Overview of the CTNote Library

Gabriel Odom, Laura Brandt, Ray Balise, and the CTN-0094 Team

2022-08-22

# Introduction and Motivation

## Background

Over 750,000 Americans have died from a drug overdose [since 1990](https://www.hhs.gov/opioids/about-the-epidemic/opioid-crisis-statistics/index.html). [Since 2018](https://www.pewtrusts.org/en/research-and-analysis/fact-sheets/2020/12/medications-for-opioid-use-disorder-improve-patient-outcomes), 2 in 3 drug overdose deaths have been from an opioid (almost 50,000 deaths). Nearly 1.3 million Americans are currently using one or more of the following medications to treat opioid use disorder (OUD): methadone, buprenorphine, and naltrexone. The [National Institute on Drug Abuse](https://nida.nih.gov/)’s [Clinical Trials Network](https://nida.nih.gov/about-nida/organization/cctn/clinical-trials-network-ctn) have funded over 100 clinical trials and clinical trial supplemental studies to study the effects of new and existing medications to treat addiction. Furthermore, opioid use disorder has in the past been predominately a disease of [middle-class white Americans](http://www.doi.org/10.1007/s11013-016-9496-5). However, these demographics have been shifting, and with this shift come exacerbated [health disparities](https://www.addictioncenter.com/news/2019/10/racial-disparities-opioid-addiction-treatment/).

Clinical trials assess the effectiveness of medication-based treatments for OUD (MOUD). A wide variety of treatment outcomes have been used in clinical trials since 1972 (Table 1 shows outcomes from 41 papers assessing MOUD). These endpoints all assess participant success or failure in treatment via the results from scheduled *urine drug screenings* (UDS) or *urine opioid screenings* (UOS). Accurately measuring the efficacy of MOUD for clinical trial participants is paramount, but no “gold standard” trial outcomes exist.

Moreover, trial outcomes commonly used in CTN trials have not been widely assessed for racial / ethnic measurement variance, but some such effects have already been described. For example, racial/ethnic bias was assessed for [OUD trial completion rates](https://www.doi.org/10.1016/j.drugalcdep.2018.06.006), and this treatment outcome was found to change disproportionately based on participant race/ethnicity in large US cities. In clinical trials for OUD, if a treatment outcome has latent racial/ethnic bias (after controlling for relevant clinical factors), then one of two scenarios will occur: (1) if minority subjects are disproportionately marked as treatment successes in error, then the observed quality of an otherwise ineffective therapeutic will be inflated for that subgroup; conversely, (2) if minority subjects are disproportionately marked as treatment failures in error, then the observed quality of an otherwise effective experimental therapeutic will be diminished. Treatment outcomes which depend on race/ethnicity will lead to inequitable over/under-treatment of minority individuals who suffer from OUD.

[1] “TABLE OF TREATMENT ENDPOINTS HERE; SEE .html VERSION”

Each row of Table 1 is an outcome used in a past clinical trial for MOUD. As preliminary work, our clinical team grouped these by clinical practice into three classes: *abstinence*, *relapse*, and substance use *reduction*. These classes were further split into subclasses based on their clinical interpretation. Of note, these metrics also differed in how they considered a scheduled but incomplete UDS (i.e. a “missing” UDS). We removed outcomes which could not be used to measure subject-specific outcomes (e.g. “proportion of subjects in treatment group with positive UDS in study week 3”); we removed outcomes which could not yield a univariate metric of subject success (e.g. “number of positive UDS per three-week window over time”); we split papers which assessed multiple outcomes (e.g. complete abstinence between weeks 5-12 AND total number of abstinent weeks), and we combined papers which used identical outcomes. At the end of our literature search, we have 53 outcomes from registered clinical trials to consider.

## Objectives

Clinically meaningful, patient-centric endpoints beyond abstinence are needed to define success in clinical trials. - Dr. Nora Vulkow, Address to CTN, April 2022

This is the first large-scale, empirical (data driven) comparison of MOUD clinical trial endpoints, and to our knowledge, this is the first racial/ethnic sensitivity analysis of multiple clinical trial outcome definitions. Furthermore, this is the only standard and code-based library of treatment outcome definitions to date, and we believe it will be useful to all future substance use disorder clinical trials in NIDA’s [Clinical Trials Network](https://nida.nih.gov/about-nida/organization/cctn/clinical-trials-network-ctn) (CTN) and in other substance use disorder research.

