**MRT Data Preparation Protocol**

The purpose of this protocol is to get the data ready for the primary MRT analyses and other analyses by our lab

1. **Create new repository and folder structure for all code on Github using Sabina’s template.**
2. **Discuss with Jess about confidentiality statements to sign regarding the data.**
3. **Create new repository on mbox (discuss with Susan) for the data.**
4. **Data Pre-processing**
   1. Organize raw data.
   2. Process raw data into data frames for planned analyses.
   3. Create an Rdata workspace file that contains all data frames for the planned analyses. (ex. csv.RData)
   4. This data can only go on mbox. Do not put this data on your computer or on others’ computers. See <https://github.com/StatisticalReinforcementLearningLab/HeartstepsV1Code/wiki/1-Mounting-Box> for instructions on how to access mbox files locally without downloading them.
5. **Initial check of availability.** 
   1. Perform descriptive analyses to understand availability such as
      1. Plot percent unavailability at each decision point.
      2. Count number of people available at each decision point.
6. **Check randomization** (have example code from Danny)
   1. Perform descriptive analyses to check to see if the fractions of time points assigned a particular treatment correspond to the randomization probabilities. Note that this is only for participants who were available.
   2. Perform descriptive analyses to check if the randomization is balanced across pre-decision point measures of the proximal outcome.
7. **Other types of descriptive analysis**
8. **Think about and make a proposal to research team about the data/participants that will be included in primary analyses.** 
   1. Do descriptive statistics concerning users who left study prior to study end
   2. Do descriptive statistics concerning users who are absent for periods during the study
   3. Make a proposal to team about which users and which data on these users will be part of primary analysis. Make a proposal about potential sensitivity analyses (see below)
9. **Summary statistics** (suggestions below; depending on the study some of these suggestions will be nonsensical)
   1. Check for odd or unusual or outlier values for all important variables, paying attention to unusual missingness which might indicate a bug, participant travel, drop-out, etc. which may have gone unnoticed in pre-processing.
   2. Mean and SD of proximal outcome for available decision points for different treatments.
   3. Plot number of days for which the proximal outcome is completely missing/zero for each participant.
   4. Plot average proximal outcome by day, user, and which intervention option was delivered, only among available times.
   5. Plot mean difference in proximal outcome between intervention option delivered by user.
   6. Plot mean different in proximal outcome between intervention options averaged over all users.

After the primary analysis (whatever that might be for the particular MRT) you need to conduct sensitivity analyses.

1. **Sensitivity Analysis**
   1. Re-run the model with varying number of users if the primary analysis concerns only a subset of users.
   2. Re-run models deleting unusual users, such as users that left the trial early, for whom there was no strong scientific justification for removal.
   3. Re-run models with potential outliers removed
   4. If there is significant decrease in the treatment effect over time, re-run the model with a subset of days that excludes later/last days of study. This may be conditional on their being fewer people participating in the study towards the end. Also it might need to be considered how participants may become more engaged with the intervention near the end anticipating their exit interview.
   5. Re-run the model with data from participants that was potentially mislabeled.