Forecasting Risk of Alcohol Lapse up to Two Weeks in Advance using Time-lagged Machine Learning Models

Kendra Wyant[aff-1], Gaylen E. Fronk[aff-1,aff-2], Jiachen Yu[aff-1], Claire E. Punturieri[aff-1], and John J. Curtin[aff-1]

[aff-1]Department of Psychology, University of Wisconsin-Madison

[aff-2]Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina

Abstract

We developed machine learning models to predict future alcohol lapses within 24-hour windows lagged 1 day, 3 days, 1 week, and 2 weeks. We engineered features from 4x daily ecological momentary assessment from individuals (N=151; 51% male; mean age=41; 87% non-Hispanic White) in early recovery over three months. We trained and evaluated models using nested cross-validation. Median posterior auROC was high (0.85–0.91) for all models but decreased modestly with increasing lag. Models performed worse for non-advantaged groups (non-White and/or Hispanic, below poverty, female). Past alcohol use, abstinence self-efficacy, and craving were the most important features, with the magnitude of importance varying meaningfully by lag. These findings demonstrate feasibility of predicting next-day lapses up to two weeks in advance. Embedding these models in a recovery monitoring support system could enable adaptive, personalized care. Improving model fairness and optimizing the delivery of model feedback to sustain engagement remain critical next steps.

Keywords: Substance use disordersPrecision mental health

Introduction

Alcohol and other substance use disorders (SUDs) are serious chronic conditions, characterized by high relapse rates [1,2], substantial co-morbidity with other physical and mental health problems [2,3], and an increased risk of mortality [4,5]. Too few individuals receive medications or clinician-delivered interventions to help them initially achieve abstinence and/or reduce harms associated with their use [3]. Moreover, this problem is even worse for subsequent continuing care during SUD recovery. Continuing care, including both risk monitoring and ongoing support, is the gold standard for managing chronic health conditions such as diabetes, asthma, and HIV [6]. Yet, continuing care for SUDs is largely lacking despite ample evidence that SUDs are chronic, relapsing conditions [3,7,8].

An important focus of continuing care during SUD recovery is to prevent lapses (i.e., single instances of goal-inconsistent substance use) and full relapse back to harmful use [9,10]. Critically, the risk factors that instigate lapses during recovery are individualized, numerous, dynamic, interactive, and non-linear [11,12]. The optimal supports to address these risk factors and encourage continued, successful recovery vary both across individuals and within an individual over time. Given this, continuing care could benefit greatly from a precision mental health approach that seeks to provide the right support to the right individual at the right time, every time [13–15]. However, such monitoring and personalized support must also be highly scalable to address the substantial unmet need for SUD continuing care.

Recent advances in both smartphone sensing [16] and machine learning [17] hold promise as a scalable foundation for monitoring and personalized support during SUD recovery. Smartphone sensing approaches (e.g., ecological momentary assessment [EMA], geolocation sensing) can provide the frequent, longitudinal measurement of proximal risk factors that is necessary for prediction of future lapses with high temporal precision. EMA may be particularly well-suited for lapse prediction because it can provide privileged access to subjective experiences (e.g., craving, affect, stress, motivation, self-efficacy) that are targets for change in evidence-based approaches for relapse prevention [9,10,18]. Furthermore, individuals with SUDs have found EMA to be acceptable for sustained measurement for up to a year with relatively high compliance [19,20], suggesting that this method is feasible for long-term monitoring throughout SUD recovery.

Machine learning models are well-positioned to use EMAs as inputs to provide temporally precise prediction of the probability of future lapses with sufficiently high performance to support decisions about interventions and other supports for specific individuals. These models can handle the high dimensional feature sets that may result from feature engineering densely sampled raw EMA over time [**wyantMachineLearningModels2023?**]. They can also accommodate non-linear and interactive relationships between features and lapse probability that are likely necessary for accurate prediction of lapse probability. Moreover, rapid advances in the tools for interpretable machine learning (e.g., Shapley values [21]) now allow us to probe these models to understand which risk features contribute most strongly to a lapse prediction for a specific individual at a specific moment in time. Interventions, supports, and/or suggested lifestyle adjustments can then be personalized to address these risks following from our understanding about relapse prevention.

Preliminary research is now emerging that uses features derived from EMAs in machine learning models to predict the probability of future alcohol use [22,23,**wyantMachineLearningModels2023?**]. This research is important because it rigorously required strict temporal ordering necessary for true prediction, with features measured before alcohol use outcomes. These studies also used resampling methods (e.g., cross-validation) that prioritize model generalizability to increase the likelihood these models will perform well with new people. Perhaps most importantly, [**wyantMachineLearningModels2023?**] demonstrated that machine learning models using EMA can provide predictions with very high temporal precision at clinically implementable levels of performance. Specifically, we developed models that predict lapses in the immediate future (i.e., the next day and even the next hour) with areas under the receiver operating characteristic curve (auROCs) of 0.91 and 0.93, respectively.

[**wyantMachineLearningModels2023?**]'s next day lapse prediction model can provide personalized support recommendations to address immediate risks for possible lapses. Features derived from past EMAs can be updated in the early morning to yield the predicted lapse probability for an individual that day. Personalized supports that target the top features contributing to that prediction can then be provided. For example, if predicted lapse probability is high due to frequent craving, the individual could be reminded about the benefits of urge surfing or distracting activities during brief periods when cravings arise. Conversely, guided relaxation techniques could be recommended if lapse probability is high due to recent past and anticipated stressors that day. Patients could also be assisted to implement any of these recommendations using videos or other tools within a digital therapeutic. Curtin and colleagues are currently evaluating outcomes associated with the use of this “smart” (machine learning guided) recovery monitoring and support system (RMSS) for patients with an alcohol use disorder (AUD) [24].

Despite the promise offered by a smart RMSS based on immediate future risks (e.g., the next day), such a system has limitations. Most importantly, recommendations must be limited to previously learned skills and/or supports that are available to implement that day. However, many risks may require supports that are not available in the moment. For example, to address lifestyle imbalances, several future positive activities may need to be planned. Time with supportive friends or an AA sponsor may require time to schedule. Similarly, work or family schedules may need to be adjusted to return to attending self-help meetings. If new recovery skills or therapeutic activities are needed to address emerging risks, patients may need to book sessions with a therapist. In all these instances, patients would benefit from advanced warning about changes in their lapse probability and the associated risks that contribute to these changes. A smart RMSS could provide this advanced warning by lagging lapse probability predictions further into the future (e.g., predicting lapse probability in a 24-hour window that begins two weeks in the future). However, we do not know if such lagged models could maintain adequate performance for clinical implementation.

In this study, we evaluated the performance of machine learning models that predict the probability of future lapses within 24-hour prediction windows that were systematically lagged further into the future. We considered several meaningful lags for these prediction windows: 1

day, 3 days, 1 week, and 2 weeks. We conducted pre-registered analyses of both the absolute performance of these lagged models and their relative performance compared to a baseline model that predicted lapse probability in the immediate next day (i.e., no lag). In addition to the aggregate performance of these models, we also evaluated algorithmic fairness by comparing model performance across important subgroups that have documented disparities in treatment access and/or outcomes. These include comparisons by race/ethnicity [25,26], income [27] and sex at birth [26,28]. Finally, we calculated Shapley values for feature categories defined by EMA items to better understand how these models generate predictions and how these features can be used to tailor personalized supports.

Methods

Transparency

We adhere to research transparency principles that are crucial for robust and replicable science. We preregistered our data analytic strategy. We reported all data exclusions. Our data, questionnaires, preregistration, and other study materials are publicly available on our OSF page (https://osf.io/xta67/). Our annotated analysis scripts and results are publicly available on our study website (https://jjcurtin.github.io/study_lag/).

Participants

We recruited 192 participants in early recovery from AUD in Madison, Wisconsin, USA for a 3-month longitudinal study. This sample size was determined based on traditional power analysis methods for logistic regression [29] because comparable approaches for machine learning models have not yet been validated. Participants were recruited through print and targeted digital advertisements and partnerships with treatment centers. We required that participants:

1. were age 18 or older,
2. could write and read in English,
3. had at least moderate AUD (>= 4 self-reported DSM-5 symptoms),
4. were abstinent from alcohol for 1-8 weeks, and
5. were willing to use a single smartphone (personal or study provided) while on study.

We also excluded participants exhibiting severe symptoms of psychosis or paranoia (Defined as scores >2.2 or 2.8, respectively, on the psychosis or paranoia scales of the Symptom Checklist–90 [30])

Of the 192 eligible participants, 191 consented to participate in the study at the screening visit, and 169 subsequently enrolled in the study at the enrollment visit, which occurred approximately one week later. Fifteen participants discontinued before the first monthly follow-up visit. We excluded data from one participant who did not maintain a goal of abstinence during their participation. We also excluded data from two participants due to evidence of careless responding and unusually low compliance. Our final sample consisted of 151 participants.

Procedure

Participants completed five study visits over approximately three months. After an initial phone screen, participants attended an in-person screening visit to determine eligibility, complete informed consent, and collect self-report measures. Eligible, consented participants returned approximately one week later for an intake visit. Three additional follow-up visits occurred about every 30 days that participants remained on study. Participants were expected to complete four daily EMAs. Other personal sensing data streams (geolocation, cellular communications, sleep quality, and audio check-ins) were collected as part of the parent grant's aims (R01 AA024391). Participants could earn up to $150/month if they completed all study visits, had 10% or less missing EMA data, and opted in to provide data for other personal sensing data streams.

Ethics

All procedures were approved by the University of Wisconsin-Madison Institutional Review Board (Study #2015-0780) and carried out in accordance with the principles of the Declaration of Helsinki. All participants provided written informed consent.

