Machine learning models for temporally precise lapse prediction in alcohol use disorder

Kendra Wyant*[1], Sarah J. Sant'Ana*[1], Gaylen E. Fronk[1], and John J. Curtin [1]

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

Author Note

*These authors contributed equally as co-first authors.

All data and materials have been made publicly available and can be accessed at https://osf.io/w5h9y/. All procedures were approved by the University of Wisconsin-Madison Institutional Review Board (Study # 2015-0780). This research was supported by grants from the NIAAA (R01 AA024391; JJC) and the NIDA (R01 DA047315; JJC). The authors wish to thank Susan E. Wanta for her role as the project administrator and for her help with data curation and Xiaojin (Jerry) Zhu for his contributions to project conceptualization and analyses. The authors also wish to thank Candace Lightheart, Jill Nagler, Kerry Keiser, and Megan Schultz for their contributions to data collection and Chris Gioia for the clinical supervision he provided to graduate students.

Correspondence concerning this article should be addressed to John J. Curtin, Department of Psychology, University of Wisconsin-Madison, 1202 W Johnson St, Madison, WI 53521, Email: jjcurtin@wisc.edu

Abstract

We developed three machine learning models that predict hour-by-hour probabilities of a future lapse back to alcohol use with increasing temporal precision (i.e., lapses in the next week, next day, and next hour). Model features were based on raw scores and longitudinal change in theoretically implicated risk factors collected through ecological momentary assessment (EMA). Participants (N=151; 51% male; mean age = 41; 87% White, 97% Non-Hispanic) in early recovery (1–8 weeks of abstinence) from alcohol use disorder provided 4x daily EMA for up to three months. We used grouped, nested cross-validation to select best models and evaluate the performance of those best models. Models yielded median areas under the receiver operating curves (auROCs) of .90, .91, and .94 in the 30 held-out test sets for week, day, and hour level models, respectively. Some feature categories consistently emerged as being globally important to lapse prediction across our week, day, and hour level models (i.e., past use, future self-efficacy). However, most of the more punctuate, time varying constructs (e.g., craving, past stressful events, arousal) appear to have greater impact within the next hour prediction model. This research represents an important step toward the development of a *smart* (machine learning guided) sensing system that can both identify periods of peak lapse risk and recommend specific supports to address factors contributing to this risk.

General scientific summary: This study suggests that densely sampled self-report data can be used to predict lapses back to alcohol use with varying degrees of temporal precision. Additionally, the contextual features contributing to risk of lapse may offer important insight for treatment matching through a digital therapeutic.

Keywords: ecological momentary assessment, digital therapeutics, alcohol use disorder

Machine learning models for temporally precise lapse prediction in alcohol use disorder

Introduction

Over 30 million adults in the United States (US) had an active alcohol use disorder (AUD) in 2021, and 23.3% reported engaging in past-month binge drinking (SAMHSA Center for Behavioral Health Statistics and Quality, 2021). Alcohol ranks as the third leading preventable cause of death in the US, accounting for approximately 140,000 fatalities (Centers for Disease Control and Prevention (CDC), n.d.) and economic costs that exceed $249 billion annually (Substance Abuse and Mental Health Services Administration (US) & Office of the Surgeon General (US), 2016).

Existing clinician-delivered treatments for AUD that were derived from Marlatt's relapse prevention model (Marlatt & Gordon, 1985) are effective when delivered (e.g., cognitive-behavioral therapy, mindfulness-based relapse prevention (Bowen et al., 2014)). Unfortunately, fewer than 1 in 20 adults with an active AUD receive any treatment (SAMHSA Center for Behavioral Health Statistics and Quality, 2021). Even more concerning, failure to access treatment is associated with demographic factors including race, ethnicity, geographic region, and socioeconomic status, which further increase mental health disparities (Office of the Surgeon General (US) et al., 2001). This treatment gap and associated disparities stem from well-known barriers to receiving clinician-delivered mental healthcare related to affordability, accessibility, availability, and acceptability (Jacobson et al., 2022).

Digital therapeutics may help to overcome these barriers associated with in-person, clinician-delivered treatments. Digital therapeutics provide evidence-based interventions and other supports via smartphones to prevent, treat, or manage a medical disorder, either independently or in conjunction with traditional treatments (Jacobson et al., 2022). They offer highly scalable, on-demand therapeutic support that is accessible whenever and wherever it is needed most. Several large, randomized controlled trials have confirmed that

digital therapeutics for AUD improve clinical outcomes (Campbell et al., 2014; Gustafson et al., 2014; Jacobson et al., 2022). Additionally, US adults (including patients with AUD (Wyant et al., 2023)) display high rates of smartphone ownership (over 85% in 2021), with minimal variation across race, ethnicity, socioeconomic status, and geographic settings (Center, 2021). Therefore, digital therapeutics may not only mitigate in-person treatment barriers but also combat associated disparities (Jacobson et al., 2022).

Improving Digital Therapeutics via Personal Sensing

Despite the documented benefits of digital therapeutics, their full potential has not yet been realized. Patients often don't engage with digital therapeutics as developers intended, and long-term engagement may not be sustained or matched to patients' needs (Hatch et al., 2018; Jacobson et al., 2022). The substantial benefits of digital therapeutics come from easy, 24/7 access to their intervention and other support modules. However, the burden falls primarily on the patient to identify the most appropriate modules for them in that specific moment during their recovery.

This difficulty is magnified by the dynamic, chronic, and relapsing nature of AUD (Brandon et al., 2007). Numerous risk and protective factors interact in complex, non-linear ways to influence the probability, timing, and severity of relapse (i.e., a goal-inconsistent return to frequent, harmful alcohol use) (Witkiewitz & Marlatt, 2007). Factors such as urges, mood, lifestyle imbalances, self-efficacy, and motivation can all vary over time. Social networks may evolve to become more protective or risky, and high-risk situations can arise unexpectedly. Consequently, both relapse risk and the factors driving that risk fluctuate over time.

Successful, continuous monitoring of risk for relapse and its contributing factors would enable patients to adapt their lifestyle, behaviors, and supports to their changing needs. Successful monitoring could also direct patients to engage with the most appropriate digital therapeutic modules, addressing the unique risks present at any given moment throughout their recovery. Such continuous monitoring is now feasible via personal

sensing (i.e., in-situ data collection via sensors embedded in individuals' day to day lives) (Bae et al., 2018; Chih et al., 2014; Epstein et al., 2020; Moshontz et al., 2021; Soyster et al., 2022; Wyant et al., 2023).

The current project focuses explicitly on using ecological momentary assessment (EMA) for monitoring risk of return to alcohol use. EMA can be easily implemented with only a smartphone. Moreover, comparable item responses can be collected consistently across different hardware and operating systems. Thus, EMA can be incorporated essentially identically into any existing or future smartphone-based digital therapeutic. EMA, like other personal sensing methods, can support the frequent, in-situ, longitudinal measurement necessary for monitoring fluctuating relapse risk. Long-term monitoring with EMA has been well-tolerated by individuals with AUD (Wyant et al., 2023). Additionally, previous research has validated the use of EMA to measure known risk and protective factors for relapse, including craving (Dulin & Gonzalez, 2017), mood (Russell et al., 2020), stressors (Wemm et al., 2019), positive life events (Dvorak et al., 2018), and motivation/efficacy (Dvorak et al., 2014). EMA offers privileged access into these and other subjective factors that may be difficult to quantify reliably through other sensing methods.

