NIDA CTN Protocol 0094

Individual Level Predictive Modeling of Opioid Use Disorder Treatment Outcome

Lead Investigators: **Sean Luo, M.D., Ph.D.**
Daniel J. Feaster, Ph.D.

Sponsor: National Institute on Drug Abuse (NIDA)
23 August 2019
Version 1.2.7

Lead Investigator (LI):	**Sean X. Luo, MD, PhD** Greater New York Node Columbia University
Co-LI:	**Daniel J. Feaster, PhD** Florida Node Alliance University of Miami
CCTN Scientific Officers:	
	Landhing Moran, PhD National Institute on Drug Abuse
Data and Statistics Center (DSC):	**Paul Van Veldhuisen, PhD** DSC Principal Investigator The Emmes Company
Clinical Coordinating Center (CCC):	**Robert Lindblad, MD** CCC Principal Investigator The Emmes Company

CONFIDENTIALITY STATEMENT

This document is confidential communication. Acceptance of this document constitutes agreement by the recipient that no unpublished information contained herein will be published or disclosed without prior approval of the Lead Investigator or other participating study leadership and as consistent with the NIDA terms of award.

TABLE OF CONTENTS

1 LIST OF ABBREVIATIONS

ASI	Addiction Severity Index
BUP	Buprenorphine
BUP/NX	Combination Buprenorphine/Naloxone
CCTN	Center for the Clinical Trials Network
CDE	Common Data Elements
CFR	Code of Federal Regulations
CoC	Certificate of Confidentiality
COWS	Clinical Opiate Withdrawal Scale
CTN	Clinical Trials Network
CV	Cross validation
DSC	Data and Statistics Center
FTND	Fagerstrom Test for Nicotine Dependence
GCP	Good Clinical Practice
HHS	Department of Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act
ICH	International Council for Harmonisation
IRB	Institutional Review Board
LI	Lead Investigator
LSTM	Long Short Term Memory
HMM	Hidden Markov Model
MET	Methadone
MOUD	Medication treatment for opioid use disorder
NIDA	National Institute on Drug Abuse
OUD	Opioid use disorder
RBS	Risky Behavior Survey
RF	Random Forest
PITE	Predicted individual treatment effects
POATS	Prescription Opioid Addiction Treatment Study (CTN0030)
PR-AUC	Precision-recall Area Under the Curve
SF-36	Medical Outcomes Study Short Form 36-item Health Survey
SNP	Single Nucleotide Polymorphism
START	Starting Treatment with Agonist Replacement Therapies (CTN0027)
SVM	Support Vector Machine
TLFB	Standard 4-week Timeline Followback of drug use
VIMP	Variable Importance
UDS	Urine Drug Screen
X:BOT	Extended-Release Naltrexone vs. Buprenorphine for Opioid Treatment (CTN0051)
XR-NTX	Extended-release naltrexone

2 STUDY SYNOPSIS

2.1 Study Objectives

High rates of Opioid Use Disorder (OUD) represent a national public health crisis, and improvement in dissemination of evidence-based OUD treatment is a central treatment and research goal. A persistent problem in the dissemination of medication treatments for opioid use disorder (MOUD) is patient dropout, and matching patients to suitable medication early has the potential to minimize dropout. However, to our knowledge, there is no scientifically validated method to match patients to their optimal treatment and estimate their risk of relapse and dropout at the outset of or early on in treatment. An individual level predictive model that calculates the risk of relapse and suggests optimal medication choice can be broadly clinically useful in the community. Secondly, such a risk stratification model can be one of the first steps in the eventual development of adaptive treatment trials that modify treatment choice intelligently and algorithmically (e.g. sequential adaptive designs (Collins et al., 2014)). The overall objective of this study is to develop and disseminate individual level risk prediction models for both relapse and drop out built using a variety of strategies, using aggregated CTN datasets from 3 opioid use disorder clinical trials: traditional statistical approaches such as linear and logistic regression and penalized regression; newer machine learning techniques, such as Deep Learning (LeCun et al., 2015) will also be applied.

2.2 Study Design and Outcomes

This is a secondary analyses and data mining study using a harmonized collection of three multi-site clinical trials from the CTN, including Starting Treatment with Agonist Replacement Therapies (START, CTN0027) (Saxon et al., 2013), Prescription Opioid Addiction Treatment Study (POATS, CTN0030) (Weiss et al., 2011) and Extended-Release Naltrexone vs. Buprenorphine for Opioid Treatment (X:BOT, CTN0051) (Lee et al., 2018). The study is a harmonization and algorithm development and validation project. We aim to develop three sets of risk prediction algorithms: 1) we wish to predict the risk for an individual with opioid use disorder to drop out of treatment at the end of 6 months of treatment (*drop-out models*), defined as someone who did not present for any of the follow-up appointments for the last 4 weeks; 2) *abstinence models*: probability of someone achieving full abstinence (negative urine toxicology for opioids) in all of the last 4 weeks of treatment); 3) *time-dependent models*: an individual level prediction of trajectory of recovery through the 6 months of treatment. For each of these sets of outcomes, we will also use two sets of predictor data: 1) the baseline assessments such as demographics and psychological assessments, severity of existing substance use disorders and other psychiatric co-morbidies, and other relevant predictors, 2) the baseline assessments plus treatment response data at the end of the first month of treatment. These models will be cross-validated through three different populations as above.

2.3 Sample Size and Study Population

We will include all the participants who were randomized in the three studies cited above. For START, n=1269 participants were randomized. For POATS, individuals (n=653) were randomized to Opioid Dependence Counseling (ODC) plus Standard Medication Management (SMM) (n=329) versus SMM (n = 324) alone in a two-phase adaptive design, and hence they can be thought of as either having BUP alone or BUP plus a specific enhanced psychotherapy program. In X:BOT n=570 individuals were randomized from inpatient treatment facilities. While the three studies aimed to recruit individuals who present in different contexts, all 3 studies have had at least 6 months of planned follow-up period, standardized follow-up schedule and urine toxicology data, as well as overlapping and standardized baseline assessment instruments.

2.4 Treatment/Assessment/Intervention and Duration

We will not enroll any participants directly for this study. Here we will summarize the interventions performed in the trials cited in the study. For START, participants were treatment-seeking individuals with opioid dependence. The START treatment was outpatient and consisted of 24 weeks of open-label treatment of 1) BUP/NX delivered as daily self-administered sublingual formulation or 2) Methadone delivered daily in either liquid or tablet formulation. The primary outcome was liver-toxicity at 24 weeks.

For POATS, participants were treatment-seeking individuals who had a prescription opioid dependence diagnosis. The POATS treatment was in two phases. In the first phase, participants received either 1) BUP/NX and standard medical management of BUP/NX for 4 weeks with a taper of BUP/NX in the last two weeks or 2) enhanced counseling+standard medical management of BUP/NX for 4 weeks with a taper of BUP/NX in the last two weeks. All participants were followed for an additional 8 weeks to assess treatment success (defined by the original trial investigators: completing week 12 with self-reported opioid use on no more than 4 days in a month, absence of 2 consecutive opioid-positive urine test results, no additional substance use disorder treatment (other than self-help), and no more than 1 missing urine sample during the 12 weeks.) Treatment failures in phase one were re-randomized in phase 2 to a similar protocol, but the active BUP/NX lasted for 16 weeks and again there are 8 follow-up weeks to assess outcomes of phase II.

For X:BOT, the treatment consisted of 24 weeks of open-label treatment of 1) XR-NTX (Vivitrol; Alkermes) delivered as monthly intramuscular injections or 2) BUP/NX (Suboxone; Indivior) delivered as daily self-administered sublingual film. The primary outcome measured was opioid relapse-free survival during 24 weeks of outpatient treatment, and urine toxicology was measured throughout the study period.

