-assisted opioid therapy is effective).

D. Study Design, Study Population, Hypothesis Testing

The study design is a cross-sectional (observational) study. There will be no repetitive analysis. The analysis will be performed on all observations in question, collectively. There will be no time-stratified analyses as all observations are considered to be from the same time point (i.e., in the same calendar year, 2018). The demographic and substance use factors that will be included in the statistical analyses were collected at the point of admission to the treatment facility.

For the first research question, hypothesis testing will be performed using a Wald test. For the second research question, hypothesis testing will be performed using a likelihood ratio test, where the model comparison will involve the model from the first research question. For the third research question, hypothesis testing will be performed using a Wald Test. For all three tests, a significance level of α = 0.05 will be used. Confidence intervals and p-values for odds ratios and hypothesis testing will be reported for all three research questions. The analysis populations are those recorded in the TEDS-A data set – people aged 12 years or older who were admitted to a publicly-funded substance abuse facility in one of 48 states (where the two excluded states are Georgia and Oregon due to incomplete data).

E. Descriptive Statistics

To understand the representativeness of the data, below are some descriptive statistics of various demographic variables:

- Sex – Female (35.6%), Male (64.3%)
- Age at admission (categorized into the following bins):
  - 12-20 (6.1%), 21-29 years (25.6%), 30-39 years (30.9%), 40-49 years (18.2%), 50 years and older years (19.2%)
- Living arrangements (categorized into the following bins):
  - Houseless (14.0%), Dependent living (16.3%), Independent living (57.6%)

F. Eligibility Criteria, Missing Data

The eligibility criteria are to have been admitted to a publicly-funded substance abuse facility in the United States. State laws require substance abuse treatment programs to report their publicly-funded admissions to the state, and the states then report these data from their state administrative systems to the Substance Abuse & Mental Health Data Archive (SAMHSA). This also means there is no issue of withdrawal from the study, as data was collected at the point of admission, and when the patient left the facility does not play a role in these analyses. However, while lost to follow-up data is not an issue, some demographic and substance use data was not collected fully, and hence there is the issue of missing data. If any value of the covariates or response variables are missing for a certain observation, the observation will be omitted from the analyses. This likely will not be a big issue due to the sheer size of the data set (1,935,541 total observations), and that the proportion of missing data for the covariates as listed above is not high enough to cause the sample size to be too small after having removed observations with missing recorded values.

G. Baseline Characteristics, Control for Confounding

The baseline characteristics are age (categorized), sex, living arrangements, whether or not the person admitted had a co-occurring mental health disorder, and whether or not the person admitted was on medication-assisted opioid therapy. These characteristics will be descriptively summarized via percentages. Note that age and sex (as well as living arrangements in the third research question) may be potential confounding variables, and hence will be included in the logistic regression models as a way to control for possible confounding.

H. Primary Analysis

For each research question, a logistic regression model will be fit. In the first question, the response is admittance for misuse of more than one substance, and the predictor of interest is co-occurring mental and substance use disorders. The other covariates to be included are age and sex. The odds ratios will be calculated and interpreted, along with confidence intervals and p-values. In the second question, the response and predictor of interest will be the same as in the first question, as will be the covariates included for adjustment purposes. However, living arrangements will also be included as a potential effect modifier in the second question. Again, odds ratios will be calculated and interpreted, along with confidence intervals and p-values. The p-value resulting from the likelihood ratio test, comparing the second model to the first model, will also be reported. In the third question, the response is admittance for opioid abuse, and the predictor of interest is medication-assisted opioid therapy. The other covariates to be adjusted for are age, sex, and living arrangements. Similar to the first two research questions, odds ratios will be calculated and interpreted, along with confidence intervals and p-values.

An important note: in logistic regression, it is assumed that the observations are independent, however in this analysis, it is possible that this assumption is not realistic due to the likely possibility that certain observations in the TEDS-A data set are from the same patients (since some people may be admitted to a treatment facility more than once during a calendar year). In this project, analyses will proceed while assuming observations are independent, as there is no recorded data to suggest the extent to which observations may be correlated. However, perhaps a future project may be to apply a newly-proposed approach to logistic regression with dependent observations that involves gradient descent[2].