Promising Preliminary Research

Preliminary research is now emerging that uses EMA responses as features in machine learning models to predict the probability of future alcohol use (Bae et al., 2018; Chih et al., 2014; Soyster et al., 2022; Walters et al., 2021). This research is important because it rigorously required strict temporal ordering necessary for true prediction, with features measured before alcohol use outcomes. It also used resampling methods (e.g., cross-validation) that prioritize model generalizability to increase the likelihood these models will perform well with new patients.

Despite this initial promise, several important limitations exist. Some prediction models have been trained using convenience samples (e.g., college students) (Bae et al., 2018; Soyster et al., 2022). Other models have been developed to predict hazardous alcohol

use in non-treatment-seeking populations (Walters et al., 2021). In both these instances, features that predict planned or otherwise intentional alcohol use among individuals not motivated to change their behavior may not generalize to patients in AUD recovery. Moreover, individuals who have not yet begun to contemplate and/or commit to behavior change regarding their alcohol use are unlikely to use digital therapeutics (Prochaska et al., 1992).

A handful of other models have been trained to predict putative precursors of substance use, such as craving (Burgess-Hull et al., 2022; Dumortier et al., 2016) and stress (Epstein et al., 2020). Although craving and stress may be associated with substance use, their relationships with relapse are complex, inconsistent, and not always very strong (Fronk et al., 2020; Sayette, 2016), making these constructs less than ideal as prediction targets.

Models that predict lapses (i.e., single instances of goal-inconsistent alcohol use) may be preferred. Lapses are clearly defined, observable, and have temporally precise onsets and offsets. Conversely, definitions of relapse vary widely (Witkiewitz & Marlatt, 2007), and it is difficult to delineate precisely when relapse begins or ends. Lapses always precede relapse and therefore may serve as an early warning sign for intervention. Finally, maladaptive responses to a lapse (e.g., abstinence violation effects; (Marlatt & Gordon, 1985)) can undermine recovery by themselves, making lapses clinically meaningful events to detect and address.

An early lapse prediction model developed by Gustafson and colleagues (Chih et al., 2014) provided the foundation on which our current project builds. Participants completed EMAs once per week for 8 months while using a digital therapeutic after discharge from an inpatient treatment program for AUD. These EMAs were used as features in a machine learning model to predict alcohol use lapses. However, the temporal precision for both the features and outcome was coarse. Model predictions were updated only once per week at most, and lapse onsets could occur anytime within the next two weeks. This coarseness

restricts the model from being used to implement *just-in-time* interventions (e.g., guided mindfulness or other stress reduction techniques, urge surfing) that are well-suited to digital therapeutics.

The Current Study

The current study addresses these limitations of previously developed prediction models. We trained our models using participants in early recovery from moderate to severe AUD who reported a goal of alcohol abstinence. We developed three separate models that provide hour-by-hour probabilities of a future lapse back to alcohol use with increasing temporal precision: lapses in the next week, next day, and next hour. Model features were engineered from raw scores and longitudinal change in responses to 4X daily EMAs. These features were derived to measure theoretically-implicated risk factors and contexts (Marlatt & Gordon, 1985) including past use, craving, past pleasant events, past and future risky situations, past and future stressful events, emotional valence and arousal, and self-efficacy. This research represents an important step toward the development of a "smart" (machine learning guided) sensing and prediction system that can be embedded within a digital therapeutic both to identify periods of peak lapse risk and to recommend specific supports to address factors contributing to this risk.

Method

Transparency and Openness

We adhere to research transparency principles that are crucial for robust and replicable science. We reported how we determined the sample size, all data exclusions, all manipulations, and all study measures. We provide a transparency report in the supplement. Finally, we made the data, analysis scripts, annotated results, questionnaires, and other study materials publicly available (https://osf.io/w5h9y/).

Our study design and analyses were not pre-registered. However, we restricted many researcher degrees of freedom via cross-validation. Cross-validation inherently includes replication; models are fit on held-in sets, decisions are made in held-out

validation sets, and final performance is evaluated on held-out test sets.

Participants

We recruited 151 participants in early recovery (1-8 weeks of abstinence) from AUD in Madison, Wisconsin, US. This sample size was determined based on traditional power analysis methods for logistic regression (Hsieh, 1989) 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 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 at least 1 week but no longer than 2 months, and
5. were willing to use a single smartphone (personal or study provided) while enrolled in the study.

We also excluded participants exhibiting severe symptoms of psychosis or paranoia.

Procedure

Participants completed five study visits over approximately three months. After an initial phone screen, participants attended an in-person screening visit for eligibility determination, informed consent, and collection of self-report measures. Eligible and 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 while on study. 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).

Measures

EMA

Participants completed four brief (7-10 questions) EMAs daily following text message reminders. All EMAs included seven items that asked about any past alcohol use; current affective state (valence and arousal); craving; and past stressful events, risky situations, and pleasant events. The first EMA each day included three additional questions about the likelihood of future risky situations, stressful events, and drinking alcohol in the upcoming week (i.e., future efficacy).

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 the participants' typical day, with at least one hour between EMAs.

Individual Differences

We collected self-report information about demographics (age, sex, race, ethnicity, education, employment, income, and marital status) and clinical characteristics (AUD milestones, number of quit attempts, lifetime AUD treatment history, lifetime receipt of AUD medication, DSM-5 AUD symptom count, and current drug use (WHO ASSIST Working Group, 2002)). Only age, sex, race, education, and marital status were used as model features.

Data Analytic Strategy

Data preprocessing, modeling, and Bayesian analyses were done in R using the tidymodels ecosystem (Kuhn & Wickham, 2020). All models were trained and evaluated using high-throughput computing resources provided by the University of Wisconsin Center for High Throughput Computing (Center for High Throughput Computing, 2006).

Lapse Labels

We predicted future lapses in three window widths that varied in their temporal precision: one week, one day, and one hour. Prediction windows were updated hourly. All classification models provide hour-by-hour predictions of future lapse probability for all three window widths.

We labeled each prediction window as *lapse* or *no lapse* using the EMA item "Have you drank any alcohol that you have not yet reported?". If participants answered yes to this question, they were prompted to enter the hour and date of the start and end of the drinking episode. Their responses were validated by study staff during monthly follow-up visits.

For more detail on the creation of our prediction windows, see Lapse Labels in the Supplemental Methods section of our Supplement.

Feature Engineering

Features were calculated using only data collected prior to the start of each prediction window. This ensured our models were making true *future predictions* versus identifying concurrent associations.

Features were derived from three sources: baseline demographic characteristics (i.e., age, sex, race, marital status, education); day of the week and the time of day (daytime vs. evening/night) of the start of the prediction window; and previous EMA responses. We scored raw min, max, median, and count features from EMA items within varying lead up times (6, 12, 24, 48, 72, and 168 hours prior to start of prediction window). We scored change EMA response features by subtracting the mean response for each feature over all data prior to the start of the prediction window from the associated raw feature.

For more detail on feature engineering steps see Feature Engineering in the Supplemental Methods section of our Supplement. We also made a sample feature engineering script (i.e., tidymodels recipe) available on our study's OSF page.

Model Training and Evaluation

Statistical Algorithm and Hyperparameters. We trained and evaluated three separate classification models: one each for week, day, and hour prediction windows. We initially considered four well-established statistical algorithms (XGBoost, Random Forest, K-Nearest Neighbors, and Elastic Net) that vary across characteristics expected to affect model performance (e.g., flexibility, complexity, and ability to handle higher-order interactions natively) (Kuhn & Johnson, 2018). However, preliminary exploratory analyses suggested that XGBoost consistently outperformed the other three algorithms. Furthermore, the Shapley Additive Explanations (SHAP) method, which we planned to use for explanatory analyses of feature importance, is optimized for XGBoost. For these reasons, we focused our primary model training and evaluation on the XGBoost algorithm only.