Measures

Ecological Momentary Assessments

Participants completed four brief (7-10 questions) EMAs daily. The first and last EMAs of the day were scheduled within one hour of participants' typical wake and sleep times. The other two EMAs were scheduled randomly within the first and second halves of their typical day, with at least one hour between EMAs. Participants learned how to complete the EMA and reviewed the meaning of each question with a member of the research team during their intake visit to ensure consistent question interpretation.

On all EMAs, participants reported dates and times of any previously unreported past alcohol use. Next, participants rated the maximum intensity of recent (i.e., since last EMA) experiences of craving, risky situations, stressful events, and pleasant events. Finally, participants rated their current affect on two bipolar scales: valence (Unpleasant/Unhappy to Pleasant/Happy) and arousal (Calm/Sleepy to Aroused/Alert). On the first EMA each day, participants also rated anticipated risky situations, stressful events, and the likelihood that they would drink alcohol in the next week (i.e., abstinence self-efficacy).

Individual Characteristics

We collected self-report information about demographics (age, sex at birth, race, ethnicity, education, marital status, employment, and income) and clinical characteristics (AUD milestones, number of quit attempts, lifetime AUD treatment history, lifetime receipt of AUD medication, DSM-5 AUD symptom count, current drug use [31], and presence of psychological symptoms [30] to characterize our sample. DSM-5 AUD symptom count and presence of psychological symptoms were also used to determine eligibility. Demographics were included as features in our models. A subset of these variables (sex at birth, race, ethnicity, and income) were used for model fairness analyses, as they have documented disparities in treatment access and outcomes. As part of the aims of the parent project, we collected many other trait and state measures throughout the study. A complete list of all measures can be found on our study's OSF page (https://osf.io/xta 67/).

Data Analytic Strategy

Data preprocessing, modeling, and Bayesian analyses were done in R (version 4.4.2) using the tidymodels ecosystem [32–34]. Models were trained and evaluated using high-throughput computing resources provided by the University of Wisconsin Center for High Throughput Computing [35].

Predictions

A *prediction timepoint* (Figure 1, Panel A) is the hour at which our model calculates a predicted probability of a lapse within a future 24-hour prediction window for any specific individual. We calculated the features used to make predictions at each prediction timepoint within a feature scoring epoch that included all available EMAs up until, but not including, the prediction timepoint. The first prediction timepoint for each participant was 24 hours from midnight on their study start date. This ensured at least 24 hours of past EMAs were available in the feature scoring epoch. Subsequent prediction timepoints for each participant repeatedly rolled forward hour-by-hour until the end of their study participation.

The *prediction window* (Figure 1, Panel B) spans a period of time in which a lapse might occur. The prediction window width for all models was 24 hours (i.e., models predicted the probability of a lapse occurring within a specific 24-hour period). Prediction windows rolled forward hour-by-hour with the prediction timepoint. However, there were five possible *lag times* between the prediction timepoint and start of the associated prediction window. A prediction window either started immediately after the prediction time point (no lag) or was lagged by 1 day, 3 days, 1 week, or 2 weeks.

Given this structure, our models provided hour-by-hour predicted probabilities of an alcohol lapse in a future 24-hour period. Depending on the model, that future period (the prediction window) might start immediately after the prediction timepoint or up to 2 weeks into the future. For example, at midnight on the 30th day of participation, the feature scoring epoch would include the past 30 days of EMAs. Separate models would predict the probability of lapse for 24-hour periods starting at midnight that day, or 24-hour periods starting 1 day, 3 days, 1 week or 2 weeks after midnight on day 30.

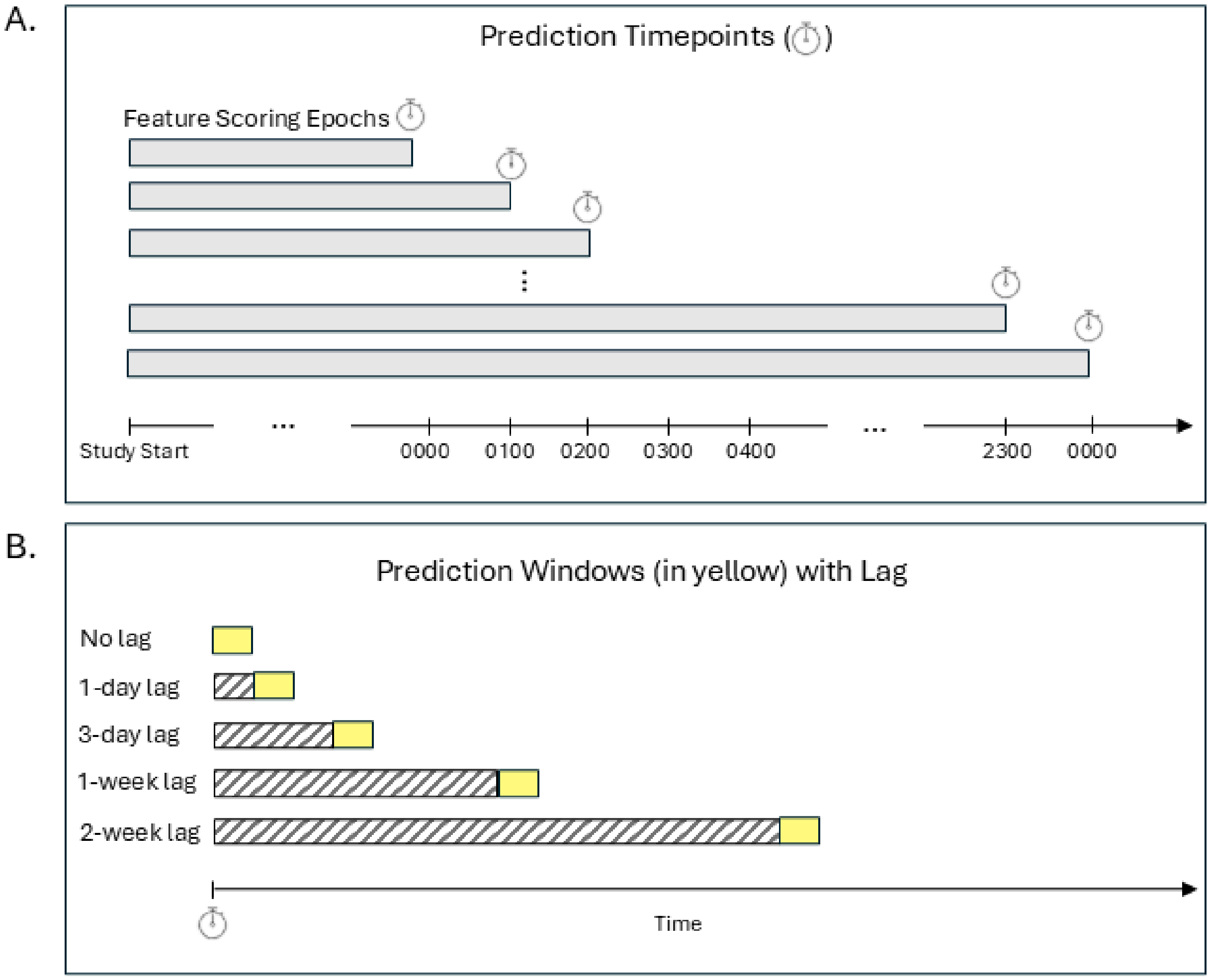

Figure 1: Panel A shows the prediction timepoints at which our model calculated a predicted probability of a lapse. All available data up until, but not including, the prediction timepoint was used to generate these predictions. Features were created for varying feature scoring epochs before the prediction timepoint (i.e., 12, 24, 48, 72, and 168 hours). Prediction timepoints were updated hourly. Panel B shows how the prediction window (i.e., window in which a lapse might occur) rolls forward hour-by-hour with the prediction timepoint. The prediction window width for all models was 24 hours. Additionally, there were five possible lag times between the prediction timepoint and start of the prediction window. A prediction window either started immediately after the prediction timepoint (no lag) or was lagged by 1 day, 3 days, 1 week, or 2 weeks.

Labels

The start and end dates and times of past drinking episodes were reported on the first EMA item. A prediction window was labeled *lapse* if the start date/hour of any drinking episode fell within that window. A window was labeled *no lapse* if no alcohol use occurred within that window +/- 24 hours. If no alcohol use occurred within the window but did occur within 24 hours of the start or end of the window, the window was excluded. We used this conservative 24-hour fence for labeling windows as no lapse (vs. excluded) to increase the fidelity of these labels. Given that most windows were labeled no lapse, and the outcome was highly unbalanced, it was not problematic to exclude some no lapse events to further increase confidence in those labels.

This method produced totals of: 274,179 labels for the baseline (no lag) model; 270,911 labels for the 1-day lagged model; 264,362 labels for the 3-day lagged model; 251,458 labels for the 1-week lagged model; and 228,420 labels for the 2-week lagged model.

Feature Engineering

Features were calculated using only data collected in feature scoring epochs before each prediction timepoint to ensure our models were making true future predictions. For the no lag model, the prediction timepoint was at the start of prediction window, so all data prior to the start of the prediction window were included. For the lagged models, the prediction timepoint was 1 day, 3 days, 1 week, or 2 weeks prior to the start of the prediction window, so the last EMA data used for feature engineering were collected 1 day, 3 days, 1 week, or 2 weeks prior to the start of the prediction window.

