Developing a Composite Endpoint for Substance Abuse Trials

Lakhvir Atwal, Janice Ferrer, Bryan Ho, Paul Lin, Melissa Myers

20 May, 2021

Abstract

The ultimate goal of a clinical study is often to demonstrate the efficacy and safety of a new treatment or medical intervention. This goal is highly dependent on the choice of a clinically relevant and valid primary endpoint. An ideal design of a clinical trial begins with the selection of a single primary end point that completely characterizes the medical outcome of interest and allows for an efficient statistical evaluation of treatment effect. Because disorders and treatments are often complex, their manifestations are often multifaceted. As such, the selection of a single primary endpoint may not be possible, and so researchers often utilize composite endpoints for characterizing treatment effectiveness. By opting to use a composite endpoint in a clinical study, researchers can achieve higher statistical precision and thus, increase the power of the study and allow for smaller sample size. This paper aims to develop a composite endpoint and assesses its use as a valid measure of efficacy in a clinical trial of buprenorphine treatment for opioid abuse.

1 Introduction

1.1 Clinical Trials: Background

Clinical research involves investigating proposed medical treatments, assessing the relative benefits of competing therapies, and establishing optimal treatment combinations. A clinical trial is an experiment testing medical treatments on human subjects. The clinical investigator controls factors that contribute to variability and bias such as the selection of subjects, application of the treatment, evaluation of outcome, and methods of analysis. If the treatment can be applied uniformly and potential biases are controlled, clinical trials are imperative to demonstrate the efficacy and safety of a medical therapy (1,2,3,6). Because clinicians have to fulfill ethical obligations to patients seeking help, researchers often employ what is called an "active control trial". Although efficacy is generally best evaluated by comparing a treatment to a placebo, active control trials are used in situations where withholding treatment from individuals by assigning them to a placebo group is unethical (30). Thus, comparison is made between the active control, usually in the form of a standard treatment, and the experimental therapy (4,5). Clinical trials that evaluate the effectiveness of drugs for treating opiate dependency have utilized "reduction in drug-taking behavior" (drug use patterns) as an endpoint. However, there is great interest in further developing clinically meaningful endpoints for use in substance abuse trials.

1.2 Substance Abuse

In a clinical trial, primary endpoints are typically used as an efficacy measure, and are used to answer the primary or most important question in the trial. They should be a direct measure of the treatment effects being studied. The paper "Determining the primary Endpoint for Stimulant Abuse Trial: Lessons Learned from STRIDE", serves as a great literature review for primary endpoints in substance use disorder (SUD) trials (9). Stimulant Reduction Intervention using Dosed Exercise (STRIDE) is a trial that is to be conducted. Stimulant abuse is a relapsing illness that does not have many efficacious treatments. A variety of endpoints are required for research to treat individuals with SUD. Although long term abstinence is the preferred clinical goal, the lack of efficacious treatments can make this an unrealistic goal. It has been observed that long-term continuous abstinence is not the best endpoint to select for a clinical trial. However,

an endpoint must be selected to determine the efficacy of available treatment. Typically in this field, chosen endpoints are believed to be clinically meaningful toward the goal of long-term abstinence. Researchers have included components such as: reduction in use and increased days of abstinence. Measurement approaches for these components vary as well including: self-reported rates of abstinence and toxicology reports. This paper also serves as a framework for selecting a primary endpoint in a SUD trial.

To choose a primary endpoint we must look at the intent of the study, as well as any existing options in literature. A thorough literature review is necessary to find any available endpoints for a specific population of interest or related to the intervention. These candidate endpoints should serve the intent of the study. And finally, we must determine whether the endpoint is clinically meaningful. This is an iterative process that requires constant reevaluation. Figure 1 depicts a flow diagram with the steps needed to decide on a primary endpoint for a clinical trial.

Following Figure 1, we must first identify the intent of the study. With the intent of the study in mind, we must review the literature for any existing endpoints in use. Upon the literature review, researchers found that there is no golden standard endpoint used in SUD trials. However, a meta-analysis of trials found that in abstinence, reported either through self-report or urine drug screens or a combination of the two, was a common endpoint. Researchers found that using the percent of days of abstinence was the most reasonable primary endpoint. The most common outcome measures utilized are urine screens and self-reported outcomes. The researchers decided to use a self-report component to assess the percent days abstinent, using a Time Line Follow Back. TLFB is an interview conducted in a nonjudgmental manner with no adverse consequences for disclosure of use. A potential disadvantage of the TLFB is inaccuracy due to unreliable memory; thus, the patients are supplemented with a substance use diary to help account for any substance abuse. The TLFB has shown to correlate well with urine drug screens, which researchers have claimed to be clinically meaningful. Establishing a clinical meaningful endpoint can potentially guide future clinical trials in similar fields.

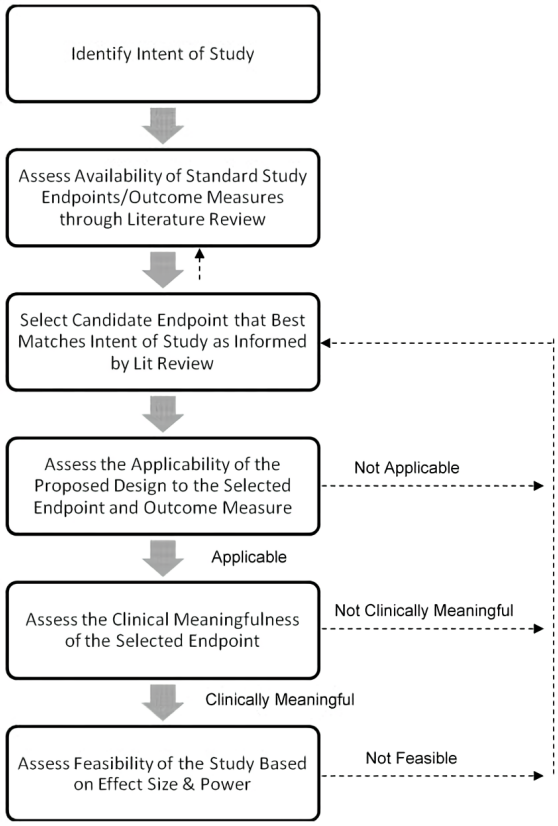


Figure 1: Flow diagram of the steps needed to decide on a primary endpoint for a clinical trial

2 Protocol Overview

2.1 Objective of the Study

This project is based on a 16-week maintenance study through a clinical study protocol set in the early 1990's (February 7, 1992). The main objective of this study is to determine the safety and effectiveness between two treatment doses for sublingual buprenorphine: 8 mg per day in comparison to a 1 mg per day in lowering the illicit use of opioids. To measure the use of opioids in patients, there are urine testings done 3 times a week, retention rates, opiate craving scores and global rating scores recorded for each patient who meet the DSM-III-R criteria for opiate dependence. The secondary objective of the study is to collect more experience with two other doses for buprenorphine: 4 mg per day and 16 mg per day, for the purpose of safety in the same population.

2.2 Science and Background of Buprenorphine in Treating Opioid Addiction

Addiction to opioids has been an increasing battle since the early 20th century, and thus there has been a quest for nonaddictive analgesic to treat opioid disorder among researchers. In the 1960's, the opiate use has even reached epidemic proportions in the United States which stimulated the search even more for a treatment for opiate dependence (15). Since then, the synthetic opiate agonist methadone hydrochloride was used to maintain opioid use disorder. This maintenance medication works by acting on opioid receptors in the brain but does so more slowly so that it does not produce euphoria. Other maintenance treatments for opiate use prescribed include LAAM (Levo alpha acetylmethadol), another opiate agonist similar to methadone, naltrexone (opiate antagonist), and buprenorphine, the focus treatment of this study.

Buprenorphine is a partial agonist analgesic that was discovered in 1966 and has been going through clinical investigation for opioid dependence treatment since then. As a partial opioid agonist, this means that it binds to the same receptors that other opioids such as heroin activate but activates them less strongly than a full agonist such as methadone does (16). When considering the route of administration for this study, it is studied that "buprenorphine is pharmacologically active when given by the subcutaneous, sublingual and oral routes of administration" (21). Out of these routes of administration, a study from Reckitt and Colman in 1987 shows that when buprenorphine is given sublingually, the drug is readily absorbed with approximately 50% of a dose absorbed systemically. There are also a few common adverse effects reported with sublingual administration of buprenorphine which include drowsiness, dizziness, nausea, and vomiting (28).

Buprenorphine has similar characteristics to methadone as it can decrease cravings and withdrawal symptoms in someone with opioid use disorder without producing euphoria. However, unlike methadone, clinical studies suggest that the blockage of opiate agonist effects by buprenorphine induces a low level of opiate physical dependence. Researcher Jasinski has even heralded buprenorphine's potential as it produced long-lasting "changes in feelings that are acceptable to addicts," and was "less toxic than methadone" (12).

In one study where 2 mg of buprenorphine was given subcutaneously to 8 patients in place of their mean daily dose of 36 mg of methadone, only mild discomfort resulted, and after 28 days of buprenorphine administration, abrupt placebo substitution resulted in a mild withdrawal syndrome of several days duration (13). In another study from Kosten and Kleber, 1988, a 30-day, outpatient, open-label clinical trial was done where sublingual doses of 2, 4, and 8 mg of buprenorphine were given to the patients. This clinical trial seemed effective in maintaining abstinence and keeping patients in treatment. Buprenorphine abruptly discontinued at the end of the trial and as a result, those maintained on the 8 mg had considerable rise in withdrawal symptoms whereas those on the 2 and 4 mg had very little withdrawal symptoms (14).

Due to the results of these studies indicating less opiate dependence and effectiveness in maintaining abstinence, it is suggested that buprenorphine could be beneficial for detoxification and maintenance. However, as these trials contain a small number of patients, the results need to be confirmed with controlled trials of larger groups. Thus, our proposed study aims to confirm these results by obtaining a larger and controlled sample size for testing. Also, it is important to note that our proposed study extends the research from a

previous study that resulted in 8 mg of buprenorphine being superior to 20 mg of methadone and similar to 60 mg of methadone in a 180-day detoxification study, which further suggests buprenorphine as a comparable treatment to methadone.

2.3 Study Methods

In this trial, there are at least 480 opiate abusers who are recruited by 10-12 study sites with a maximum of 60 patients each site. After eligibility screening, patients will be randomly assigned to one of four treatment groups: buprenorphine 1, 4, 8, or 16 mg/day for 16 weeks. Patients will be administered their buprenorphine dose daily which dispense per week. To be an eligible patient, one should satisfy the inclusion criteria specified in the study admission form. In this trial, the investigators strive to have at least a third of the patients in their studies be female, but require that at least 25% of study patients are female.

The randomization will be accomplished by assigning patients to pre-coded medication supplies. To randomize a patient, the investigators will telephone the Data Coordinating Center, who will randomly assign the patient to a non sequential patient number. All staff at participating sites will be blind to patients' buprenorphine doses which will be supplied to the study sites in pre packaged unit doses consisting of 1 ml liquid containing 1, 2, 4, 8, 12 or 16 mg of buprenorphine. Patients assigned to 1 mg of buprenorphine will receive 1 mg per day for the trial duration of 16 weeks. Those patients assigned to more than 1 mg of buprenorphine will be inducted by receiving 2 mg on day one, 4 mg on day two, 8mg on day three, 12 mg on day four, and 16 mg per day for the rest of the study, which depends on which dosage of buprenorphine patients are assigned. Patients who miss more than four sequential days of dosing will be reinducted on buprenorphine. Patients' medication supplies will contain supplies for 3 reinduction cycles. The reinduction dose schedule is the same as the initial induction. Patients are in the study voluntarily and can terminate their participation at any time. The following are possible reasons a patient can be terminated: medication toxicity; missing 7 consecutive days of dosing; buprenorphine toxicity, if patient requires a fourth induction; intercurrent illness or medical complications, administrative termination, and pregnancy.

2.4 Study Measures

There are four major efficacy measures, which are: urine samples for drugs of abuse, days of retention in treatment, opiates and cocaine craving scores, and global rating scores. There also are other measures which may relate to the major attributes to provide the additional evidence to illustrate the result, which are serum plasma levels of buprenorphine, missed doses, adverse experiences, a complete physical examination, the amount of psychosocial treatment, pregnancy testing, laboratory data.

First for major measures: urine samples will be collected under observation on Monday, Wednesday, and Friday for each patient. If the patient fails to give a sample on due, it will be recorded as missing. Urine samples will be sent to a central laboratory (University of Utah) to be analyzed for morphine and cocaine or metabolites. An opiate or cocaine positive will be morphine or cocaine greater than 300 ng/ml. Next, days of retention in treatment is an important measure of treatment effectiveness. If a patient is terminated for any reason, the date of the last buprenorphine dose will be considered the termination date. Next is opiate and cocaine craving scores, the patient will be instructed to record the peak craving that has occurred during the past 7 days. And the last one global rating scores are an overall rating of the patient's status and compare to both his/her status at the previous rating. These scores will be completed by both the patients and the staff.

Secondary measures such as serum plasma levels of buprenorphine will be collected at weeks 2 and 8; Each missed dose will be recorded by each study site. Adverse experiences will be evaluated at screening and weekly. A complete physical examination will be done at baseline and termination and recorded on the medical monitoring form.

2.5 Study Data Components & Wrangling

In this section we introduce the relevant data that was collected on each patient that was randomized into the study's 4 treatment groups ($n = 736$). The data was initially recorded on a series of paper case report forms and later translated into `.csv` files.

Although patient ID's are used throughout the files to identify and track information for each patient, the patient ID's do not come pre-packaged with their respective treatment group for each data set. Instead, a single file `TXGRP.csv` contains the patient ID treatment group assignment. In order to have this information for each data set, we created an `add_group` function in `R` that adds a column of corresponding treatment groups using `TXGRP.csv` to any inputted data. The function also removes columns with all `NA` values from the input data.

The following briefly describes each data set and their usage in the clinical study. Of the 33 `.csv` files obtained from the protocol, we introduce 13 data sets relevant to the 16-week maintainence treatment of interest.

2.5.1 Background Information

This data set is comprised of demographic information of the sample collected. There are categorical variables including race, gender, education level, work history, marital status and living arrangements. Age and income are the only two quantitative variables in this data set. Data is for screening only.