Candidate XGBoost model configurations differed across sensible values for the hyperparameters mtry, tree depth, and learning rate using grid search. All configurations used 500 trees with early stopping to prevent over-fitting. All other hyperparameters were set to defaults established by the tidymodels packages. Candidate model configurations also differed on outcome resampling method (i.e., up-sampling and down-sampling of the outcome using majority/no lapse to minority/lapse ratios ranging from 1:1 to 5:1). We calibrated predicted probabilities using the beta distribution to support optimal decision-making under variable outcome distributions (Kull et al., 2017).

Performance Metric. Our primary performance metric for model selection and evaluation was area under the Receiver Operating Characteristic Curve (auROC) (Kuhn & Johnson, 2018). auROC indexes the probability that the model will predict a higher score for a randomly selected positive case (i.e., lapse) relative to a randomly selected negative case (i.e., no lapse). This metric was selected because it 1) combines sensitivity and specificity, which are both important characteristics to consider for clinical implementation; 2) is an aggregate metric across all decision thresholds, which is important because optimal

decision thresholds may differ across settings and goals; and 3) is unaffected by class imbalance, which is important for comparing models with differing window widths and levels of class imbalance.

Cross-validation. We used participant-grouped, nested cross-validation for model training, selection, and evaluation with auROC. 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.

Nested cross-validation uses two nested loops for dividing and holding out folds: an outer loop, where held-out folds serve as *test sets* for model evaluation; and inner loops, where held-out folds serve as *validation sets* for model selection. Importantly, these sets are independent, maintaining separation between data used to train the models, select best models, and evaluate those best models. Therefore, nested cross-validation removes optimization bias from the evaluation of model performance in the test sets and can yield lower variance performance estimates than single test set approaches (Jonathan et al., 2000).

We used 1 repeat of 10-fold cross-validation for the inner loops and 3 repeats of 10-fold cross-validation for the outer loop. Best model configurations were selected based on the median auROC across the 10 *validation sets*. Final performance evaluation of those best model configurations was based on the median auROC across the 30 *test sets*. For completeness, we report median auROC for our best model configurations for each model (week, day, and hour) separately from both the validation and test sets. In addition, we report other key performance metrics for the best model configurations including sensitivity, specificity, balanced accuracy, positive predictive value (PPV), and negative predictive value (NPV) from the test sets (Kuhn & Johnson, 2018).

Bayesian Estimation of auROC and Model Comparisons

We used a Bayesian hierarchical generalized linear model to estimate the posterior probability distributions and 95% Bayesian confidence intervals (CIs) for auROC for the

three best models (i.e., week, day, and hour). To determine the probability that these models' performance differed systematically from each other, we regressed the auROCs (logit transformed) from the 30 test sets for each model as a function of window width. Following recommendations from the tidymodels team (Kuhn, n.d., 2022), we set two random intercepts: one for the repeat, and another for the fold within repeat (folds are nested within repeats for auROCs collected with 3x10-fold cross-validation). We report the 95% (equal-tailed) Bayesian CIs from the posterior probability distributions for our models' auROCs. We also report 95% (equal-tailed) Bayesian CIs for the differences in performance among the three models. For more detail on these analyses see Bayesian Analyses in Supplemental Methods section of the Supplement.

Shapley Additive Explanations for Feature Importance

We computed Shapley Values (Lundberg & Lee, 2017) to provide a consistent and objective explanation of the importance of categories of features (based on EMA items) across our three models. Shapley values are model-agnostic and possess several useful properties including: Additivity (Shapley values for each feature can be computed independently and summed); Efficiency (the sum of Shapley values across features must add up to the difference between predicted and observed outcomes for each observation); Symmetry (Shapley values for two features should be equal if the two features contribute equally to all possible coalitions); and Dummy (a feature that does not change the predicted value in any coalition will have a Shapley value of 0). We calculated Shapley values from the 30 test sets using the SHAPforxgboost package that provides Shapley values in log-odds units for binary classification models. We averaged the three Shapley values for each observation for each feature across the three repeats to increase their stability. To calculate the local (i.e., for each observation) impact of categories of features (e.g., all features associated with the EMA craving item), we added Shapley values across all features in a category, separately for each observation. To calculate global importance for categories of features, we averaged the absolute value of the Shapley values of all

features in the category across all observations.

Results

Demographic and Clinical Characteristics

One hundred ninety-two participants were eligible for enrollment. Of these, 191 consented to participate, and 169 subsequently enrolled in the study. Fifteen participants discontinued prior to 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 (see Figure S1 for more detail on enrollment and disposition).

The final sample included approximately equal numbers of men (N=77; 51%) and women (N=74; 49%) who ranged in age from 21 - 72 years old. The sample was majority White (N=131; 87%) and non-Hispanic (N=147; 97%). Participants self-reported a mean of 8.9 DSM-5 symptoms of AUD (SD=5.8; range=4-11) and a mean of 5.5 previous quit attempts (SD=5.8, range=0-30). Most participants (N=84; 56%) reported one or more lapses during their participation. The mean number of lapses per participant during the study period was 6.8 (SD=12.0; range=0-75). Table 1 provides more detail on demographic and clinical characteristics of the sample.

EMA Compliance, Features, and Prediction Window Labels

Participants on average completed 3.1 (SD=0.6) of the four daily EMAs each day (78% compliance overall). Participants completed at least one EMA on 95% of days. Across individual weeks in the study, EMA compliance percentages ranged from 75% to 87% completion for all of the 4x daily EMAs and from 92% - 99% for at least one daily EMA completed (see Figure S3).

Using these EMA reports, we created datasets with 270,081, 274,179, and 267,287 future prediction windows for the week, day, and hour window widths, respectively. Each dataset contained 286 features and an outcome labeled as *lapse* or *no lapse*. These datasets

were unbalanced with respect to the outcome such that lapses were observed in 68,467 (25.3%) week windows, 21,107 (7.7%) day windows, and 1,017 (0.3%) hour windows.

Model Performance

auROC

Best model configurations were selected via *validation set* performance. The median auROCs for the best configurations were high for the week (median=0.90, IQR=0.02, range=0.88-0.92), day (median=0.91, IQR=0.01, range=0.89-0.93), and hour (median=0.94, IQR=0.01, range=0.93-0.95) prediction windows.

Best model configurations were evaluated via *test set* performance. The median auROC across the 30 test sets remained high for the week (median=0.89, IQR=0.04, range=0.78-0.96), day (median=0.90, IQR=0.05, range=0.79-0.97), and hour (median=0.93, IQR=0.05, range=0.85-0.97) prediction windows. The left panel of Figure 1 displays the ROC curves by model (i.e., window width) derived by aggregating predicted lapse probabilities across all test sets. Figure S4 presents the individual ROC curves from each test set.

The right panel of Figure 1 displays posterior probability distributions for the auROC separately by model. The median auROCs from these posterior distributions were 0.89, 0.90, and 0.93 for the week, day, and hour models, respectively. 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 (i.e., chance performance) for any of the three window widths: week [0.88-0.91], day [0.89-0.92], hour [0.92-0.94].

We used these posterior probability distributions for the auROCs to formally compare the differences in performance of these models. The median increase in auROC for the hour vs. the day model was 0.02 (95% CI=[0.02-0.03], yielding a probability of 1.000 that the hour model had superior performance relative to the day model. The median increase in auROC for the hour vs. the week model was 0.03 (95% CI=[0.03-0.04],

yielding a probability of 1.000 that the hour model had superior performance relative to the week model. The median increase in auROC for the day vs. the week model was 0.01 (95% CI=[0.00-0.02], yielding a probability of 0.981 that the day model had superior performance relative to the week model. Figure S5 presents histograms of the posterior probability distributions for these model contrasts on auROC.