Our scientific questions are as follows: (1) When applied to the same clinical trials data, do these outcomes follow the same clinically intuitive clusters (abstinence, relapse, use reduction)? (2) When applied to the same clinical trials data, *and* controlling for clinically relevant confounders, are some of these outcomes more sensitive to race/ethnicity than others?

# Methods

## The Clinical Trials Data

The [CTN-0094 Project](http://ctndisseminationlibrary.org/protocols/ctn0094.htm) research team harmonized data from 3 large-scale MOUD clinical trials: [CTN-0027](http://ctndisseminationlibrary.org/protocols/ctn0027.htm): “Starting reatment with Agonist Replacement Therapies (START)”; [CTN-0030](http://ctndisseminationlibrary.org/protocols/ctn0030.htm): “Prescription Opiate Abuse Treatment Study (POATS)”; and [CTN-0051](http://ctndisseminationlibrary.org/protocols/ctn0051.htm): “Extended-Release Naltrexone vs. Buprenorphine for Opioid Treatment (X:BOT)”. These three are nationally representative, prospective clinical trials with over 3600 combined participants. These trials treated participants for 16-24 weeks, and collected rich baseline substance use data, demographics, and medical/psychiatric evaluations.

We have 3560 fully de-identified subjects with demographics and weekly UDS results for the months they were in treatment, 2492 of whom completed randomization to a treatment arm. This harmonized data will be released in the R packages ctn0094data and [ctn0094DataExtra](https://github.com/CTN-0094/ctn0094DataExtra). To simplify our analysis of endpoint sensitivity to race/ethnicity, we coded race and ethnicity conjointly as “Non-Hispanic White” (), “Non-Hispanic Black” (), “Hispanic” (), and “Other” (). In order to summarize complex substance use patterns over the weeks of the multiple clinical trials, we standardized all dates to the day of trial consent, and we coded the urinalysis results for each day/week as a “word”.

## Substance Use Pattern “Word”

As described in detail in [Odom et al, (2022)](ctn0094DataExtra%20Vignette%20/%20Publication%20here), we create a [sufficient statistic](https://doi.org/10.1098/rsta.1922.0009) of all weekly UDS results. For a substance of interest (opioids in our case), this statistic is a compact representation of the full pattern of substance use for an individual participant. To be of use, this summary statistic must have the following properties:

1. It can be directly parsed by a computer
2. It can be directly interpreted by a human
3. It represents all of the same information about participant use of a substance or group of substances that would be present in a medical summary of the participant

### Pattern Definition

In order to compactly summarize a participant’s pattern of use for a particular substance over time, we first define the following legend:

* **+**: positive for the substance(s) in a specified window of time (a day, week, month, etc.) by urine screen (or self report, if such data are of interest)
* **–**: negative for the substance(s)
* **o**: subject failed to provide a urine sample
* \_: no specimens required (weekends, holidays, pre-randomization period, alternating visit days/weeks)
* \*: inconclusive results or mixed results (e.g. subject provided more than one urine sample in the time interval, and they did not agree)

### Use Pattern Examples

We will first observe the recorded opioid use data via urine screen for subject 2089 after randomization. [1] “TABLE OF SUBSTANCE USE FOR EXAMPLE PARTICIPANT HERE; SEE .html VERSION”

While we observe that this participant used various substances throughout the clinical trial, it’s difficult to make clinical judgement using the data in this form. On the one hand, this data can be clearly parsed by a computer, and it represents all of the opioid use for this subject. However, data in this form is not easy to directly interpret by a clinician.