2.5.2 Drug Use History

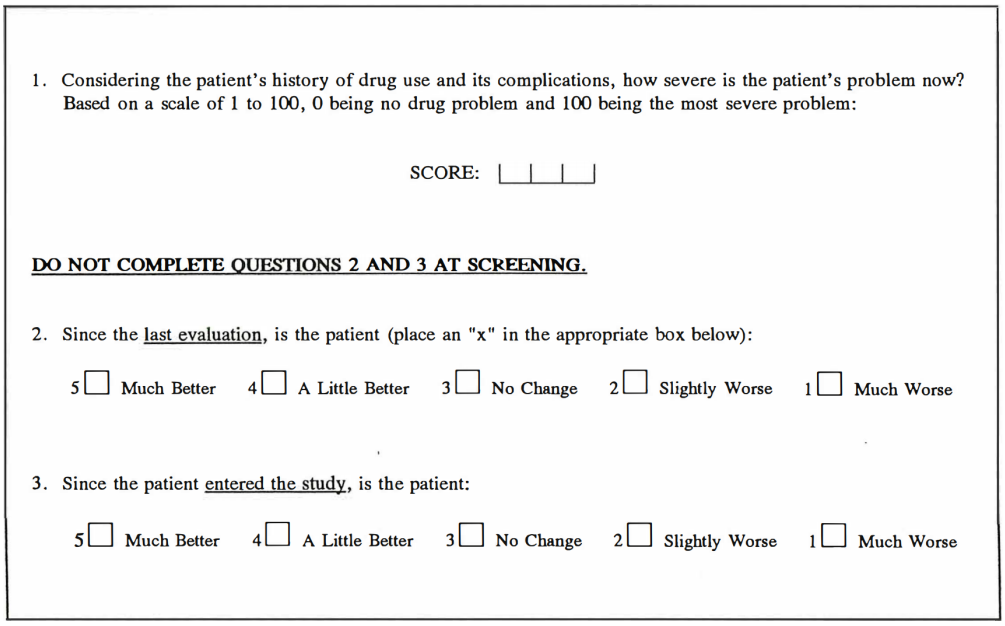
Patients' use of 7 different types of drugs and their mode of use (i.e. smoke or injection) is recorded here as well as the time frame of use (quantitative) of each drug. Data is for screening only.

2.5.3 DSM-III-R Criteria for Diagnosis of Opidate Dependence

The DSM-III-R data includes conditions for diagnosis of opiate dependence. There are a total of ten conditions related to the dependence of the opiate. Some conditions included are: opiates are taken in larger amounts or over longer periods than the person intended, a desire for the drug persists, or the patient has made one or more successful efforts to cut down or to control opioid use. There are 8 more conditions in the form. Depending on whether or not the condition applies to the patient they will either mark yes or no for each condition. Note a minimum of three conditions must apply to the patient for them to be eligible for the study.

2.5.4 Global Rating Scale: Patient & Staff Reported

The global rating scores are one of the major efficacy measures and represent a general rating of the status of the patient in terms of how severe their opiate drug abuse is. These rating scores are collected from both patient and staff at the time of the rating and are compared to both his or her status from the rating of the last completed study and upon entering the study (screening). The current status portion of the forms which asks for the patient's severity on the day of visit is based on a scale from 0 to 100 with 0 defined as an "absence of drug problems" and 100 defined as the "worst ever case". Then for the comparison status portion, which compares the severity from either previous rating period or from screening, is based on a Likert scale from 1 to 5 where 1 is "much worse" and 5 is "much better". The global rating scores are also collected every 4 weeks in a span of the 16-week study. Below are the questions asked on the status of the patient for every rating period (screening, Weeks 4, 8, 12, and 16).



1. Considering the patient's history of drug use and its complications, how severe is the patient's problem now? Based on a scale of 1 to 100, 0 being no drug problem and 100 being the most severe problem:

SCORE: | | |

DO NOT COMPLETE QUESTIONS 2 AND 3 AT SCREENING.

2. Since the last evaluation, is the patient (place an "x" in the appropriate box below):

5☐ Much Better 4☐ A Little Better 3☐ No Change 2☐ Slightly Worse 1☐ Much Worse

3. Since the patient entered the study, is the patient:

5☐ Much Better 4☐ A Little Better 3☐ No Change 2☐ Slightly Worse 1☐ Much Worse

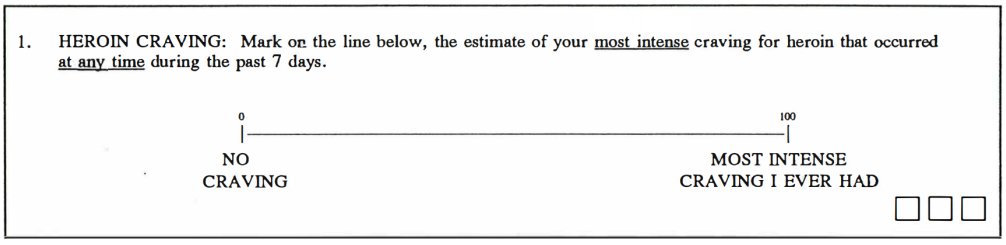
Figure 2: Global Rating Scale (GRS) questionnaire for clinician

2.5.5 Medical History and Status

General health of patients is assessed across 12 different body systems as yes/no questions for abnormalities (12 categorical). Current and ongoing problems and medications are noted (3 categorical). Data is recorded at screening only.

2.5.6 Craving Scale

Another major efficacy measure, the craving scale data is composed of Visual Analogue Scale (VAS) responses taken directly from the patients in the trial. Most commonly utilized in questionnaire forms, the VAS is a psychometric response scale which attempts to measure a patient's attitude towards a treatment on a continuous scale. In this protocol, patients were asked to mark the peak craving for heroin that they had experienced within the last 7 days on a 100mm line once at screening, and then for every week they remained within the study. A mark at the start of the line designates that they felt no craving, whereas a mark at the end of the line designates that they experienced maximum craving. These marked responses were then measured along the 100mm line to quantify a patient's craving for heroin on a 0-100 scale.



1. HEROIN CRAVING: Mark on the line below, the estimate of your most intense craving for heroin that occurred at any time during the past 7 days.

0 — 100

NO CRAVING — MOST INTENSE CRAVING I EVER HAD

☐☐☐

Figure 3: Visual Analogue Scale (VAS) questionnaire for patient

2.5.7 Laboratory Report

The lab report records 29 different specific clinical values for hematology, blood chemistry and urinalysis. These findings are summarised in a final question regarding clinically significant abnormalities (1 categorical). Data is recorded at 2 week intervals from screening to week 16.

2.5.8 Physical Exam

Vital signs (6 quantitative) and external physical health (11 categorical) are recorded for each patient. Pregnancy status and birth control method are recorded for female patients. Data is recorded every 4 weeks from screening through week 16.

2.5.9 Electrocardiogram

Twenty-nine categorical variables recording the presence or absence of conditions are recorded. Overall normal/abnormal results are noted (1 categorical). Data is recorded at screening, week 4 and week 16.

2.5.10 Study Admission

The study admission data includes the inclusion/exclusion criteria required to enter the study. The inclusion criteria includes whether or not a potential patient meets the following: DSM-III-R, availability throughout the duration of the study, mentally competent to give consent, within commuting distance of one of the clinics, and 18 years of age or older. The patient is considered ineligible for the study if they answer no to any of the inclusion criteria. The exclusion criteria includes: pregnant or nursing female, female of childbearing potential who refuses birth control, acute hepatitis, alcohol dependence, daily use of anticonvulsants, enrollment in methadone maintenance, previous subject for a buprenorphine trial, enrollment in another research project. The patient is not eligible for the study if they answer yes to any of the aformentioned exclusion criteria. If the patient is eligible for the study, the date randomized and date of first dose is also included in this data set.

2.5.11 Weekly Self-Report of Drug Use

The weekly self-report collects the frequency of drug use on weekdays and weekends, the dollar amount spent on drugs, and the primary mode of drug use. The categories of drugs contain heroin, cocaine, methamphetamine, alcohol, tranquilizers, marijuana, and PCP.

DRUG	USED DRUG?	WHAT DAYS OF THE WEEK DID YOU USE DRUGS AND HOW MANY TIMES?				TOTAL DOLLAR AMOUNT SPENT	Primary Mode of Abuse 1=Oral 2=I.V. 3=Snorting 4=Smoking 5=Sublingual 6=Other
		Fri. Sat. Sun.	If Yes: # of Times	Mon. thru Thurs.	If Yes: # of Times		
1. Heroin or other opiate	1 ☐ Yes 2 ☐ No	1 ☐ Yes 2 ☐ No	☐☐	1 ☐ Yes 2 ☐ No	☐☐	$ ☐☐☐☐	☐

Figure 4: Drug Use questionnaire for patient

2.5.12 Adverse Events

Any adverse events for each patient are recorded each week for the duration of the study. General health and health concerns are noted. Then specific details of any events are recorded. Events are classified in a

multidimensional manner including: type of report, study relatedness, severity, action taken, and outcome. Specific events, body systems affected, date of onset, and duration are recorded. Medical event codes are recorded but are outdated and unusable. Over 15,000 events are recorded in this data set.

2.5.13 Termination

The termination data lists every patient's date of termination from the study, the primary reason for being terminated, and a Likert scale from much better to much worse of the patient's condition since they entered the study. Note that once a patient completes the 16 week protocol, they are considered terminated from the study. For patients who were not able to complete the full 16 week period, the primary reason for their termination is recorded as well. Termination dates for these patients are considered to be the number of days since randomization into a treatment group to their last dosage of buprenorphine.

3 Exploratory Data Analysis

We will begin our exploratory data analysis by looking over the demographics of all of our patients. This study recruited opioid addicts that were seeking treatment at 12 different clinics in the United States. The final sample contained 736 patients who were successfully randomized into a treatment group. After looking over the baseline characteristics of the patients in the study, we continue with an exploration of the data sets relevant to constructing a composite endpoint.

3.1 Background and Drug Use History

Looking at Figure 5, the racial distribution was approximately 49% Caucasian, 22% African American, and 28% Hispanic. The majority of patients are male, with around a third being female. It is important to note that the proportion of females in each treatment group stems from the requirement of the protocol having at least 25% of their patients to be female. The majority of patients reported completing high school (32%) or some college (31%). Only about 10% were college graduates with the rest completing less than 12 years of school. In the past three years nearly 30% had been unemployed. Approximately 26% were currently married, 30% divorced/separated and 40% had never married. Looking at patient ages, we have categorized the ages in four groups from 18-30, 31-42, 43-54, and those older than 55. We see that a majority of the patients are mid-career (31-42) and only few patients are over the age of 55. The four dosage groups were not significantly different on any of these characteristics.

Overall, the majority of patients have a background of being primarily white, male, graduated from high school, never been gainfully employed, have mean annual family income of approximately $20,000, have never been married, and claim to not use heroin or cocaine in their household. Additionally, we compared the drug use history of each of the four treatment groups (1 Mg, 4 Mg, 8 Mg, and 16 Mg). We can see that all of the patients in the study are opioid users which is essential in observing the effectiveness of buprenorphine for maintenance of opioid abuse disorder. The majority of opioid addicted patients in this study used heroin and other opioids for over 10 years, cocaine for about 7 years, consumed alcohol for about 15 years, and used marijuana for about 12 years of their life.The majority of them also do not typically use methamphetamine and PCP and about half of them have used tranquilizers. Other drug use observed in these patients such as cocaine, methamphetamine, etc. can serve as an additional guide for the inclusion exclusion criteria in obtaining eligible patients for the study. For example, those who may meet the DSM-III-R diagnosis of current alcohol dependence or sedative-hypnotics dependence are excluded from the study. For specific numbers on drug use history and family income, please see the tables in the Appendix.

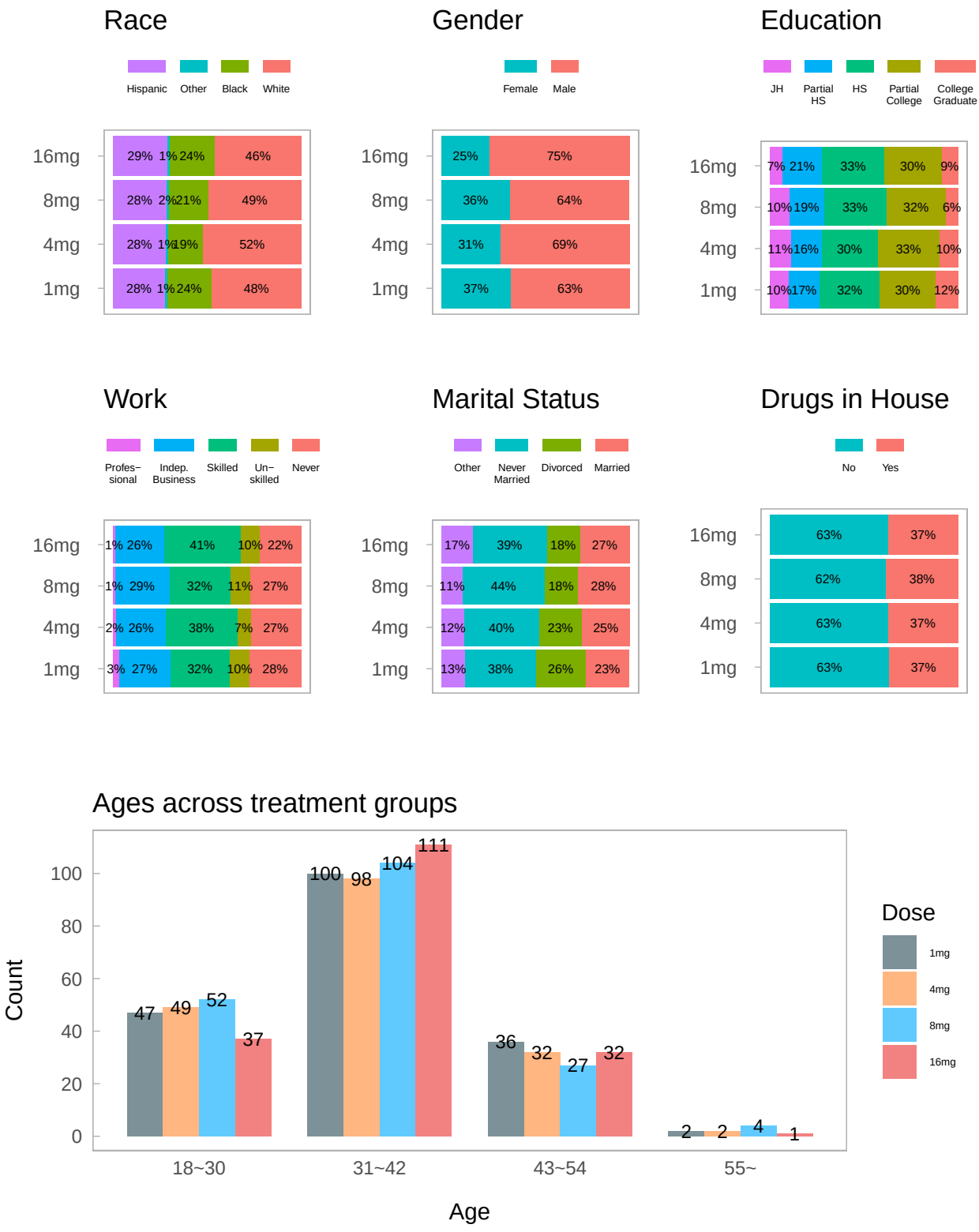


Figure 5: Patient Background Data

3.2 Global Rating Scores

Global Rating Scores are commonly used in clinical research. They are designed to quantify the worsening or improvement of a patient over time. The scale is considered "global" because it allows patients to decide for themselves what they consider important. This is different from outcome measures that are designed to one specific dimension of the patient's health (22). For our study the Global Rating Scale consists of two components: patient reported-outcomes and staff-reported outcomes. Patient reported outcomes are a report of the patient's status without any interpretation of a clinician or anyone else (23). In this study staff will also provide a score for the patient's status, which is the clinician reported outcome.