Other Performance Metrics

Figure S6 displays histograms for the predicted probabilities of lapse for all observations in the 30 *test sets* separately by model and true outcome. We evaluated the sensitivity, specificity, balanced accuracy, PPV, and NPV when these predicted lapse probabilities were used for binary classification (*lapse* vs. *no lapse*) with decision thresholds identified by Youden's Index (see Table 2).

We created Precision-Recall curves by concatenating predicted lapse probabilities across the 30 test sets to evaluate the trade-off between PPV (i.e., precision) and sensitivity (i.e., recall) across decision thresholds (see Figure 2). The PPV of any model can be increased by increasing the decision threshold; however, increasing the decision threshold will also lower the model's sensitivity. For example, the dotted lines in Figure 2 depict the sensitivities (0.72, 0.47, and 0.33 for week, day, and hour models, respectively) associated with decision thresholds that yield 0.70 PPV for each model.

Feature Importance

We display the global importance (mean |Shapley value|) for feature categories for each of the three models in Panel A of Figure 3. These feature categories are ordered by their aggregate global importance (i.e., total bar length) across the three models. The importance of each feature category for specific models is displayed separately by color.

We display local Shapley values that quantify the influence of feature categories on individual observations (i.e., a single prediction window for a specific participant) for each model in Panels B-D of Figure 3.

Discussion

Model Performance

All three of our models performed exceptionally well, yielding auROCs of .90, .91, and .94 for week, day, and hour level models, respectively. auROCs above .9 are generally described as having "excellent" performance, meaning that the model will correctly assign a higher probability to a positive case (e.g., lapse) than a negative case 90% of the time (Mandrekar, 2010). This confirms that EMA can be used in our models to predict future alcohol lapses in the next week, next day, and next hour with high levels of sensitivity and specificity for new patients that were not used to train these models.

This study addressed several important limitations of previous research to advance us toward robust sensing and prediction models that can be embedded within digital therapeutics. First, our models were trained on a relatively large, treatment-seeking sample of adults in early recovery from AUD that more closely matches the individuals most likely to benefit from such models within a digital therapeutic. Second, we explicitly predicted episodes of goal-inconsistent alcohol use (i.e., lapses) because features that predict goal-inconsistent use likely differ from those that predict other types of alcohol use (e.g., episodes of binge drinking among college students, intentional instances of drinking among people not in recovery). Third, we measured EMA features and alcohol use with sufficient frequency and granularity to train well-performing models with high temporal resolution - specifically hour-by-hour predicted probabilities for lapses in the next week, day, and hour. Fourth, we collected features and outcomes over a clinically meaningful duration (up to three months) during a high risk period (initial remission (Hagman et al., 2022) from AUD). Fifth, we used cutting edge resampling methods (grouped, nested, k-fold cross-validation) to provide valid estimates of how our models would perform with new participants that were not used to train these models. Finally, we used methods from interpretable machine learning (i.e. SHAP (Lundberg & Lee, 2017; Molnar, 2022)) to better understand how our models made predictions globally and locally for specific

participants at discrete moments in time.

Understanding & Contextualizing Model Performance

As noted, we used SHAP to explore how key relapse prevention model constructs (represented by categories of features) contributed to predicted lapses. Some constructs consistently emerged as globally important across week, day, and hour level models. Unsurprisingly, the largest contribution to lapse prediction was past use. An individual who reported lapsing frequently was more likely to lapse at any given observation in the future. This is consistent with decades of research on relapse precipitants and our understanding of human behavior more generally (i.e., past behavior predicts future behavior) (Marlatt & Gordon, 1985). Additionally, decreases in self-efficacy were also strongly associated with increased probability of future lapses across all three models.

The contribution of some constructs differed depending on the width of the prediction window. For example, many of the more punctuate, time-varying constructs (e.g., craving, past stressful events, arousal) had greater impact on predicted lapse probabilities in the next hour model. The next hour model was also better able to exploit features for the time and day of the week for the start of the prediction window. Of course, prediction window start time and start day were not useful features in the next week model because its associated prediction window spanned all days and times. The increased global importance for these categories of features to immediate lapse risk likely contributed to the next hour model outperforming the day and week models. These important global differences in next hour lapse risk also highlight the need for just-in-time interventions that can address these imminent but short-lived risks.

The individual, local Shapley values also shed light on the multidimensional and heterogeneous nature of lapse risk in our sample. Sina plots of local Shapley values (Figure 3) display meaningful ranges of scores for most feature categories. This means that even feature categories with lower global importance (e.g., past pleasant events, future stressful events) still consequentially impacted lapse probability predictions for some individuals at

specific times. This variability in locally important features speaks to the need for personalized recommendations about optimal interventions and other supports to address the unique lapse risks for that person at that moment in time.

The demographic variables included in our models did not display either high global importance or local importance for specific predictions. Despite the diversity in socioeconomic status, gender, and age in our sample, these features did not significantly contribute to lapse prediction. While this does not rule out these features' predictive utility, it does suggest that other EMA feature categories (e.g., past use, future efficacy, craving) may be more relevant for lapse prediction than these characteristics. Race and ethnicity also did not emerge as globally or locally important features. However, the limited representation of participants of color in our sample warrants caution in drawing conclusions about the predictive utility of race and ethnicity at this time.

Considerations for Clinical Implementation

Smart Digital Therapeutics

We believe these models may be most effective when embedded in a "smart" digital therapeutic that can guide patients toward optimal, adaptive engagement to address their ongoing and momentary risks. These models can provide the patient's predicted future lapse probability and the features that meaningfully contribute to that predicted probability. We consciously selected EMA items to map onto known risk factors from the Relapse Prevention model. Consequently, these outputs can be used to recommend specific intervention and support modules that are risk-relevant for each patient - much like a clinician would do if they were available in the moment. For example, during sensed periods of high stress, modules that can lower stress (e.g., guided mindfulness, guided body scans) could be recommended. If increased time with risky people or locations is driving lapse risk, the digital therapeutic can support patients to find and attend AA or other support meetings. They could also be encouraged to participate in the in-app discussion board to build a healthy community there.

These module recommendations can also be tuned more precisely using the patient's current lapse probability. If increased craving yields a high predicted lapse probability, stimulus control modules would be recommended (e.g., immediate removal of drinking cues from environment, leave unsafe environments). Conversely, if craving is detected but lapse probability is lower, urge management modules that permit coping with the craving in-place could be recommended (e.g., urge surfing, distracting activities/games).

Of course, we must first determine how best to provide module recommendations such that patients trust and follow the recommendation. Increasing the interpretability and transparency of otherwise "black box" machine learning prediction models can improve perceptions of them, but providing complex or otherwise unnecessary information may instead undermine trust in these models (Molnar, 2022). Therefore, additional research using appropriate research designs is needed to optimize recommendation messaging to increase adherence and associated clinical outcomes (Collins, 2018).

A smart digital therapeutic can potentially improve clinical outcomes in multiple ways. First, feedback from the prediction model could improve patient insight and self-monitoring by connecting their daily experiences to changes in their lapse risk. Second, it can remove patient uncertainty about how to use the digital therapeutic, which could otherwise present a barrier to engagement for some due to the substantial content available. Third, a smart digital therapeutic could encourage risk-relevant engagement. Rather than simply trying to increase overall time using the digital therapeutic, patients could be guided to engage with the interventions and other supports that specifically target their personal risk factors at that moment in time. Finally, like all digital therapeutics, a smart digital therapeutic can address issues of availability, affordability, accessibility, and acceptability that affect clinician-delivered mental healthcare and contribute to health disparities. Thus, smart digital therapeutics are well-positioned to pursue the precision mental health goal to "provide the right treatment to the right patient at the right time, every time" (Kaiser, 2015).