A total of 285 features were derived from three data sources:

1. *Prediction window*: We dummy-coded features for day of the week at the start of the prediction window.
2. *Demographics*: We created quantitative features for age (in years) and personal income (in dollars), and dummy-coded features for sex at birth (male vs. female), race/ethnicity (non-Hispanic White vs. non-White and/or Hispanic), marital status (married vs. not married vs. other), education (high school or less vs. some college vs. college degree), and employment status (employed vs. unemployed).
3. *Previous EMA responses*: We calculated raw and change features using EMAs in varying feature scoring epochs (i.e., 12, 24, 48, 72, and 168 hours) before the prediction timepoint for all EMA items. Raw features included min, max, and median scores for each EMA item across all EMAs in each epoch for a given participant. We calculated change features by subtracting each participant's baseline mean score for each EMA item from their raw feature. These baseline mean scores were calculated using all of a participant's EMAs collected from the start of participation until the prediction timepoint. We also created raw and change features based on the most recent response for each EMA question and raw and change rate features from previously reported lapses and number of completed EMAs.

Features had missing values if the participant did not respond to the relevant EMA question during the associated scoring epoch. The proportion of missing values across features and models was low (median = .02, range = 0 - .13). We imputed missing data using median imputation for numeric features and mode imputation for nominal features. We selected coarse median/mode methods for handling missing data due to the computational costs associated with more advanced forms of imputation (e.g., KNN imputation, multiple imputation). Importantly, our imputation calculations are done using only held-in data and can be applied to any new observation.

Other generic feature engineering steps included removing zero and near-zero variance features as determined from held-in data. A sample feature engineering script (i.e., tidymodels recipe) containing all feature engineering steps is available on our OSF study page.

Model Training and Evaluation

Model Configurations. We trained and evaluated five separate classification models: one baseline (no lag) model and one model for 1-day, 3-day, 1-week, and 2-week lagged predictions. We considered four well-established statistical algorithms (elastic net, XGBoost, regularized discriminant analysis, and single layer neural networks) that vary across characteristics expected to affect model performance (e.g., flexibility, complexity, handling higher-order interactions natively) [36]. Candidate model configurations differed across sensible values for key hyperparameters. Configurations also differed on outcome resampling method (i.e., no resampling and up-sampling and down-sampling of the outcome using majority/no lapse to minority/lapse ratios ranging from 5:1 to 1:1).

Cross-validation. We used participant-grouped, nested cross-validation for model training, selection, and evaluation with auROC. auROC indexes the probability that the model will predict a higher score for a randomly selected positive case (lapse) relative to a randomly selected negative case (no lapse). Grouped cross-validation assigns all data from a participant as either held-in or held-out to avoid bias introduced when predicting a participant's data from their own data. Folds were stratified on a between-subject variable of low vs. high lapsers (low lapsers reported fewer than 10 lapses while on study, and high lapsers reported 10 or more lapses while on study). We used 2 repeats of 5-fold cross-validation for the inner loops (i.e., *validation* sets) and 6 repeats of 5-fold cross-validation for the outer loop (i.e., *test* sets). Best model configurations were selected using median auROC across the 10 validation sets. Final performance evaluation of those best model configurations used median auROC across the 30 test sets.

Bayesian Model. We used a Bayesian hierarchical generalized linear model to estimate the posterior probability distributions and 95% Bayesian credible intervals (CIs) from the 30 held-out test sets for our five best models. Following recommendations from the rstanarm team and others [37,38], we used the rstanarm default autoscaled, weakly informative, data-dependent priors that take into account the order of magnitude of the variables to provide some regularization to stabilize computation and avoid over-fitting. Priors were set as follows: residual standard deviation ~ normal(location=0, scale=exp(2)), intercept (after centering predictors) ~ normal(location=2.3, scale=1.3), the two coefficients for window width contrasts ~ normal (location=0, scale=2.69), and covariance ~ decov(regularization=1, concentration=1, shape=1, scale=1). We set two random intercepts to account for our resampling method: one for the repeat, and another for the fold nested within repeat. We specified two sets of pre-registered contrasts for model comparisons. The first set compared each lagged model to the baseline no lag model (1-day lag vs. no lag, 3-day lag vs. no lag, 1-week lag vs. no lag, 2-week lag vs. no lag). The second set compared adjacently lagged models (3-day lag vs. 1-day lag, 1-week lag vs. 3-day lag, 2-week lag vs. 1-week lag). auROCs were transformed using the logit function and regressed as a function of model contrast.

From the Bayesian model we obtained the posterior distribution (transformed back from logit) and Bayesian CIs for auROCs for all five models. To evaluate our models' overall performance, we report the median posterior probability for auROC and Bayesian CIs. This represents our best estimate for the magnitude of the auROC parameter for each model. If the CIs do not contain .5 (chance performance), this provides strong evidence (> .95 probability) that our model is capturing signal in the data.

We then conducted Bayesian model comparisons using our two sets of contrasts - baseline and adjacent lags. For both model comparisons, we determined the probability that the models' performances differed systematically from each other. We also report the precise posterior probability for the difference in auROCs and the 95% Bayesian CIs.

Fairness Analyses. Using the same 30 held-out test sets, we calculated the median posterior probability and 95% Bayesian CI for auROC for each model separately by race/ethnicity (non-White and/or Hispanic vs. non-Hispanic White), income (below poverty line vs. above poverty line, and sex at birth (female vs. male). We used course race/ethnicity groupings due to the limited diversity of these demographics in our sample. The poverty cutoff was defined from the 2024 federal poverty line for the 48 contiguous United States. Participants at or below $15,060 annual income were categorized as below poverty. We conducted Bayesian group comparisons to assess the likelihood that each model performs differently by group. We summarize the differences in posterior probabilities for auROC across models. Individual Bayesian fairness contrasts for all five models are available in the supplement.

Model Characterization

To further characterize and understand our models, we used our inner resampling procedure (2 repeats of 5-fold cross validation grouped on participant and stratified by high/low lapsers) on the full data set to select a single best model configuration for each classification model (no lag, 1-day, 3-day, 1-week, and 2-week lag). The final configuration selected for each model represents the most reliable and robust configuration for deployment. We can better understand our final models by looking at the calibration of the predicted probabilities and the most important features contributing to those predictions.

Model Calibration. The best model configuration for each classification model was fit on the full data set. We fit this configuration using single 5-fold cross-validation. This method allowed us to obtain a single predicted probability for each observation, while still using separate data for model training and prediction. We calibrated our probabilities using Platt scaling [39]. We calculated Brier scores to assess the accuracy of our raw and calibrated probabilities for the no lag and 2-week lagged models. Brier scores range from 0 (perfect accuracy) to 1 (perfect inaccuracy). A table of Brier scores for all five models is available in the supplement. We provide calibration plots for the no lag and 2-week lagged models (calibration plots for all five models are available in the supplement).

Global Feature Importance. We used the same single 5-fold cross-validation procedure to calculate raw Shapley values for observations in our held-out folds. Raw Shapley values index the importance of any feature (or set/category of features as described below) to any single prediction for a specific observation (i.e., for a specific 24-hour window for a specific participant), which indicates the “local importance” of that feature [21]. More precisely, the magnitude of the raw Shapley value for any feature indicates how much the feature score for that observation adjusted the prediction (in log-odds units) for that observation relative to the mean prediction across all observations. Positive Shapley values indicate that the feature score increased the prediction for that observation and negative values indicate that the feature score decreased the prediction.

Raw Shapley values are additive across features for an observation, with their sum across features equal to the total adjustment of the predicted value for that observation vs. the mean predicted value across all observations. This property allows raw Shapley values to be added together across features within a category to index the importance of that feature category. We created feature categories by summing raw Shapley values for all features associated with specific EMA items. In three instances, we combined features across two similar EMA items (i.e., past and anticipated risky situations, past and anticipated stressful events, and affective valence and arousal) to yield seven feature categories for distinct constructs assessed by the original 10 EMA items. Specifically, we calculated Shapley values for past use, craving, affective state, past/anticipated risky situations, past/anticipated stressful events, past pleasant events, and abstinence self-efficacy.

Shapley values can be aggregated across observations to describe the global importance of any feature (or feature category) across all predictions (i.e., for all 24-hour windows for all participants) in the dataset. Global feature importance is calculated by averaging the absolute value of the Shapley values for a feature across all observations. A large mean absolute Shapley value indicates that the feature makes big contributions to the predictions across the dataset. Global feature importance is a descriptive statistic that indicates the importance of the feature for predictions in a specific dataset, rather than a hypothetical population of observations. We provide a descriptive plot of the relative ranking of feature categories by their global feature importance for the no lag and 2-week lagged models. Global feature importance plots for all five models are available in the supplement.

Results

Demographic and Lapse Characteristics

Table 1 provides a detailed breakdown of the demographic and clinical characteristics of our sample (N = 151).

Table 1: Demographic and Clinical Characteristics

	N	%	M	SD	Range
Age			41	11.9	21-72
Sex at Birth					
Female	74	49.0			
Male	77	51.0			
Race					
American Indian/Alaska Native	3	2.0			
Asian	2	1.3			
Black/African American	8	5.3			
White/Caucasian	131	86.8			
Other/Multiracial	7	4.6			
Hispanic, Latino, or Spanish origin					
Yes	4	2.6			
No	147	97.4			
Education					
Less than high school or GED degree	1	0.7			
High school or GED	14	9.3			
Some college	41	27.2			
2-Year degree	14	9.3			
College degree	58	38.4			
Advanced degree	23	15.2			
Employment					
Employed full-time	72	47.7			
Employed part-time	26	17.2			
Full-time student	7	4.6			
[illegible] than	[illegible]	[illegible]	[illegible]298	[illegible]807	[illegible]0,000
N = 151					
[illegible] Two participants reported 100 or more quit attempts. We removed these outliers prior to calculating the mean (M), standard deviation (SD), and range.					

