Project Part 1

Sara Altman & Hershel Mehta

October 25, 2017

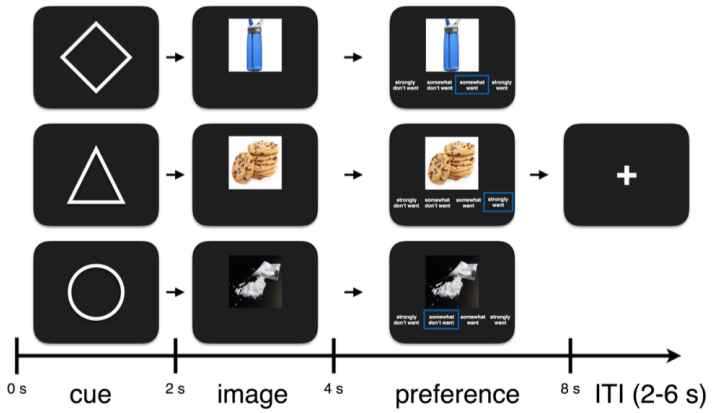
Dataset Description / Data Cleaning

Trial-level Brain Data

Stimulant-dependent veteran patients (n=39) were recruited from a 28-day in-patient rehabilitation program during weeks 2-4 of their program, and control subjects (n=39) were recruited from the community. So, in total, we have 78 subjects. For this portion of the project, we have chosen to focus solely on our drug population, since our interest is in predicting stimulant-dependent veterans who relapse from those who do not relapse.

Subjects were scanned using function magnetic resonance imaging (fMRI) while completing a cue reactivity task where they saw one of three types of images: drugs (reward stimulus), food (reward control), and neutral objects (stimulus control).

Each subject saw 18 images from each type (3 x 18 = 54 trials total), each of which consisted of a cue period (2s), image period (2s), choice period (2s), and outcome period (2s). See the figure below for a visualization of the task (top row = neutral; middle row = food; bottom row = drug):



We acquired whole-brain images from the during each 2s period (called "TR"). For this portion of the project, we focused on a part of the brain that deals with dopamine (called the Nucleus Accumbens, or for short, "nacc"). In the future, we hope to add in different regions of interest (called "roi"), which should give us even more data to work with.

Subject-level Behvaioral Data.

In addition to our trial-level brain data, we also collected subject-level behavioral data. These consist of useful demographic informaiton like "age", but also clinically relevant information, such as "years of use" or "lifetime depression". Also crucially, for our response variables, we followed-up with each patient for assessments of sobriety and relapse up to one year after scanning. These gave us a few interesting response variables:

- Continuous Response Variable: One possibility for a continuous response variable is the variable `obstime` which is a measure of how long the subject was sober in days after the program. We're a bit skeptical

about this metric since it's often difficult to reach these patients, so the metric is really a measure of the time at which we learned how many days they were sober.

- Binary Response Variable: For a binary response variable, one possibility is the variable `relapse`, which is 1 if the subject relapsed and 0 otherwise. This would be a useful response variable since we're interested in being able to predict if someone will relapse or not given certain characteristics, as well as inferring what types of characteristics are associated with relapse.We also might get data for healthy controls as well. In that case, we could also use whether or not someone is a drug user as a binary response variable.

Questions to Answer

We want to know what characteristics (e.g., time sober, mental illness history, other drug use history, education level, age) are important predictors of methamphetamine use relapse, and if we can accurately predict relapse using these characteristics. We also want to know if NAcc response during a fMRI task can predict relapse well.

This dataset is exciting because it can help us learn about what predicts relapse, as well as a potential way to predict if a new patient will relapse. This would be useful for people working with past and present methamphetamine users, as well as a addition to the scientific literature on addiction and relapse.

Data Prep / Exploration

- Test / Training Split: We split 20% of our subjects from our subject-level data. We then removed all of the trial-level data from those same subjects from our trial-level brain data. Note that we needed to split at the subject-level since our data
- Brain data preprocessing: Brain data can often be very noisy due to many factors: subjects can move in the scanner, there is scanner draft, slight differences for when each slice of data was collected. We corrected for some of these errors in brain data by preprocessing our data using the following transformation: slice time correction, motion correction, spatial smoothing (4 mm blurring kernel), and a highpass filter. We also normalized our signal to percent signal change. We used visual checks of motion to ensure our subjects were not consistent movers throughout their scan.
- Behavioral data preprocessing: We examined the spread of our potential response variables and excluded rows with "NA" for relapse, since we don't know if they relapsed or not. If a row does not have an "NA" value but has other "NA" values, we may still work with the remaining columns, since our dataset is limited. In addition, we checked for collinearity between our chosen response variables and other potential candidates (such as `days2sober`, `relapse6mo`), and eliminated covariates that were redundant and had many "NA"s.
- Correlation plots:

Since this dataset was challenging to work with, we spent much of our time in the process of acquiring, wrangling, and preprocessing the data. We are excited to have a dataset clean and ready for analyses!