Categorical Lapse Predictions

Our models natively provide quantitative predictions of lapse probabilities. These lapse probabilities can also be used to make specific categorical predictions (lapse vs. no-lapse) for the relevant future prediction windows. This is accomplished by applying a decision threshold to the quantitative predicted lapse probabilities; i.e., when the probability exceeds the decision threshold, a lapse is predicted.

We observed high sensitivity and specificity for these categorical predictions at a decision threshold selected to balance these two performance metrics. However, the PPV (i.e., proportion of predicted lapses that were true lapses) of these categorical predictions was moderate to very low at this threshold (ranging from .60 down to .02 across models). For this reason, we believe that these categorical predictions should be provided to patients with extreme caution (if at all for the models with lower PPV). Instead, we favor use of the quantitative lapse probabilities as a risk indicator as proposed above to guide intervention and support recommendations.

If categorical predictions are necessary, PPV can be improved by raising the decision threshold, but this comes at the cost of reduced sensitivity. We explored this trade-off in the precision-recall curves displayed in Figure 2. From these curves, it is clear decision thresholds that yield higher PPV (e.g., .70) exist for all three models, but the associated sensitivity will be lower (e.g., 0.72, 0.47, and 0.33 for the week, day, and hour models, respectively, at this threshold). Clinical implementation of categorical predictions will require selecting an optimal decision threshold after weighing the cost of missing true lapses (low sensitivity) vs. predicting lapses that subsequently do not occur (low PPV). Different thresholds could be used depending on the purpose, context, available resources, or even patient preference.

Additional Limitations and Future Directions

Successful clinical implementation of our models will require several important steps to address limitations in our work to date. To start, we need to enrich the training data for

these models to include diversity across race, ethnicity, and geographic region. The prediction models in our study may not work well with people of color or people from rural communities. Prediction models must be trained on diverse samples of individuals. Otherwise, their use may exacerbate rather than reduce existing mental healthcare disparities. We must also collect data from individuals in later stages of recovery beyond initial remission because the features that predict lapses may differ in these later periods as individuals become more stable. We are intentionally addressing both of these issues in a current NIH protocol that recruits for demographic and geographic diversity across the US and follows these participants for up to 1.5 years into their recovery (Moshontz et al., 2021).

The chronic nature of AUD may require sustained use of a sensing and prediction system. However, this means that the burden of using such systems must be considered. Participants with AUD find three months of 4x daily EMA to be generally acceptable and report that they could hypothetically sustain this for at least a year if there were clinical benefits to them (Wyant et al., 2023). However, they also report that 1x daily EMA may be more feasible still (Wyant et al., 2023). We plan to develop future prediction models that use only the single morning EMA. This would allow us to contrast the assessment burden vs. model performance trade-off between our current models and putatively lower burden models using only 1x daily EMA. We also plan to train models that use features based on passively sensed geolocation and cellular communications data-streams (i.e., meta-data from call and text messages; text message content) that were also collected from our participants. It may be that these passively sensed signals are sufficient as inputs to an exceptionally low burden prediction model. Alternatively, they can be added to models that also include EMA to increase model performance further and/or to reduce the frequency or length of the EMA surveys while maintaining comparable performance.

Our current models predict probability of imminent lapses. The next hour and day models are well-positioned to identify and recommend just-in-time interventions to address these immediate risks. However, the next week model may not have sufficient temporal

specificity to recommend immediate patient action. Instead, its clinical utility may be improved if we shifted this coarser window duration into the future. For example, we could train a model to predict the probability of lapse at any point during a week window that began two weeks in the future. This "time-lagged" model could provide patients with increased lead time to implement supports that might not be immediately available to them (e.g., schedule an appointment with a therapist, request support from an AA sponsor or appropriate family and friends).

In this study, we have demonstrated that sensing and prediction systems can now be developed to predict future lapses with high temporal resolution. Important steps still remain before these systems can be embedded within smart digital therapeutics and delivered to patients. However, the necessary steps are clear and, when completed, these smart digital therapeutics hold promise to advance us toward precision mental health solutions that may reduce both barriers and disparities in the treatment of AUD.

References

Bae, S., Chung, T., Ferreira, D., Dey, A. K., & Suffoletto, B. (2018). Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. *Addictive Behaviors*, *83*, 42–47. https://doi.org/10.1016/j.addbeh.2017.11.039

Bowen, S., Witkiewitz, K., Clifasefi, S. L., Grow, J., Chawla, N., Hsu, S. H., Carroll, H. A., Harrop, E., Collins, S. E., Lustyk, M. K., & Larimer, M. E. (2014). Relative Efficacy of Mindfulness-Based Relapse Prevention, Standard Relapse Prevention, and Treatment as Usual for Substance Use Disorders. *JAMA Psychiatry*, *71*(5), 547–556. https://doi.org/10.1001/jamapsychiatry.2013.4546

Brandon, T. H., Vidrine, J. I., & Litvin, E. B. (2007). Relapse and relapse prevention. *Annual Review of Clinical Psychology*, *3*(1), 257–284. https://doi.org/10.1146/annurev.clinpsy.3.022806.091455

Burgess-Hull, A. J., Panlilio, L. V., Preston, K. L., & Epstein, D. H. (2022). Trajectories of craving during medication-assisted treatment for opioid-use disorder: Subtyping for early identification of higher risk. *Drug and Alcohol Dependence*, *233*, 109362. https://doi.org/10.1016/j.drugalcdep.2022.109362

Campbell, A. N. C., Nunes, E. V., Matthews, A. G., Stitzer, M., Miele, G. M., Polsky, D., Turrigiano, E., Walters, S., McClure, E. A., Kyle, T. L., Wahle, A., Van Veldhuisen, P., Goldman, B., Babcock, D., Stabile, P. Q., Winhusen, T., & Ghitza, U. E. (2014). Internet-delivered treatment for substance abuse: A multisite randomized controlled trial. *The American Journal of Psychiatry*, *171*(6), 683–690. https://doi.org/10.1176/appi.ajp.2014.13081055

Center for High Throughput Computing. (2006). *Center for high throughput computing*. Center for High Throughput Computing. https://doi.org/10.21231/GNT1-HW21

Center, P. R. (2021). *Mobile Fact Sheet*. Pew Research Center.

Centers for Disease Control and Prevention (CDC). (n.d.). Annual Average for United

States 2011–2015 Alcohol-Attributable Deaths Due to Excessive Alcohol Use, All Ages. In *2022 Alcohol Related Disease Impact (ARDI) Application Website*. https://nccd.cdc.gov/DPH_ARDI/Default/Default.aspx.