Model Evaluation

Figure 2 presents the full posterior probability distributions for auROC for each model (no lag, 1-day, 3-day, 1-week, and 2-week lag). The median auROCs from these posterior distributions were 0.91 (no lag), 0.89 (1-day lag), 0.88 (3-day lag), 0.87 (1-week lag), and 0.85 (2-week lag). These values represent our best estimates for the magnitude of the auROC parameter for each model. The 95% Bayesian CI for the auROCs for these models were relatively narrow and did not contain 0.5: no lag [0.90-0.92], 1-day lag [0.88-0.90], 3-day lag [0.87-0.90], 1-week lag [0.85-0.88], 2-week lag [0.83-0.87].

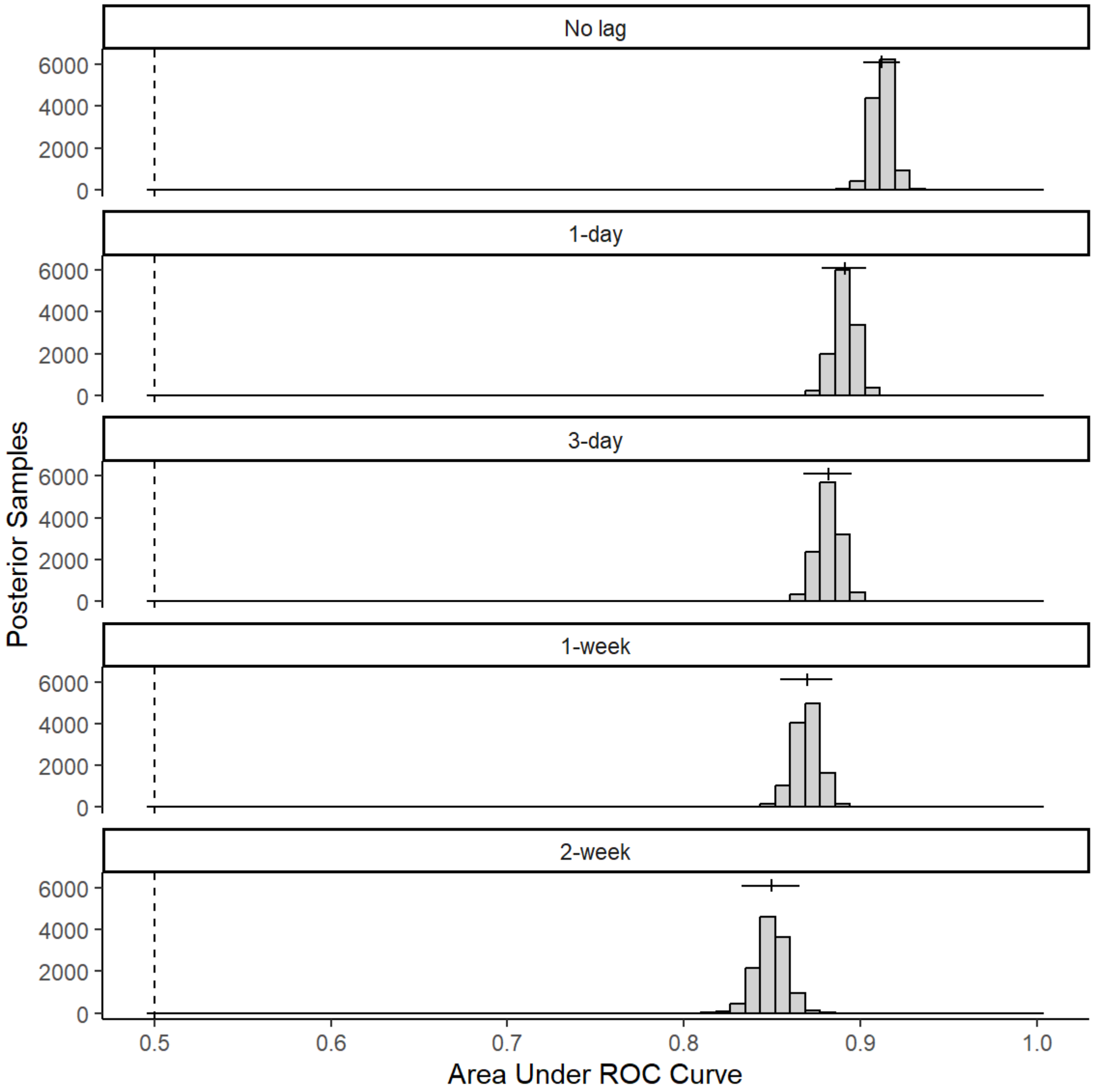

Figure 2: Posterior probability distributions for area under ROC curve (auROC) for each model (no lag, 1-day, 3-day, 1-week, and 2-week lag). Each distribution reflects 12,000 posterior samples (4 chains × 3,000 samples) from a Bayesian hierarchical generalized linear model. Horizontal lines depict 95% Bayesian credible intervals (CI) and vertical solid lines depict median posterior probability for auROC. Vertical dashed line represents expected performance from a random classifier (.5 auROC).

Model Comparisons

Table 2 presents the median difference in auROC, 95% Bayesian CI, and posterior probability that that the auROC difference was smaller than 0 for all baseline and adjacent lag contrasts. Median auROC differences less than 0 indicate the more lagged model, on average, performed worse than the more immediate model (e.g., 1-day lag – no lag, 3-day lag – 1-day lag). There was strong evidence (probabilities = 1) that the lagged models performed worse than the baseline (no lag) model, with average drops in auROC ranging from 0.02-0.06, and the previous adjacent lagged model, with average drops in auROC ranging from 0.01-0.02.

Table 2: Median difference in auROC, 95% Bayesian credible interval (CI), and posterior probability that that the auROC difference was smaller than 0 for all baseline and adjacent lag contrasts.

Contrast	Median	Bayesian CI	Probability
Baseline Contrasts			
1-day vs. No lag	−0.021	[−0.025, −0.017]	1
3-day vs. No lag	−0.03	[−0.035, −0.025]	1
1-week vs. No lag	−0.042	[−0.048, −0.037]	1
2-week vs. No lag	−0.062	[−0.07, −0.056]	1
Adjacent Contrasts			
3-day vs. 1-day	−0.009	[−0.013, −0.005]	1
1-week vs. 3-day	−0.012	[−0.017, −0.008]	1
2-week vs. 1-week	−0.02	[−0.026, −0.015]	1
Median auROC differences less than 0 indicate the more lagged model, on average, performed worse than the more immediate model (e.g., 1-day lag - no lag, 3-day lag - 1-day lag). Bayesian CI represents the range of values where there is a 95% probability that the true auROC difference lies within that range. Probability indicates the posterior probability that this difference is smaller than 0 (i.e., the models are performing differently).			

Fairness Analyses

Table 3 presents the median difference in auROC, 95% Bayesian CI, and posterior probability that the auROC difference was smaller than 0 for the three fairness contrasts: race/ethnicity (non-White and/or Hispanic; *N* = 20 vs. Non-Hispanic White; *N* = 131), sex at birth (female; *N* = 74 vs. male; *N* = 77), and income (below poverty line; *N* = 49 vs. above poverty line; *N* = 102). Median auROC differences less than 0 indicate the model, on average, performed worse for the non-advantaged group (female, non-White and/or Hispanic, below poverty line) compared to the advantaged group (male, non-Hispanic White, and above poverty line). In Table 3 we present fairness analyses for our baseline model (no lag) and for our longest lagged model (2-week lag), as this is likely the most clinically useful lagged model for providing advanced warning of lapse risk. Fairness analyses for all five models are available in the supplement.

There was strong evidence (probabilities > .84) that our models performed worse for the non-advantaged groups compared to the advantaged groups. On average, across all five models, there was a median decrease in auROC of 0.13 (range 0.13-0.17) for participants who were non-White and/or Hispanic compared to participants who were non-Hispanic White. On average, across all five models, there was a median decrease in auROC of 0.05 (range 0.04-0.10) for female participants compared to male participants. On average, across all five models, there was a median decrease in auROC of 0.02 (range 0.01-0.04) for participants below the federal poverty line compared to participants above the federal poverty line.

The proportion of positive lapse labels over all labels (lapse and no lapse) for each demographic subgroup were relatively consistent across groups: race/ethnicity (6%, non-White and/or Hispanic vs. 8%, non-Hispanic White), income (12%, below poverty line vs. 7%, above poverty line), sex at birth (9%, female vs. 7%, male).

Table 3: Median difference in auROC, 95% Bayesian credible interval (CI), and posterior probability that that the auROC difference was smaller than 0 for fairness contrasts for the no lag and 2-week lagged models.

Contrast	Median	Bayesian CI	Probability
Fairness Contrasts (No Lag)			
female vs. male	−0.043	[−0.059, −0.028]	1
non-White and/or Hispanic vs. non-Hispanic White	−0.131	[−0.222, −0.057]	0.999
below poverty line vs. above poverty line	−0.012	[−0.033, 0.007]	0.848
Fairness Contrasts (2-week Lag)			
female vs. male	−0.098	[−0.125, −0.073]	1
non-White and/or Hispanic vs. non-Hispanic White	−0.13	[−0.208, −0.058]	0.998
below poverty line vs. above poverty line	−0.039	[−0.073, −0.008]	0.98
Median auROC differences less than 0 indicate the model, on average, performed worse for the disadvantaged group (female, non-White and/or Hispanic, income below poverty line) compared to the advantaged group (male, non-Hispanic White, income above poverty line). Bayesian CI represents the range of values where there is a 95% probability that the true auROC difference lies within that range. Probability indicates the posterior probability that this difference is smaller than 0 (i.e., the models are performing differently for fairness subgroups).			

Model Calibration

The raw probabilities produced by our final models were not well calibrated. Consequently, we used Platt scaling to improve calibration. Platt scaling showed excellent improvement to the no lag model with a Brier score of .043. Calibration also improved probability accuracy for the 2-week lagged model with a Brier score of .063. For comparison, raw probability scores yielded Brier scores of .071 and .077 for the no lag and 2-week lagged models, respectively.