Now observe the opioid use pattern summary for this participant:

## [1] "++++---+--------------o-"

Using the basic pattern definition above, we can clearly see that this subject had a challenging first month (++++), improved in the second month (---+), and then remained abstinent from opioids for the remainder of the clinical trial (--------------o-). Participant substance use pattern data in this form *is* easy to directly interpret by a clinician. Moreover, this symbolic representation of substance use can be parsed and summarized by a computer. Finally, notice that this pattern represents the entire course of treatment for subject 2089 without loss of any clinically relevant information about their opioid addiction/dependence.

### Use Pattern Limitations

While this substance use pattern summary satisfies the three conditions necessary to be a “sufficient” statistic mentioned previously, there are three main limitations:

The first limitation surrounds the idea of poly-drug use. These use pattern summaries are substance or substance group specific. Notice that, while this participant does appear to curb their weekly opioid and heroin use during the course of treatment, their cocaine use does *not* decrease over the same time interval. However, the **opioid** use pattern summary can not display information about concurrent **cocaine** use. While this is a current limitation, one of our current areas of research involves expanding the support of substance use pattern “words” to include poly-substance use in a meaningful way (including thee ability to preserve cross-substance correlations).

The second limitation is one of parsimony. In order to ensure that use pattern summaries work regardless of a computer’s locale or operating system, we limit the symbols in the “word” to proper symbols from the [American Standard Code for Information Interchange](https://www.ascii-code.com/) (ASCII) list. Technically speaking, there are 128 [7-bit ASCII](https://www.ibm.com/docs/en/aix/7.1?topic=support-ascii-characters) symbols, but only 96 are visible (printable) characters. Put simply, these 96 characters are all the symbols that a traditional North American computer keyboard can make. Therefore, we could define a substance use pattern “word” with any combination of these 96 printable symbols. However, such a visual summary would be incredibly challenging to interpret, consequently failing to meet the requirement that our summary be easily read by a clinician. In the interest of simplicity, we chose the five symbols mentioned above (+, -, o, \*, and \_). We recognize that different clinical trial designs may necessitate the introduction of additional symbols to this use pattern summary, but we strongly recommend (due to human [memory constraints](https://doi.org/10.1037/h0043158)) that such lists be kept to seven symbols or fewer. As an example, we discussed the benefit of adding a special symbol for a “missing but excused” clinic visit, but we ultimately did not because such instances were rare in our data.

The third limitation is due to the cold and unfeeling logic of a computer. The computer cannot understand the many extenuating circumstances that pervade our human experience. Worth repeating is the old maxim of computing: “computers give you what you ask for, not what you want.” Consider a concrete example: assume a participant was supposed to visit the clinic for a weekly urine test on Thursday, but they came on Wednesday instead due to a conflict at their work; their urine sample from Wednesday came back “clean”. What would a nurse at the substance use clinic do in this case? Most people are reasonable; they would count the urine screen for that week as negative for opioids. What would a computer do? Without the direct input of a human to override the clinical trial protocol, a computer would count the negative urine on Wednesday but still mark the participant as a “failure to appear” / “missing urine screen” on Thursday. In many clinical trial protocols, missing urine screens are imputed as “positive” for the substance of interest, so now this participant’s use pattern summary is marked \* (mixed results) instead of - (negative results) for that week.

## Computing Treatment Outcomes with CTNote

We mentioned above that these use pattern summaries are “computable” (that a computer can parse them). The software package CTNote:: for the R computing language is one such tool to enable computers to parse these use patterns. This software package contains the following groups of routines (also known as functions):

* Handle missing UDS data with recode\_missing\_visits() and impute\_missing\_visits()
* Account for the study observation design or trial visit protocol with collapse\_lattice() and view\_by\_lattice()
* Detect a substance use pattern with detect\_subpattern() and detect\_in\_window()
* Measure the longest periods of consecutive behavior with measure\_retention() and measure\_abstinence\_period()
* Count the substance use events with count\_matches()

By executing various combinations of these routines, we were able write algorithms to define each of the treatment outcome definitions in Table 1. In our supplemental material, we include a “library” of algorithms to calculate each definition listed in Table 1. The input of each algorithm is a set of substance use pattern summaries for all clinical trial participants; the output of each algorithm is a calculated treatment endpoint for each participant included in the input.