Chih, M.-Y., Patton, T., McTavish, F. M., Isham, A. J., Judkins-Fisher, C. L., Atwood, A. K., & Gustafson, D. H. (2014). Predictive modeling of addiction lapses in a mobile health application. *Journal of Substance Abuse Treatment*, *46*(1), 29–35. https://doi.org/10.1016/j.jsat.2013.08.004

Collins, L. M. (2018). *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: The Multiphase Optimization Strategy (MOST)*. Springer International Publishing. https://doi.org/10.1007/978-3-319-72206-1

Dulin, P. L., & Gonzalez, V. M. (2017). Smartphone-based, momentary intervention for alcohol cravings amongst individuals with an alcohol use disorder. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *31*(5), 601–607. https://doi.org/10.1037/adb0000292

Dumortier, A., Beckjord, E., Shiffman, S., & Sejdi, E. (2016). Classifying smoking urges via machine learning. *Computer Methods and Programs in Biomedicine*, *137*, 203–213. https://doi.org/10.1016/j.cmpb.2016.09.016

Dvorak, R. D., Pearson, M. R., & Day, A. M. (2014). Ecological Momentary Assessment of Acute Alcohol Use Disorder Symptoms: Associations With Mood, Motives, and Use on Planned Drinking Days. *Experimental and Clinical Psychopharmacology*, *22*(4), 285–297. https://doi.org/10.1037/a0037157

Dvorak, R. D., Stevenson, B. L., Kilwein, T. M., Sargent, E. M., Dunn, M. E., Leary, A. V., & Kramer, M. P. (2018). Tension reduction and affect regulation: An examination of mood indices on drinking and non-drinking days among university student drinkers. *Experimental and Clinical Psychopharmacology*, *26*(4), 377–390. https://doi.org/10.1037/pha0000210

Epstein, D. H., Tyburski, M., Kowalczyk, W. J., Burgess-Hull, A. J., Phillips, K. A.,

Curtis, B. L., & Preston, K. L. (2020). Prediction of stress and drug craving ninety minutes in the future with passively collected GPS data. *Npj Digital Medicine*, *3*(1), 26. https://doi.org/ghqvcw

Fronk, G. E., Sant'Ana, S. J., Kaye, J. T., & Curtin, J. J. (2020). Stress Allostasis in Substance Use Disorders: Promise, Progress, and Emerging Priorities in Clinical Research. *Annual Review of Clinical Psychology*, *16*(1), 401–430. https://doi.org/10.1146/annurev-clinpsy-102419-125016

Gustafson, D. H., McTavish, F. M., Chih, M.-Y., Atwood, A. K., Johnson, R. A., Boyle, M. G., Levy, M. S., Driscoll, H., Chisholm, S. M., Dillenburg, L., Isham, A., & Shah, D. (2014). A smartphone application to support recovery from alcoholism: A randomized clinical trial. *JAMA Psychiatry*, *71*(5), 566–572. https://doi.org/10.1001/jamapsychiatry.2013.4642

Hagman, B. T., Falk, D., Litten, R., & Koob, G. F. (2022). Defining Recovery From Alcohol Use Disorder: Development of an NIAAA Research Definition. *The American Journal of Psychiatry*, *179*(11), 807–813. https://doi.org/10.1176/appi.ajp.21090963

Hatch, A., Hoffman, J. E., Ross, R., & Docherty, J. P. (2018). Expert Consensus Survey on Digital Health Tools for Patients With Serious Mental Illness: Optimizing for User Characteristics and User Support. *JMIR Mental Health*, *5*(2), e46. https://doi.org/10.2196/mental.9777

Hsieh, F. (1989). Sample size tables for logistic regression. *Statistics in Medicine*, *8*, 795–802.

Jacobson, N. C., Kowatsch, T., & Marsch, L. A. (Eds.). (2022). *Digital Therapeutics for Mental Health and Addiction: The State of the Science and Vision for the Future* (1st edition). Academic Press.

Jonathan, P., Krzanowski, W. J., & McCarthy, W. V. (2000). On the use of cross-validation to assess performance in multivariate prediction. *Statistics and Computing*, *10*(3), 209–229. https://doi.org/10.1023/A:1008987426876

Kaiser, J. (2015). Obama gives East Room rollout to Precision Medicine Initiative. In *Science*. https://www.science.org/content/article/obama-gives-east-room-rollout-precision-medicine-initiative.

Kuhn, M. (n.d.). Bayesian Analysis of Resampling Statistics — perf_mod. In *TidyModels.org*. https://tidyposterior.tidymodels.org/reference/perf_mod.html.

Kuhn, M. (2022). *Tidyposterior: Bayesian Analysis to Compare Models using Resampling Statistics*.

Kuhn, M., & Johnson, K. (2018). *Applied Predictive Modeling* (1st ed. 2013, Corr. 2nd printing 2018 edition). Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*.

Kull, M., Filho, T. M. S., & Flach, P. (2017). Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, *11*(2), 5052–5080. https://doi.org/10.1214/17-EJS1338SI

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Marlatt, G. A., & Gordon, J. R. (Eds.). (1985). *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors* (First edition). The Guilford Press.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Independently published.

Moshontz, H., Colmenares, A. J., Fronk, G. E., Sant'Ana, S. J., Wyant, K., Wanta, S. E., Maus, A., Jr, D. H. G., Shah, D., & Curtin, J. J. (2021). Prospective Prediction of

Lapses in Opioid Use Disorder: Protocol for a Personal Sensing Study. *JMIR Research Protocols*, *10*(12), e29563. https://doi.org/10.2196/29563

Office of the Surgeon General (US), Center for Mental Health, S. (US)., & (US), H. (2001). *Mental Health: Culture, Race, and Ethnicity*. Substance Abuse and Mental Health Services Administration (US).

Prochaska, J. O., DiClemente, C. C., & Norcross, J. C. (1992). In search of how people change: Applications to addictive behaviors. *American Psychologist*, *47*(9), 1102–1114. https://doi.org/10.1037/0003-066X.47.9.1102

Russell, M. A., Linden-Carmichael, A. N., Lanza, S. T., Fair, E. V., Sher, K. J., & Piasecki, T. M. (2020). Affect Relative to Day-Level Drinking Initiation: Analyzing Ecological Momentary Assessment Data with Multilevel Spline Modeling. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *34*(3), 434–446. https://doi.org/10.1037/adb0000550

SAMHSA Center for Behavioral Health Statistics and Quality. (2021). 2021 NSDUH Detailed Tables | CBHSQ Data. In *Substance Abuse and Mental Health Services Administration*. https://www.samhsa.gov/data/report/2021-nsduh-detailed-tables.

Sayette, M. A. (2016). The Role of Craving in Substance Use Disorders: Theoretical and Methodological Issues. *Annual Review of Clinical Psychology*, *12*, 407–433. https://doi.org/10.1146/annurev-clinpsy-021815-093351

Soyster, P. D., Ashlock, L., & Fisher, A. J. (2022). 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*, *36*(3), 296–306. https://doi.org/10.1037/adb0000666

Substance Abuse and Mental Health Services Administration (US), & Office of the Surgeon General (US). (2016). *Facing Addiction in America*. US Department of Health and Human Services.

Walters, S. T., Businelle, M. S., Suchting, R., Li, X., Hébert, E. T., & Mun, E.-Y. (2021). Using machine learning to identify predictors of imminent drinking and create tailored messages for at-risk drinkers experiencing homelessness. *Journal of Substance Abuse Treatment*, *127*, 108417. https://doi.org/10.1016/j.jsat.2021.108417

Wemm, S. E., Larkin, C., Hermes, G., Tennen, H., & Sinha, R. (2019). A day-by-day prospective analysis of stress, craving and risk of next day alcohol intake during alcohol use disorder treatment. *Drug and Alcohol Dependence*, *204*, 107569. https://doi.org/10.1016/j.drugalcdep.2019.107569

WHO ASSIST Working Group. (2002). The Alcohol, Smoking and Substance Involvement Screening Test (ASSIST): Development, reliability and feasibility. *Addiction (Abingdon, England)*, *97*(9), 1183–1194.