Figure 3 shows the calibration plots for the raw and calibrated probabilities for the no lag and 2-week lagged model. It also includes a histogram of raw probabilities that demonstrates our models produced variable predicted probabilities, spanning nearly the entire 0 - 1 range. Calibration plots and Brier scores for all 5 models are available in the supplement.

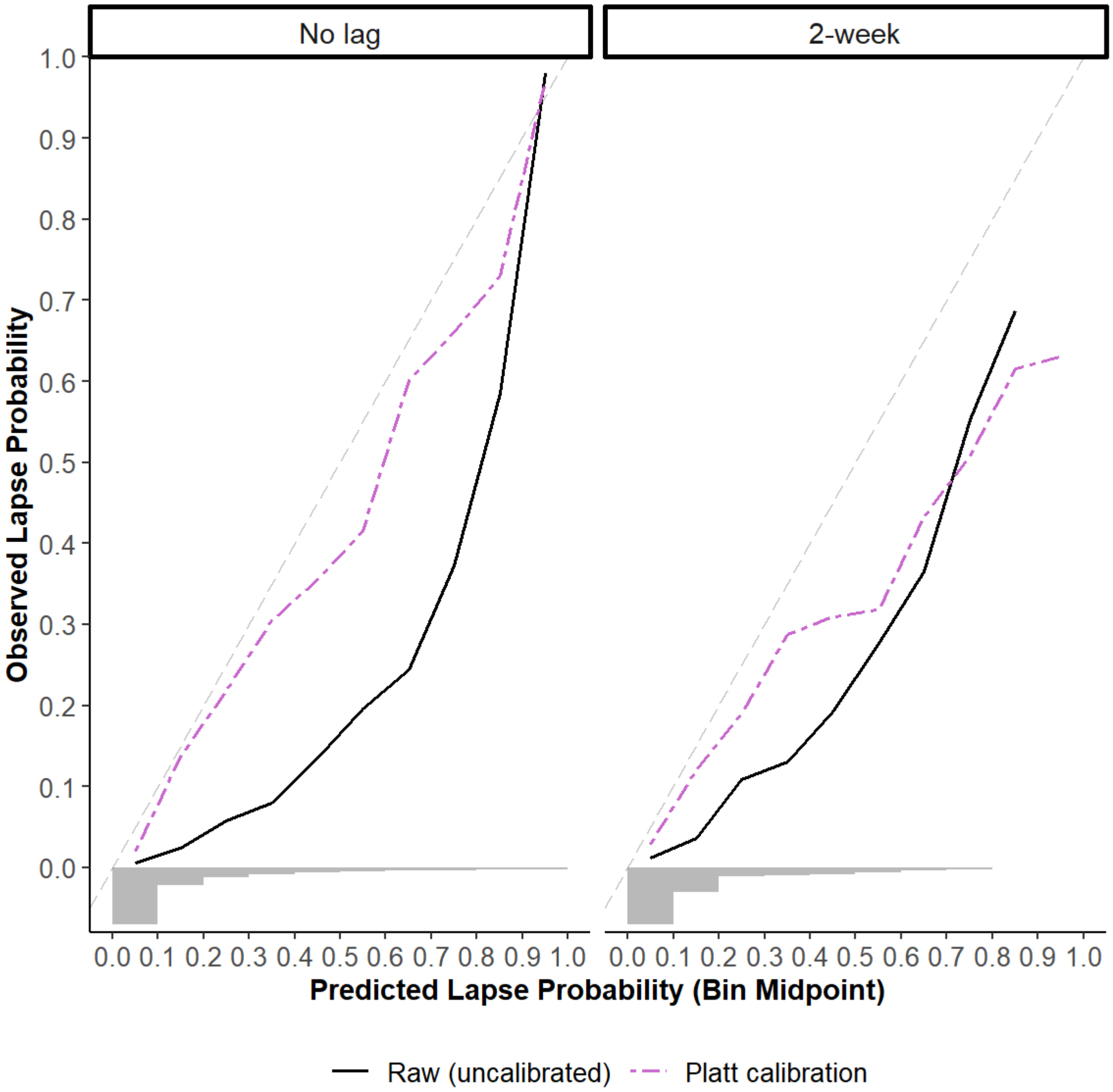

Figure 3: Calibration plots of raw and calibrated lapse probabilities for the baseline (no lag) and 2-week lagged models. Predicted probabilities (x-axis) are binned into deciles. Observed lapse probability (y-axis) represents the proportion of actual lapses observed in each bin. The dashed diagonal represents perfect calibration. Points below the line indicate overestimation and points above the line indicate underestimation. Raw probabilities are depicted as solid black curves. Platt calibrated probabilities are depicted as pink dashed curves. The grey histogram along the bottom of the plot represents the proportion of raw probabilities in each bin.

Feature Importance

Global feature importance is an indicator of how important a feature category was to the model's predictions, on average (i.e., across all participants and all observations). The top globally important feature category (i.e., highest mean |Shapley value|) for all models was past use. Future efficacy was a strong predictor for more immediate model predictions (i.e., no lag), but its importance diminished as lag time increased. On the other hand, as lag time increased, past/future risky situations increased in importance. Craving was consistently important in magnitude across all models. Figure 4 shows the relative ranking of feature categories for the no lag and 2-week lagged models. A plot of global feature importance for each feature category as a function of lag time is available in the supplement. These findings were also consistent across demographic subgroups (plots of global feature importance by demographic group are available for the no lag and 2-week lagged models in the supplement).

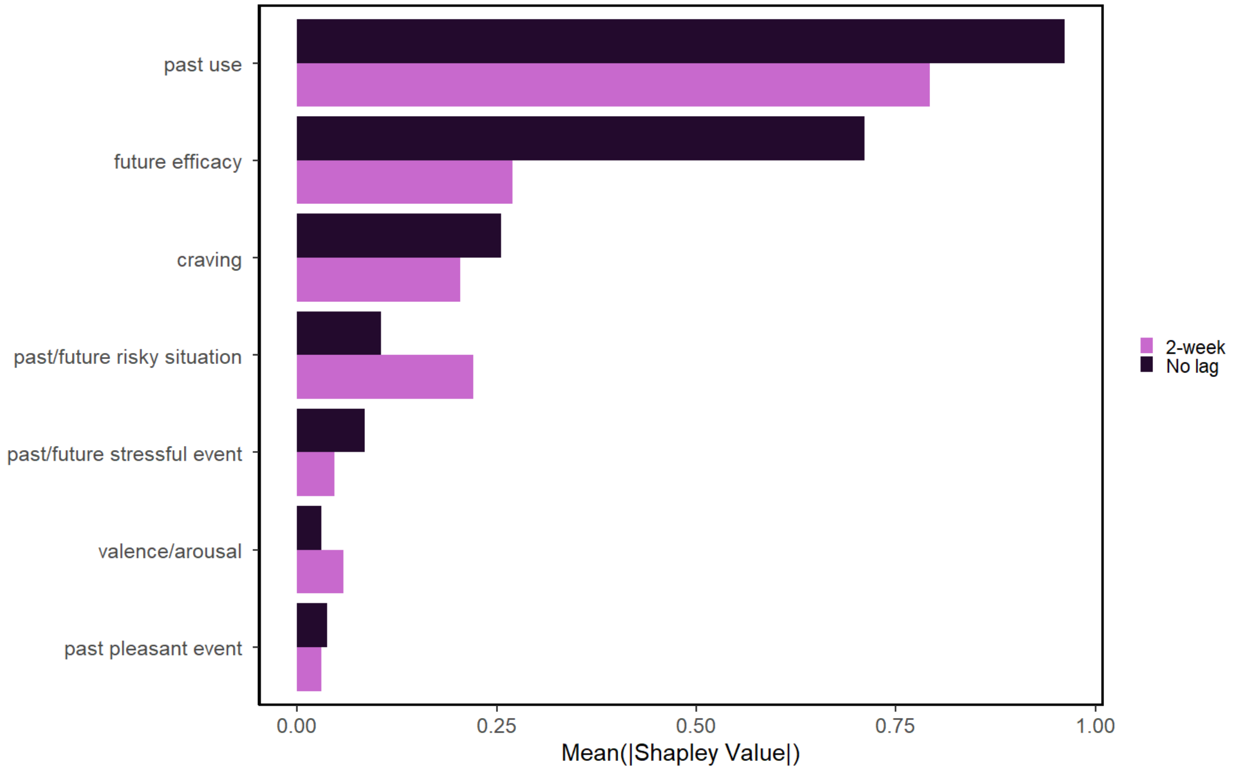

Figure 4: Global feature importance (mean |Shapley value|) for feature categories for the no lag and 2-week lagged models. Feature categories are ordered by their aggregate global importance. The importance of each feature category for each model is displayed separately by color.

Discussion

Model Evaluation and Comparisons

All the models that we evaluated performed exceptionally well. The no lag model, which predicts the probability of an immediate (i.e., within 24 hours) lapse back to alcohol use, had a .91 median posterior probability for auROC. Our 2-week lagged model, which made the most distal predictions, had a .85 median posterior probability for auROC, suggesting lagged models can be used to shift a 24-hour prediction window meaningfully into the future.

Across models (no lag, 1 day, 3 days, 1 week, and 2 weeks), model performance systematically decreased as models predicted further into the future. All lagged models had lower performance compared to the no lag baseline model and to the preceding adjacent lag model. This is unsurprising given what we know about prediction and substance use. Many important relapse risk factors are fluctuating processes that can first emerge and/or change day-by-day, if not more frequently. As lag time increases, features become less proximal to the start of the prediction window. Still, we wish to emphasize that our lowest auROC (.85) is still quite good, and the benefit of advanced notice (i.e., 2 weeks) likely outweighs the modest cost to model performance.