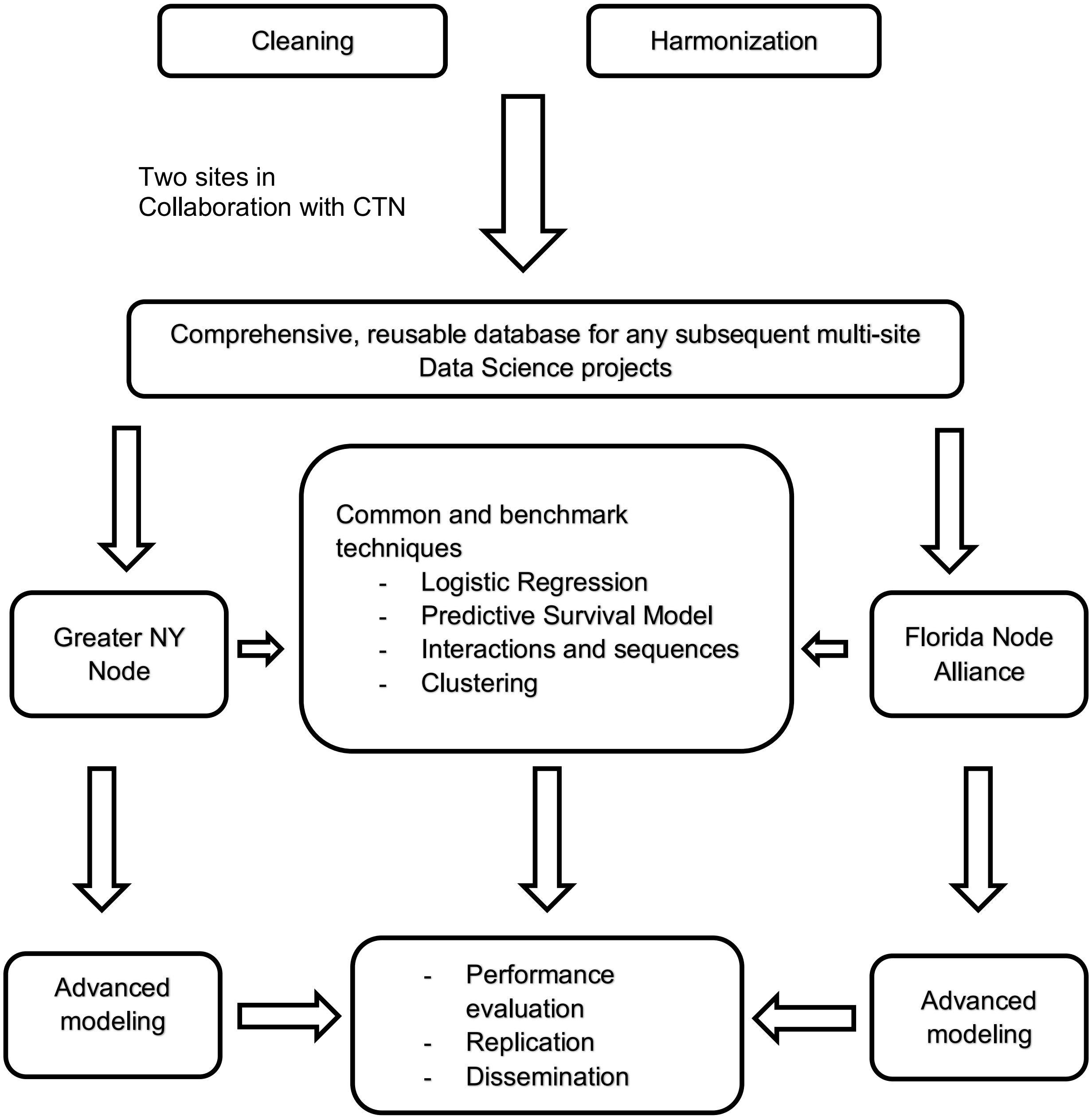
2.5 Safety Reporting

This is not applicable for this study.

2.6 Analyses

This study is designed to develop a set of predictive models that would allow risk stratification of an individual’s risk of drop-out or relapse at the start of or very early on in treatment. Each of the predictive models will have a similar input-output structure, though the underlying modeling methodology would differ. We will therefore analyze and evaluate the models using conventional metrics of predictive performance: cross validation and associated Precision-Recall Area under the Curve (PR-AUC). PR-AUC is a better metric of predictive performance for datasets with unbalanced outcomes, which would account for differences in different treatment groups having different dropout rates. The performance measures will be tabulated. Secondarily, we will analyze the relative importance of specific predictors in the models using a standard non-parametric measure (VIMP), with the caveat that some models are internally complex (black box) and incorporate possibly nested layers of predictor interactions and non-linearities.

3 STUDY SCHEMA



4 INTRODUCTION

4.1 Background and Significance to the Field

While high prevalence of OUD is a public health crisis, there are existing treatments that are effective and based on sound scientific evidence. Three large comparative effectiveness trials conducted through the CTN have shown that all three treatments (BUP, MET, XR-NTX) are comparably effective and reduce relapse rates from up to 90% to approximately 50%. While this is an encouraging result, one might wonder why the remaining 50% or drop out early in treatment trials or otherwise cannot achieve abstinence. Secondly, within the 50% who do achieve abstinence at the end of the trial, they do so with distinct patterns and trajectories: some achieve abstinence early, others gradually (Hser et al., 2017). These divergent patterns of treatment response at the individual level present a challenge and an opportunity for precision medicine: is it possible to predict, on an individual basis, who will respond to treatment and maintain abstinence, and who will relapse or drop out? Currently, assignment of patients to medications is mainly arbitrary, based on convenience or patient/provider opinions (Matusow et al., 2013). Quantitative models might allow for better standardization of care delivery and better identification of higher-risk subgroups. This information could also be useful for patients and family members in every day clinical practice. Predictive models would have a role in encouraging patient-centered care in the field of substance abuse treatment, similar to their current roles in cancer care (Lasonos et al., 2008).

Study Rationale

This would be the first study to apply state of the science machine learning techniques to any large multisite randomized trial in opioid use disorder, and our effort may generate a gold standard risk scoring instrument for both relapse and dropout in the treatment of opioid use disorder. We conceptualize the study as both applying new techniques as well as developing new methods. In particular, conventional subgroup analyses are prone to type-I errors (Pocock et al., 2002) and cannot compare treatment-covariate interactions by calculating how well they perform in prediction or whether matching treatment improves outcome. Rather, conventional methods only provide a table of p-values (Rothwell et al., 2005). The proposed approach, on the other hand, focuses on predictability rather than statistical significance, and might lead to easy-to-calculate risk stratification scores similar to other scores such as TIMI (Antman et al., 2000a) and CHADS2 (Henriksson et al., 2010) scores that are successfully applied in other areas of medicine. Secondly, newer and more nonlinear methods have been applied to complex behavioral health phenotypes and gene X environment interactions, though with a mix of success (Bzdok and Meyer-Lindenberg, 2018), but have not been applied to OUD treatment data.

5 OBJECTIVES

5.1 Primary Objective

The primary objective of this study is to develop a risk score that would provide a quantitative and qualitative measure of specific clinical outcomes (i.e. drop out/achieving abstinence) in the treatment of OUD. Our hypothesis is that risk scores that incorporates baseline and early treatment response, including predictor by treatment interactions, will outperform conventional linear models using baseline covariates obtained from traditional subgroup analyses. An associated hypothesis is that a more complex model incorporating nonlinearities (i.e. Random Forest) will result in better overall predictive performance.

5.2 Secondary Objective(s)

The secondary objectives of this study include:

1. Develop a comprehensive, harmonized database of treatment data, including baseline covariates, instruments and measures obtained during the treatment, and collected outcome data. This serves as a comprehensive gold-standard CTN database for OUD treatment that can be available for future scientific studies.

2. Estimate the relative importance of predictors in the best predictive models obtained above and assess whether future studies of OUD in the CTN can use a substantially reduced set of baseline covariate measures (i.e. can a reduced dataset capture most of the predictive information?). This set of covariates may constitute future Common Data Elements (CDE) for future OUD studies.

3. Create a replicable data pipeline which can be used to process future Opioid Data to both validate and update the results of the future protocol.

5.3 Exploratory Objective(s)

Given the richness in the enumerated CTN datasets, we will also explore the following objectives:

1. Assess whether a limited set of genetic information, primarily common polymorphisms in the opioid receptor genes, might affect predictive performance on an individual level. This dataset is available for START and X:BOT studies.

2. Assess whether baseline characteristics can be used to construct a predictive model for long-term outcome of OUD (up to 5 years). This dataset is available for START and POATS studies, and may become available for X:BOT in the future.

6 STUDY DESIGN

6.1 Overview of Study Design

This study is conceptualized as a domain-specific scalable predictive modeling project, and hence contains several common phases (see Section 3.0, Study Schema). First, data from several sources with common elements, such as the same baseline screening instruments and measures collected during the trial (TLFB, etc.) collected at the same study times would be merged into a single, complete database. Measures that are only collected in one study will be treated as missing values for the other studies. This step also requires data cleaning, verification and replication for quality assurance. Secondly, conventional predictive models will be developed using standard statistical methods (predictive survival models, predictive regression models), and benchmark performances will be measured and replicated at the two study sites. Model comparison will be conducted using cross-validation using both data from one study and across 3 different studies. Thirdly, each of the sites will then generate advanced prediction models incorporating nonlinearities and interactions. This phase will include development of new predictive modeling approaches tailored to OUD clinical trial data, as detailed in the subsequent sections. Finally, the models constructed will be compared using standard model comparison methods. The most successful models/risk scoring instrument will be presented as gold-standard risk predictive models and planned for dissemination: a web-based application will be developed where clinicians may enter patient level data and receive an estimate of risk of drop-out or treatment success.

6.2 Study Timeline

The overall timeline for this study is approximately 24 months. We anticipate the following duration for each component of the study:

<u>Data harmonization and cleaning</u>:12-13 months

1. Initial phase (6 months): in the first 6 months, we expect to harmonize the data sources and create a new library of common data elements within the existing dataset, including basic demographic information, previous history of substance use disorders and other psychiatric diagnoses, baseline urine toxicology data, as well as elements in outcomes data such as longitudinal urine drug screens, timeline followback data, and derived outcome data.