I. Statistical Software

R will be used to perform the aforementioned statistical analyses.

[2] Daskalakis, C.; Dikkala, N.; Panageas, I. (2019). Regression from dependent observations. *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of* Computing, Association for Computing Machinery: Phoenix, AZ, USA, 881-889.

V. Results

Note: after removing all instances of missing data, the total number of observations was reduced from 1,935,541 to 1,209,722.

Research Question 1

Our fitted model is the following:

$$\widehat{Odds(p_\iota)} = \exp\{-0.138 + 0.428 * 1_{\{psyprob\}} - 0.003 * 1_{\{male\}} + 0.496 * 1_{\{age:21-29\}} \\ + 0.424 * 1_{\{age:30-39\}} + 0.184 * 1_{[age:40-49]} - 0.108 * 1_{\{age:50+\}}\}$$

where the variables represent the following:

- p_i: probability of being admitted for use of more than one substance (1 = Yes, 0 = No)
- $1_{\{psyprob\}}$: an indicator variable for co-occurring mental and substance use disorder status, where 1 = Yes, 0 = No
- $1_{\{male\}}$: an indicator variable for sex, where 1 = Male, 0 = Female
- $1_{\{age:21-29\}}$: an indicator variable for the age bin 21-29 years old, where 1 = Yes, 0 = No
- $1_{\{age:30-39\}}$: an indicator variable for the age bin 30-39 years old, where 1 = Yes, 0 = No
- $1_{[age:40-49]}$: an indicator variable for the age bin 40-49 years old, where 1 = Yes, 0 = No
- $1_{\{age:50+\}}\}$: an indicator variable for the age bin 50+ years old, where 1 = Yes, 0 = No
- $e^{\widehat{\beta_0}} = 0.871$: the intercept represents a female aged 12-20 years old who does not have a co-occurring mental and substance use disorder

Below is a table of all odds ratios, 95% confidence intervals (C.I.'s), and p-values. Note that a signficance level of α = 0.05 is used to deduce statistical significance.

Variable	Odds Ratio	95% C.I.	P-Value
Intercept	0.871	(0.857, 0.855)	**< 0.001**
Co-occurring mental and substance use disorders	1.534	(1.522, 1.545)	**< 0.001**
Sex	0.997	(0.989, 1.005)	0.415
Age: 21-29 years	1.642	(1.615, 1.669)	**< 0.001**
Age: 30-39 years	1.529	(1.504, 1.554)	**< 0.001**
Age: 40-49 years	1.202	(1.181, 1.222)	**< 0.001**
Age: 50+ years	0.898	(0.883, 0.913)	**< 0.001**

Table 1: Research Question 1 Reported Statistics

Note: All covariates have a p-value of < 0.001 and hence are statistically significant, except for sex, where the p-value was 0.415.

When comparing two populations of the same sex and age group but that differ in co-occurring mental and substance use disorders status, we estimate that the odds of being admitted to a substance abuse facility for the use of more than one substance is 1.534 times greater (95% C.I.: 1.522, 1.545) for the population that has co-occurring mental and substance use disorders compared to the population that does not have co-occurring mental and substance use disorders. We reject the null hypothesis that there is no association between odds of being admitted for use of more than one substance and co-occurring mental and substance use disorders (p < 0.001).

Research Question 2

Our fitted model is the following:

$$\begin{aligned}\widehat{Odds}(p_i) = \exp\{&0.161 + 0.417 * 1_{\{psyprob\}} - 0.020 * 1_{\{male\}} + 0.502 * 1_{\{age:21-29\}} \\ &+ 0.422 * 1_{\{age:30-39\}} + 0.173 * 1_{\{age:40-49\}} - 0.124 * 1_{\{age:50+\}} \\ &- 0.259 * 1_{\{dependent\ living\}} - 0.353 * 1_{\{independent\ living\}} \\ &+ 0.113 * 1_{\{psyprob\}} * 1_{\{dependent\ living\}} \\ &- 0.027 * 1_{\{psyprob\}} * 1_{\{independent\ living\}}\}\end{aligned}$$

where the variables represent the following:

- p_i: probability of being admitted for use of more than one substance (1 = Yes, 0 = No)
- $1_{\{psyprob\}}$: an indicator variable for co-occurring mental and substance use disorder status, where 1 = Yes, 0 = No
- $1_{\{male\}}$: an indicator variable for sex, where 1 = Male, 0 = Female
- $1_{\{age:21-29\}}$: an indicator variable for the age bin 21-29 years old, where 1 = Yes, 0 = No
- $1_{\{age:30-39\}}$: an indicator variable for the age bin 30-39 years old, where 1 = Yes, 0 = No
- $1_{\{age:40-49\}}$: an indicator variable for the age bin 40-49 years old, where 1 = Yes, 0 = No
- $1_{\{age:50+\}}$: an indicator variable for the age bin 50+ years old, where 1 = Yes, 0 = No
- $1_{\{dependent\ living\}}$: an indicator variable for the living arrangements level being dependent living, where 1 = Yes, 0 = No
- $1_{\{independent\ living\}}$: an indicator variable for the living arrangements level being independent living, where 1 = Yes, 0 = No
- $1_{\{psyprob\}} * 1_{\{dependent\ living\}}$: the interaction term for having co-occurring mental and substance use disorders, and experiencing dependent living
- $1_{\{psyprob\}} * 1_{\{independent\ living\}}$: the interaction term for having co-occurring mental and substance use disorders, and experienceing independent living

- $e^{\widehat{\beta_0}} = 1.174$: the intercept represents a houseless female aged 12-20 years old who does not have co-occurring mental and substance use disorders

Below is a table of all odds ratios, 95% confidence intervals (C.I.'s), and p-values. Note that a signficance level of α = 0.05 is used to deduce statistical significance.

Variable	Odds Ratio	95% C.I.	P-Value
Intercept	1.174	(1.150, 1.199)	**< 0.001**
Co-occurring mental and substance use disorders	1.518	(1.490, 1.547)	**< 0.001**
Sex	0.981	(0.973, 0.988)	< **0.001**
Age: 21-29 years	1.651	(1.624, 1.679)	< **0.001**
Age: 30-39 years	1.525	(1.500, 1.550)	< **0.001**
Age: 40-49 years	1.188	(1.168, 1.209)	**< 0.001**
Age: 50+ years	0.883	(0.868, 0.899)	**< 0.001**
Dependent living	0.772	(0.759, 0.785)	**< 0.001**
Independent living	0.703	(0.693, 0.713)	**< 0.001**
Co-occurring mental and subustance use disorders * Dependent living	1.120	(1.091, 1.149)	**< 0.001**
Co-occurring mental and subustance use disorders * Independent living	0.973	(0.953, 0.994)	**0.011**

Table 2: Research Question 2 Reported Statistics

Note: all but one covariates have a p-value of < 0.001. The p-value for the second interaction term ($1_{\{psyprob\}} * 1_{\{independent\ living\}}$) is 0.011. At a significance level of α = 0.05, all covariates are statistically significant.

When comparing people in dependent living to houseless people of the same sex and age group, the estimated odds ratio of being admitted for more than one substance for two populations that differ in co-occurring mental and substance use disorders status is 1.120 (95% C.I.: 1.091, 1.149), with the dependent living folks experiencing a higher odds of admittance for more than one substance.

When comparing people in independent living to houseless people of the same sex and age group, the estimated odds ratio of being admitted for more than one substance for two populations that differ in co-occurring mental and substance use disorders status is 0.973 (95% C.I.: 0.953, 0.994), with the independent living folks experienceing a smaller odds of admittance for more than one substance.

To determine whether or not including the effect modifier allows for a better-fitting model, a likelihood ratio test was performed, where the model in Question 1 was used as the nested model for comparison, and the model in Question 2 was used as the more complex model. The likelihood ratio test gave a p-value of < 0.001. Thus, at a significance level of α = 0.05, we conclude that the more complex model, which includes living arrangements as an effect modifier, is a better fit.