Collectively, these results suggest we can achieve clinically meaningful performance up to two weeks out. Our rigorous resampling methods (grouped, nested, k-fold cross-validation) make us confident that these are valid estimates of how our models would perform with new individuals. Furthermore, it should be noted that both the no lag and 2-week lagged models can be combined in a complementary fashion that allows both for highly accurate immediate lapse prediction and advanced warning about future lapse risk.

Model Fairness

In recent years, the machine learning field has begun to recognize the need to evaluate model fairness when algorithms are used to inform important decisions (e.g., healthcare services offered, eligibility for loans, early parole). Algorithms that perform favorably for only majority group members may exacerbate existing disparities in access to resources and important clinical outcomes [40]. In this study, we assessed model fairness by comparing model performance across important subgroups with known disparities in substance use treatment access and/or outcomes - race/ethnicity (non-White and/or Hispanic vs. non-Hispanic White), income (below poverty line vs. above poverty line), and sex at birth (female vs. male).

All models performed worse for people who were non-White and/or Hispanic, and for people who had an income below the poverty line. These findings should be interpreted cautiously given the lack of diversity in our training data. Participants of color were severely underrepresented in our training data (*N* = 20, 13%). Individuals below the poverty line were also underrepresented, though to a lesser degree (*N* = 49, 32%). Nevertheless, these findings are concerning and the fact that we do not have adequate sample sizes for these comparisons highlight a large limitation of our study.

An obvious solution to this problem involves intentional recruitment for diversity in training data when developing prediction models. For example, we are now working to increase the racial, ethnic, and income diversity of our training data for alcohol lapse prediction while simultaneously optimizing feedback from these models for implementation purposes [24]. In a separate project, we developed a national recruitment method that enabled us to recruit for racial, ethnic and income diversity while also focusing on much needed diversity across geographic location (e.g., rural vs. urban; [20]). We expect geographic diversity in the training data may also be crucial to develop fair models because the features that predict lapse in urban and suburban settings may differ from those that predict lapse in rural environments. If rural participants are not used to train models, the implementation of these models may compound existing disparities in SUD treatment in these communities [41,42].

Future research can also explore potential computational solutions to mitigate performance disparities that emerge when subgroups are poorly represented in available training data. For example, training data from under-represented subgroups could be up-sampled (e.g., using the synthetic minority oversampling technique), or the cost functions used by the learning algorithms could be adjusted to differentially weigh prediction errors based on participant characteristics. In another vein, modeling approaches that yield idiographic, person-specific models [43–46] may reduce performance disparities across subgroups. For example, we have begun to develop state space models whose parameters can be initialized with priors derived from existing training data but then adjusted over time to fit patterns present within a specific individual's time-series [47]. Such models may mitigate issues of unfairness to a large degree because they will weigh the individual's own data more heavily than group level estimates over time as more data accrue.

Of note, problems with model fairness can emerge even when subgroups are well-represented in the training data. Our models performed less well for women compared to men despite the fact that women were well-represented in the training data (*N* = 74, 49%). Instead, this differential performance may have resulted from more fundamental problems with the features available to the model. We chose our EMA items using domain expertise from decades of research on the factors that predict relapse. However, prior to the 1993 National Institute of Health Revitalization Act [48] that mandated the inclusion of minorities and women in research, women were mostly excluded from substance use treatment research due to their childbearing potential [49]. As a result, it is possible that our theories about the causes and contributors to relapse are biased toward constructs that are more relevant for men than women. If true, features derived from EMA items that tap these constructs would be expected to under-perform when predicting lapses for women. More research may be needed to identify relapse risk factors for women (e.g., interpersonal relationship problems [50], hormonal changes [51]), and other groups under-represented in the literature before we can fully address these performance disparities.

In the meantime, data-driven (bottom-up) approaches can be used to engineer high-dimensional feature sets that are not explicitly grounded in existing, and potentially biased, theories. For example, we have begun to explore the application of natural language processing techniques (e.g., LIWC; topic modeling; BERT [52–54]) to text messages and other social media activity by our participants to engineer features that may predict future lapses. Such features may or may not align with existing theories about relapse, but because they are anchored to participants' own words, they may serve as reliable indicators of lapse risk for certain individuals, particularly when used within learning algorithms that employ feature selection, regularization, or other techniques to address the bias-variance trade-off with high-dimensional feature sets. Furthermore, emerging techniques for interpreting machine learning models [55] can be applied to models that perform well to bootstrap the identification of new lapse risk constructs based on these novel features.

Beyond issues of training data representation and lacunae or outright biases in our theories, it is also true that historically marginalized groups that have experienced systemic racism, exclusion, or other stigma around substance use (e.g., societal expectations for women regarding attractiveness, cleanliness and motherhood [56]) may feel less trusting in disclosing substance use [57]. These experiences could prompt some individuals in these subgroups to under-report lapses and/or risk factors, which could also degrade performance and evaluation of our models for these subgroups. We observed relatively comparable percentages of lapses reported among disadvantaged compared to advantaged groups. However, comparable lapse rates do not necessarily confirm comparable reporting accuracy because it is possible that there were systematic differences in lapse rates across groups that were masked by issues of trust.

Model Characterization

Calibration

After applying Platt scaling to our predicted probabilities, our models were generally well calibrated with increasing monotonic relationships between calibrated model output and lapse event rates. Well-calibrated probabilities indicate that the predicted probability aligns closely with the true likelihood of an outcome (i.e., a lapse). Our no lag model had excellent calibration. However, the calibration plots suggest that with a longer lag time of 2 weeks, the model tends to over-predict the likelihood of lapses when predicted probabilities were higher.

This pattern may not necessarily be problematic. Research suggests that people often struggle to interpret probabilistic feedback, especially when it’s provided in raw numerical form [58–60]. As a result, it may be more effective to communicate risk using coarser categories (e.g., low, medium, or high risk) or through relative changes in risk (e.g., “Your risk of lapse is higher this week compared to last week"). These forms of feedback may be less sensitive to small miscalibrations at the extremes as long as the relationship between predicted probabilities and the observed event rate is monotonic.

Feature importance

The relative ordering of top global features remained somewhat consistent across the no lag and 2-week lagged models. Past use was the most important feature in both models in our dataset. This is not surprising given that our outcome was lapse, and past behavior is often the best predictor of future behavior. This finding also supports decades of clinical research on relapse prevention, where lapses (i.e., single instances of goal inconsistent alcohol use) are seen as powerful precursors to relapse (i.e., full return to harmful drinking; [9]). Abstinence self-efficacy emerged as the second most important feature in both models in our dataset, indicating that participants had reasonably accurate insight into their near-term success with maintaining alcohol abstinence. Craving was also an important predictor in both models, suggesting that it may be an important target for intervention to support early recovery efforts.

Several feature categories displayed sizeable differences in global importance by lag time. The importance of abstinence self-efficacy dropped by more than 50% in the 2-week lagged model relative to the no lag model. This may indicate that self-efficacy during early recovery is unstable even across shorter periods of time such that their current self-efficacy does not strongly predict abstinence success even two weeks into the future. In fact, craving and risky situations become as important as self-efficacy when predicting two-week lagged lapses. It may be that these other experiences are shaping and changing the individual's self-efficacy rapidly in early recovery. This also suggests that more frequent clinical assessments of self-efficacy as a target for intervention may be needed rather than assuming stability in this construct if initial assessment suggests it is high. Also, our study cannot determine if this differential importance of self-efficacy for immediate vs. lagged lapses persists beyond early recovery (where people may be encouraged to take a "one day at a time" mindset). Self-efficacy may become a more stable predictor of future abstinence success after longer periods of recovery, but our sample was limited to participants in early recovery (<= 8 weeks of abstinence at intake).

Past use was less important for the 2-week lagged model compared to the no lag model. This indicates that the predictive strength of a lapse on the likelihood of subsequent lapses diminishes to some degree over a relatively short period of time. This is good news and reinforces that single lapses do not always mark a return to consistent patterns of frequent, and potentially harmful, alcohol use. Despite this reduction in importance of past use as a predictor of lagged alcohol use, past use did remain the most important category for two-week lagged lapses. Lapses may provide "teachable moments" that can be used to reinforce recovery motivation, better understand risks, and develop skills to address those risks [11]. Conversely, lapses should not be ignored because they remain strong predictors of further use.

Surprisingly, past and anticipated risky situations were more important in the 2-week lagged vs. no lag model, suggesting that the impact of these situations on lapses back to use may be delayed. It may be that persistent exposure to risks is necessary to undermine an abstinence goal and lead to return to alcohol use. Alternatively or additionally, people may also be better able to anticipate future risky situations (e.g., vacations, anniversaries of significant dates) than future acute stressors or even future self-efficacy. Regardless, the increased importance of risky situations for predicting lagged lapses provides an opportunity to intervene prior to the lapse, particularly if the individual is encouraged to assess future risks and/or makes use of a recovery monitoring prediction model.

We were also surprised that stressful events, pleasant events, and affective state features did not make more important contributions to predictions across models. These constructs are highlighted in numerous theories about addiction and relapse [9,10,61] and represent targets for intervention in many existing treatments [18,62–64]. It may be that their impact is subsumed within other more powerful features (i.e., past use and self-efficacy). However, this seems unlikely given that the methodology underlying Shapley values allows for a fair distribution of importance among the relevant predictive features even when those features are correlated [55]. Alternatively, we may need to more carefully consider the nuanced roles that these constructs play (e.g., within the context of individual coping strategies, social support or environmental factors) in the return to alcohol use during recovery [65].