The first question on the form asks for a score from 0 to 100 for the severity of their drug problem, 0 being no drug problem and 100 being the most severe. This question will be answered at screening as well as weeks: 4, 8, 12, and 16 of the study. Below we have the line graphs of the means of both the staff and patient reported global rating scores. The graphs represent the staff and patient reports, respectively.

Based on Figure 6, the staff (left) reported that the patients in the 16mg group had the lowest drug problem severity score throughout the study followed by the 8mg group, the 4mg group, and finally the 1mg group. In all groups the score drastically dropped from the initial screening to the fourth week. Looking at the patient reported graph (right), the 1mg group retains the highest drug problem severity throughout each week. However, with the patient reported outcomes, the 8mg group had the lowest drug problem severity. In both the patient reported outcome and staff reported outcome, we see a strong drop in the screening score to the fourth week. However, unlike the staff reported mean GRS scores, the patient reported mean GRS scores fluctuate between treatment groups. (i.e. higher dosage does not consistently correspond to lower drug problem severity).

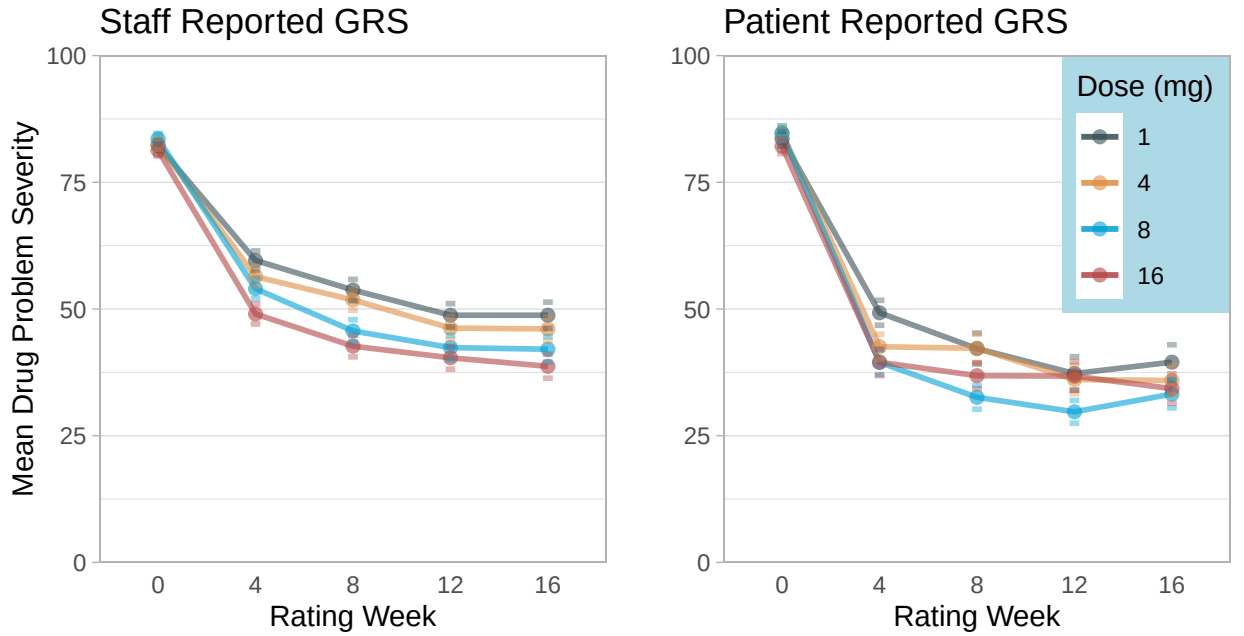


Figure 6: Global Rating Score Data

In Figure 7 we have the percent of patients in each category of the Likert scale at termination with respect to how the staff thought they were doing at baseline. We can see that a majority of our patients have some type of improvement with the combination of blue (a little better) and purple (much better). However, if we take a look at our much better category there are far fewer patients in our active control group in comparison to the other three treatments groups.

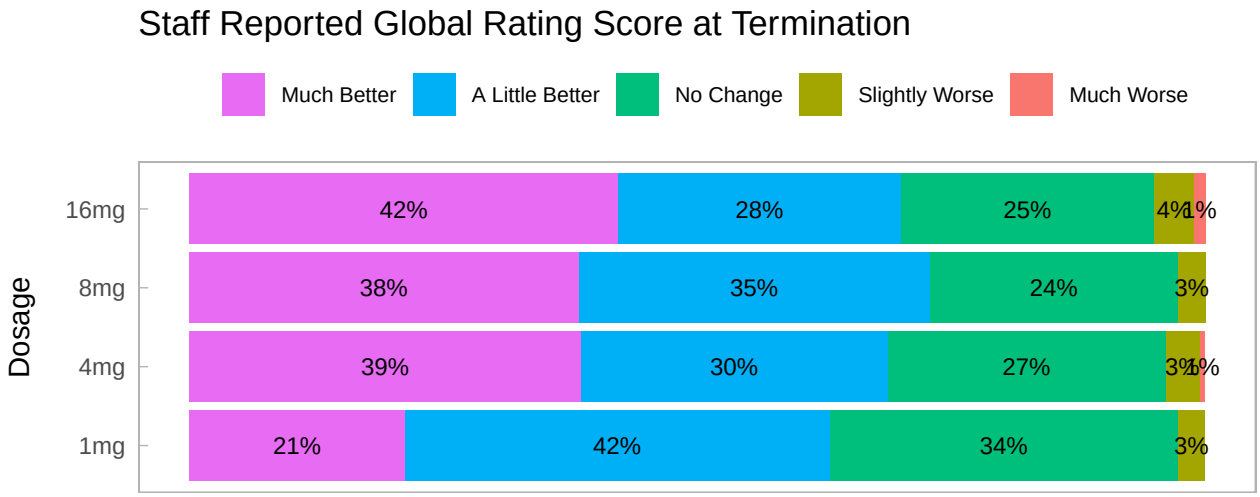


Figure 7: GRS at Termination

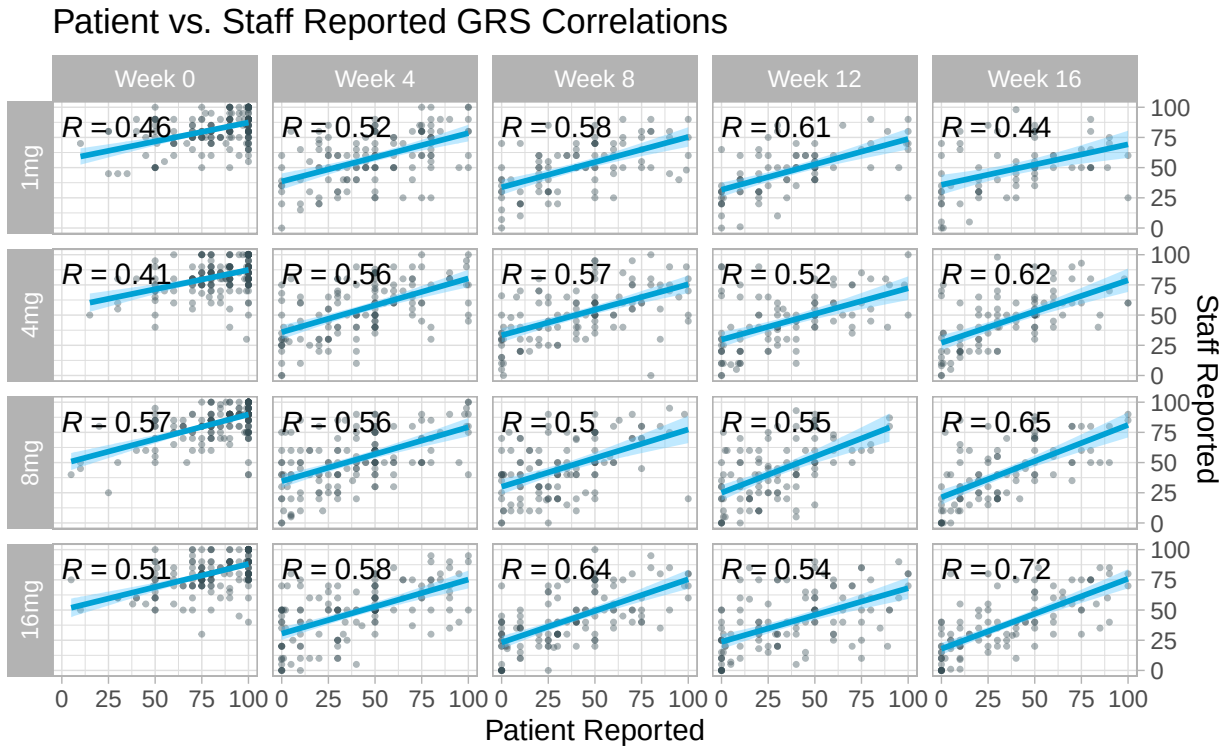


Figure 8: GRS Correlations

We now look at the correlations between the patient and staff reported drug problem severities shown in Figure 8. We observe that the correlations remain similar across treatment groups as well as over time. Although the patients and staff are not necessarily in strong agreement, they still have a moderately positive relationship. Within the graph, we observe that the pairwise observations start out close to (100,100) at screening (week 0) for all dosage groups. Over time, the treatment groups show different movement in the pairwise scores where in general the higher dosages correspond to more decreases in drug problem severity compared to baseline, indicating a decrease in their dependence on opiates. Note that patients are dropping out between each rating period, and we can see visually that the active control/1mg group seems to have sparser pairwise observations by the end of the study than the higher dosage groups. Next, we look more closely at these dropouts in the form of patient retention.

3.3 Patient Retention

In Figure 9 we display the patient retention and the number of patients remaining on the 16-week study at 1 week intervals per treatment group. We note that the active control/1mg group appears to have the worst retention with less than 40% of the original cohort remaining by the end of the study. For the 4mg, 8mg, and 16mg groups, retention remained relatively similar until about day 35, where the 16mg group seems to diverge from the 4mg & 8mg groups by keeping higher retention.

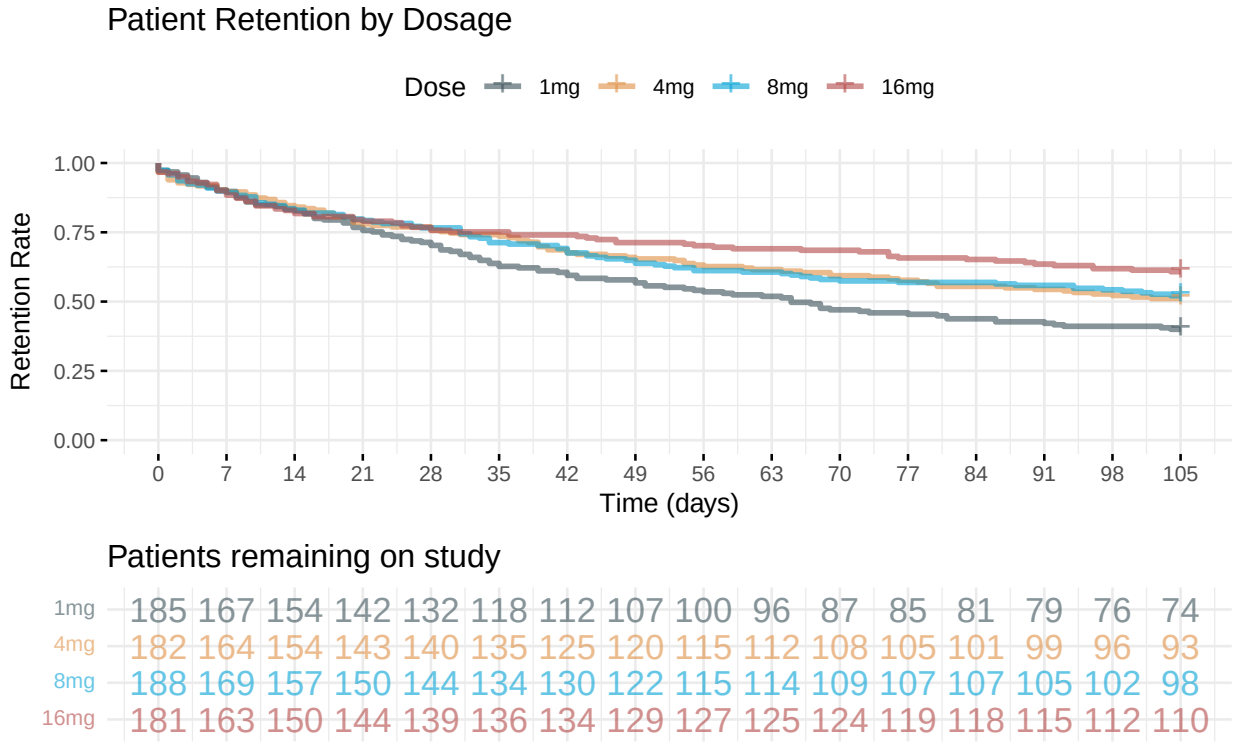


Figure 9: Patient Retention

Of the 736 participants enrolled 375 (51%) completed the study, where completion is defined as remaining in the study for 16 weeks. Completion percentages by treatment group were 40% for 1 mg, 51% for 4 mg, 52% for 8 mg and 61% for the 16 mg group. Figure 10 displays the trial flow of patients in the study and cites reasons for termination. Enrollment failures were patients that were randomized but did not receive any dose of medication.