Milestone	Time	Deliverable
Database consolidation	1 month	Cleaned investigator databases obtained from DSC
Variable cleaning	2 months	Generation of a multi-study variable dictionary, matching study participant data from one study to another
Database harmonization	2 months	Harmonization of three databases through repeated query and re-entry of existing datasets
Final data cleaning	1 month	Cleaned and annotated combined dataset

2. In the second phase of harmonization, more complex strategies will be developed for resolving the issue that different instruments were used for the same characteristics in different studies, with several studies having substantial missing values due to variations in practice at different sites. We will therefore budget another 6-7 months where instruments for assessing similar characteristics will be compared through a focused literature search and calibrated to one or more standard derived item. Examples of this include Self-Reported Opioid Withdrawal Scale (SOWS) vs. Clinical Opioid Withdrawal Scale (COWS), which can be summarized into a single opioid withdrawal severity value. Modeling using the simplified harmonization in step 1 will occur concurrent to extended harmonization in step 2.

Milestone	**Time**	**Deliverable**
Consolidation of different inventories in different datasets through review of literature	2 months	Plan for harmonization of different instruments
Modeling imputation of different instruments	2 months	Models for migrating results from one type of instrument(s) to another type of instrument(s) collected in different studies
Database harmonization	2 months	Inclusion of finalized standard derived measures in the new database
Final data cleaning	1 month	Cleaned and annotated combined dataset

Model development, performance evaluation: 12 months

Milestone	**Time**	**Deliverable**
Traditional predictive modeling and risk scores	3 months	Risk score based on Generalized Linear Models (GLMS): logistic regression, survival models
Adaptive predictive modeling	3 months	Predictive models and performances calculated using Ensemble Learning and Deep Learning methods.
Assessment of variable importance and establishing Common Data Elements (CDE)	3 months	Calculation and tabulation of standard VIMP scores for predictors incorporated into the models

Final data cleaning and reporting	3 months	Combined reports of performances of variable predictive models

Dissemination and reporting: 6 months

Risk score dissemination	3 months	Calculation and tabulation of a OUD treatment drop-out score
Web-based predictive model dissemination	3 months	Building a predictive model interface for clinician access

7 OUTCOME MEASURES

7.1 Primary Outcome Measure

There will be two primary outcomes examined:

1. Probability and estimated time to drop out of treatment at the end of 6 months of treatment, defined as someone who did not present for any of the follow-up appointments for the last 4 weeks. This is the standard binary outcome that has been used for OUD dropout at a particular time (Saxon et al., 2013; Weiss et al., 2011).

2. Probability of someone achieving full abstinence (negative urine toxicology for opioids in all of the last 4 weeks of treatment). While there are several reportable definitions of abstinence, we here operationalize a more stringent definition (Hughes et al., 2003) that is more easily interpretable in clinical practice.

7.2 Secondary Outcome Measure(s)

For modeling, there will be several secondary outcomes of interest:

1. Dynamic modeling of trajectory of urine toxicology data through the 6 months of treatment. Evaluation of thse time-dependent models with be through time-dependent ROC analysis.

2. Dynamic modeling trajectory of self-reported patterns of use, through the 6 months of the trial.

Corresponding to our secondary objectives, milestones for each of the objectives in Section 5.2:

1. A harmonized, de-identified database for all participants who were randomized into the 3 cited CTN trials, with annotation derived from the investigators database.

2. A list of standardized predictors variables by prediction importance.

3. An informatics schema for future CTN clinical trials to import trial data into a standard format.

7.3 Other Outcome Measures

We will also explore whether baseline characteristics can be used to construct a predictive model for long-term outcome of OUD (up to 5 years) for the subset of data for which this is available. These models will be exploratory in nature. We will also explore the potential of incorporating genetics predictors in our model

Outcomes will include:

1. Indicator of ongoing MOUD treatment during the period of long-term follow-up.

2. Indicator of successful recovery at long-term follow-up.

3. Predictive performance for outcomes 7.1 and 7.2 with additional genetics information incorporated.

8 STUDY POPULATION

We aim to develop a harmonized dataset that includes all randomized participants from the three studies cited in Section 2.2. The harmonized dataset will therefore be a heterogeneous population of individuals meeting criteria for opioid use disorder who were recruited in a variety of pragmatic settings. The details of the harmonization procedure will be elaborated in sections below. In particular, the three studies have different settings with respect to study population. In START, START, n=1269 participants were randomized, with 740 randomized to buprenorphine-naloxone (BUP/NX) and 529 randomized to methadone (MET), with a good representation of individuals who remained in treatment (n=340 for BUP and n=391 for MET) vs. missed more than 14 days of treatment (n=251 for BUP and n=94 for MET). For POATS, individuals (n=653) were randomized to Opioid Dependence Counseling (ODC) plus Standard Medication Management (SMM) (n=329) versus SMM alone (n = 324) in a two-phase adaptive design, and hence they can be thought of as either having BUP alone or BUP plus a specific enhanced psychotherapy program. In X:BOT, n=570 individuals were randomized from inpatient treatment facilities with n=283 randomized to XR-NTX and n=287 randomized to BUP and followed-up as outpatients.

9 SITE SELECTION

9.1 Number of Sites

This analytics study will utilize two sites to develop, replicate, and disseminate risk scores for MOUD.

9.2 Site Characteristics

Greater New York Node

The CTN Greater New York Node has established a team of researchers to focus on developing individual level predictive models in substance use disorder. We have a team of treatment researchers (Nunes, Luo) as well as an ongoing collaborative relationship with the Mental Health Data Science (MHDS) group at New York State Psychiatric Institute headed by Dr. Melanie Wall. The MHDS provides statistical collaboration, data analytic methodological development, and data management for psychiatric and mental health research conducted within the Department of Psychiatry, Research Foundation for Mental Hygiene (RFMH), and the New York State Psychiatric Institute (NYSPI). This site can serve as a model for a CTN Data Science platform for future projects in other substance use disorders.

Florida Node Alliance

The CTN Florida Node Alliance has a team of researchers (led by Dr. Feaster) that have focused on methods for predicting individual level treatment effects (Lamont et al., 2018; Lu et al., 2018). The Biostatistics Division of the Department of Public Health Sciences at the University of Miami Miller School of Medicine has numerous faculty and students working on related research; this environment will be conducive to successful completion of CTN-0094.

9.3 Rationale for Site Selection

A successful collaborative Data Science research project requires multiple types of expertise within a single site. Given the extraordinary complexity of OUD clinical trial data and the potential methodological challenges, it is essential that such projects are led by teams composed of translational researchers who have both domain expertise in addiction treatment research and methodological expertise. There are only a few such sites (such as the proposed two sites) within the CTN that has such expertise. In addition, the proposed sites have been engaged in delivering treatment in several of OUD effectiveness trials, and the Greater NY Node was the lead site for X:BOT.

9.4 Leadership Plan and Team Structure

The study will be led by the **Lead Investigator (Dr. Luo)** and **Co-Lead Investigator (Dr. Feaster)** who will each lead a team at their respective site consisting of a masters level statistician who is responsible for data cleaning and harmonization, a research assistant, and a PhD level biostatistician who will be responsible for building predictive models, performance evaluation, collaborative coding, and web-based dissemination. Drs. Luo and Feaster will reach executive decisions through consensus, and in the unlikely event of disagreement, will seek advisory through Dr. Edward V Nunes, who is the co-Lead Investigator of the CTN Greater New York Node. The two teams will conduct regular (approximately once every two weeks) team meetings to discuss study task list delegation, model building progress and performance comparison, and other related tasks.

In addition, given that the goal of this study is to produce clinically valuable models that may assist risk stratification, we aim to utilize the significant clinical expertise of a composite team of consultants who are all CTN-affiliated Principal Investigators for the three CTN OUD trials, including Dr. Andrew Saxon, Dr. Roger Weiss, and Dr. John Rotrosen. In addition to Dr. Luo, all three of the PI are board certified and practicing addiction psychiatrists, who also have had ongoing and significant leadership and administrative roles on the clinical side of OUD treatment. Regular team meetings will occur in discussions structured around optimizing model development and dissemination for clinical practice. Once the team feels we have a sufficiently developed model and prior to our final report, we will plan a presentation to the wider audience of CTN clinicians to obtain feedback on the models and ways to maximize the clinical utility of the results.

10 STUDY PROCEDURES

10.1 Data Harmonization

Data harmonization is defined as a process where data from different sources are combined in a way that would allow users to conduct analyses that exploit the commonalities between different datasets. In this study, the first step involved in developing a predictive model for treatment response in OUD is to create a large database of all randomized individuals in all three studies. Each individual will be indexed with a unique CTN ID and identified by the study with which he/she has participated. This allows the database to provide an easy way to query information from 3 studies together as well as separately. Furthermore, this database will allow us to query data for individuals in different studies but at similar times of the treatment (e.g., buprenorphine 1 week post induction, regardless of whether the participant was inpatient or outpatient, heroin user or prescription pill user). We therefore conceptualize the harmonization process being anchored by time, as all three studies have a roughly six-month follow-up process. Secondly, other than baseline instruments, many follow-up data points are missing due to participant no-shows. Harmonizing data sources will also provide a principled way to impute missing values for subsequent predictive modeling work. Missing values can be imputed within study or across studies. In the final harmonized dataset, the studies have a common set of outcomes using Urine Drug Screen (UDS) tests despite differences in measures of outcomes that we will detail below. The harmonization will therefore address the differences in how UDS and other outcome measures were conducted.