### Example Abstinence Outcome

The Schottenfeld et al. (2008) treatment endpoint definition is “the length of the longest period of consecutive opioid abstinence”, where missing clinic visits are imputed to represent a urine screen positive for the substance of interest. Therefore, the algorithm to calculate this definition for subject 2089 would be

# The participant's use pattern summary; the %>% symbol is read "and then"  
"++++---+--------------o-" %>%   
 # change all missing visits to positive  
 recode\_missing\_visits(missing\_becomes = "+") %>%   
 # find the length of the longest abstinent period, in weeks  
 measure\_abstinence\_period()

### Example Use Reduction Outcome

The Haight et al. (2019) treatment endpoint definition is the “percentage of negative UOS from week 5 to week 24”; because participants included in the CTN-0094 harmonized data set were not all followed for 24 weeks, we truncate this observation window to 15 weeks of study. Therefore, the code to calculate this definition for subject 2089 would be

count\_matches(  
 use\_pattern = "++++---+--------------o-",  
 match\_is = "-",  
 # Mixed results weeks count as half of a negative week  
 mixed\_results\_are = "\*", mixed\_weight = 0.5,  
 # The end-of-protocol for our trials is 15 weeks  
 start = 5, end = 15,  
 # Return the proportion instead of the count  
 proportion = TRUE  
)

### Example Relapse Outcome

The Krupitsky et al. (2006) treatment endpoint has relapse defined as “3 consecutive positive UOS”, where missing clinic visits are imputed to represent a urine screen positive for the substance of interest. Therefore, the algorithm to calculate this definition for subject 2089 would be

"++++---+--------------o-" %>%   
 recode\_missing\_visits(missing\_becomes = "+") %>%   
 # change all mixed-results visits to positive  
 recode\_missing\_visits(missing\_is = "\*", missing\_becomes = "+") %>%   
 # detect if the subject had 3 weeks of positive or missing UDS in a row  
 detect\_subpattern(subpattern = "+++")

If we were interested in relapse from a “time-to-event” or reliability perspective, the algorithm above changes only in the last line:

"++++---+--------------o-" %>%   
 recode\_missing\_visits(missing\_becomes = "+") %>%   
 recode\_missing\_visits(missing\_is = "\*", missing\_becomes = "+") %>%   
 # detect if the subject had 3 weeks of positive or missing UDS in a row, AND  
 # measure the number of weeks until this pattern \*starts\*  
 detect\_in\_window(  
 # Look at all possible 3-week windows, and  
 window\_width = 3L,  
 # check if any of them have all 3 positive UDS  
 threshold = 3L  
 )

# Results

Now that we have an opioid-specific use pattern for each participant in these three harmonized clinical trials, and we also have a suite of computer functions to calculate each treatment endpoint, we can answer further questions about measuring treatment success or failure. Note that by treatment failure, we mean that the treatment failed to help the participant, not that the participant themselves is a failure in treatment. There are 53 composite outcomes, but a few of these outcomes involve more than one component. For example, a “time to relapse” metric would involve both the length of time from induction to relapse and an indicator if a relapse was observed. While this type of data is structurally similar to survival data, we note that ([unlike standard survival analysis](https://doi.org/10.4103/0970-0218.66859)) the censoring indicator will rarely be independent of the time-to-event measure.

## Clustering Treatment Outcomes

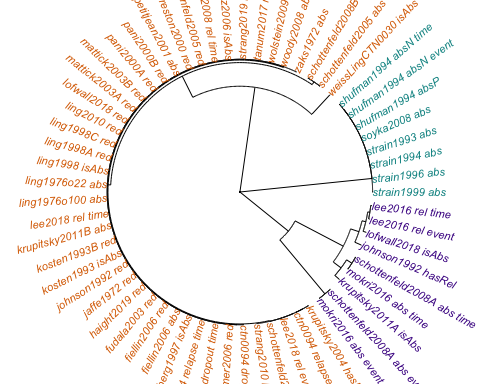
### Hierarchical Clustering Process Details