Research Question 3

Our fitted model is the following:

$$\begin{aligned}\widehat{Odds(p_\imath)} = \ \exp\{&-1.707 + 3.353 * 1_{\{opioid\ therapy\}} - 0.133 * 1_{\{male\}} + 1.576 * 1_{\{age:21-29\}} \\ &+ 1.519 * 1_{\{age:30-39\}} + 1.084 * 1_{[age:40-49]} + 0.792 * 1_{\{age:50+\}} \\ &- 0.011 * 1_{\{dependent\ living\}} - 0.228 * 1_{\{independent\ living\}}\}\end{aligned}$$

where the variables represent the following:

- p_i: probability of being admitted for opioid misuse (1 = Yes, 0 = No)
- $1_{\{opioid\ therapy\}}$: an indicator variable for whether or not a patient is on medication-assisted opioid therapy, where 1 = Yes, 0 = No
- $1_{\{male\}}$: an indicator variable for sex, where 1 = Male, 0 = Female
- $1_{\{age:21-29\}}$: an indicator variable for the age bin 21-29 years old, where 1 = Yes, 0 = No
- $1_{\{age:30-39\}}$: an indicator variable for the age bin 30-39 years old, where 1 = Yes, 0 = No
- $1_{\{age:40-49\}}$: an indicator variable for the age bin 40-49 years old, where 1 = Yes, 0 = No
- $1_{\{age:50+\}}$: an indicator variable for the age bin 50+ years old, where 1 = Yes, 0 = No
- $1_{\{dependent\ living\}}$: an indicator variable for the living arrangements level being dependent living, where 1 = Yes, 0 = No
- $1_{\{independent\ living\}}$: an indicator variable for the living arrangements level being independent living, where 1 = Yes, 0 = No
- $e^{\widehat{\beta_0}} = 0.181$: the intercept represents a houseless female aged 12-20 years old who is not on medication-assisted opioid therapy

Below is a table of all odds ratios, 95% confidence intervals (C.I.'s), and p-values. Note that a signficance level of α = 0.05 is used to deduce statistical significance.

Variable	Odds Ratio	95% C.I.	P-Value
Intercept	0.181	(0.177, 0.186)	**< 0.001**
Medication-assisted opioid therapy	28.575	(28.065, 29.094)	**< 0.001**
Sex	0.875	(0.868, 0.883)	**< 0.001**
Age: 21-29 years	4.836	(4.725, 4.950)	**< 0.001**
Age: 30-39 years	4.569	(4.464, 4.676)	**< 0.001**
Age: 40-49 years	2.957	(2.887, 3.030)	**< 0.001**
Age: 50+ years	2.207	(2.154, 2.262)	**< 0.001**
Dependent living	0.989	(0.975, 1.002)	0.103
Independent living	0.796	(0.787, 0.805)	**< 0.001**

Table 3: Research Question 3 Reported Statistics

Note: all but one covariates have a p-value of < 0.001. The p-value for the covariate $1_{\{dependent\ living\}}$ is 0.103. Hence, at a significance level of $\alpha = 0.05$, all covariates except for $1_{\{dependent\ living\}}$ are statistically significant.

When comparing two populations of the same sex, age group, and living arrangmenets but that differ in medication-assisted opioid therapy status, we estimate that the odds of being admitted for opioid abuse is about 28.575 times greater (95% C.I.: 28.065, 29.094) for the population that is on medication-assisted opioid therapy compared to the population that is not on medication-assisted opioid therapy. We reject the null hypothesis that there is no association between odds of being admitted for opioid abuse and medication-assisted opioid therapy ($p < 0.001$).

When comparing two populations of the same sex, age group, and medication-assisted opioid therapy status but that differ in living arrangements (either independent living or houseless), we estimate that for the independent living population, the odds of being admitted for opioid abuse is about 0.796 times (95% C.I.: 0.787, 0.805) that of the houseless population. We reject the null hypothesis that there is no association between odds of being admitted for opioid abuse and whether or not someone experiences independent living or is houseless ($p < 0.001$).