The harmonized database will have a two-dimensional structure indexed by time, as measured by visit week, and participant. Each measured instrument will be a sub-dataset derived from the original Emmes dataset, but with entries harmonized using the reference determined through consensus and separately documented by the data management team to account for possible discrepancies in the entry and numbering of the instruments in different studies. In particular, START and X:BOT, while having differences in enrolled population, have similar designs, with a parallel group randomization, pragmatic dosing adjustment during maintenance, and a six-month follow-up. POATS has a different design, since there is in particular two medication taper-phases. However, we can consider the taper dosing as a pragmatic dosing schedule for BUP treatment for prescription opioid users. Hence the two phases, totaling 6 months, map well onto the six-month follow-up schedule of the other two studies.

10.1.1 Baseline Harmonization

At baseline, as stipulated by the protocol, the following instruments were collected by the START investigators: demographics, Clinical Opiate Withdrawal Scale (COWS), DSM-IV checklist, Risk Behavior Survey (RBS), vitals, Medical Outcomes Study Short Form 36-item Health Survey (SF-36), Fagerstrom Test for Nicotine Dependence (FTND), Standard 4-week Timeline Followback (TLFB), as well as medical and psychiatric assessments, including physical exam and psychiatric evaluations, and standard laboratory measurements. A small number of individuals also had Addiction Severity Index (ASI) recorded. The UDS was standardized to 10 substances. For POATS, a similar set of instruments was collected: demographics, DSM diagnoses, vitals, labs, COWS, TLFB, SF-36, RBS, FTND, and ASI. In addition, POATS also collected Visual analogue scale for craving (VAS), Beck Pain Inventory and Beck Depression Inventory. POATS used the Composite International Diagnostic Interview (CIDI) as opposed to the DSM-IV checklist. We propose to convert both to a standard list of binary diagnostic flags for major psychiatric diagnoses: major depressive disorder, bipolar disorder, post-traumatic stress disorder (PTSD), generalized anxiety disorder, panic disorder, schizophrenia or other psychotic disorders or personality disorders. In X:BOT, at baseline individuals similarly received labs, TLFB, UDS, ASI-Lite, and FTND. X:BOT additionally collected Hamilton Depression Scale (which can be converted to a corresponding Beck Depression Scale for depression severity). Further, X:BOT used a different instrument for assessing for risky behavior called Risk Assessment Battery, which will be converted to RBS. Furthermore, since X:BOT enrolled inpatients, there was a record of data on detoxification such as number of days and medications used. For quality of life measures, X:BOT uses EQ-5D (EuroQol Inventory 5-Dimensional Measure) rather than SF-36, and while an explicit conversion algorithm would be used, it has unfortunately been shown that these two measures are not tightly coupled (regressing component scores using polynomial spine can have an R-square of ~0.60) (Franks et al., 2004). Finally, X:BOT used Subjective Opioid Withdrawal Scale (SOWS) rather than COWS to measure withdrawal severity. For harmonization and cleaning, we will include all non-overlapping instruments, which means that many participants in the database will have missing values for instruments that were measured in one study but not another. Manipulation and creation of derived variables will appear during the model development stage.

10.1.2 Outcome Harmonization

Outcomes measures of the three studies were similar: all obtained UDS and TLFB on the weekly visits. Secondarily, all three studies used standard dose logs and medical management logs to record adjunct medication uses (if any). POATS, due to its specific goal relating to prescription opioid use, had a more extensive record of subjective pain measures. POATS also had more missing UDS between weeks 17-24 as BUP taper in the second phase occurs between weeks 13-16. Because of its multi-phase design, POATS enrolled two groups of participants: the first group completed the study in phase I or dropped out. The second group was randomized in phase II. We will harmonize the two participant groups separately: the first group (a very small group, 6.3% of the total population, or n=43) will have 12 weeks of data by design as they were successful after a very short course of treatment. A total of n=163 were lost to follow-up and never randomized in the second phase. These individuals will also be included, but 12-24 week data will be recorded in the database as missing. Finally, the 360 individuals randomized to the second phase will have full 24-week treatment records, and we will record their data as starting from the second phase, since they were re-inducted to BUP, and their outcome data would follow the same timeline as the other two studies.

10.1.3 Follow-up Instrument Harmonization

While TLFB and UDS were universal outcomes, other secondary outcomes were also collected in the three studies. For example, both POATS and X:BOT had longitudinal VAS craving data, and START had more extensive compliance records. For the initial harmonization, we will include all the intermediate inventories in the database, indexed by time from treatment initiation.

Because the studies were designed differently, there were also follow-up data available after the 24-week treatment phase. For example, START had TLFB, UDS and dosing records available at week 32; X:BOT had records of these data at week 36. We anticipate that as we harmonize the datasets, a significant portion of the data will be missing. We will continue to index all recorded data in the comprehensive database by time from initiation of treatment. The harmonized data will be comprehensively documented using an indexed data dictionary as well as a combined assessment inventory table and time line.

10.2 Individual Level Prediction Models

10.2.1 Data Pre-Processing

Data pre-processing will consist of obtaining derived values from the comprehensive database. We will provide here a set of standard variables that will be obtained from the harmonized and cleaned database, which can be also be released as part of the public resource. The standard set of predictors will include both itemized instruments as well as the total scores for the specific dimensional inventories (e.g., Hamilton Depression Scale) as well as subsection scores.

We will obtain the following standard outcome measures: 1) drop-out at week 24. This is defined as missing all 4 visits between weeks 20-24. 2) Complete abstinence: attending all scheduled clinic visits between 20-24 and providing negative UDS for opioids. 3) Total number of negative opioid UDS in the weeks 20-24.

While modeling each of the predictors as a univariate item has the advantage of easy interpretability, for a dataset of this type, significant dimensionality reduction will likely be needed. In particular, both outcomes and predictors can be thought of complex multi-dimensional objects; while we may simplistically define an individual's status at week 24 to be either dropped out or not dropped out, the reality is that the majority of individuals present to some visits but not others and provide a mixture of positive and negative UDS. It may be advantageous to model directly the high-dimensional statistical structure. There are numerous statistical methods to conduct dimensionality reduction with high-dimensional data (typically called unsupervised learning in machine learning literature) (Dy and Brodley, 2004), and there is no "off-the-shelf" recipe for feature selection and feature design. We propose to use several strategies to model the predictor dataset. First, we propose to use two sets of predictors: baseline predictors and "early treatment phase" predictors. It has been recognized that the efficacy of treatment of OUD at 6 months may be more easily and precisely predicted after the patient received treatment for a short period of time (typically one month). The status of the patient before and shortly after the initiation of treatment would therefore significantly inform us in making the most accurate predictions.

10.2.2 Drop-Out Models

We will build the following benchmark models for a binary predictive model of whether an individual patient will drop-out of treatment:

1) Logistic regression with univariate predictors at baseline
2) Logistic regression with univariate predictors and treatment assignment interaction
3) Logistic regression with multivariate predictors using step-wise variable selection (largest 10 effect sizes with p-values < 0.05)
4) Logistic regression with early treatment response of week 1-4
5) Logistic regression with early treatment response of week 1-4 and their interaction with treatment group assignment

After obtaining benchmark models, we will aim to test the performance of several more complex models. We will test the following methodological approaches: Support Vector Machine (SVM), Random Forest (RF), Penalized regression (LASSO and group LASSO), Vector Quantization, and Ensemble Learning.

10.2.3 Abstinence Models

Similar to Drop-Out Models, we will build benchmark models for a binary predictive model of whether an individual will achieve full abstinence of opioid use at the end of six months. We will use similar modeling strategy as in developing Drop-Out Models.

10.2.4 Time Dependent Models

In the clinic, it is useful to estimate an individual's probability of drop-out or failure to achieve abstinence at any particular time following the (re-)initiation of treatment. It is therefore useful to provide an interpolated treatment response trajectory—an estimate of probability of attending clinic and providing a negative urine as a function of time, as modified by existing known characteristics of the individual. This problem is classically a scalar-on-function regression: the outcome variable is a function (in time). There is also a large literature for methodological approaches (Goldsmith et al., 2011) to solve the problem that both the predictor dataset and the functional outcome are high-dimensional and incomplete, and therefore requires regularization. We propose to use standard penalized spline regression with polynomial splines to obtain the probability trajectory for an individual receiving OUD treatment. We will again use the same set of benchmark predictors as listed above to generate these models.

We will also test several new methods recently developed for the construction of time-dependent predictive models. Typically, these models are probabilistic graphical models: the structure of the model can be expressed as nested conditional probabilities between random variables, though the fitting procedures and the structure of the models are divergent. Whether these methods are useful for specific applications typically depends on if the underlying statistical structure is readily detectable by the method. One of the secondary goals of this study is to assess whether the current OUD clinical trial datasets exhibit such structures and whether known algorithms could detect these structures. In particular, we will test the following approaches: coupled Hidden Markov Models (HMM), Bayesian Belief Network (BBN), Long short-term Memory (LSTM), and assess whether these methods would produce reliable predictive models. In particular, these models would also generate binary outcomes as defined above and can be compared against binary classification models as enumerated above.