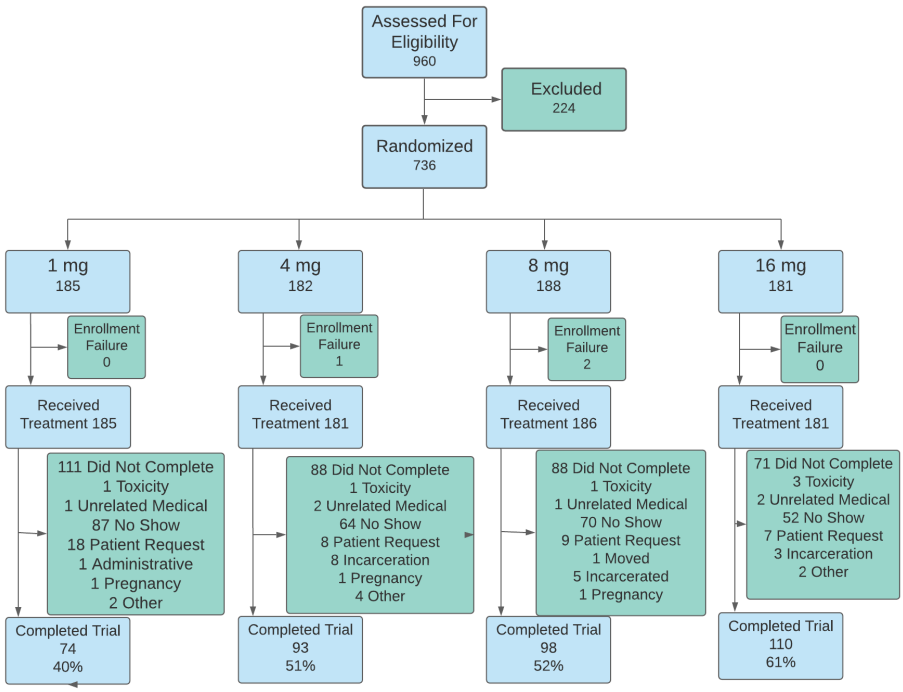


Figure 10: Participant flow chart for clinical trial of Buprenorphine for the treatment of opioid dependence

3.4 Craving Scores

We now attempt to understand the center and spread of the craving scores across groups and over time through Figure 11 below. Note the dramatic drop in craving scores from screening to week 4, as well as the similarity between the groups at week 0. Over time we can see that the 4mg, 8mg, and 16mg groups separate themselves from the active control group consistently throughout the study, with the 4mg & 8mg groups showing consistent decreases. On the other hand, the 16mg group shows the most dramatic decrease, but then does not show much of an improvement as the study continues, even under-performing the 4mg & 8mg groups near the end of the study.

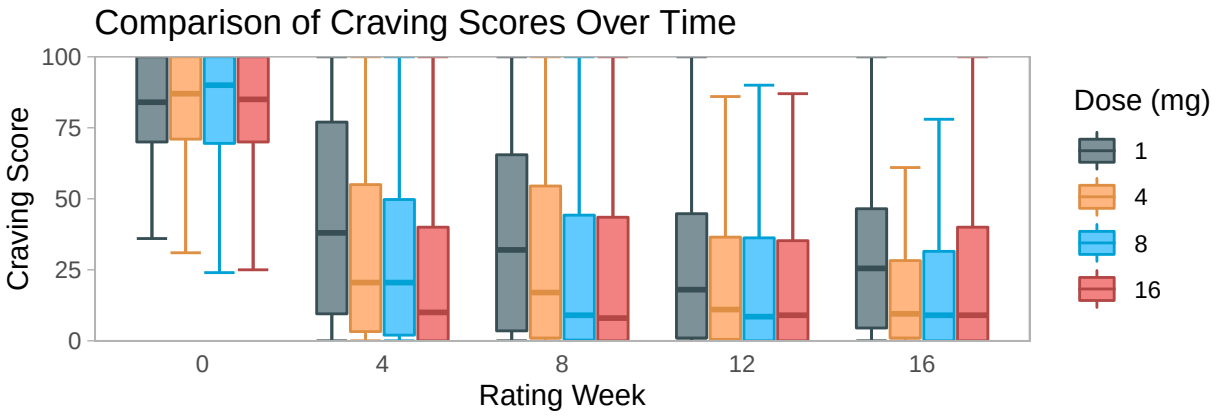


Figure 11: Craving Scores: Boxplot

Looking at the mean craving scores at each week of the study in figure 12, we confirm what we had seen in the boxplot. This alternative view displays that the mean craving scores for the 4mg, 8mg, and 16mg groups are all within each others standard errors, but show clear separation from the active control group.

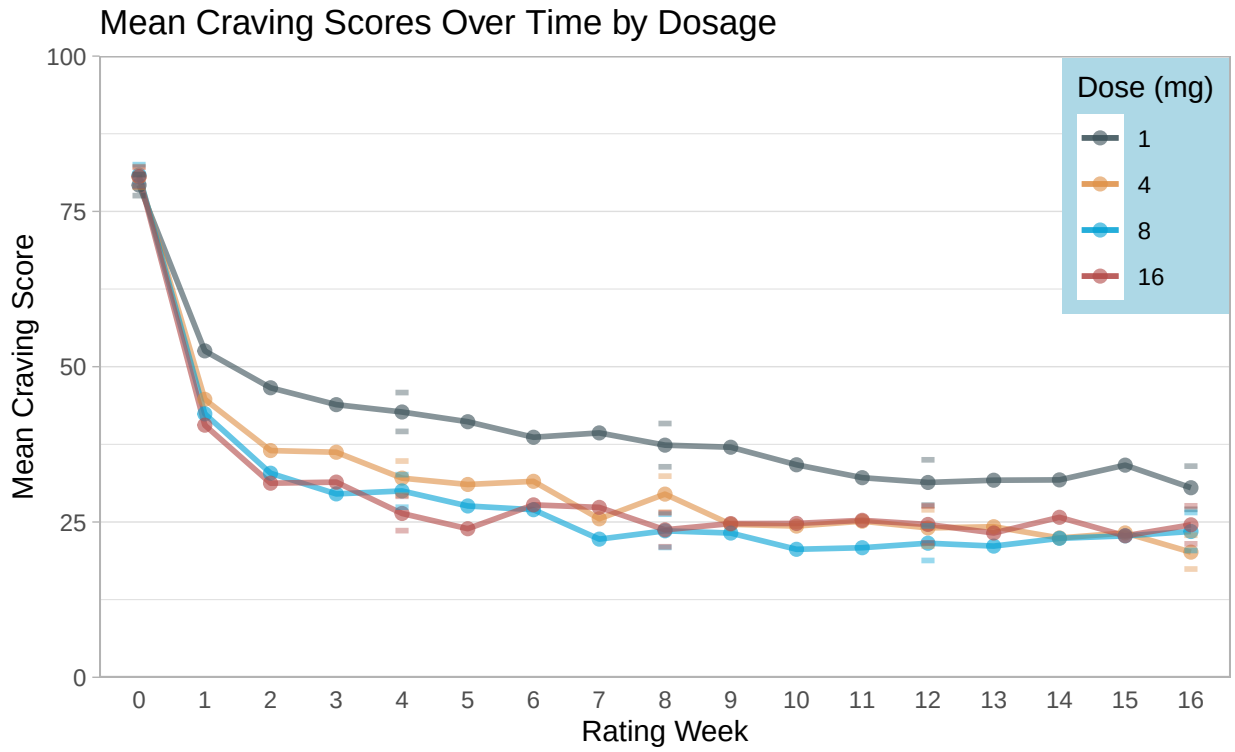


Figure 12: Craving Scores: Means

3.5 Adverse Events

An adverse event is any unexpected medical occurrence associated with the use of a drug in humans, whether or not it is considered drug related (40). This can be any unfavorable and unintended sign, symptom, or disease temporarily associated with the use of a drug, without any judgment about causality or relationship to the drug. An adverse event can arise from any use of the drug and from any route of administration, formulation, or dose, including an overdose. Pre-existing conditions that are ongoing during the clinical trial and concomitant medications taken prior to participation in the clinical trial are considered baseline characteristics and not reported in the adverse events.

An adverse event or suspected adverse reaction is considered serious if it results in any of the following outcomes: death, a life-threatening adverse event, inpatient hospitalization or prolongation of existing hospitalization, a persistent or significant incapacity or substantial disruption of the ability to conduct normal life functions.

In this study no deaths were reported. There were over 15,000 recorded adverse events ranging from headaches, depression, accidents and toxicity. Most adverse events were expected withdrawal symptoms such as headaches (31%), insomnia (25.8%) and constipation (24%). Constipation is a possible dose related event, occurring more frequently in the 8mg group than in the 1mg group ($p = 0.043$). Elevated liver function occurred in 14 patients, spread evenly across all 4 groups (12, 1mg; 13, 4mg; 14, 8mg; 12, 16mg). But only 6 of these resulted in toxicity. Figure 13 summarizes the top 5 serious adverse events.

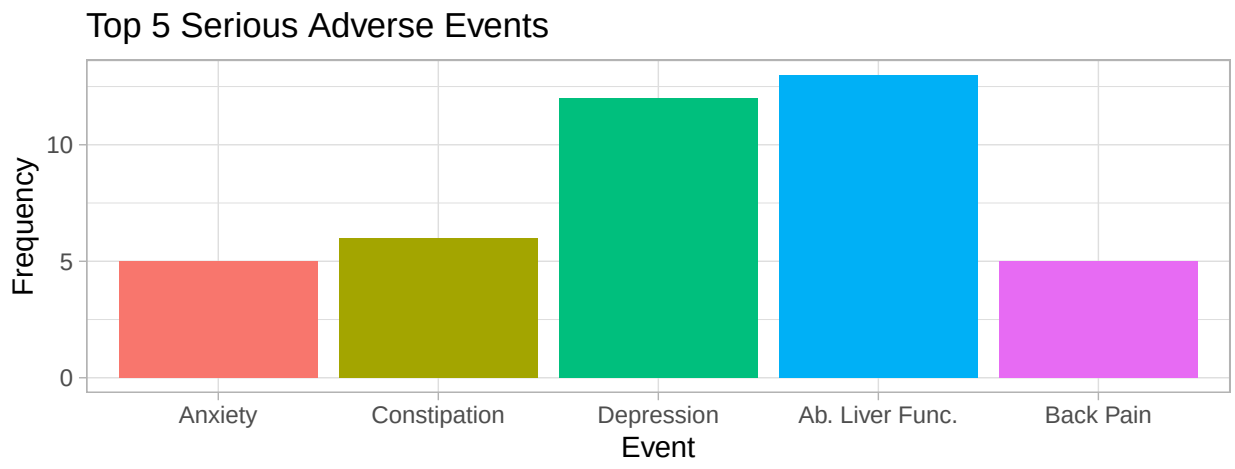


Figure 13: Adverse Events

4 Composite Endpoints

We begin this section by introducing the concept of a composite endpoint and discuss the rationale for its use in clinical studies. In an ideal clinical study, a single primary endpoint is available which completely characterizes the medical outcome of interest and allows for an efficient evaluation of treatment efficacy. This is often not the case, as diseases and the effects of a medical intervention on any given patient are usually multifaceted. Thus, a clinically meaningful treatment benefit is not always able to be achieved by evaluating a single event endpoint, especially when the event of interest occurs with low frequency. Moreover, there are some diseases for which more than one clinical outcome is considered clinically important, and all such outcomes are expected to be affected by the treatment being studied.

The composite endpoint takes these points into consideration by combining several endpoints of interest into a single outcome variable. The constituent endpoints are referred to as the components of the composite endpoint. Several types of composite endpoints are considered within the literature of clinical trials, and distinguish themselves from each other based on the underlying scale of measure of the components. Most often, the composite endpoint is defined as a binary indicator variable, a time-to-event variable, or a clinical score variable.

We first consider the case where the component endpoints are given by binary event indicators. An event in this context refers to one out of several predefined event types within a clinical study, such as death or hospitalization. The composite endpoint is then defined to be an encompassing event indicator that equals 1 whenever at least one of the component indicators is 1. In other words, the composite outcome is considered to have occurred when at least one of the component events of interest manifest within a clinical trial. Thus, the composite indicator equals 0 when all of the component indicators equal 0. Such a composite endpoint is referred to as a *composite binary endpoint*. A variation of this type of composite endpoint is known as a clinical response variable, where a patient is considered to be a *responder* if some specified number of underlying binary event indicators are simultaneously 1. A standard example of this type of composite endpoint is the American College of Rheumatology ACR-20/ACR-70 for assessing rheumatoid arthritis.

Another type of composite endpoint is the *time-to-first-event endpoint*, where the underlying components are given by event times during a study. Again, events refer to some predefined events of clinical interest. This type of composite endpoint is widely used in clinical trials where several types of clinical events that describe the condition of the patient are of interest. In these studies, patients are monitored for the occurrence of any of the component event times until either the event of interest manifests or until the observational period is over. Patients who do not experience any of the predefined event types or patients who are lost to follow-up within the duration of the study are treated as censored observations. Composite time-to-event endpoints

are commonly evaluated by survival analysis techniques within the literature. A closely related type of this composite endpoint is one that is defined to be an event rate during an observational period, where the occurrence of any one event from a set of possible events is considered to be a positive outcome.

The last type of composite endpoint, referred to as a clinical score, is a scale or index based on a rating system developed by the clinical researcher, and is essentially a weighted composite endpoint. This type of composite endpoint usually attempts to combine continuous or ordinal variables, and sometimes even differently scaled variable components. Many clinical scores are defined to be some combined measure of several rating scales that assess varying aspects of the endpoints of the study. A classic example of such a composite endpoint is the Hamilton Depression Rating Scale. However, evaluating such composite endpoints is generally more difficult than the previous two types we have mentioned.