Witkiewitz, K., & Marlatt, G. A. (2007). Modeling the complexity of post-treatment drinking: It's a rocky road to relapse. *Clinical Psychology Review*, *27*(6), 724–738. https://doi.org/10.1016/j.cpr.2007.01.002

Wyant, K., Moshontz, H., Ward, S. B., Fronk, G. E., & Curtin, J. J. (2023). Acceptability of Personal Sensing Among People With Alcohol Use Disorder: Observational Study. *JMIR mHealth and uHealth*, *11*(1), e41833. https://doi.org/10.2196/41833

Table 1

Demographics and clinical characteristics

	N	*%*	*M*	*SD*	*Range*
Age			41	11.9	21-72
Sex					
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			
Homemaker	1	0.7			
Disabled	7	4.6			
Retired	8	5.3			
Unemployed	18	11.9			
Temporarily laid off, sick leave, or maternity leave	3	2.0			
Other, not otherwise specified	9	6.0			
Personal Income			$34,298	$31,807	$0-200,000
Marital Status					
Never married	67	44.4			
Married	32	21.2			
Divorced	45	29.8			
Separated	5	3.3			
Widowed	2	1.3			
Alcohol Use Disorder Milestones					
Age of first drink			14.6	2.9	6-24
Age of regular drinking			19.5	6.6	11-56
Age at which drinking became problematic			27.8	9.6	15-60
Age of first quit attempt			31.5	10.4	15-65
Number of Quit Attempts*			5.5	5.8	0-30
Lifetime History of Treatment (Can choose more than 1)					
Long-term residential (6+ months)	8	5.3			
Short-term residential (< 6 months)	49	32.5			
Outpatient	74	49.0			

Individual counseling	97	64.2			
Group counseling	62	41.1			
Alcoholics Anonymous/Narcotics Anonymous	93	61.6			
Other	40	26.5			
Received Medication for Alcohol Use Disorder					
Yes	59	39.1			
No	92	60.9			
DSM-5 Alcohol Use Disorder Symptom Count			8.9	1.9	4-11
Current (Past 3 Month) Drug Use					
Tobacco products (cigarettes, chewing tobacco, cigars, etc.)	84	55.6			
Cannabis (marijuana, pot, grass, hash, etc.)	66	43.7			
Cocaine (coke, crack, etc.)	18	11.9			
Amphetamine type stimulants (speed, diet pills, ecstasy, etc.)	15	9.9			
Inhalants (nitrous, glue, petrol, paint thinner, etc.)	3	2.0			
Sedatives or sleeping pills (Valium, Serepax, Rohypnol, etc.)	22	14.6			
Hallucinogens (LSD, acid, mushrooms, PCP, Special K, etc.)	14	9.3			
Opioids (heroin, morphine, methadone, codeine, etc.)	16	10.6			

Reported 1 or More Lapse During Study Period					
Yes	84	55.6			
No	67	44.4			
Number of reported lapses			6.8	12	0-75

Note:

N = 151

Two participants reported 100 or more quit attempts. We removed these outliers prior to calculating the mean (M), standard deviation (SD), and range.

Table 2

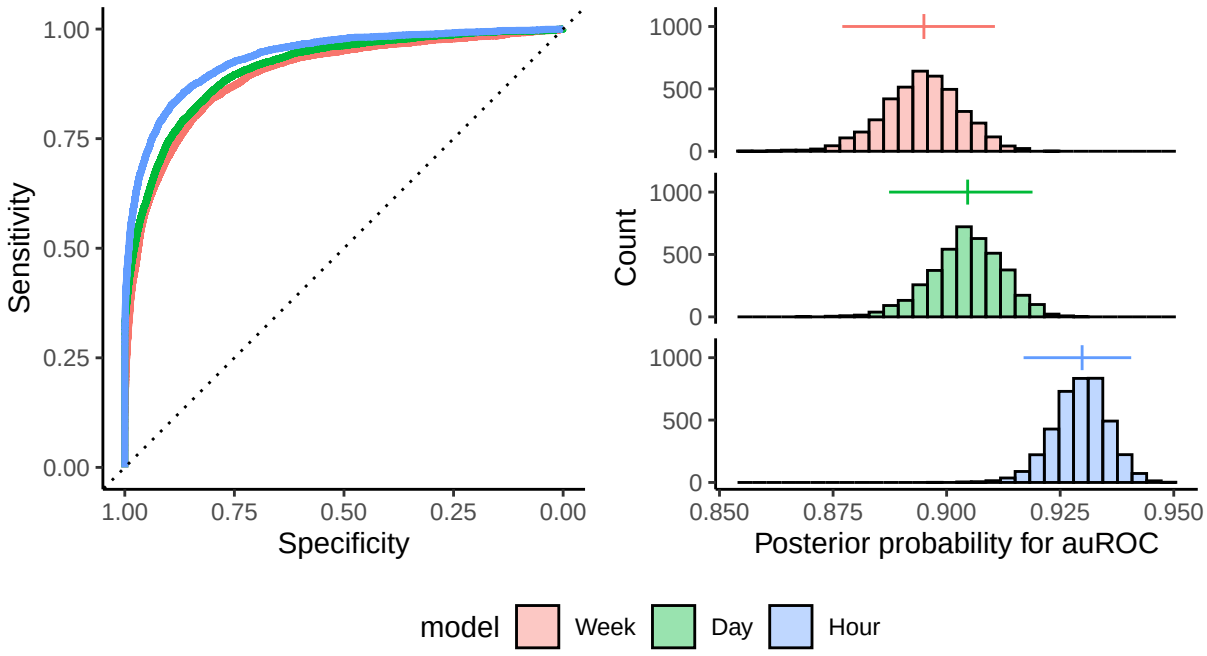
Performance Metrics by Model

Metric	Week	Day	Hour
auROC	0.90	0.91	0.94
sensitivity	0.83	0.85	0.86
specificity	0.81	0.81	0.87
balanced accuracy	0.82	0.83	0.86
positive predictive value	0.60	0.27	0.02
negative predictive value	0.93	0.98	1.00

Note: Areas under the receiver operating characteristic curves (auROCs) summarize the model's sensitivity and specificity over all possible decision thresholds. Sensitivity, specificity, balanced accuracy, positive predictive value, and negative predictive value are performance metrics calculated at a single decision threshold for each model determined with Youdens index.