10.3 Models of Secondary Objectives

10.3.1 Predicted Individual Treatment Effect (PITE)

It is necessary to get predicted individual treatment effects (PITE, Lamont, et al. 2018) to determine which treatment will have the best outcomes for an individual. The use of PITEs allows the comparison of an individual's response to different treatments. We will use the assigned treatment assignment in the various clinical trials to create these PITES. We use the generic term, machine, for a learner that is used to make this prediction. These machines could be based on a deep-learning approach, random forest, or Lasso, for examples.

Counterfactual machines. To describe our approach for creating PITEs we begin by introducing some notation. Let $\{(T_1,\mathbf{X}_1,Y_1),\ldots,(T_n,\mathbf{X}_n,Y_n)\}$ denote the data where $\mathbf{X}_i$ is the p-dimensional covariate (feature, independent variable) for patient i and Y_i is the outcome. For this illustration, we assume that Y_i is a binary outcome, $Y_i \in \{0,1\}$, such as drop-out within 6-months. Note that this methodology applies to general outcomes; for example, it applies to multiclass (categorical) outcomes $Y_i \in \{C_1,\ldots,C_J\}$ and continuous outcomes. Variables T_i record the treatment for patient i. For simplicity, we assume for the moment that the treatment is one of two values, $T_i \in \{0,1\}$. Each patient i is administered one of the two treatments. Thus $T_i = 0$ or $T_i = 1$, however we would like to know what the predicted probability for Y_i is under both treatment regimens, even though we know very well that the patient can only experience one treatment within the trial. That is, we would like to predict $p_{i,0} = \mathrm{P}\{Y_i = 1 | T_i = 0, \mathbf{X}_{i,}\}$ and $p_{i,1} = \mathrm{P}\{Y_i = 1 | T_i = 1, \mathbf{X}_i\}$.

To do so, we create two machines under each treatment type and use one for counterfactual inference. A machine for treatment $T = 0$ is constructed by running a specific learner (or ensemble of learners) for classification, given that our example is binary) using only the data for patients with treatment $T_i = 0$. Call this machine M_0. Likewise, a machine for treatment $T = 1$, denoted by M_1 is constructed by running the classification machine using only data with $T_i = 1$.

Now given a patient *i* with $T_i = 0$ we obtain *i*'s predicted value using M_0. To obtain *i*'s counterfactual probability, we assume there is a clone of *i* that is identical to *i* in all ways except that the clone has received the alternate treatment. Thus, the clone has an identical $\mathbf{X}_i$ but differs because $T_i = 1$. Then to obtain the clone's estimated outcome, we simply drop the clone down M_1 and obtain the predicted probability which represents *i*'s counterfactual probability estimate. The value, represents the estimate of treatment effect for *i.* Note that when $T_i = 1$ a treatment effect estimate is obtained in an analogous fashion using M_0 as the counterfactual machine.

10.3.2 Variable Importance and Model Selection

We will create PITE estimates (Y*) comparing the three MOUD treatments using the various estimators we will examine (e.g. Support Vector Machine (SVM), Random Forest (RF), Penalized regression (LASSO and group LASSO), Vector Quantization, and Ensemble Learning). As noted above, these PITE estimates will be the difference in predictions of response on two different (pair-wise) treatment comparisons at the individual level. The predictors of this difference will vary from the predictors of the individual treatments. We will therefore utilize both Random Forest and penalized regression on these different PITE estimates to examine which variables are most predictive of the difference in the performance of pairs of treatments.

10.3.3 Selection of Minimally Sufficient Common Data Elements

We will examine the performance of our PITE estimates utilizing the full feature sets and compare these to the performance of PITES estimates with the most important features included. This model selection will be done using both an RF model of the PITES and on the penalized regression model selection of the PITES. RF provides a rapidly computable internal measure of variable importance (VIMP) that can be used for ranking variables and for variable selection, which is especially useful for high-dimensional as well as low-dimensional data. VIMP refers to the extent of increase in prediction error when that variable is permuted. In order to calculate a variable's importance, the given variable is randomly permuted, and the permuted data are dropped down the trees in the forest (i.e. all the decision-rules of the forest are applied to the new data, which has the one variable permuted, to get a new prediction error). This new prediction error is then compared to previous prediction error from the un-permuted data. This difference is the VIMP of the variable. The larger the VIMP of a variable, the more predictive the variable. It therefore indicates a more important variable compared to others (Breiman, 2001). Utilizing both RF and penalized regression will ensure that we have a parsimonious set of predictors, but we do not miss important non-linearities and higher order interactions that may be important for prediction.

10.4 Exploratory Analyses

10.4.1 Models Incorporating Genetics

Genetic variation may influence treatment response in MOUD (Crist et al., 2018a). While the current dataset precludes a more definitive incorporation and development based on genetics data, we propose to use the existing genetics data from the two CTN studies to explore the effect of incorporating genetics into the predictive models that we will build. For the START (CTN-0027) study, single nucleotide polymorphisms (SNPs) were sequenced using a method to maximize genotyping coverage of candidate genes including the gene encoding the delta-opioid receptor (*OPRD1*) and the mu-opioid receptor (*OPRM1*). For *OPRD1* it was found that 6 SNPs captured 67% of SNPs in the region with a minor allele frequency cutoff of 10% using the International HapMap Project. Several additional variants were sequenced due to existing evidence suggesting that the variant was related to opioid use disorder. Altogether, 36 alleles were sequenced for a subsample of n=664 individuals for that study (Crist et al., 2013). For *OPRM1*, 4 common haplotype blocks were identified by four SNPs, but only in European-American participants (n=582), as such blocks were not able to be identified due to lack of statistical power in the sample (Crist et al., 2018b).

We will code the SNP data as additional categorical predictors obtained at baseline, and build additional predictive models using the same model development plan as above. Given the limitations of the existing sample and the heterogeneity in the genetics data, our aim for this exploratory analysis is to demonstrate feasibility of incorporating these data sources into a conventional clinical trial dataset. In addition, genetics samples acquired from X:BOT (CTN-0051), which are currently being processed, may be incorporated into this database in the future once the analytic framework is completed.

10.4.2 Models of Long Term Outcomes

A long-term follow-up dataset exists for the START study: all randomized study participants were followed-up for approximately 2-8 years, with a mean of 4.5 years post-randomization. 2 sites were dropped due to logistics and difficulty in staffing, but out of 1080 targeted participants, 89.4% were located and 797 were interviewed, with n=795 individuals providing long term follow-up Timeline Followback (TLFB) data, including death, opioid use in the past 30 days, and treatment participation (Hser et al., 2016). A similar study exists for the POATS study with n = 375 individuals followed-up to up to 42 months after main trial enrollment (Weiss et al., 2015). While such dataset does not exist for X:BOT, the CTN-0051 study team is currently planning on collecting similar data for this study.

Harmonization of long-term treatment outcome from existing START and POATS follow-up would provide an opportunity to study several important questions, including the potential to predict whether someone who was successfully treated initially could safely discontinue MOUD after treatment. For the proposed protocol, we aim to 1) collect the long-term follow-up data from the study investigators and harmonize the TLFB and demographic data in the follow-up samples, and 2) explore the performance of predictive models using long-term abstinence as the outcome of interest.

10.5 Model Dissemination

10.5.1 Risk Scores

Using models obtained from Section 10.2.2. we will obtain a simple risk score (i.e. a standard CTN-OUD Treatment Risk Score) that gives an estimate of an individual remaining in treatment at the end of six months. The risk score will have several characteristics: it will be easy to calculate and can be done with pen and paper; it will be easily interpretable through a table look-up and level stratification (high, medium, low risk groups), and can be actionable in the future in either designing novel implementation strategies or designing new treatment trials.

There are several different strategies to develop risk scores with competing risks (Austin et al., 2016). We will initially use a simple approach to use the variables in Section 10.2.2 with combined robust locally weighted smoothing of categorical and numerical predictor variables, which convert them to factor variables (Zhang et al., 2017). We will then refit the factors to a logistic regression model to calculate the score for each of the factors. The original logistic regression equation will used to estimate the probability of drop out for each level of risk.

Generating risk scores from complex machine learning models is an active area of research. In particular, these models are largely nonlinear and non-convex, and the predictions often do not relate to interpretable trends in the predictors. Risk scoring, on the other hand, requires the resulting score to be monotonic and retain intuition in proportionality. We propose here that if a number of more complex methods in fact achieve substantial predictive performance enhancements, a competing "model-based" risk score be generated imposing a soft-max output layer, which imposes a probability interpretation of a categorical output (Bridle, 1990). This method can be applied to any number of algorithms involving ensemble learning or neural networks.