The motivation behind using composite endpoints is often two-fold: an increase in statistical efficiency and the ability to hollistically evaluate the benefit of a medical intervention. We mentioned that an ideal primary endpoint allows for an efficient evaluation of treatment efficacy. In the context of a clinical trial, the power often depends on the total number of observed events. If the event type of interest corresponds to a relatively rare event, utilizing such an event as a primary endpoint may result in low power, or the need to increase the sample size substantially. This may not be feasible nor affordable in most cases. Because the binary and time-to-first-event composite endpoints both combine different event types of clinical interest, they are able to increase the number of expected event occurrences. As a result, composite endpoints can increase power and remove the need to have large sample sizes in situations where the clinical event of interest can be considered rare.

Another way in which composite endpoints increase statistical efficiency is by removing the need to adjust for type I error from multiple testing. Rather than testing a system of multiple primary endpoints in a single study, the composite endpoint "addresses the multiplicity problem without requiring adjustment to the type I error." [18] Furthermore, it avoids the need to choose a single primary endpoint when many clinical outcomes may be of equal importance.

On the other hand, composite endpoints are not without their limitations in that they are prone to misinterpretation based on their outcome. For example, a positive result of a composite does not imply that there was a positive outcome for each component, nor does it imply that they share the same magnitude. Moreover, individually significant components may have been hidden or watered-down by nonsignificant ones in the composite. In this regard, it is imperative to comprehensively analyze the components of the composite separately. Unfortunately, a composite endpoint can be developed poorly when it combines multiple clinical objectives by including separate effects for the sake of statistical expediency. Montori et al gives three recommendations for assessing the validity of composite endpoint in a clinical trial [19]:

1. Are the component end points of similar importance to patients?

   When all components of a composite endpoint are of equal importance to the patient, it will not be misleading to assume that a positive composite outcome implies the component outcomes are similar. That being said, the components themselves should all be clinically relevant for the objective of the trial.

2. Did the more and less important endpoints occur with similar frequency?

   If the frequencies of the individual component outcomes greatly differ, the composite outcome becomes more difficult to interpret.

3. Are the component endpoints likely to have similar relative risk reductions?

   When all the components have similar treatment effects, we can be confident in a composite endpoint. Researchers should construct a composite endpoint for which the biology would lead us to expect similar effects across components.

In developing a composite endpoint, it is useful to consider these recommendations in order to avoid its distinct limitations and maximize statistical efficiency without being expedient. Unfortunately, the recommendations laid out above are usually hard or perhaps even unrealistic to fulfill in practice. Again, in the case where all the recommendations are not all met, it is important to evaluate the individual components in addition to the composite. We will now consider the case of a binary responder-type composite endpoint.

5 Selection of Potential Component Endpoints

After completing our EDA, we identified our five potential component endpoints: craving score, global rating score (staff), global rating score (patient), retention rate, and self-report of drug use. For the type of composite endpoint we will use, we will construct a binary composite endpoint meaning that the component endpoints will be given by binary event indicators. We will also use the variation of this type of composite endpoint: clinical response variable. Therefore, we will be implementing a responder analysis on the primary endpoints. For the construction of a responder endpoint, we define a threshold value for which a patient is considered to be a responder on the basis of each component (38). This is appealing in that it simplifies several (potentially complex) pieces of information into a single responder/non-responder variable.

Our next steps include delving into the literature for potential thresholds of success for each of these endpoints. However, through our research we have found that neither the patient reported global rating scores or the retention rate is a reliable measure for efficacy. A report found as a guidance for industry on endpoints for efficacy specifically for opioid use disorder mentions that retention rate is not recommended as a stand-alone endpoint (37). Patient reported global rating scores have been questioned in their reliability in patients estimates of their previous health status which is illustrated as the problem of "recall bias". Due to this issue of possible inaccurate recall of ones previous health status, there is criticism that a patient's score is more influenced by the current health status of the patient rather than measuring the transition in health status (31). As such, we have decided not to use either global rating score (patient) or retention rate as a component of our composite endpoint.

5.1 Craving Score

Several studies use a craving scale as a measure of efficacy for clinical trials. In a similar trial studying alcohol craving, researchers used the Penn Alcohol Craving Scale (PACS). Researchers proposed a cutoff of a score greater than 20 to indicate a positive alcohol craving symptom (32). We will implement a similar measure for our component. Craving scores are recorded weekly; thus, we want to assure that the patients are improving from their baseline craving score (i.e. initial craving score at the beginning of the study). In order for a patient to be considered a responder in the craving score component, patients must have a decrease in score from baseline as well as have a score of less than 20 at some point in the study.

5.2 Staff Reported Global Rating Score

Global rating scores or scales reported by staff are commonly known today as Clinical Global Impression Scales of improvement (CGI-I). These scales are widely accepted endpoints that measures change in severity. Typically, they come in the form of a 7-point Likert Scale with a score of "1" indicating very improved since initiation of treatment and "7" indication very much worse since the initiation of treatment (33,34). It is important to note here the score analyzing change from baseline as we will apply the following method similarly to the global rating score as a component of the endpoint. For the CGI-I, scores of 1 or 2 both show improvement from baseline; therefore, researchers have determined that a score of 1 or 2 will be considered a success (35). The global rating score in our study is in the form of a 5-point Likert Scale, in which a score of 4 or 5 shows improvement. We will be analyzing the change from screening of each patient to their termination score, those patients that have a score of a 4 or 5 will be considered a responder.

5.3 Self-Report of Drug Use

Our last efficacy measure is self-report of drug use. This measure is commonly used in substance abuse trials to track a patient's abstinence. The Society for Research on Nicotine and Tobacco recommend the following thresholds to denote the length of time abstinent. The researchers determined that trials that are longer than 3 months in duration should demonstrate efficacy with a report of prolonged abstinence of 4 or more weeks. They have also suggested a 2-week grace period (36). We will consider a patient responder in this component, if for any four consecutive weeks in the program they have continuous abstinence.

5.4 Validation of Craving Score Threshold

In the selection of potential component endpoints for a clinical responder endpoint, it is necessary to specify clinically meaningful thresholds so that efficacy can be most optimally measured. As specified previously, we found that a craving score of 20 and below was proposed as being clinically meaningful in an alcohol dependency trial. Since this is not specific to opioid use disorder, we wish to validate this particular threshold with the protocol data using an anchor-based approach based on staff reported Global Rating Scales at termination. This type of post hoc analysis will utilize the pooled patient data and will derive clinically meaningful thresholds through an ROC analysis. (39)

We define the independent variable of interest to be the minimum craving score achieved throughout each patient's participation in the study. For the dependent variable, we utilize the staff reported Global Rating Score at termination where an improving score (4, A Little Better and 5, Much Better) is considered to be clinically meaningful. The ROC curve is obtained by calculating the TPR (sensitivity) and FPR (1-specificity) for every possible threshold for the craving score (0-100). We define the clinically meaningful threshold to be the cutpoint which maximizes the sum of sensitivity and specificity. In Figure 14 we display the results of the analysis.

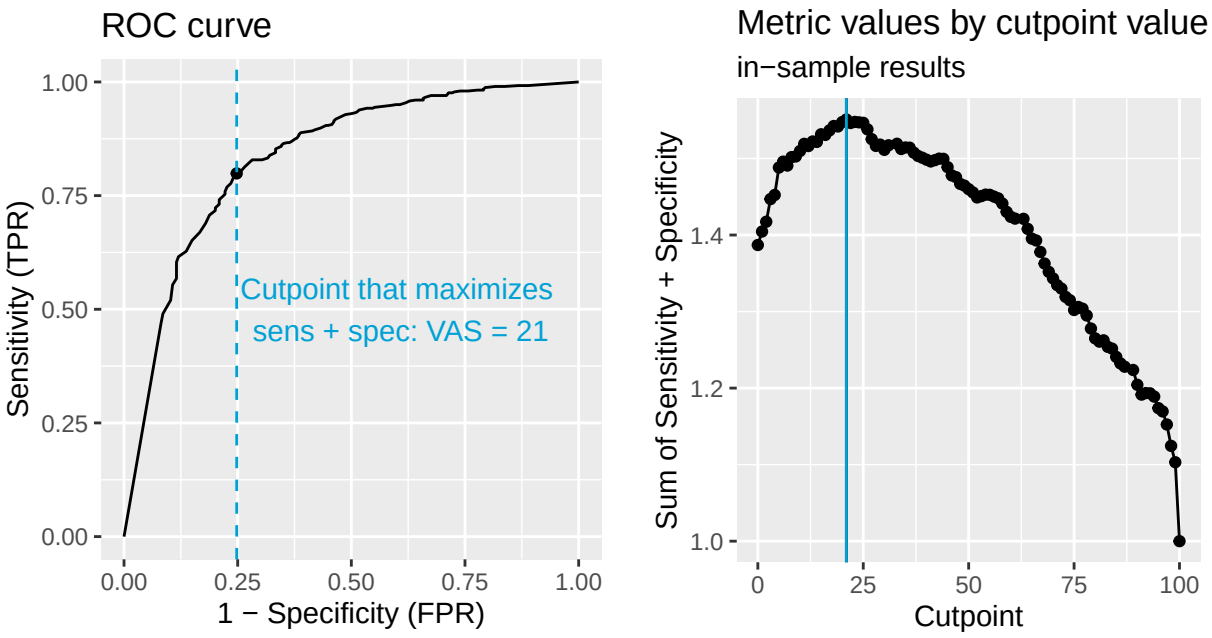


Figure 14: ROC Derived Threshold Analysis

We observe that maximizing the selected metric of the sum of sensitivity and specificity, we obtain a clinically meaningful threshold of ≤ 21, which agrees with the threshold proposed in the alcohol dependency trial. Moving forward, we utilize the optimized threshold for dichotomizing craving scores.

5.5 Component Domains and Responder Definitions

The primary endpoint that evaluates treatment efficacy should encompass one or more of the important features of a disorder and should be clinically meaningful. There may be circumstances when no single endpoint adequately serves this purpose, as in the case of Opioid Use Disorder (OUD). Clinical trials evaluating effectiveness of drugs for treating OUD have historically used abstinence of drug use as an endpoint. At the same time, the FDA has expressed great interest in expanding the primary and secondary endpoints used in clinical trials of drugs for treating OUD, including other outcome measures important to patients and their families, clinicians, and the public. Efficacy is commonly evaluated on the observed effect of drug abstinence (urinalysis). Patient-important outcomes are not represented in such trials. This deficit is of substantial concern to the growing evidence base in opioid use disorder. In developing a composite endpoint then, it is important to draw on different outcome domains in order to best represent the patient's response to treatment.

The FDA has recommended reduction in drug use patterns as a primary endpoint, rather than strict abstinence. To develop a composite endpoint that measures this change, we utilize component endpoints from three different domains. The Global Rating Scale (clinician reported) represents a measure of change in disease status using diagnostic criteria for OUD as outlined by the FDA. This measurement assesses a patient's overall health from baseline to termination. The Self Report of Drug Use outcome demonstrates a change in drug use pattern by defining a responder as a patient that abstains from the use of opioids for 4 consecutive weeks. This is a threshold known to be associated with clinical benefit. Craving Score is a patient reported outcome that reflects the patient's perspective on their treatment. By defining a responder to treatment across these three domains, we aim to develop a measure of treatment efficacy that incorporates multiple dimensions of opioid use disorder.

Table 1 summarizes the responder definitions informed by the literature and post-hoc analysis that will be used to dichotomize the component endpoints.

Table 1: Responder Thresholds for each Component

Component	Threshold
Craving Score	Score of 21 or less at any point in the study AND less than at screening
Global Rating Score	Score of 5 (Much Better) at Termination
Self Report of Drug Use	4 consecutive weeks of abstinence during enrollment

6 Methods

In order to begin the modeling process, each component endpoint must be dichotomized. In general, we consider the jth component endpoint EP_{ij} for patient i to be defined as a binary indicator variable that takes the value 1 if patient i is considered a responder, and 0 if they are considered a non-responder:

$$EP_{ij} = \begin{cases} 1 & \text{if patient } i \text{ is considered a responder for the } j\text{th component} \\ 0 & \text{otherwise} \end{cases}$$

In the context of the study protocol data, we will create a new column for global rating score, where each patient that was considered a responder for this endpoint receives a 1 and each patient that was considered a failure for this endpoint receives a 0. Both craving score and self-report of drug use will also have columns designated to the success/failure of their outcome. The primary endpoint is defined to a binary variable that indicates the success of the buprenorphine treatment depending on the outcome of the individual or composite endpoints. We will then look at 7 different potential primary endpoints, where A denotes Craving Scale, B denotes Global Rating Scores, and C denotes Self-Report of Drug Use:

$$A_i = \begin{cases} 1 & \text{if patient } i \text{ is considered a responder for Craving Scale} \\ 0 & \text{otherwise} \end{cases}$$

$$B_i = \begin{cases} 1 & \text{if patient } i \text{ is considered a responder for Global Rating Scale} \\ 0 & \text{otherwise} \end{cases}$$

$$C_i = \begin{cases} 1 & \text{if patient } i \text{ is considered a responder for Self-Report} \\ 0 & \text{otherwise} \end{cases}$$

$$AB_i = \begin{cases} 1 & \text{if } A_i = 1 \cap B_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$AC_i = \begin{cases} 1 & \text{if } A_i = 1 \cap C_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$BC_i = \begin{cases} 1 & \text{if } B_i = 1 \cap C_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$ABC_i = \begin{cases} 1 & \text{if } A_i = 1 \cap B_i = 1 \cap C_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

Table 2 below summarizes the proportions of responders for each treatment group in all 7 cases. We can see a clear dose response for cases C, AC, BC, and ABC; as dosage increases the number of responders increases as well. On the other hand, we observe a clear separation in proportion from the active control group, but not as good of a dose response among the higher dosages for the remaining cases. The cases including the Self-Report of Drug Use as a component have fewer responders in comparison to the Craving Score and Global Rating Scores; as such, it is our most stringent endpoint and drives the composite endpoint when included.

Table 2: Proportion of Responders by Primary Endpoint

DOSE	A	B	C	AB	AC	BC	ABC
1mg	0.470	0.211	0.157	0.168	0.151	0.086	0.086
4mg	0.643	0.379	0.297	0.368	0.286	0.203	0.203
8mg	0.681	0.383	0.340	0.367	0.335	0.245	0.245
16mg	0.680	0.420	0.464	0.392	0.431	0.320	0.309

6.1 Logistic Regression

In searching for a potential primary endpoint, it is often of interest to utilize one that shows good treatment effect separation. In order to evaluate our potential primary endpoints, we utilize a logistic regression model with dosage as a fixed predictor and the binary outcome of 1 of the 7 potential primary endpoints as the response:

$$\log\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 X_{\text{Dose}}$$

where Y is the probability of being a responder for a potential primary endpoint and X_{Dose} is a vector of treatment dosages (4mg, 8mg, 16mg) being compared to the active control (1mg). Note that the coefficient β_1 is a coefficient vector of length three as well.