Additional Limitations and Future Directions

We believe our lapse prediction models will be most effective when embedded in an RMSS designed to deliver adaptive and personalized continuing care. This system could send daily, weekly, or less frequent messages to users with personalized feedback about their risk of lapse and provide support recommendations tailored to their current recovery needs. This study provides initial support that immediate and lagged prediction models can be trained to high accuracy using EMA for recovery monitoring. Furthermore, locally important features from these models can be used to identify the specific factors that contribute to each lapse risk prediction.

The no lag model can be used to guide individuals to take immediate, actionable steps to maintain their recovery goals and support them to implement these steps within the RMSS. For example, the RMSS can recommend an embedded urge surfing activity when someone's immediate risk is driven by strong craving whereas a guided relaxation video can be provided to the user when they report stressful events. Similarly, the RMSS can encourage (and explicitly support) the user to reflect on recent past successes and/or skills they have developed when their self-efficacy is low.

The 2-week lagged model provides individuals with advanced warning of their lapse risk. This model is well-suited to support recovery needs that cannot be addressed immediately within an RMSS app, such as scheduling positive or pleasant activities, increasing social engagement, or attending a peer-led recovery meeting. To be clear, we do not believe an RMSS app alone will be sufficient to deliver continuing care. We expect individuals will require additional support throughout their recovery from a mental health provider (e.g., motivational enhancement, crisis management, skill building), a peer (e.g., sponsor, support group), or family member. Importantly, these types of supports take time to set up, highlighting the value of this lagged 2-week model.

At this point, it is still unclear the best way to provide risk and support information from our models to people. For an RMSS to be successful, users must trust the system, consistently engage with the system over time, and find the system beneficial. We have recently launched an

NIAAA funded project to optimize daily support messages by examining the impact of several key message components (e.g., lapse probability, locally important features, a risk-relevant recovery activity recommendation, the linguistic style and tone of the message) on engagement, trust, clinical outcomes [24].

For a system using lagged models, we can imagine that lags longer than two weeks (i.e., more advanced warning) would be better still. In the present study, we could not train models with lags longer than two weeks because participants only provided lapse reports for up to three months. With the 2-week lagged model, we had approximately 17% fewer labeled observations for training because the first two weeks (out of 12 weeks) of labels for each participant were discarded. This data loss may be one factor that contributed to the decreases in model performance with increases in lag time and we believed that greater data loss (e.g., 25% for a 3-week lag) would not be tenable. We have recently completed data collection on a NIDA funded project where participants provided EMA and other sensed data for up to 12 months [20]. These data will allow us to train models with longer lags and to better evaluate the impact of data loss on model performance because lag time can be increased substantially with proportionally less data loss given 52 weeks of labeled observations per participant.

While the number of observations are large and support the complexity of a machine learning pipeline, we are limited by the number of participants. Successful machine learning models must generalize well to new data. Our analyses only speak to how well our models generalize to a mostly homogenous sample of 151 participants. We made the most efficient use of our sample (i.e., using nested cross-validation to maximize the amount of data used for both selection and evaluation); however, there is reason to worry about the generalizability to other populations, such as individuals living in rural locations and members of other diverse groups not adequately represented in our training data.

Additionally, our use of features from 4x daily EMA as model inputs may raise concerns about measurement burden. We confirmed that participants can comply with such EMA schedules over this time period and that they find it acceptable given its potential benefits to them [19,see also 66]. However, frequent daily surveys may become too burdensome within an RMSS intended for use over many, many months to years for long-term continuing care. We have begun to address this concern by training no lag models with fewer EMAs (1x daily) and have found comparable performance [47]. Additionally, reinforcement learning could potentially be used for adaptive EMA sampling. For example, each day the algorithm could make a decision to send out an EMA or not based on inferred latent states of the individual based on previous EMA responses and predicted probability of lapse.

We have also begun to explore how we can supplement our models with data from lower burden sensing methods. Geolocation, which can be passively sensed, could compliment EMA well [67]. First, it could provide insight into information not easily captured by self-report without lengthy surveys. For example, the amount of time spent in risky locations, or changes in routine (e.g., loss of job; move to new city) that could indicate life stressors can be detected in movement patterns. Second, the near-continuous sampling of geolocation could offer risk-relevant information that would otherwise be missed in between the discrete sampling periods of EMA. Furthermore, potentially powerful features can be engineered by combining geolocation data with contextual information available in public sources (e.g., census data, alcohol outlet density) [68,69] or collected from the user directly (e.g., self-evaluated riskiness of a given location) [20].

Conclusions

This study suggests it is possible to accurately predict alcohol lapses both immediately and up to two weeks into the future using lagged machine learning prediction models. The no lag model could guide users to engage with a smart RMSS that provides daily recovery activities that are personalized to their lapse risk and the factors contributing to that risk. The 2-week lagged model could enable patients to seek out and implement recovery support that is not immediately available to them within the RMSS. Several important steps remain prior to implementing the no lag and 2-week lagged models within a smart RMSS. Feedback and support messages from these models should be optimized to sustain system engagement, trust, and clinical outcomes. Passive sensing of model inputs may allow assessment of a broader range of risk factors with less burden for system users. Perhaps most important, model fairness must be improved by decreasing disparities in performance for less privileged groups. We remain optimistic about the potential to implement these models within a smart RMSS because these barriers, while challenging, are surmountable.

References

1. McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness: Implications for treatment, insurance, and outcomes evaluation. JAMA. 2000;284: 1689–1695. doi:10.1001/jama.284.13.1689
2. Dennis M, Scott CK. Managing Addiction as a Chronic Condition. Addiction Science & Clinical Practice. 2007;4: 45–55.
3. Substance Abuse and Mental Health Services Administration. 2023 NSDUH Detailed Tables CBHSQ Data. https://www.samhsa.gov/data/report/2023-nsduh-detailed-tables;
4. Hedegaard H, Miniño AM, Spencer MR, Warner M. Drug overdose deaths in the United States, 1999–2020. 2021.
5. Centers for Disease Control and Prevention (CDC). Annual Average for United States 2011–2015 Alcohol-Attributable Deaths Due to Excessive Alcohol Use, All Ages. 2022 Alcohol Related Disease Impact (ARDI) Application Website. https://nccd.cdc.gov/DPH_ARDI/Default/Default.aspx;
6. Wagner EH, Austin BT, Davis C, Hindmarsh M, Schaefer J, Bonomi A. Improving Chronic Illness Care: Translating Evidence Into Action. Health Affairs. 2001;20: 64–78. doi:10.1377/hlthaff.20.6.64
7. Stanojlović M, Davidson L. Targeting the Barriers in the Substance Use Disorder Continuum of Care With Peer Recovery Support. Substance Abuse: Research and Treatment. 2021;15: 1178221820976988. doi:10.1177/1178221820976988
8. Socías ME, Volkow N, Wood E. Adopting the "cascade of care" framework: An opportunity to close the implementation gap in addiction care? Addiction. 2016;111: 2079–2081. doi:10.1111/add.13479

9. Marlatt GA, Gordon JR, editors. Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors. First edition. New York: The Guilford Press; 1985.
10. Witkiewitz K, Marlatt GA. Relapse prevention for alcohol and drug problems: That was zen, this is tao. American Psychologist. 2004;59: 224–235. doi:10.1037/0003-066X.59.4.224
11. Witkiewitz K, Marlatt GA. Modeling the complexity of post-treatment drinking: It's a rocky road to relapse. Clinical Psychology Review. 2007;27: 724–738. doi:10.1016/j.cpr.2007.01.002
12. Brandon TH, Vidrine JI, Litvin EB. Relapse and relapse prevention. Annual Review of Clinical Psychology. 2007;3: 257–284. doi:10.1146/annurev.clinpsy.3.022806.091455
13. Bickman L, Lyon AR, Wolpert M. Achieving Precision Mental Health through Effective Assessment, Monitoring, and Feedback Processes. Administration and Policy in Mental Health and Mental Health Services Research. 2016;43: 271–276. doi:10.1007/s10488-016-0718-5
14. DeRubeis RJ. The history, current status, and possible future of precision mental health. Behaviour Research and Therapy. 2019;123: 103506. doi:10.1016/j.brat.2019.103506
15. Kranzler HR, McKay JR. Personalized Treatment of Alcohol Dependence. Current Psychiatry Reports. 2012;14: 486–493. doi:10.1007/s11920-012-0296-5
16. Mohr DC, Zhang M, Schueller SM. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annual Review of Clinical Psychology. 2017;13: 23–47. doi:10.1146/annurev-clinpsy-032816-044949
17. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. New York, NY: Springer; 2009.