10.5.2 Web-based Predictive Instruments

We will present the final model using a user friendly format following other risk scores for primary and secondary prevention in medicine such as the Framingham Score (Lloyd-Jones et al., 2004) and TIMI score (Antman et al., 2000b). In particular, in the case of Framingham Score, a point-system was developed for ease of clinical use (Sullivan et al., 2004). In psychiatry, an individual level risk calculator has been built and disseminated online (Cannon et al., 2016). We will build a similar software front end for the underlying risk prediction models and disseminate the model for clinical use.

11 STUDY ASSESSMENTS

In this section, we will delineate in detail the performance and cross-validation methods and procedures we will use in this study for the models that would be developed in our study.

11.1 Within-Study Cross Validation Procedure

To ensure we get appropriate estimates of prediction error and that our models do not suffer from overfitting, we will utilize two methods of cross-validation, k-fold cross-validation or for methods which utilize bootstrapping (e.g. random forests) estimation of prediction error on out-of-bag samples. K-fold cross-validation randomly breaks the sample up into k-subsamples. Then (k-1)/k of the samples are used as the training data set and the remaining or left-out 1/k of the original sample is used to create a truly out-of-sample prediction error. The estimation of the predictive model is repeated k times, each time with a different 1/kth partition left out so that there will be K models; each person in the sample will have one model estimated in which their data was not included. The CVk estimate of the prediction error is the average across the entire sample of the individual's loss function evaluated using the model estimated when that person was not in the training data.

Random forests, for example, uses bootstrapping. In Random forest, a number of regression trees are estimated, each on a bootstrapped sample of the original sample. The average number of unique observations in a bootstrapped sample of size N is 63.2% of N. This means that for any one tree in the forest 36.8% of the sample will not be included in its calculation. If there are 1000 trees in the forest that would imply that each person would be left out of about 368 trees. The individual is considered to be out-of-bag for these trees. To get a truly out-of-sample prediction error, we calculate our prediction error only on predictions from trees that are out-of-bag for that individual. Similar methods will be used for neural networks and ensemble learning methods.

11.2 Evaluation of Generalization Performance

We will use our estimates of prediction error to guide model selection. Most of the methods we will be involve the choice of tuning parameters. To ensure that our methods do not over fit and provide the best predictions given the data, we will use our cross-validated estimate of prediction error (be it k-fold based or out-of-bag based) to choose the optimal tuning parameters to minimize true out-of-sample prediction error.

11.3 Sensitivity Analyses and Estimation of Prediction Error

We will estimate the predictive performance sensitivity to different clinic populations using Monte Carlo methods. In particular, we will exploit the fact that CTN trials are conducted on clinical sites that have different features. We will calculate the expected error and variance on the error across different clinic sites, as well as across different demographic covariates of interest, including sex, age, ethnicity. For each specific model type, sensitivity analyses of individual predictors will be subsumed under analysis of variable importance as described above.

11.4 Sex as a Biological Variable and Related Considerations

To comply with NIH requirements that CTN studies include plans for inclusion of women and addressing aspects of study generalizability to other special populations, we plan to include the following additional specific analyses as an extension to the general sensitivity analyses as outlined above. We propose to obtain predictive performance measures specifically for a subsample of male vs. female study participants with the best performing traditional and newer models and report the differences in model performance, including both difference in mean accuracy measures as well as variances in accuracies. We will also conduct that the same analyses for ethnic groups. We will then separately report an estimate of error in the predictive model for the highest vs. lowest proportion of sites with female participants as a measure of the influence on the sampling population on predictive performance. While we do not anticipate effects relating to biological sex or ethnic groups in our predictive models, specified analysis of this type might uncover any potential systematic bias in the model.

12 TRAINING REQUIREMENTS

There are no training requirements for this study. Study personnel have the requisite skills and training to perform the planned analysis. However, we do hope to develop a training program to disseminate the methods used and describe the data pipeline created as a part of CTN-0094 so that this work will initiate a CTN Data Science platform for future methodology development and analytics studies carried out within the CTN.

13 STATISTICAL DESIGN AND ANALYSES

This study is a secondary analysis study and therefore study design consists of a detailed statistical design and analysis plan. Each component analysis segment will have a detailed statistical analysis plan created prior to initiation to describe fully the procedures to be undertaken. Here we give the general outline of approach within each component. Each component analysis will be assigned to one of the research sites and the primary lead, but be replicated by the other site as described. Components will be organized by the families of machine learning approaches to be utilized, for example Random Forests. Throughout we will use appropriate cross-validation procedures, as described in section 11.

The first stage will be to generate predictions for each of the outcomes, drop-out (and time to drop-out) and abstinence, long-run (5 year follow-up outcome) and each of the interventions, BUP/NX, XR-NTX, and Methadone. Within BUP/NX we will examine the appropriateness of combining predictions across trials (and potentially across the different counseling interventions) by comparing both individual predictions from combined and separate models and comparing resulting prediction error. The final step of this stage will create a model selection procedure within each of these approaches when selection is not already a part of the method (some approaches, such as regularized regression, have integrated model selection), to assess the minimally necessary set of covariates for predication.

The next step in the process would be calculation of PITEs as described in 10.3.1 on each possible pair-wise comparison of treatments. PITES will then be assessed for variable importance using VIMP from random forests and the results of LASSO and gradient boosting on the estimated PITES.

The final step in the analyses process is to examine whether predictions and important variables differ by gender or by race/ethnicity. Across the trials, 34% of participants are female (n=838), 23% (n=569) are minorities, and 13.5% (336) are Hispanic.

Our final summarization within a component will include prediction error, important variables from the model selection exercise and description of the differences in disaggregated (study specific) and aggregated predictions (on the harmonized data). Note that important variables for prediction of outcome within a treatment and important variables for the PITES will not generally be the same set of variables.

14 REGULATORY COMPLIANCE, REPORTING and MONITORING

14.1 Statement of Compliance

This study will be conducted in accordance with the current version of the protocol, in full conformity with the ethical principles outlined in the Declaration of Helsinki, the Protection of Human Subjects described in the International Council for Harmonization Good Clinical Practice (GCP) Guidelines, applicable United States (US) Code of Federal Regulations (CFR), the NIDA Terms and Conditions of Award, and all other applicable state, local, and federal regulatory requirements. The Lead Investigator and Co-LI will assure that no deviation from, or changes to the protocol will take place without prior agreement from the Sponsor and documented approval from the Institutional Review Board (IRB).

14.2 Single Institutional Review Board

Prior to initiating the study, participating site investigators will obtain written approval from the Ethics Review Committee (ERC) or Institutional Review Board (IRB) to conduct the study at their respective site, which will include approval of the study protocol. If changes to the study protocol become necessary, protocol amendments will be submitted in writing by the investigators for IRB approval prior to implementation. Because this is a secondary analysis study, there are no consent forms, recruitment materials, or any materials given to participants. IRB continuing review will be performed annually, should the determination be made that an IRB review is necessary. Each site principal investigator is responsible for maintaining copies of all current IRB approval or exemption notices and approval for all protocol modifications. These materials must be received by the investigator prior to the initiation of research activities at the site and must be available at any time for audit. Unanticipated problems involving risk to study participants, for example any breach of confidentiality, will be promptly reported to and reviewed by the IRB of record, according to its usual procedures.

For this particular study, the New York State Psychiatric institute (NYSPI) IRB will be the single IRB of record for the protocol and will provide study oversight in accordance with 45 CFR 46. Participating institutions have agreed to rely on NYSPI-IRB and have entered into reliance/authorization agreements for Protocol CTN-0094. NYSPI-IRB will follow written procedures for reporting its findings and actions to appropriate officials at each participating institution.

14.3 Informed Consent

This is a study based on existing clinical trials data from studies that have already been completed. For this reason, we are requesting a waiver of informed consent from the IRB and will provide the approval once obtained. The IRB of record will evaluate the proposal on existing federal regulations (45 CFR 46.116(f)(3)) using the following criteria, all of which are satisfied by the proposal:

- The research involves no more than minimal risk to the subjects;
- The waiver or alteration will not adversely affect the rights and welfare of the subjects;
- The research could not practicably be carried out without the waiver or alteration, since it would be impossible to locate all of the participants of these existing trials; and

- Whenever appropriate, the subjects will be provided with additional pertinent information after participation.

The study does not preempt any applicable federal, state, or local laws which require additional information to be disclosed in order for informed consent to be legally effective. It is in conformance with 42 CFR 2.52, which allows for research-related provisions with regard to the disclosure of substance use disorder patient identifying information in the absence of the informed consent process and HIPAA authorization. In particular this study will use data that are already released in the public domain, and therefore will not include any identifiable information such as SSN, patient name, and other demographic information such address or phone number.

14.4 Quality Assurance Monitoring

In accordance with federal regulations, the study sponsor is responsible for ensuring proper monitoring of an investigation and ensuring that the investigation is conducted in accordance with the protocol. The study lead investigators will monitor conformity to the protocol by the associated personnel. This includes but is not limited to appropriate protection and maintenance of the data integrity, documentation and sharing of computer code, and sharing and dissemination of study results. Non-conformity with protocol and federal regulations will be reported as a protocol deviation and submitted to the study sponsor and study IRB for further review.