Figure 1

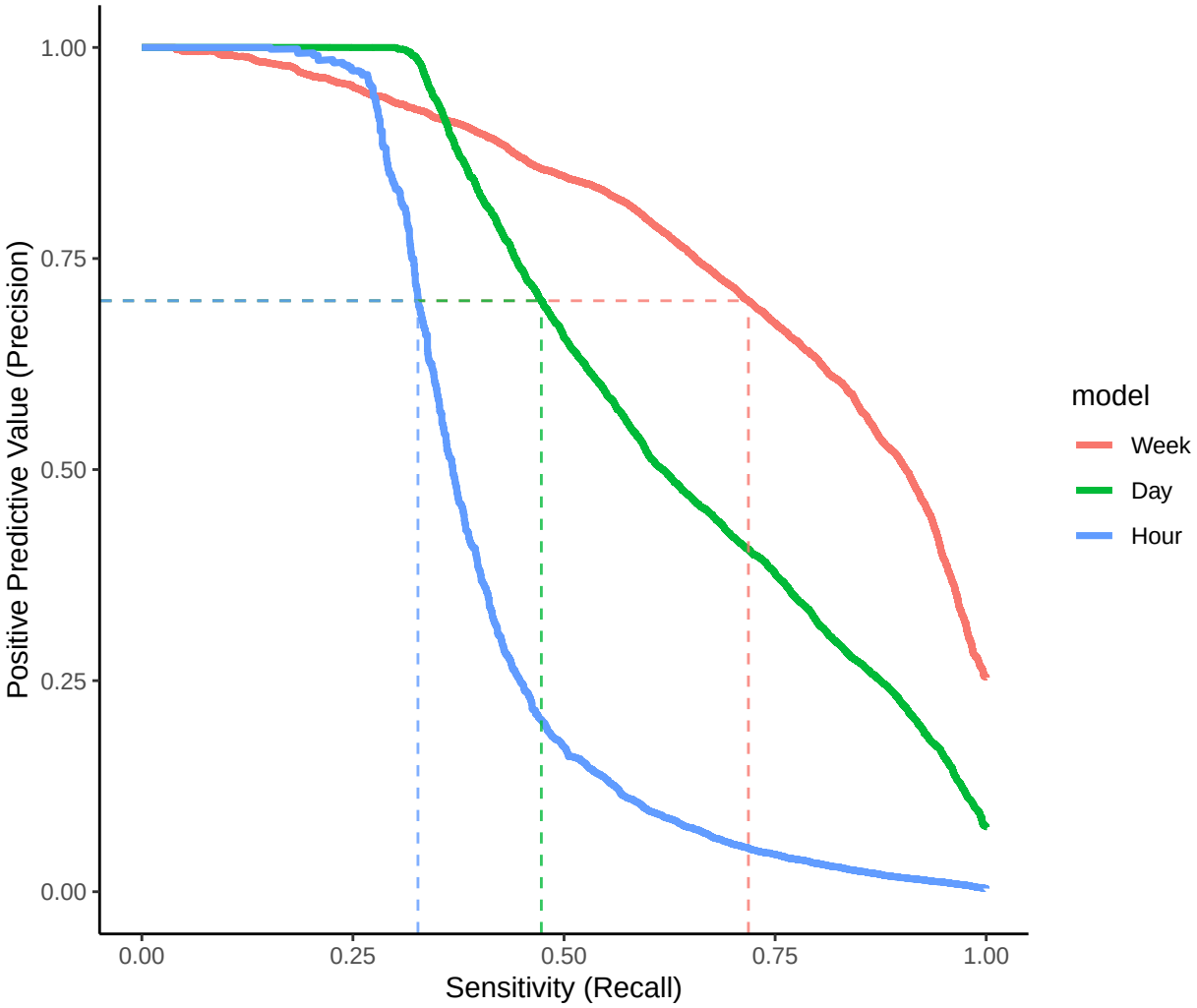
ROC curves and posterior probabilities for auROCs by model



Note. The left panel depicts the aggregate receiver operating characteristic (ROC) curve for each model, derived by concatenating predicted lapse probabilities across all test sets. The dotted line represents the expected ROC curve for a random classifier. The histograms on the right depict the posterior probability distribution for the areas under the receiver operating characteristic curves (auROCs) for each model. The vertical lines represent the median posterior probability and the horizontal line represents the boundaries 95% CI.

Figure 2

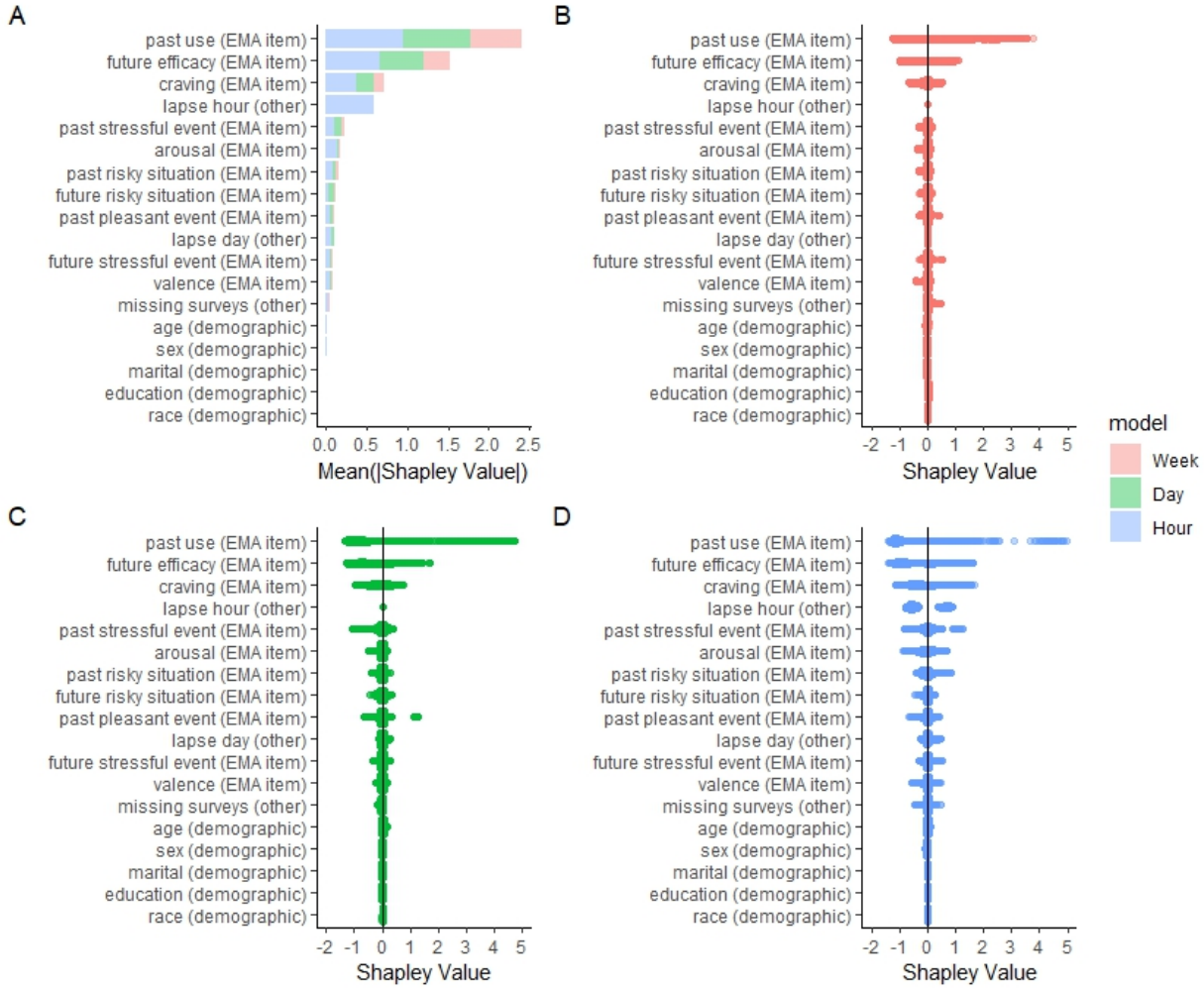
Precision-recall curves by model



Note. The plot depicts the aggregate precision-recall curves for each model, derived by concatenating predicted lapse probabilities across all test sets. The dotted lines depict the sensitivities (0.72, 0.47, and 0.33 for week, day, and hour models, respectively) associated with decision thresholds that yield 0.70 positive predictive value for each of those models.

Figure 3

Feature importance (Shapley values) by model



Note. Panel A displays the global importance (mean |Shapley value|) for feature categories for each model. Raw EMA features are grouped into categories by the original item from the EMA. Features from demographics and the day and hour for the start of the prediction window are also included. Feature categories are ordered by their aggregate global importance (i.e., total bar length) across the three models. The importance of each feature category for specific models is displayed separately by color. Panels B-D display local Shapley values that quantify the influence of feature categories on individual observations (i.e., a single prediction window for a specific participant) for each model.