We use this simple model for all 7 potential primary endpoints in order to get the estimated log odds β coefficients. After exponentiating the log odds, we plot the estimated odds ratios and 95% confidence intervals in the Figure 15.

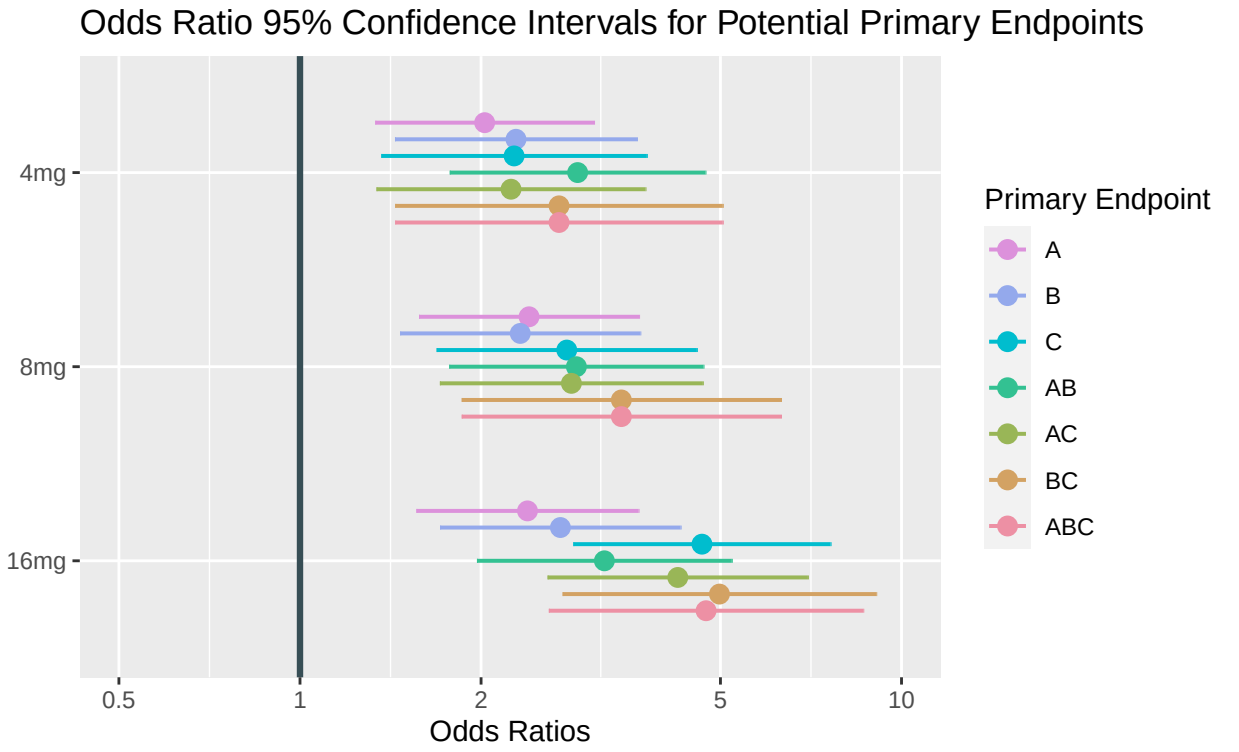


Figure 15: Odds Ratios from Logistic Regression

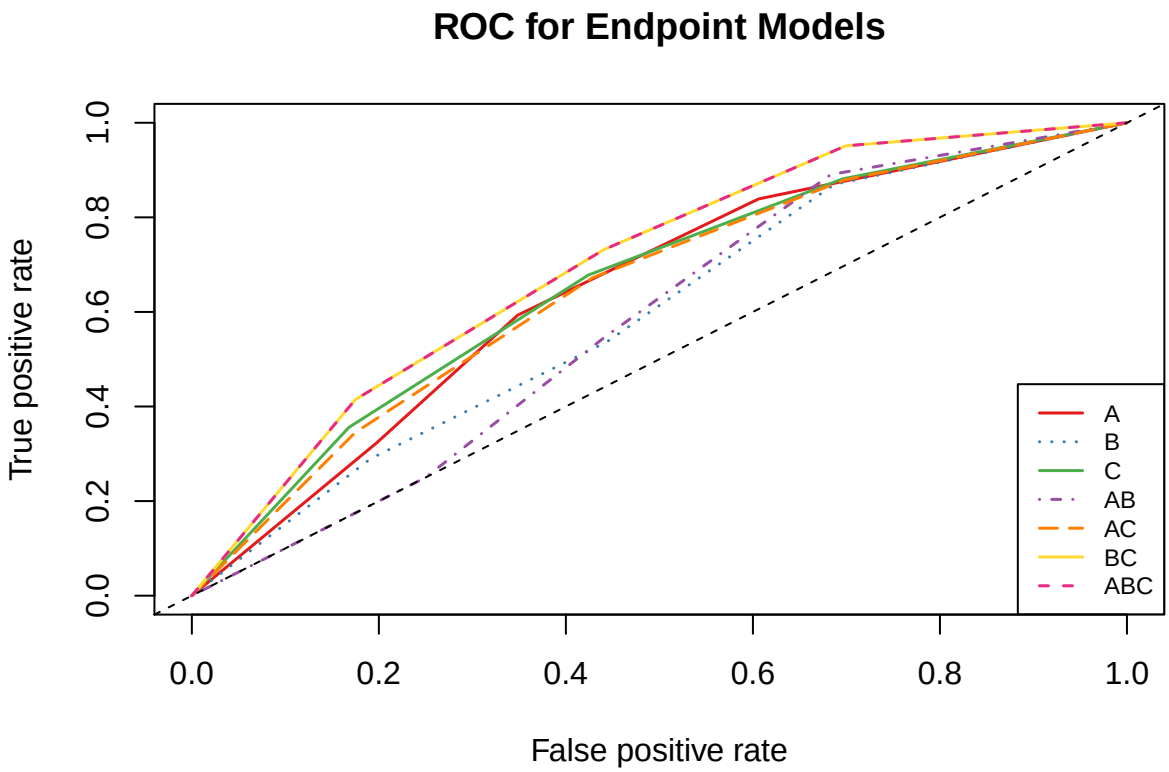
We see that all of our potential primary endpoints are showing clear separation from the 1mg dose group, which served as the reference level in our logistic regression models, as indicated by their confidence intervals not touching the OR = 1 line. Focusing on the standalone components as potential primary endpoints, we observe that each of the components (A, B, C) is sufficient in demonstrating treatment efficacy on its own.

6.2 Endpoint Evaluation: Which Endpoint is the Best?

To further evaluate each of the seven logistic regression models consisting of each of the endpoints as our response, we will conduct resampling methods involving randomly dividing the data set of observations into a training set and a test set or validation set. Each of the endpoint models are fitted onto the training set, and we use those fitted models to predict responses (predicted probabilities of patient being "Responder" or not) for observations in the test set. In conducting the training-test split, we randomly sample 75% of observations of the data set to create the training set and the remaining 25% of observations held out of those randomly sampled will be the test set. Then, we can assess each of these fitted models by either obtaining a test set error rate or by plotting a receiver operating characteristics (ROC) curve, which compares sensitivity (true positive rate) versus specificity (false positive rate) over a set of values for the ability to predict the binarized response. Amongst all the endpoint models, the model which is considered to perform the best is the one with highest area under the curve (AUC). Additionally, we will also assess each of the fitted models on the training set by examining the Akaike information criterion (AIC) and Bayesian information criterion (BIC) in which the best model has the lowest AIC and/or BIC. In Table 3 we have the following results from each of our fitted endpoint models:

Table 3: Endpoint Model Performance

	AIC	BIC	AUC
A	733.93	751.18	0.65
B	700.50	717.75	0.60
C	662.85	680.11	0.66
AB	676.21	693.46	0.58
AC	652.01	669.26	0.65
BC	557.21	574.46	0.70
ABC	**553.59**	**570.84**	**0.70**



We can see that based on AUC, both models with BC and *ABC* as the response performed the best with an AUC of 0.70. However, according to AIC and BIC, *ABC* performs the best overall with an AIC of 553.59 and 570.84. Therefore, to conduct a covariate analysis, we will use the logistic regression model with *ABC* as the response.

6.3 Covariate Analysis

Now that we have obtained the best endpoint as the response, we must now add in the baseline characteristics (i.e. demographic, medical, and other relevant information collected at the beginning of the study) to include in the model as potential covariates. Baseline characteristics is an important feature for research and analyses in clinical trials in order to confirm proper conduct of randomization among patients, whether or not there is imbalance in baseline characteristics that may cause chance bias, and whether analyses should be adjusted for baseline variables (42). In our analysis, the ultimate question lies in which subset of these potential

covariates should be included in the final model? To answer this question, we must choose a method to select the appropriate baseline variables for our logistic regression model. Figure 16 below is an illustration of the process used to perform a variable selection method to choose baseline characteristics as covariates.

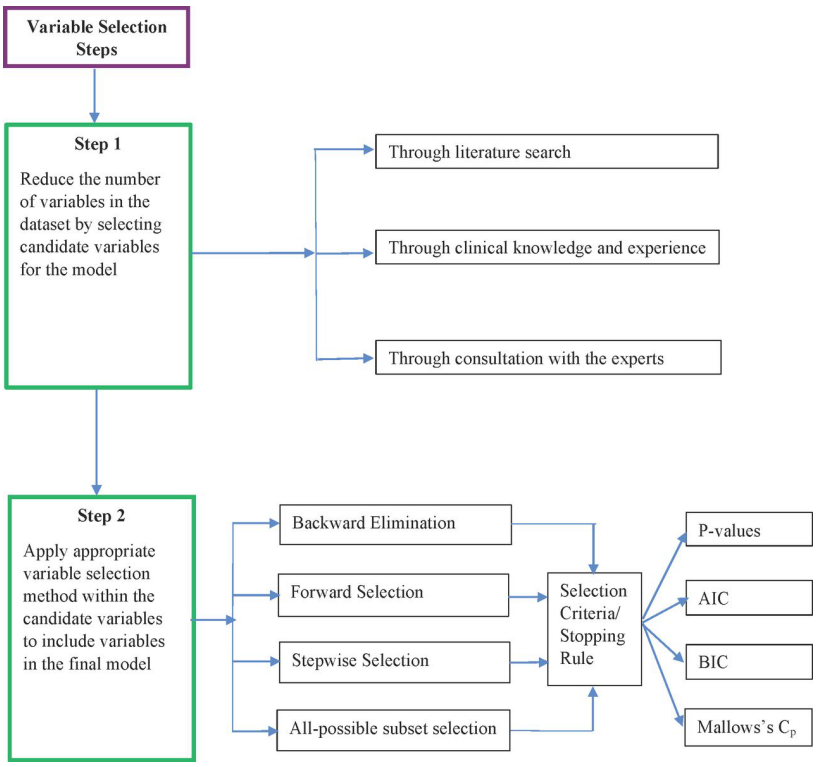


Figure 16: Variable Selection Flow chart. Adapted from BMJ Journals Family Medicine and Community Health 2020.(41)

Variable selection helps determine a set of variables that will produce the best fit for the model that would make accurate predictions.Through literature search and consultation, the best variable selection method adapted for this analysis is a best subset (all possible) selection algorithm and choosing the best model with baseline covariates using AIC and BIC criteria. It is important to note that for this clinical trial, a confirmatory clinical trial that is used to confirm prior results that Buprenorphine might be beneficial as a treatment for patients, data-dependent selection methods such as the best subset selection algorithm is usually not advised and can lead to misleading inferences (43). It is suggested that "choice of baseline characteristics by which an analysis is adjusted should be determined by prior knowledge of an influence on outcome" (42). But, since we do not have a prespecified response or endpoint for the analyses to determine influence from baseline characteristics, a variable selection method can still be adapted for a covariate analysis.

In beginning the covariate analysis, we have implemented the best subset selection algorithm amongst all 2^k possible models for $k = 12$ covariates. The covariates considered in the model include treatment dose, race, gender, education (i.e. Highest Level of Education Attained), work, income, marital status, living status, drugs in house (i.e. Is there Heroin or Cocaine use in the household where you live?), jail (i.e. Are you presently awaiting charges, trial or sentence that is likely to result in your going to jail during the next 6 months?), age, and baseline heroin craving score. Many of these variables in the data are categorical with the exception of income and baseline craving score with multiple factor levels. Therefore, before adding in the baseline characteristics into the selection algorithm, we have dichotomized or reduced the number of levels for each of the categorical variables to the following:

1. Race: White, Black, Hispanic/Other
2. Education: High school and below, Minimum High School Educated
3. Work: Unemployed, Employed
4. Marital Status: Married, Unmarried
5. Living Status: Not Alone, Alone/Unstable
6. Drugs In House: Yes, No/Don't Know
7. Age: less than or equal to 35 years, greater than 35 years

Dichotomizing or reducing the levels of each of these categorical variables will allow for easier interpretation of coefficients after fitting each of the models. After conducting the best subset selection algorithm for all possible models, we have obtained the final results for best models with criteria AIC and BIC:

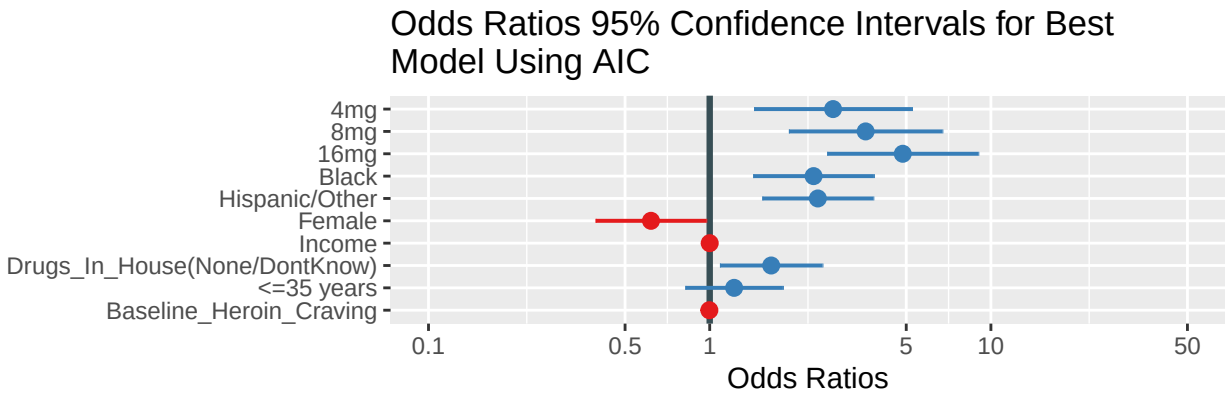


Figure 17: Model with Lowest AIC

The best model obtained using lowest AIC as criteria resulted in covariates: treatment dose, race, gender, income, drugs in house, age, and baseline heroin craving score. To determine the relationship with the responder variable ABC, we observe whether the 95% confidence intervals of the odds ratios for each covariate are greater than one or not. We see that higher treatment dose, race, and drugs in house seem to have a positive association with ABC, which means that higher treatment doses, being black or hispanic/other as opposed to white, and not having drugs at home translates into greater odds of the patient being a responder. In contrast, the negative association between gender and responder ABC indicates that the patient being female as opposed to male puts the patient at lower odds of being a responder.