14.5 Participant and Data Confidentiality

Participant confidentiality and privacy are strictly held in trust by the participating investigators, their staff, and the sponsor(s) and funding agency, and will be maintained in accordance with all applicable federal regulations and/or state/Commonwealth law and regulations. By signing the protocol signature page, the investigator affirms that information furnished to the investigator by NIDA will be maintained in confidence and such information will be divulged to the IRB/Privacy Board, Ethical Review Committee, or similar expert committee; affiliated institution; and employees only under an appropriate understanding of confidentiality with such board or committee, affiliated institution and employees.

The original studies collecting the data to be analyzed had a Certificate of Confidentiality (CoC) which prevented them from being compelled to release data on the individual participants. Because CTN-0094 personnel will not have access to any identifying information, it would not be possible for them to provide data on any particular person if requested, therefore a CoC for CTN-0094 is not required.

CTN-0094 staff will protect the confidentiality of study data by only analyzing the data behind the firewalls of their respective institutions and storing the data on password-protected, HIPAA-compliant storage systems.

14.6 Investigator Assurances

Each site must have on file a Federalwide Assurance (FWA) with the HHS Office for Human Research Protection setting forth the commitment of the organization to establish appropriate policies and procedures for the protection of human research subjects in alignment with 45 CFR 46, Subpart A, with documentation sent to NIDA or its designee. Research covered by these regulations cannot proceed in any manner prior to NIDA receipt of certification that the research has been reviewed and approved by the IRB provided for in the assurance (45 CFR 46.103). Prior to initiating the study, the principal investigator at each study site will sign a protocol signature page, providing assurances that the study will be performed according to the standards stipulated therein.

14.6.1 Financial Disclosure/Conflict of Interest

All investigators will comply with the requirements of 42 CFR Part 50, Subpart F to ensure that the design, conduct, and reporting of the research will not be biased by any conflicting financial interest. Everyone with decision-making responsibilities regarding the protocol will confirm to the sponsor annually that they have met their institutional financial disclosure requirements.

14.7 Clinical Monitoring

Due to the nature of this study, clinical monitoring will be limited. Qualified node personnel (Node QA monitors) or other designated party(ies) will provide site management for each site during the trial. Node QA staff or other designated party(ies) will audit study files, including the site regulatory binder. This will take place as specified by the local protocol team, node PI or lead team and will occur as often as needed to help prevent, detect, and correct problems at the study sites. Node QA personnel will verify that study procedures are properly followed and that site personnel are trained and able to conduct the protocol appropriately. If the node personnel's review of study documentation indicates that additional training of site study personnel is needed, node QA personnel will undertake or arrange for that training.

14.8 Inclusion of Women and Minorities

The study data included both women and minorities. Across the trials, 34% of participants are female (n=838), 23% (n=569) are minorities, and 13.5% (336) are Hispanic.

14.9 Regulatory Files

The regulatory files should contain all required regulatory documents, study-specific documents, and all important communications. Regulatory files will be checked at each participating site for regulatory document compliance prior to study initiation, throughout the study, as well as at study closure.

14.10 Records Retention and Requirements

Research records (including deidentified data sets and study regulatory files) are to be maintained by the investigator in a secure location for a minimum of 3 years after the study is completed and closed. These records are also to be maintained in compliance with IRB, state and federal requirements, whichever is longest. The sponsor and Lead Investigator must be notified in writing and acknowledgment must be received by the site prior to the destruction or relocation of research records.

14.11 Reporting to Sponsor

The site principal investigator agrees to submit accurate, complete, legible and timely reports to the Sponsor, as required. These include, but are not limited to, reports of any changes that significantly affect the conduct or outcome of the trial or increase risk to study participants. Safety reporting will occur as previously described. At the completion of the trial, the Lead Investigator will provide a final report to the Sponsor.

14.12 Audits

The Sponsor has an obligation to ensure that this trial is conducted according to good clinical research practice guidelines and may perform quality assurance audits for protocol compliance.

The Lead Investigator and authorized staff from the Greater New York Node or Florida Node Alliance; the National Institute on Drug Abuse Clinical Trials Network (NIDA CTN, the study sponsor); NIDA's contracted agents, monitors or auditors; and other agencies such as the Department of Health and Human Services (HHS), the Office for Human Research Protection (OHRP) and the Institutional Review Board of record may inspect research records for verification of data, compliance with federal guidelines on human participant research, and to assess participant safety (as applicable).

14.13 Study Documentation

Each participating site will maintain appropriate study documentation for this project, in compliance with ICH E6 R2 and regulatory and institutional requirements for the protection of confidentiality of participants. Study documentation includes research data sets, monitoring logs, sponsor-investigator correspondence, and signed protocol and amendments, and Ethics Review Committee or Institutional Review Board correspondence. As part of participating in a NIDA-sponsored study, each site will permit authorized representatives from NIDA and regulatory agencies to examine (and when permitted by law, to copy) study records for the purposes of quality assurance reviews, audits, and evaluation of the study progress and data validity.

Source documents include all recordings of observations or notations of clinical activities and all reports and records necessary for the evaluation and reconstruction of the research study. Whenever possible, the original recording of an observation should be retained as the source document; however, a photocopy is acceptable provided that it is a clear, legible, and exact duplication of the original document.

Because this project is a secondary analysis on existing data the only risk to study participants would be to confidentiality. Although our data will be de-identified, project staff will protect the data from re-identification by following the procedures described in section 15, and any potential breaches will be documented and reported.

14.14 Safety Monitoring

Study teams at both study sites will have no direct participant contact. In addition, each of the researchers and the associated scientists have worked with sensitive healthcare data in the past. Our team has an understanding and commitment to protecting patient information. All HIPAA guidelines are understood and will be followed. We are utilizing a publicly released de-identified data set. No patient is being contacted. Only members of the research team will be reviewing and manipulating the data. Manipulated databases are being kept on the secure drive within a secure building and shared, when appropriate, using a standard 128-bit encryption over the web. All source code and results will be similarly stored.

14.15 Data and Safety Monitoring Board (DSMB)

Given the nature of this secondary data analysis, a DSMB will not be convened for this study.

15 DATA MANAGEMENT

15.1 Design and Development

This protocol will utilize the publicly released data from a Data and Statistics Center (DSC) utilized in the original studies. The DSC will assist in understanding the data dictionary and structure but will not be directly involved in harmonization and predictive modeling.

15.2 Site Responsibilities

The sites will collaboratively maintain the resulting harmonized database as well as the code base for the duration of the project.

15.3 Privacy and Confidentiality

The data safety monitoring plan will be enforced to ensure that participant privacy is protected:

- All staff will be required to complete formal training regarding patient privacy and confidentiality of clinical information. Provisions for disciplinary action (including termination of employment) in case of unauthorized disclosure are included as policy and are covered in the formal training process.
- All HIPAA guidelines will be fully understood (through mandatory training) and followed. The following precautions will be implemented to maintain the confidentiality of identifiable information:
    - All data are backed up regularly and backups are stored on a server in a separate location from the original server.
    - All computers (workstations and servers) and storage devices are secured in locked offices.

15.4 Data Sharing

The data being used for this secondary analysis are currently available on the NIDA datashare. We will share code, data pipelines, and procedures that we create as a result of this research. This should enable researchers to replicate and extend our findings.

Pursuant to the HEAL Inititive Public Access and Data Sharing Policy, the following steps will be implemented:

1. Any preprint or in press publications, including technical documents, will be deposited in a public repository such as PubMed Central and indexed accordingly.

2. The resultant harmonized de-identified databases would be hosted by NIDA-CTN publicly as detailed in the Dissemination Plan.

3. Annotated code and libraries for porting new CTN clinical trial data as well as for predictive modeling will be uploaded to NIDA-CTN and eventually managed by the CTN Data Science Task Force/Committee.

4. Public facing release of results (i.e. web-based packages for bed-side predictions) will be disseminated through the NIDA-CTN as appropriately evaluated by the CTN Steering Committee, once the models are completed and validated.

15.5 Data Safety Management Plan

General Data Security Precautions: Identifiable data used in this study will be stored on limited-access devices located in a locked server room on the grounds of the New York State Psychiatric Institute or University of Miami office. There will be no electronic transfer of protected data via the Internet or e-mail; Remote access to the data is limited to secure VPN or VPN and Citrix.

15.6 Web Access and Sensitive Patient Data

The data, applications and virtual computing sessions that will be used by the study team all reside behind the firewall at each of the study node. Researchers access virtual computing sessions from their local workstation using Microsoft at no time in the data analysis process is data downloaded to local servers or workstations. Sensitive patient data may appear on the local screen, but it is not stored on the local workstation.

16 PUBLICATIONS AND OTHER RIGHTS

Per NIH policy, the results of the proposed trial are to be made available to the research community and to the public at large. The planning, preparation, and submission of publications will follow the policies of the Publications Committee of the CTN.

Software and algorithm designed and developed for this study will be released as a public resource to be used free of charge, under a GNU General Public License (GPL). The resulting risk prediction models as disseminated will be maintained initially by the project lead nodes, and eventually transitioned to a CTN managed central resource. Code and data will be freely distributed and maintained according to current CTN data sharing standards.