VI. Assumptions, Limitations, Future Suggestions

Statistical Model Assumptions

1. Independent Observations

Note that in each multiple logistic regression model, it is assumed that observations are independent. However, it is very important to note that this assumption realistically may not be met due to the inherent nature of the data set. The data set includes admittance occurences, without noting specific individuals. Hence, some observations within the same calendar year may be the same person, as some people may be admitted to a substance abuse facility more than once in the same year. Unfortunately, there is no way of knowing which observations may in fact be correlated, and hence it is impossible to verify whether or not the independent observations assumption is valid. Note that failure to meet this assumption does not impact the validity of the coefficient estimates in the logistic regression models. Instead, this may lead to inaccurate standard error calculations, which could affect confidence interval calculations. However, it is also possible that this assumption may not be too problematic if a small proportion of observations are actually correlated.

2. No Multicollinearity

The logistic regression model also assumes that there is no multicollinearity between the predictors in a given model. Note that in each statistical model, age and sex were included as potential confounders that are generally not correlated (in other words, there are people of all ages that are both male and female). In the third model, where living arrangements is included as another confounder (but not as an effect modifier, as in the second model), living arrangements may be correlated with age, as younger people are generally more likely to be in dependent living rather than in independent living. To see whether or not multicollinearity is an issue in this model, the Variance Inflation Factor (VIF) was calculated for all predictors in this model, and the following table shows those results:

Predictor	VIF
Medication-Assisted Opioid Therapy	1.02
Age	1.04
Sex	1.01
Living Arrangements	1.04

Table 4: VIF Values for Model 3

We see that all VIF values are very close to 1, and hence it is unlikely that multicollinearity is a concern.

3. Sufficiently Large Sample Size

Another assumption that is needed in order to provide correct p-values is that the sample size is large enough. It is likely that the sample size in this project of 1,209,722 is large enough to result in accurate p-values and hence accurate hypothesis testing conclusions.

4. True Model for Data is Correct

This assumption implies that the true data generating model involves binomially-distributed observations with the correct model for the probability of any of the three events that were investigated. Note that the true data generating model can never be known, and hence it is incredibly unlikely that this assumption could ever be verified. However, it is still important to note that this should not take away from the utility of fitting a multiple logistic regression model (or any statistical model, for that matter). Hence, this assumption is not so concerning and does not lend itself to a large limitation of the analyses performed in this project.

5. No Other Confounding Variables

In a multiple logistic regression model, it is assumed that all potential confounding variables have been included. However, this is almost never the case in reality, and these analyses are no exception. This is of particular relevance in the third statistical model, where a very large odds ratio is reported for comparing populations who are or are not on medication assisted opioid therapy and their probabilities of being admitted to a substance abuse facility for opioid use. This large odds ratio does not necessarily mean that medication-assisted opioid therapy is ineffective! There are certain counounders at play that were not included in the model, such as medication-assisted opioid therapy *start date*. It is likely that a person who has just started therapy is at a much higher risk of being admitted, as they are still in need of aid, while a person who is near the end of their therapy is at a lower risk of being admitted. While this result turns out to be rather uninformative, the unrealistic nature of this assumption still should not take away from the utility of fitting a regression model, or take away from the results and interpretations. While these results were not helpful in asnwering the third research question, this multiple logistic regression model was very useful in determining a second target population, which was another primary aim. Moreover, as Dr. George Box says, "all models are wrong"[3], but that should not dissuade us from attempting to approximate the truth!

Limitations

1. Missing Data

The biggest limitation in this project is the issue of missing data. Note that 725,819 observations were omitted from the analyses, which is 37.5% of the total number of observations. While the subset of observations used was still a sufficiently large number, this does not mean that the 725,819 omitted observations were not consequential. Additionally, 37.5% is a moderately large proportion to be omitting, and hence omitting these observations may lead to the results not adequately representing the entire study population.