Given this library of MOUD treatment endpoints (53 composite, including both “time” and “event” metrics; 61 distinct), we applied each endpoint to all 2492 participants who had been randomized to a treatment. This process enables us to empirically contrast endpoints via hierarchical clustering. Our process is as follows:

1. We calculated all 61 distinct endpoints for each subject, resulting in a table of outcome metric values.
2. We scaled all “time” and “count” endpoints to be ratios of their maximum value. For example, if an endpoint measures the total number of negative UDS, then we divide this count for each participant by the maximum observed count across all participants.
3. We re-framed all outcomes into a “larger is better” interpretation: for outcomes where a larger value would indicate a worse treatment outcome.
4. To reduce the influence of a single participant’s use pattern on the clustering results, we calculate the Pearson correlation matrix (using pairwise complete observations) on a bootstrap sample of the rows in Step 1.
5. We perform hierarchical clustering on the bootstrap correlation matrix from Step 3 using complete linkage, and cut the resulting tree into clusters.
6. For each number of clusters , we generate a network connection matrix. The values of this matrix are 1 if treatment endpoints and are in the same cluster, 0 otherwise.
7. We repeat this process 10,000 times for each , and sum these matrices. In the end, a value of means that in every random bootstrap sample, the treatment endpoints and are in the same cluster. Similarly, a value of means that the treatment endpoints and are never in the same cluster for any random bootstrap sample.
8. We also performed *k*-means clustering for each number of clusters. We found that the optimal number of clusters was (by observing an “elbow” in the total within-cluster Sum of Squares plot).

### Hierarchical Clustering Results

The bootstrap clustering results for are shown in the accompanying figure.



The teal cluster on the top right of the wheel, including Shufman et al. and Strain et al. endpoints all have the following trait in common: they ignore or remove missing UDS from their calculation. Notice that this teal cluster has 0 subsplits; in every single bootstrap sample, these metrics are all clustered together. We believe that these endpoints are far too lenient, because missing UDS in opoiod trials is commonly a negative prognostic indicator [CITE THE LEADING CTN-0094 PAPER HERE]. In contrast, the purple cluster on the bottom right of the wheel has the most variability and highest count of sub-clusters. These clusters all depend heavily on the UDS results from a single short interval. For example, the Lofwall et a. (2018) endpoint requires a negative UDS in week 12 (with missing UDS counted as positive), otherwise the treatment will be counted as a failure. This means that if a participant was present in the clinic with negative UDS for 11 weeks in a row, but missed the visit on week 12, the treatment would have been counted as a failure. Similarly, the Krupitsky et al. (2011) endpoint requires complete abstinence from study week 5 to study week 24, and missing UDS are imputed to be positive. The endpoints in this cluster are far too harsh to be meaningful in practice, which is why the clustering results from these endpoint have so much variability (the most number of subclusters).

The larger orange cluster on the left side of the wheel is what we have termed the “Goldilocks” cluster: these outcomes are not too lenient by ignoring missing UDS, and they are not too harsh by measuring success or failure of an entire course of treatment based on a single week or pair of weeks. Of note, our *a priori* hypothesis that these endpoints would naturally cluster into “abstinence, relapse, and use reduction” did **not** hold true. Instead, our “Goldilocks” cluster contains abstinence, relapse, and use reduction endpoints. We believe that current and future MOUD clinical trials should use outcomes from this “Goldilocks” cluster, or–if researchers have the strong desire to create a set of bespoke outcomes–new outcomes should adhere to these considerations: do not ignore missing UDS, and do not make treatment failure too sensitive to a single week’s UDS.

## Assessing Racial/Ethnic Sensitivity in Outcomes

Another question that has arisen related to MOUD treatment endpoints is one of “fairness”. Now a true discussion of fairness is far beyond the scope of our work, so we sought to answer a simpler question: when applied to the same set of participants, and controlling for clinically relevant covariates, are some of these outcomes more sensitive to race/ethnicity than others? This sensitivity alone does not necessarily mean that an endpoint is “unfair”: it could be that a “sensitive” endpoint is better at detecting true differences in treatment for certain racial/ethnic subgroups. Such undue sesnsitivity to race/ethnicity could also mean that the endpoint in question is highlighting an “unfair” aspect of the design of the clinical trial itself. We leave thorough discussion such complex–but highly important–questions to future manuscripts.