17 PROTOCOL SIGNATURE PAGE

SPONSOR'S REPRESENTATIVE (CCTN SCIENTIFIC OFFICER OR DESIGNEE)

Printed Name	**Signature**	**Date**

ACKNOWLEDGEMENT BY INVESTIGATOR:

- I am in receipt of version X of the protocol and agree to conduct this clinical study in accordance with the design and provisions specified therein.
- I agree to follow the protocol as written except in cases where necessary to protect the safety, rights, or welfare of a participant, an alteration is required, and the sponsor and IRB have been notified prior to the action.
- I will ensure that the requirements relating to obtaining informed consent and institutional review board (IRB) review and approval in 45 CFR 46 are met.
- I agree to personally conduct or supervise this investigation at this site and to ensure that all site staff assisting in the conduct of this study are adequately and appropriately trained to implement this version of the protocol and that they are qualified to meet the responsibilities to which they have been assigned.
- I agree to comply with all the applicable federal, state, and local regulations regarding the obligations of clinical investigators as required by the Department of Health and Human Services (HHS), the state, and the IRB.

SITE'S PRINCIPAL INVESTIGATOR

Printed Name	**Signature**	**Date**

Clinical Site Name ______________________________

Node Affiliation ______________________________

18 REFERENCES

Antman, E.M., Cohen, M., Bernink, P.J., McCabe, C.H., Horacek, T., Papuchis, G., Mautner, B., Corbalan, R., Radley, D., and Braunwald, E. (2000a). The TIMI risk score for unstable angina/non–ST elevation MI: a method for prognostication and therapeutic decision making. Jama *284*, 835–842.

Antman, E.M., Cohen, M., Bernink, P.J., McCabe, C.H., Horacek, T., Papuchis, G., Mautner, B., Corbalan, R., Radley, D., and Braunwald, E. (2000b). The TIMI risk score for unstable angina/non–ST elevation MI: a method for prognostication and therapeutic decision making. Jama *284*, 835–842.

Austin, P.C., Lee, D.S., D'Agostino, R.B., and Fine, J.P. (2016). Developing points-based risk-scoring systems in the presence of competing risks. Stat. Med. *35*, 4056–4072.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Bridle, J.S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Neurocomputing, (Springer), pp. 227–236.

Bzdok, D., and Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. Biol. Psychiatry Cogn. Neurosci. Neuroimaging *3*, 223–230.

Cannon, T.D., Yu, C., Addington, J., Bearden, C.E., Cadenhead, K.S., Cornblatt, B.A., Heinssen, R., Jeffries, C.D., Mathalon, D.H., and McGlashan, T.H. (2016). An individualized risk calculator for research in prodromal psychosis. Am. J. Psychiatry *173*, 980–988.

Collins, L.M., Nahum-Shani, I., and Almirall, D. (2014). Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assignment, randomized trial (SMART). Clin. Trials *11*, 426–434.

Crist, R.C., Clarke, T.-K., Ang, A., Ambrose-Lanci, L.M., Lohoff, F.W., Saxon, A.J., Ling, W., Hillhouse, M.P., Bruce, R.D., Woody, G., et al. (2013). An intronic variant in OPRD1 predicts treatment outcome for opioid dependence in African-Americans. Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol. *38*, 2003–2010.

Crist, R.C., Reiner, B.C., and Berrettini, W.H. (2018a). A review of opioid addiction genetics. Curr. Opin. Psychol.

Crist, R.C., Doyle, G.A., Nelson, E.C., Degenhardt, L., Martin, N.G., Montgomery, G.W., Saxon, A.J., Ling, W., and Berrettini, W.H. (2018b). A polymorphism in the OPRM1 3′-untranslated region is associated with methadone efficacy in treating opioid dependence. Pharmacogenomics J. *18*, 173.

Dy, J.G., and Brodley, C.E. (2004). Feature selection for unsupervised learning. J. Mach. Learn. Res. *5*, 845–889.

Franks, P., Lubetkin, E.I., Gold, M.R., Tancredi, D.J., and Jia, H. (2004). Mapping the SF-12 to the EuroQol EQ-5D Index in a national US sample. Med. Decis. Making *24*, 247–254.

Goldsmith, J., Bobb, J., Crainiceanu, C.M., Caffo, B., and Reich, D. (2011). Penalized functional regression. J. Comput. Graph. Stat. *20*, 830–851.

Henriksson, K.M., Farahmand, B., Johansson, S., \AAsberg, S., Terént, A., and Edvardsson, N. (2010). Survival after stroke—the impact of CHADS2 score and atrial fibrillation. Int. J. Cardiol. *141*, 18–23.

Hser, Y.-I., Evans, E., Huang, D., Weiss, R., Saxon, A., Carroll, K.M., Woody, G., Liu, D., Wakim, P., and Matthews, A.G. (2016). Long-term outcomes after randomization to buprenorphine/naloxone versus methadone in a multi-site trial. Addiction *111*, 695–705.

Hser, Y.-I., Huang, D., Saxon, A.J., Woody, G., Moskowitz, A.L., Matthews, A.G., and Ling, W. (2017). Distinctive Trajectories of Opioid Use Over an Extended Follow-up of Patients in a Multisite Trial on Buprenorphine+Naloxone and Methadone. J. Addict. Med. *11*, 63–69.

Hughes, J.R., Keely, J.P., Niaura, R.S., Ossip-Klein, D.J., Richmond, R.L., and Swan, G.E. (2003). Measures of abstinence in clinical trials: issues and recommendations. Nicotine Tob. Res. *5*, 13–25.

Iasonos, A., Schrag, D., Raj, G.V., and Panageas, K.S. (2008). How to build and interpret a nomogram for cancer prognosis. J. Clin. Oncol. *26*, 1364–1370.

Lamont, A., Lyons, M.D., Jaki, T., Stuart, E., Feaster, D.J., Tharmaratnam, K., Oberski, D., Ishwaran, H., Wilson, D.K., and Van Horn, M.L. (2018). Identification of predicted individual treatment effects in randomized clinical trials. Stat. Methods Med. Res. *27*, 142–157.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

Lee, J.D., Nunes Jr, E.V., Novo, P., Bachrach, K., Bailey, G.L., Bhatt, S., Farkas, S., Fishman, M., Gauthier, P., and Hodgkins, C.C. (2018). Comparative effectiveness of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X: BOT): a multicentre, open-label, randomised controlled trial. The Lancet *391*, 309–318.

Lloyd-Jones, D.M., Wilson, P.W., Larson, M.G., Beiser, A., Leip, E.P., D'Agostino, R.B., and Levy, D. (2004). Framingham risk score and prediction of lifetime risk for coronary heart disease. Am. J. Cardiol. *94*, 20–24.

Lu, M., Sadiq, S., Feaster, D.J., and Ishwaran, H. (2018). Estimating individual treatment effect in observational data using random forest methods. J. Comput. Graph. Stat. *27*, 209–219.

Matusow, H., Dickman, S.L., Rich, J.D., Fong, C., Dumont, D.M., Hardin, C., Marlowe, D., and Rosenblum, A. (2013). Medication assisted treatment in US drug courts: Results from a nationwide survey of availability, barriers and attitudes. J. Subst. Abuse Treat. *44*, 473–480.

Pocock, S.J., Assmann, S.E., Enos, L.E., and Kasten, L.E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. Stat. Med. *21*, 2917–2930.

Rothwell, P.M., Mehta, Z., Howard, S.C., Gutnikov, S.A., and Warlow, C.P. (2005). From subgroups to individuals: general principles and the example of carotid endarterectomy. The Lancet *365*, 256–265.

Saxon, A.J., Ling, W., Hillhouse, M., Thomas, C., Hasson, A., Ang, A., Doraimani, G., Tasissa, G., Lokhnygina, Y., Leimberger, J., et al. (2013). Buprenorphine/Naloxone and methadone effects on laboratory indices of liver health: a randomized trial. Drug Alcohol Depend. *128*, 71–76.

Sullivan, L.M., Massaro, J.M., and D'Agostino Sr, R.B. (2004). Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat. Med. *23*, 1631–1660.

Weiss, R.D., Potter, J.S., Fiellin, D.A., Byrne, M., Connery, H.S., Dickinson, W., Gardin, J., Griffin, M.L., Gourevitch, M.N., and Haller, D.L. (2011). Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence: a 2-phase randomized controlled trial. Arch. Gen. Psychiatry *68*, 1238–1246.

Weiss, R.D., Potter, J.S., Griffin, M.L., Provost, S.E., Fitzmaurice, G.M., McDermott, K.A., Srisarajivakul, E.N., Dodd, D.R., Dreifuss, J.A., and McHugh, R.K. (2015). Long-term outcomes from the national drug abuse treatment clinical trials network prescription opioid addiction treatment study. Drug Alcohol Depend. *150*, 112–119.

Zhang, Z., Zhang, H., and Khanal, M.K. (2017). Development of scoring system for risk stratification in clinical medicine: a step-by-step tutorial. Ann. Transl. Med. *5*, 436.