Predicting alcohol lapses from contextualized cellular communication logs

Kendra Wyant and John Curtin

September 28, 2021

Study Aims

1. Train and evaluate the best performing machine learning model to predict alcohol lapse from cellular communication logs and social context information about these communications.
2. Employ a model comparison approach to compare models that use all available features (both passive signals from communications logs and actively measured social context) with models that are restricted to only passive signals.
3. Evaluate the importance of feature sets within the top performing model to inform potential in situ treatments.

Data

This project performs the initial analyses on a subset of data collected from 2017 – 2019 as part of a larger grant funded through the National Institute of Alcohol Abuse and Alcoholism (R01 AA024391). A full list of measures used in the parent project can be found at https://osf.io/brs68/.

Preregistration

Registration after data cleaning, but before any main analyses. The authors of this study have seen the data prior to this preregistration for purposes of data cleaning and developing a machine learning framework for modeling. No feature engineering or model comparisons related to the study’s aims have been conducted.

Research Transparency

We value the principles of research transparency that are essential to the robustness and reproducibility of science [1]. Consequently, we will maximize transparency through several complementary methods. First, we will report how we determined our sample size, all data exclusions, all manipulations, and all available measures in the study [2]. Second, we will complete a transparency checklist [3]. Third, we will make the data, analysis scripts and annotated results, self-report surveys, and other study materials associated with this study publicly available on our Open Science Framework page [https://osf.io/brs68/].

Method

Participants

We recruited a community sample of people in initial stages of recovery for Alcohol Use Disorder from the Madison area. Avenues of recruitment included referrals from clinics, self-help groups, Facebook, radio, and television. Those who were interested in participating in our three-month study were given a brief description (i.e., told that our research focused on learning how mobile health technology can be used to provide individual support to anyone recovering from alcohol addiction) a short phone screen to determine initial eligibility (i.e., At least 18 years old, ability to read and write in English, and an eligible smartphone with existing cellphone plan).

Two hundred sixteen participants passed the initial phone screen and came in for a more in-depth screening session. Of the 216 interested participants, 199 enrolled in the study (i.e., they consented and were eligible to participate). We excluded potential participants if they did not meet the criteria for moderate or severe Alcohol Use Disorder (as defined by the DSM-5), did not have a goal of long-term abstinence, had not abstained from alcohol for at least one week, already had over two months of abstinence, or had severe symptoms of psychosis or paranoia. Of the 199 participants enrolled in the study, 154 provided at least one month of data and make up the sample used in our analyses.

Procedure

Our study involved five in-person visits (screening, enrollment, and three follow-up visits). It is also required completion of daily EMA surveys to document alcohol lapses, and access to non-deleted text message and call logs (i.e., cellular communication logs). All procedures were approved by the University of Wisconsin-Madison Institutional Review Board.

During the screening session we obtained informed consent, determined eligibility, and documented basic demographic information. Participants that consented and were deemed eligible came back for a second enrollment visit. During enrollment, participants were briefed on how to delete log entries they did not want to share with us and completed a practice EMA survey. Participants also reported contacts they frequently communicated with and answered a series of questions documenting contextual information about their interactions with each contact (type of relationship with contact, whether they drank with contact in past, drinking status of contact, whether contact would drink in their presence, whether the contact is in recovery, the level of supportiveness the contact provides, and the pleasantness of their interactions with the contact). Participants returned for three follow-up visits, each one month apart.

At each follow-up visit, we downloaded participants' cellular communication logs. These logs included the phone number of the other party, whether the call or message was incoming or outgoing, the duration of the call, and the date and time of the call or message. Participants also provided context information for newly identified frequent contacts (i.e., at least two communications in the past month). At the third follow-up visit, participants were debriefed and thanked for their participation.

Data Analysis Plan

We will conduct all analyses in R version 4.1.1 [4] using RStudio [5] and the tidyverse [6] and tidymodels [7] ecosystem of packages.

Our first study aim is to train and evaluate the best performing machine learning model to predict alcohol lapse from contextualized cellular communication data. We will build, train, and evaluate models with several statistical learning algorithms including penalized parametric linear classification algorithms (LASSO, ridge regression, glmnet), non-parametric classification algorithms (k nearest neighbor), and ensemble methods (random forest). Candidate statistical learning algorithms will be trained on a subset of the data (training sample) using combinations of features derived from participants cellular communications and social context information.

We will use grouped 10x10-fold resampling to select the top performing model (statistical algorithm and combination of features). All folds will be grouped by participant ID so that a single participant's data will not being used to predict future data by the same participant. We will evaluate expected model performance by our top performing model on new data (i.e., participants not used to train models) with an independent held-out test sample. We will use balanced accuracy and the area under the receiver operating characteristic curve (AUC; i.e., measure of sensitivity vs. specificity across classification thresholds) as our performance metrics for model selection and evaluation.

We will use various feature engineering methods to build groups of feature sets to maximize model performance. The first distinction between features will be to discriminate between features derived from passive only measures (i.e., communication logs) and those derived from more active measures (i.e., context information). Our study's second aim is to compare models that use all available features (both passive signals from communications logs and actively measured context) with models restricted to only passive signals. To do this we will employ a model comparison approach. Through this relative comparison, we can quantify any performance benefit from adding the active component of context. The incremental benefit in performance will be inferentially evaluated by comparing balanced accuracy and AUC values.

Additionally, we will use domain knowledge to engineer features from the raw data using averages, proportions, sums, interactions, and more. For example, dates and times from the communication logs can be dummy-coded to represent features such as weekends, happy hour, and business hours. We will also engineer our raw data so that we can recognize patterns and changes over time (e.g., more activity during happy hour this week).

Finally, we will use various period durations from which to engineer our features (e.g., two weeks of data vs. one month of data) and different lapse onset times for prediction (e.g., predicting lapse one hour out vs. three days out).

Our third aim of the present study is to evaluate the importance of feature sets within the top performing model. The most predictive features will be identified based on feature important indices. We will also combine model comparison and feature ablation methods to remove subsets of context features in order to test their predictive utility.

---

[1] http://www.researchtransparency.org/who-we-are/ ↩

[2] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2160588 ↩

[3] https://www.nature.com/articles/s41562-019-0772-6 ↩

[4] https://www.r-project.org/ ↩

[5] https://www.rstudio.com/ ↩

[6] https://www.tidyverse.org/ ↩

[7] https://www.tidymodels.org/ ↩