### Mathematical Challenges

One challenge is how to measure “sensitivity”. Consider two statistical models using the same predictors:

$$ \textbf{y} = a\_0 + a\_1 \textbf{x}\_1 + \ldots + a\_p \textbf{x}\_p + \textbf{e} \qquad (1) \\ \textbf{z} = b\_0 + b\_1 \textbf{x}\_1 + \ldots + b\_p \textbf{x}\_p + \boldsymbol{\epsilon} \qquad (2) $$

Let measure a composite of race and ethnicity, are the other relevant predictors and covariates, represent outcome 1, and represent outcome 2. With this setup, we can find a test statistic and -value to measure the statistical significance of race/ethnicity on outcomes 1 and 2 **independently**. However, we can only compare models (1) and (2) if the distributions of and come from the same family (both Normal, both Binomial, etc.). If the two outcomes share the same metric space, we could use this framework to state if outcome 1 is more sensitive to race/ethnicity than outcome 2 (or vice versa).

However, our outcomes do **not** all share the same metric space. The error vectors could have univariate Bernoulli (success or failure), Poisson (number of negative UDS), Negative Binomial (number of positive UDS), Beta (ratio outcomes scaled between 0 and 1), or even multivariate (time to event and event indicator). Therefore, in order to ensure that the distribution of the model residuals shares the same family (so we can compare -values / regression coefficients), we **invert** the linear models above:

$$ \textbf{x}\_1 = a\_0 + \textbf{a}\_1\textbf{Y} + \ldots + a\_p \textbf{x}\_p + \textbf{e} \qquad (1^\*) \\ \textbf{x}\_1 = b\_0 + \textbf{b}\_1\textbf{Z} + \ldots + b\_p \textbf{x}\_p + \boldsymbol{\epsilon} \qquad (2^\*) $$

Notice the differences. We are no longer using race/ethnicity to predict the trial outcomes, but we are *using the trial outcomes to predict race/ethnicity*. The distributions of and are now both univariate *Multinomial* with composite race and ethnicity partitioned as Non-Hispanic White (NHW), Non-Hispanic Black (NHB), Hispanic, and Non-Hispanic Other (Other). Because time-to-event outcomes (survival; 2-dimensional) are included, and are now *matrices* of predictor information and and are *vectors* of regression coefficients.

### Assessing Racial/Ethnic Information in Outcomes

We fit equations and via feed-forward neural-network-based multinomial log-linear models (using the multinom() function from the [nnet package](https://CRAN.R-project.org/package=nnet) with defaults). Now that we can build comparable multinomial regression models (with NHW as the reference group), we compare them as follows:

1. Create a “null” model (no outcomes at all); fit without . This includes covariates for age, sex, study, treatment medication, and diagnoses of mental illness.
2. Compare model to the null model using a Likelihood Ratio Test and the Akiake Information Criterion.
3. Compare model to using the Likelihood Ratio Test.

Because we are fitting 53 regression models, we adjust the model -values from Step 3 using a false discovery rate (FDR) correction. The following table shows the results of these model comparisons against baseline (Non-Hispanic White) for each treatment endpoint, arranged by FDR.

[1] “TABLE OF MODEL PERFORMANCE BY RACE/ETHNICITY HERE; SEE .html VERSION”

Recall that all outcomes were transformed so to range from 0 to 1 and have larger values indicate better treatment prognosis. Notice that the “beta” column (the effect) has nearly uniform negative signs: these negative values for these endpoints are saying “if you succeeded in this treatment, you are less likely to be a minority”. [NOTE: these effects are disappear when we include the “site masked” variable. I think that site may be a proxy for race and ethnicity here. Help!]

# Discussion

We have a standard library of MOUD treatment outcomes. We have clustered these outcomes empirically, and we found an entire cluster only containing outcomes which ignored missing UDS–further giving credence to the theory that missing UDS are highly informative for opioid use treatment. We have also ascertained that participant race and ethnicity are informative predictors when measuring MOUD efficacy.