18. Bowen S, Chawla N, Grow J, Marlatt GA. Mindfulness-Based Relapse Prevention for Addictive Behaviors: A Clinician’s Guide. Second edition. New York: The Guilford Press; 2021.
19. Wyant K, Moshontz H, Ward SB, Fronk GE, Curtin JJ. Acceptability of Personal Sensing Among People With Alcohol Use Disorder: Observational Study. JMIR mHealth and uHealth. 2023;11: e41833. doi:10.2196/41833
20. Moshontz H, Colmenares AJ, Fronk GE, Sant’Ana SJ, Wyant K, Wanta SE, et al. Prospective Prediction of Lapses in Opioid Use Disorder: Protocol for a Personal Sensing Study. JMIR Research Protocols. 2021;10: e29563. doi:10.2196/29563
21. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. pp. 4768–4777.
22. Soyster PD, Ashlock L, Fisher AJ. Pooled and person-specific machine learning models for predicting future alcohol consumption, craving, and wanting to drink: A demonstration of parallel utility. Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors. 2022;36: 296–306. doi:10.1037/adb0000666
23. Walters ST, Businelle MS, Suchting R, Li X, Hébert ET, Mun E-Y. Using machine learning to identify predictors of imminent drinking and create tailored messages for at-risk drinkers experiencing homelessness. Journal of Substance Abuse Treatment. 2021;127: 108417. doi:10.1016/j.jsat.2021.108417

24. Wyant K, Sant'Ana SJ, Punturieri CE, Yu J, Fronk GE, Maggard CM, et al. Maximizing engagement, trust, and clinical benefit of AI-generated recovery monitoring and support messages for alcohol use disorder: Protocol for an optimization study. under review.
25. Pinedo M. A current re-examination of racial/ethnic disparities in the use of substance abuse treatment: Do disparities persist? Drug and Alcohol Dependence. 2019;202: 162–167. doi:10.1016/j.drugalcdep.2019.05.017
26. Kilaru AS, Xiong A, Lowenstein M, Meisel ZF, Perrone J, Khatri U, et al. Incidence of Treatment for Opioid Use Disorder Following Nonfatal Overdose in Commercially Insured Patients. JAMA Network Open. 2020;3: e205852. doi:10.1001/jamanetworkopen.2020.5852
27. Olfson M, Mauro C, Wall MM, Choi CJ, Barry CL, Mojtabai R. Healthcare coverage and service access for low-income adults with substance use disorders. Journal of Substance Abuse Treatment. 2022;137: 108710. doi:10.1016/j.jsat.2021.108710
28. Greenfield SF, Brooks AJ, Gordon SM, Green CA, Kropp F, McHugh RK, et al. Substance abuse treatment entry, retention, and outcome in women: A review of the literature. Drug and Alcohol Dependence. 2007;86: 1–21. doi:10.1016/j.drugalcdep.2006.05.012
29. Hsieh F. Sample size tables for logistic regression. Statistics in Medicine. 1989;8: 795–802.
30. Derogatis, L.R. Brief Symptom Inventory 18 - Administration, scoring, and procedures manual. Minneapolis: NCS Pearson; 2000.
31. WHO ASSIST Working Group. The Alcohol, Smoking and Substance Involvement Screening Test (ASSIST): Development, reliability and feasibility. Addiction (Abingdon, England). 2002;97: 1183–1194.
32. Kuhn M, Wickham H. Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. 2020.

33. Kuhn M. Tidyposterior: Bayesian Analysis to Compare Models using Resampling Statistics. 2022.
34. Goodrich B, Gabry J, Ali I, Brilleman S. Rstanarm: Bayesian Applied Regression Modeling via Stan. 2023.
35. Center for High Throughput Computing. Center for high throughput computing. Center for High Throughput Computing; 2006. doi:10.21231/GNT1-HW21
36. Kuhn M, Johnson K. Applied Predictive Modeling. 1st ed. 2013, Corr. 2nd printing 2018 edition. New York: Springer; 2018. doi:10.1007/978-1-4614-6849-3
37. RStudio Team. RStudio: Integrated Development for R. Boston, MA: RStudio, Inc; 2020.
38. Gabry J, Goodrich B. Prior Distributions for rstanarm Models. CRAN R-Project. https://cran.r-project.org/web/packages/rstanarm/vignettes/priors.html; 2023.
39. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. MIT Press; 1999. pp. 61–74.
40. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: How informatics interventions can worsen inequality. Journal of the American Medical Informatics Association: JAMIA. 2018;25: 1080–1088. doi:10.1093/jamia/ocy052
41. Lee JH, Wheeler DC, Zimmerman EB, Hines AL, Chapman DA. Urban–Rural Disparities in Deaths of Despair: A County-Level Analysis 2004–2016 in the U.S. American journal of preventive medicine. 2023;64: 149–156. doi:10.1016/j.amepre.2022.08.022
42. Lister JJ, Weaver A, Ellis JD, Himle JA, Ledgerwood DM. A systematic review of rural-specific barriers to medication treatment for opioid use disorder in the United States. The American Journal of Drug and Alcohol Abuse. 2020;46: 273–288. doi:10.1080/00952990.2019.1694536

43. Fisher AJ. Toward a dynamic model of psychological assessment: Implications for personalized care. Journal of Consulting and Clinical Psychology. 2015;83: 825–836. doi:10.1037/ccp0000026
44. David SJ, Marshall AJ, Evanovich EK, Mumma GH. Intraindividual Dynamic Network Analysis - Implications for Clinical Assessment. Journal of Psychopathology and Behavioral Assessment. 2018;40: 235–248. doi:10.1007/s10862-017-9632-8
45. Roche MJ, Pincus AL, Rebar AL, Conroy DE, Ram N. Enriching Psychological Assessment Using a Person-Specific Analysis of Interpersonal Processes in Daily Life. Assessment. 2014;21: 515–528. doi:10.1177/1073191114540320
46. Wright AGC, Hallquist MN, Stepp SD, Scott LN, Beeney JE, Lazarus SA, et al. Modeling Heterogeneity in Momentary Interpersonal and Affective Dynamic Processes in Borderline Personality Disorder. Assessment. 2016;23: 484–495. doi:10.1177/1073191116653829
47. Pulick E, Curtin J, Mintz Y. Idiographic Lapse Prediction With State Space Modeling: Algorithm Development and Validation Study. JMIR Formative Research. 2025;9: e73265. doi:10.2196/73265
48. Studies I of M(US)C on E and LIR to the I of W in C, Mastroianni AC, Faden R, Federman D. NIH Revitalization Act of 1993 Public Law 103-43. Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies: Volume I. National Academies Press (US); 1994.
49. Vannicelli M, Nash L. Effect of Sex Bias on Women's Studies on Alcoholism. Alcoholism: Clinical and Experimental Research. 1984;8: 334–336. doi:10.1111/j.1530-0277.1984.tb05523.x

50. Walitzer KS, Dearing RL. Gender differences in alcohol and substance use relapse. Clinical Psychology Review. 2006;26: 128–148. doi:10.1016/j.cpr.2005.11.003
51. McHugh RK, Votaw VR, Sugarman DE, Greenfield SF. Sex and gender differences in substance use disorders. Clinical Psychology Review. 2018;66: 12–23. doi:10.1016/j.cpr.2017.10.012
52. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology. 2010;29: 24–54. doi:10.1177/0261927X09351676
53. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3: 993–1022.
54. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. doi:10.48550/arXiv.1810.04805
55. Molnar C. Interpretable Machine Learning: A Guide For Making Black Box Models Explainable. Munich, Germany: Independently published; 2022.
56. Meyers SA, Earnshaw VA, D'Ambrosio B, Courchesne N, Werb D, Smith LR. The intersection of gender and drug use-related stigma: A mixed methods systematic review and synthesis of the literature. Drug and Alcohol Dependence. 2021;223: 108706. doi:10.1016/j.drugalcdep.2021.108706
57. Marwick AE, Boyd D. Privacy at the Margins Understanding Privacy at the Margins—Introduction. International Journal of Communication. 2018;12: 9.
58. Zikmund-Fisher BJ. The right tool is what they need, not what we have: A taxonomy of appropriate levels of precision in patient risk communication. Medical care research and review: MCRR. 2013;70: 37S–49S. doi:10.1177/1077558712458541

59. Fagerlin A, Ubel PA, Smith DM, Zikmund-Fisher BJ. Making numbers matter: Present and future research in risk communication. American Journal of Health Behavior. 2007;31 Suppl 1: S47–56. doi:10.5555/ajhb.2007.31.supp.S47

60. Zipkin DA, Umscheid CA, Keating NL, Allen E, Aung K, Beyth R, et al. Evidence-based risk communication: A systematic review. Annals of Internal Medicine. 2014;161: 270–280. doi:10.7326/M14-0295

61. Rawson RA, Shoptaw SJ, Obert JL, McCann MJ, Hasson AL, Marinelli-Casey PJ, et al. An intensive outpatient approach for cocaine abuse treatment. The Matrix model. Journal of Substance Abuse Treatment. 1995;12: 117–127. doi:10.1016/0740-5472(94)00080-b

62. McHugh RK, Hearon BA, Otto MW. Cognitive-Behavioral Therapy for Substance Use Disorders. The Psychiatric clinics of North America. 2010;33: 511–525. doi:10.1016/j.psc.2010.04.012

63. Liese BS, Beck AT. Cognitive-Behavioral Therapy of Addictive Disorders. First edition. New York: The Guilford Press; 2022.

64. Center for Substance Abuse Treatment. Counselor's Treatment Manual: Matrix Intensive Outpatient Treatment for People With Stimulant Use Disorders. Rockville, MD: Substance Abuse and Mental Health Services Administration; 2006.

65. Fronk GE, Sant'Ana SJ, Kaye JT, Curtin JJ. Stress Allostasis in Substance Use Disorders: Promise, Progress, and Emerging Priorities in Clinical Research. Annual Review of Clinical Psychology. 2020;16: 401–430. doi:10.1146/annurev-clinpsy-102419-125016

66. Jones A, Remmerswaal D, Verveer I, Robinson E, Franken IHA, Wen CKF, et al. Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. Addiction (Abingdon, England). 2019;114: 609–619. doi:10/gfsjzg

67. Bae SW, Suffoletto B, Zhang T, Chung T, Ozolcer M, Islam MR, et al. Leveraging Mobile Phone Sensors, Machine Learning and Explainable Artificial Intelligence to Predict Imminent Same-Day Binge Drinking Events to Support Just-In-Time Adaptive Interventions: A Feasibility Study. JMIR formative research. 2023. doi:10.2196/39862
68. Huang L, Li Q, Yue Y. Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks. New York, NY, USA: Association for Computing Machinery; 2010. pp. 27–30. doi:10.1145/1867699.1867704
69. Xie K, Deng K, Zhou X. From trajectories to activities: A spatio-temporal join approach. Proceedings of the 2009 International Workshop on Location Based Social Networks. New York, NY, USA: Association for Computing Machinery; 2009. pp. 25–32. doi:10.1145/1629890.1629897