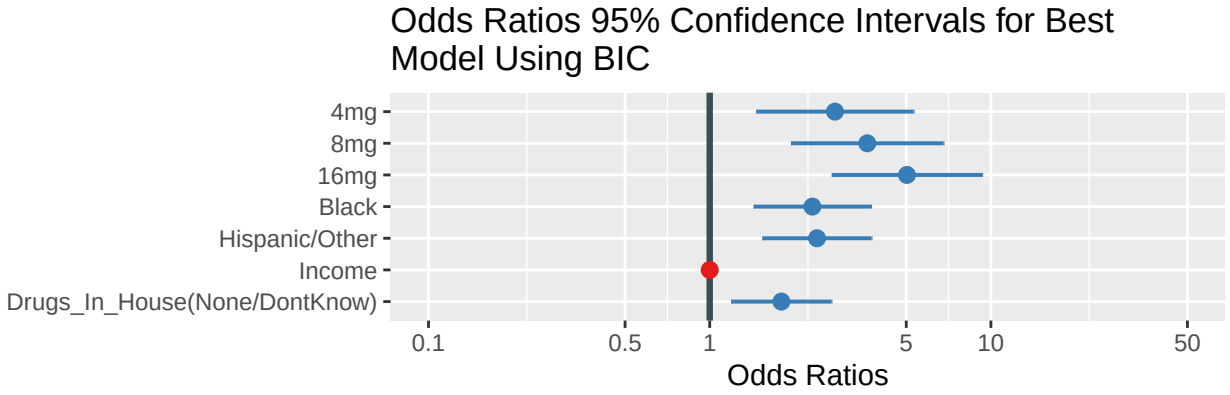
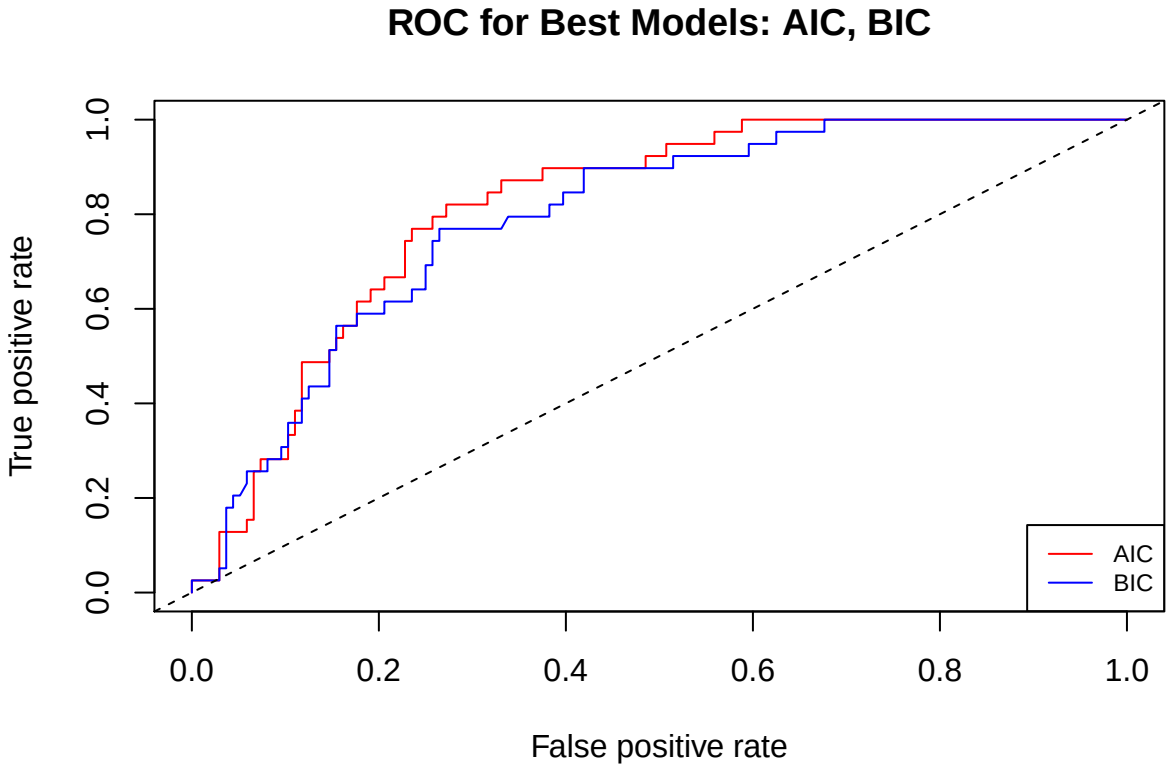


Figure 18: Model with Lowest BIC

The best model obtained using lowest BIC as criteria resulted in fewer covariates: treatment dose, race, income, and drugs in house. Again, we see that higher treatment dose, race, and drugs in house seem to have a positive association with the responder varialble ABC, which means that higher treatment doses, being black or hispanic/other as opposed to white, and not having drugs at home translates into greater odds of the patient being a responder.

In order to decide which of these two models selected using lowest AIC and lowest BIC should be the final model, we perform a test-train split of the data that includes the baseline characteristics, where the training data is obtained by randomly sampling 75% of the data and the test data is the held out 25% of the data. Then we plot ROC curves to see which model has the highest AUC.



In the above graph, we have the resulting ROC curves after fitting the models on the training data and acquiring the predicted probabilities of the patient being "Responder" or not for observations in the test set. The best model using lowest AIC as criteria had the highest AUC of approximately 0.8145 whereas the best model using lowest BIC as criteria had an AUC of 0.7917. Therefore, using AUC as the final verdict, the final model obtained from our covariate analysis is:

$$\log\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 X_{\text{Dose}} + \beta_2 X_{\text{Race}} + \beta_3 X_{\text{Gender}} + \beta_4 X_{\text{Income}} + \beta_5 X_{\text{Drugs}} + \beta_6 X_{\text{Age}} + \beta_7 X_{\text{HeroinCrave}}$$

where Y is the binary responder variable ABC denoting probability of patient being responder or not, β_0 is the intercept coefficient, and the coefficients $\beta_1, \beta_2, ..., \beta_7$ are associated with each of the resulting covariates obtained. Note that β_1 and β_2 are vectors consisting of multiple coefficients depending on the factor level observed and multiplied by their corresponding predictor variable as a dot product. For example, there are three coefficients within β_1 for each of the treatment doses (4 mg, 8 mg, and 16 mg) with 1 mg as the reference level.

Final Logistic Regression Model Results

	Response: ABC		
Predictors	*Odds Ratios*	*CI (95%)*	*P-Value*
Intercept	0.06	0.02 – 0.15	**<0.001**
4mg	2.75	1.46 – 5.36	**0.002**
8mg	3.59	1.95 – 6.89	**<0.001**
16mg	4.86	2.67 – 9.27	**<0.001**
Black	2.34	1.43 – 3.84	**0.001**
Hispanic/Other	2.42	1.54 – 3.83	**<0.001**
Female	0.62	0.39 – 0.97	**0.038**
Income	1.00	1.00 – 1.00	0.371
Drugs In House (None/Don't Know)	1.65	1.09 – 2.54	**0.019**
Age less than or equal to 35 years	1.22	0.82 – 1.82	0.326
Baseline Heroin Craving	1.00	0.99 – 1.01	0.505
Observations	719		
AIC	691.913		

Figure 19: Odds Ratios Table of the Final Model on the full data

In Figure 19, we have an odds ratios table of the final model with associated p-values. Note that the model is fitted on the full data set with missing values deleted and has an AIC of 691.913. We can see that the baseline covariates (predictors) treatment dose (4 mg, 8 mg, 16 mg with 1 mg as reference level) and race (Black and Hispanic/Other with White as reference level) in our final model have extremely low p-values, and thus deemed to be statistically significant. Gender and Drugs in House can also be considered significant when using p-value less than 0.05 to express statistical significance. From the odd ratios for each of the predictors (covariates), just like with the forest plots, we can determine the relationship with the responder ABC by observing whether the 95% confidence intervals of the odds ratios for each covariate are greater than one or not. Note that is a tabular format of the odd ratios and their confidence intervals seen in Figure 3. Again, higher treatment doses, being black or hispanic/other as opposed to white, and not having drugs at home places the patient at greater odds of being a responder whereas the patient being female as opposed to male puts the patient at lower odds of being a responder. We can interpret the odds ratios for patients in the 16 mg group as having 4.86 times the odds of being a responder compared to the 1 mg group and with 95% confidence the true odds ratio lies in the range of 2.67-9.27. Thus, odds ratios for significant covariates for all treatment groups, race (Black and Hispanic/other), gender (Female), and Drugs in House (No/Don't Know) can be interpreted in a similar fashion. For a full summary of responder proportions by covariate in each treatment group with p-values, see the summary descriptives table included in the Appendix.

7 Discussion

In this report we aimed to develop a responder-type composite endpoint utilizing a 1992 study on buprenorphine for treating Opioid Use Disorder. Due to the missing urinalysis data, we utilized available secondary measures that captured different domains of treatment efficacy in order to construct a viable primary endpoint. Although we found that the potential primary endpoints BC and ABC were highly competitive with eachother, we chose to use ABC in our final model because it includes the patient's perspective on how the treatment is effecting them. This is potentially valuable information that otherwise would not have been included in a primary endpoint that utilizes only an objective measure such as urinalysis. Because OUD is a complex disease with multifaceted psychosomatic manifestations, a composite endpoint may be an effective way to measure treatment efficacy.

As future work, one possibility is to explore the extension of this study, as it would be interesting to see how the four treatment groups perform past the initial sixteen-week study. In addition, adverse events can be included as a component that captures both safety and efficacy of a treatment in a composite endpoint. Furthermore, the composite endpoint presented in this report can potentially be improved by incorporating a component from another treatment domain. Another possibility is to construct a time-to-first-event endpoint that was introduced in this report, and conduct survival analysis methods based on Cox proportional hazards seen in the statistical methods of the protocol.

References

1. Understanding Clinical Studies, National Institute of Health: https://www.nih.gov/about-nih/what-we-do/science-health-public-trust/perspectives/understanding-clinical-studies

2. Doig, Gordon S, and Simpson, Fiona. "Understanding Clinical Trials: Emerging Methodological Issues." Intensive Care Medicine 40.11 (2014): 1755-757. Web.

3. Zivin, Justin A. "Understanding CLINICAL TRIALS." Scientific American, vol. 282, no. 4, 2000, pp. 69–75. JSTOR, www.jstor.org/stable/26058674. Accessed 12 Mar. 2021.

4. Stewart, Derek, National Translational Cancer Research Network, and Emergency Care Research Institute. Clinical Trials Explained: A Guide to Clinical Trials in the NHS for Healthcare Professionals. Malden, MA: Blackwell Pub., 2006. Web.

5. O'Kelly, Michael, and Bohdana Ratitch. Clinical Trials with Missing Data : A Guide for Practitioners : A Guide for Practitioners, John Wiley & Sons, Incorporated, 2014. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/fullerton/detail.action?docID=1636082.

6. Hackshaw, Allan K. A Concise Guide to Clinical Trials. Oxford: Wiley-Blackwell, 2009. Web.

7. McCoy CE. Understanding the Use of Composite Endpoints in Clinical Trials. West J Emerg Med. 2018;19(4):631-634. doi:10.5811/westjem.2018.4.38383

8. Opiod Use Disorder: Endpoints for Demonstrating Effectiveness of Drugs for Treatment, FDA, October 2020, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/opioid-use-disorder-endpoints-demonstrating-effectiveness-drugs-treatment-guidance-industry. Accessed 12 Mar. 2021

9. Higgins ST, Budney AJ, Bickel WK, Hughes JR, Foerg F, Badger G. Achieving cocaine abstinence with a behavioral approach. Am J Psychiatry. 1993; 150 (5):763–769. [PubMed: 8480823]

10. Miller WR, Manuel JK. How large must a treatment effect be before it matters to practitioners? An estimation method and demonstration. Drug Alcohol Rev. 2008; 27 (5):524–528. [PubMed: 18608445]

11. Trivedi, Madhukar, et al. "Determining the Primary Endpoint for a Stimulant Abuse Trial: Lessons Learned from STRIDE (CTN 0037)." Taylor & Francis, www.tandfonline.com/doi/abs/10.3109/00952990.2011.598589

12. Campbell, Nancy D, and Lovell, Anne M. "The History of the Development of Buprenorphine as an Addiction Therapeutic." Annals of the New York Academy of Sciences 1248.1 (2012): 124-39. Web.

13. Jasinski DR, Henningfield JE, Hickey JE, et al.: Progress report of the NIDA addiction research center. The Committee on Problems of Drug Dependence, NIDA Monograph Series 1982;42:92-98.

14. Kosten TR, Kleber HD: Buprenorphine detoxification from opioid dependence: A pilot study. Life Sciences 1988;42:635-641.

15. National Consensus Development Panel on Effective Medical Treatment of Opiate Addiction. "Effective Medical Treatment of Opiate Addiction." JAMA : The Journal of the American Medical Association 280.22 (1998): 1936-943. Web.

16. "Opioid Overdose Crisis." National Institute on Drug Abuse, National Institute on Drug Abuse, 11 Mar. 2021, www.drugabuse.gov/drug-topics/opioids/opioid-overdose-crisis.

17. Rauch G, Schüler S, Kieser M. Planning and Analyzing Clinical Trials with Composite Endpoints. Springer Series in Pharmaceutical Statistics. 2017.