2. Extrapolation

There are two main issues with extrapolating the results. First, note that the data set only includes data from publicly-funded substance abuse facilities, not privately-funded institutions. Hence, the results cannot be extrapolated to the entire population in the United States who is admitted to a substance abuse facility, since it is not appropriate to equate admitted people at publicy-funded institutions to admitted people at privately-funded institutions. The latter may be harder to access by the most at-risk populations, and hence the

[3] Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association, 71*(356), 8.

results in this analysis likely do not accurately represent admissions at a privately-funded substance abuse facility. The second issue ties back to the improbability of independent observations. Since some observations are likely highly correlated (since multiple admissions may represent the same patient), this means certain patients and their demographics and substance use history may carry more weight in the statistical analyses and consequently skew our results, not painting as accurate of a picture between certain response variables and predictors. This impacts extrapolation because certain patient attributes may be disproportionately represented, and hence the results may not apply to as wide of a range of people as initially desired when trying to answer these three research questions.

Future Suggestions

1. Method for Correlated Data

It would be interesting to fit a different model that does not assume that observations are independent for each research question, and compare the results to those in this project to see if there are any drastic differences. Using a method for correlated data would likely give more accurate inferences; however if the results are not so different from the results given by the multiple logistic regression models, then it is likely safe to assume that the independent observations assumption is not of big concern.

2. Further Investigate Missing Data

There are still many questions surrounding the degree of severity that the issue of missing data presents. It would be beneficial to investigate whether or not the demographics and substance use characteristics are the same or different between those with missing data and those without missing data. For instance, are there large differences between the cases with missing living arrangements data and the cases that report this data? How about for medication-assisted opioid therapy? For co-occurring mental and substance use disorders? If the cases are generally not so different, then perhaps the issue of missing data is not so bothersome. However, if the cases differ to a considerable extent, then this raises more of a concern in the validity of the results.

VII. Conclusion and Discussion

The main takeaway from this project is the two target populations that were deduced from the statistical inference. Firstly, those with co-occurring mental and substance use disorders should be targeted when giving aid and support. From the statistical analyses in Question 2, we see that indeed, living arrangements is a significant effect modifier in the relationship between co-occurring mental and substance use disorders and being admitted to a facility for the use of more than one drug. Hence, when we think about giving aid to those who experience a co-occurring disorder, we must approach populations differently depending on their living situation. How we give aid to someone in houseless living who has a co-occurring mental and substance use disorder should be different from how we give aid to someone in independent living who, likewise, has a co-occurring mental and substance use disoder.

Additionally, while the results from Question 3 were unable to answer the original research question regarding the effectiveness of medication-assisted opioid therapy, when investigating those being admitted for opioid use, a second target population was identified: people in houseless living. This is especially important when considering the current state of the opioid crisis in the United States, where deaths from overdose involving opioids have increased over six times since 1999[4]. While this is very much a data-driven project, it is important to remember that behind these numbers are real, at-risk people, who are dying at increasingly alarming rates. Aid is a necessity, and hopefully this project can help future policy or grassroots organizing in focusing efforts towards populations that are in need of more support.

[4] Centers for Disease Control and Prevention. (2021, March 25). *Data Overview*. Centers for Disease Control and Prevention. https://www.cdc.gov/drugoverdose/data/index.html.

References

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association, 71*(356), 8.

Centers for Disease Control and Prevention. (2021, March 25). *Data Overview*. Centers for Disease Control and Prevention. https://www.cdc.gov/drugoverdose/data/index.html.

Daskalakis, C.; Dikkala, N.; Panageas, I. (2019). Regression from dependent observations. *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of* Computing, Association for Computing Machinery: Phoenix, AZ, USA, 881-889.

NIDA. (2005, June 1). Drug Abuse and Addiction: One of America’s Most Challenging Public Health Problems. https://archives.drugabuse.gov/publications/drug-abuse-addiction-one-americas-most-challenging-public-health-problems.

SAMHDA (2018). Treatment Episode Data Set: Admissions (TEDS-A).