18. ICH. (1998). Statistical principles for clinical trials - E9. ICH. https://database.ich.org/sites/default/files/E9_Guideline.pdf

19. Montori, V. M., Permanyer-Miralda, G., Ferreira-González, I., Busse, J. W., Pacheco-Huergo, V., Bryant, D. et al. (2005). Validity of composite end points in clinical trials. British Medical Journal, 330, 594.

20. FDA. (2017). Guidance for industry - Multiple endpoints in clinical trials. Draft. FDA. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM536750.pdf.

21. A Multicenter Clinical Trial of Buprenorphine in Treatment of Opiate Dependence.

22. Editors, Health Catalyst. "The Power of Patient-Reported Outcome Measures." Health Catalyst, 17 Nov. 2020, www.healthcatalyst.com/insights/unlocking-the-power-of-patient-reported-outcome-measures-proms/.

23. Kamper, Steven J, et al. "Global Rating of Change Scales: a Review of Strengths and Weaknesses and Considerations for Design." The Journal of Manual & Manipulative Therapy, Journal of Manual & Manipulative Therapy, Inc., 2009, www.ncbi.nlm.nih.gov/pmc/articles/PMC2762832/.

24. Cook, R. J., Lawless, J. F. (1997). Marginal analysis of recurrent events and a terminating event. Stat. Med. 16:911–924.

25. Nelson, W. B. (2003). Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications. American Statistical Association and the Society for Industrial and Applied Mathematics.

26. O'Neill, R. T. (1987). Statistical analyses of adverse event data from clinical trials, special emphasis on serious events. Drug Information J. 21:9–20.

27. Rosenkranz, G. (2006). Analysis of adverse events in the presence of discontinuations. Drug Information J. 40(01):79–87.

28. Lange WR, Fudala PJ, Dax EM, Johnson RE: Safety and side-effects of buprenorphine in the clinical management of heroin addiction. Drug and Alcohol Dependence 1990;26:19-28.

29. Reckitt and Colman: Buprenorphine Brochure for Investigators. November, 1987.

30. "APA Dictionary of Psychology." American Psychological Association, American Psychological Association, dictionary.apa.org/active-control-trial.

31. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. J Man Manip Ther. 2009;17(3):163-170. doi:10.1179/jmt.2009.17.3.163

32. Hartwell, Emily E., et al. "Convergence between the Penn Alcohol Craving Scale and Diagnostic Interview for the Assessment of Alcohol Craving." Addictive Behaviors Reports, Elsevier, 18 June 2019, www.sciencedirect.com/science/article/pii/S2352853218301950#bb0035.

33. Kolevzon, Alexander, et al. "Development of an Adapted Clinical Global Impression Scale for Use in Angelman Syndrome." Journal of Neurodevelopmental Disorders, BioMed Central, 4 Jan. 2021, jneurodevdisorders.biomedcentral.com/articles/10.1186/s11689-020-09349-8.

34. Busner, Joan, and Steven D Targum. "The Clinical Global Impressions Scale: Applying a Research Tool in Clinical Practice." Psychiatry (Edgmont (Pa. : Township)), Matrix Medical Communications, July 2007, www.ncbi.nlm.nih.gov/pmc/articles/PMC2880930/.

35. "Interpreting ADHD Rating Scale Scores ." Primary Psychiatry, Primary Psychiatry, addadult.com/wp-content/uploads/2010/03/Goodman-DW-et-al-2010-Interpreting-ADHD-Rating-Scale-Scores-Linking-ADHD-RS-to-CGI-in-Two-LDX-Trials.pdf.

36. Hughes JR;Keely JP;Niaura RS;Ossip-Klein DJ;Richmond RL;Swan GE; "Measures of Abstinence in Clinical Trials: Issues and Recommendations." Nicotine & Tobacco Research : Official Journal of the Society for Research on Nicotine and Tobacco, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/12745503/.

37. Opioid Use Disorder: Endpoints for Demonstrating Effectiveness of Drugs for Treatment; Guidance for Industry; Availability. (2020). In The Federal Register / FIND (Vol. 85, Issue 192, p. 62305). Federal Information & News Dispatch, LLC

38. Snapinn, S.M., Jiang, Q. Responder analyses and the assessment of a clinically relevant treatment effect. Trials 8, 31 (2007). https://doi.org/10.1186/1745-6215-8-31

39. Gammaitoni, A., et al. Defining a Minimal Clinically Important Difference in Seizure Frequency Using Data from a Phase 3 Clinical Study of Add-On, Low-Dose Fenfluramine HCl Oral Solution in Dravet Syndrome Patients Receiving an Antiepileptic Drug Regimen Containing Stiripentol. www.zogenix.com/news-releases/scientific-posters-publications/

40. Adverse Event Reporting: Improving Human Subject Protection. U.S. Department of Health and Human Services Food and Drug Administration, January 2009

41. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Family Medicine and Community Health 2020;8:e000262. doi: 10.1136/fmch-2019-000262

42. Roberts C, Torgerson DJ. Understanding controlled trials: baseline imbalance in randomised controlled trials. BMJ. 1999;319(7203):185. doi:10.1136/bmj.319.7203.185

43. Raab GM, Day S, Sales J. How to Select Covariates to Include in the Analysis of a Clinical Trial. Controlled Clinical Trials. 2000;21(4):330-342. doi:10.1016/s0197-2456(00)00061-1

	1mg	4mg	8mg	16mg
Background				
Race (n ,%)				
White	88 (48%)	95 (52%)	93 (50%)	83 (46%)
Black	44 (24%)	34 (19%)	39 (21%)	44 (24%)
Hispanic	51 (27%)	51 (28%)	53 (27%)	52 (29%)
Other	2 (1%)	2 (1%)	3 (2%)	2 (1%)
Age (mean)	36	35	35	36
Gender (n ,%)				
Male	117 (63%)	125 (69%)	120 (64%)	135 (75%)
Female	68 (37%)	57 (31%)	68 (36%)	46 (25%)
Education (n ,%)				
College Graduate	22(12%)	18 (9%)	12(6%)	16 (9.5%)
Partial college	55 (30%)	60 (33%)	60 (31%)	55 (31%)
HS Graduate	59 (31%)	54 (30%)	62 (33%)	60 (33%)
Partial HS	31 (17%)	30 (17%)	35 (19%)	38 (20%)
Junior High	18(10%)	18 (11%)	19 (11%)	12(6.5%)
Work (n ,%)				
Never Employed	51 (28%)	49 (27%)	51 (27%)	40 (22%)
Unskilled Employee	19 (10%)	13 (7%)	20 (11%)	18 (10%)
Machine Operator	20 (11%)	25 (14%)	23 (12%)	23 (13%)
Skilled Employee	39 (21%)	44 (24%)	38 (20%)	51 (28%)
Sales	38 (21%)	33 (18%)	40 (21%)	31 (17%)
Admin.	12 (6%)	15 (8.5%)	14 (8%)	16 (9%)
Business Manager	4 (2%)	2 (1%)	1 (.5%)	1 (.5%)
Higher Executive	2 (1%)	1 (.5%)	1 (.5%)	1 (.5%)
Income (mean)	19,529	20,458	20,458	20,938
Marital Status (n ,%)				
Married	43 (24%)	46 (25%)	52 (27.5%)	48 (27%)
Remarried	23(13%)	22 (12%)	21 (10.5%)	29(16%)
Divorced	48 (25%)	42 (23%)	33 (18%)	32 (18%)
Never Married	69 (38%)	72 (40%)	82 (44%)	71 (39%)
Drugs in House (n ,%)				
Yes	68 (35%)	68 (37%)	72 (38%)	67 (37%)
No	117 (65%)	114 (63%)	116 (62%)	113 (63%)

	1mg	4mg	8mg	16mg
Drug Use History				
Heroin (opiate) (n,%)				
Yes	185 (100%)	182 (100%)	188 (100%)	181(100%)
No	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Years of use mean)	11.11	10.72	11.65	12.91
Cocaine (n ,%)				
Yes	159 (86%)	150 (82%)	147 (78%)	146 (80%)
No	26 (14%)	32 (18%)	41 (22%)	34 (20%)
Years of use (mean)	6.8	6.6	7.8	7.7
Methamphetamine (n ,%)				
Yes	70 (38%)	64 (35%)	56 (30%)	56 (31%)
No	115 (62%)	118 (65%)	132 (70%)	125 (69%)
Years of use (mean)	3.7	2.5	3	5
Alcohol(n ,%)				
Yes	134 (74%)	130 (72%)	119 (63%)	119 (66%)
No	51 (26%)	52 (28%)	69 (37%)	62 (34%)
Years mean)	15.44	13.85	15.68	13.53
Tranquilizers(n ,%)				
Yes	82 (44%)	91 (50%)	84 (45%)	77 (43%)
No	103 (56%)	91 (50%)	104 (55%)	103 (57%)
Years of use (mean)	6	5.5	8.1	5.8
Marijuana(n ,%)				
Yes	136 (74%)	138 (76%)	131 (70%)	133 (73%)
No	49 (26%)	44 (24%)	57 (30%)	48 (27%)
Years of use (mean)	12.46	10.64	11.35	12.15
PCP(n ,%)				
Yes	29 (16%)	32 (18%)	41 (22%)	33 (18%)
No	156 (84%)	150 (82%)	147 (78%)	148 (82%)
Years of use (mean)	2.8	1.5	3.1	3.4

Summary Descriptives of Responder Proportions by Treatment Group

	1 Mg			4 Mg			8 Mg			16 Mg		
	0	1	p-value	0	1	p-value	0	1	p-value	0	1	p-value
	N=169	*N=16*		*N=145*	*N=37*		*N=142*	*N=46*		*N=125*	*N=56*	
RACE:			0.118			0.013			0.089			<0.001
White	84 (95.5%)	4 (4.55%)		81 (85.3%)	14 (14.7%)		76 (81.7%)	17 (18.3%)		68 (81.9%)	15 (18.1%)	
Black	38 (86.4%)	6 (13.6%)		21 (61.8%)	13 (38.2%)		25 (64.1%)	14 (35.9%)		31 (70.5%)	13 (29.5%)	
Hispanic/Other	47 (88.7%)	6 (11.3%)		43 (81.1%)	10 (18.9%)		41 (73.2%)	15 (26.8%)		26 (48.1%)	28 (51.9%)	
GENDER:			0.196			0.220			0.030			1.000
Male	104 (88.9%)	13 (11.1%)		96 (76.8%)	29 (23.2%)		84 (70.0%)	36 (30.0%)		93 (68.9%)	42 (31.1%)	
Female	65 (95.6%)	3 (4.41%)		49 (86.0%)	8 (14.0%)		58 (85.3%)	10 (14.7%)		32 (69.6%)	14 (30.4%)	
SCHOOL:			0.373			0.890			0.789			0.147
Minimum High School Educated	126 (92.6%)	10 (7.35%)		106 (80.3%)	26 (19.7%)		100 (74.6%)	34 (25.4%)		95 (72.5%)	36 (27.5%)	
Below High School	43 (87.8%)	6 (12.2%)		39 (78.0%)	11 (22.0%)		42 (77.8%)	12 (22.2%)		30 (60.0%)	20 (40.0%)	
WORK:			0.563			0.544			0.709			0.663
Unemployed	48 (94.1%)	3 (5.88%)		41 (83.7%)	8 (16.3%)		40 (78.4%)	11 (21.6%)		26 (65.0%)	14 (35.0%)	
Employed	121 (90.3%)	13 (9.70%)		104 (78.2%)	29 (21.8%)		102 (74.5%)	35 (25.5%)		99 (70.2%)	42 (29.8%)	
INCOME	14560 [7926;25000]	7730 [3882;19500]	0.053	15000 [7573;30000]	10800 [7690;26000]	0.356	11000 [5796;25000]	16000 [9120;25000]	0.215	14400 [8000;30000]	12000 [5828;21000]	0.090
MARITAL:			0.536			0.363			0.183			0.834
Unmarried	129 (92.1%)	11 (7.86%)		111 (81.6%)	25 (18.4%)		106 (78.5%)	29 (21.5%)		88 (68.2%)	41 (31.8%)	
Married	38 (88.4%)	5 (11.6%)		34 (73.9%)	12 (26.1%)		36 (67.9%)	17 (32.1%)		37 (71.2%)	15 (28.8%)	
LIVING:			0.502			0.773			0.620			0.962
Not Alone	138 (92.0%)	12 (8.00%)		129 (80.1%)	32 (19.9%)		117 (74.5%)	40 (25.5%)		98 (69.5%)	43 (30.5%)	
Alone/Unstable	30 (88.2%)	4 (11.8%)		16 (76.2%)	5 (23.8%)		25 (80.6%)	6 (19.4%)		27 (67.5%)	13 (32.5%)	
DRUGS:			0.454			0.100			0.074			0.682
Yes	64 (94.1%)	4 (5.88%)		59 (86.8%)	9 (13.2%)		60 (83.3%)	12 (16.7%)		48 (71.6%)	19 (28.4%)	
No/Dont_Know	105 (89.7%)	12 (10.3%)		86 (75.4%)	28 (24.6%)		82 (70.7%)	34 (29.3%)		77 (67.5%)	37 (32.5%)	
JAIL:			1.000			1.000			1.000			0.646
Yes	4 (100%)	0 (0.00%)		3 (75.0%)	1 (25.0%)		2 (100%)	0 (0.00%)		3 (60.0%)	2 (40.0%)	
No	164 (91.1%)	16 (8.89%)		142 (79.8%)	36 (20.2%)		140 (75.3%)	46 (24.7%)		122 (69.3%)	54 (30.7%)	
AGE:			0.954			0.851			0.759			0.571
greater than 35 years	88 (90.7%)	9 (9.28%)		69 (78.4%)	19 (21.6%)		74 (74.0%)	26 (26.0%)		72 (71.3%)	29 (28.7%)	
less than or equal to 35 years	81 (92.0%)	7 (7.95%)		75 (80.6%)	18 (19.4%)		67 (77.0%)	20 (23.0%)		53 (66.2%)	27 (33.8%)	
HEROIN_CRAVE	84.0 [70.0;100]	86.0 [67.2;100]	0.964	89.0 [73.0;100]	79.0 [51.0;98.0]	0.121	90.0 [66.0;100]	88.0 [79.0;100]	0.419	85.0 [67.0;100]	84.5 [75.0;100]	0.662

[1] Note. Medians, First and Third Quartiles reported for Quantitative Variables and Exact Fisher Test used for Categorical Variables