Using smartphone sensing and machine learning for personalized daily lapse risk prediction in a national sample of people with opioid use disorder

Kendra Wyant[aff-1] and John J. Curtin[aff-1]

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

Abstract

Enter abstract here.

Keywords: Substance use disordersPrecision mental health

Introduction

Methods

Transparency and Openness

We adhere to research transparency principles that are crucial for robust and replicable science. First, we published this study's protocol as a registered report (International Registered Report Identifier [IRRID]: DERR1-10.2196/29563) during the initial enrollment of pilot participants (Moshontz et al., 2021). Second, we followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis extension for Artificial Intelligence (TRIPOD+AI) guidelines (Collins et al., 2024). A TRIPOD+AI checklist is available in the supplement. Finally, our features, labels, questionnaires, and other study materials are publicly available on our OSF page (https://osf.io/zvm7s/overview) and our annotated analysis scripts and results are publicly available on our study website (https://jjcurtin.github.io/study_risk2/).

For transparency, we documented the changes made to the registered report below:

Sensing Data Streams

Our candidate models included features derived from a subset of intake self-report measures, daily EMA, and geolocation sensing data (see Measures section). Consequently, we excluded cellular communication data, daily video check-ins, app usage data, and certain self-report measures. These decisions were informed by (1) our group's personal sensing work with alcohol use disorder, (2) technological constraints, and (3) the desire to balance feature diversity for capturing lapse complexity and minimizing participant burden and computational cost.

1. In our alcohol use disorder research, EMA and geolocation sensing have shown moderate to excellent predictive signal (aurocs .72-.91) (Wyant et al., 2024; Wyant et al., under review,). Cellular communication sensing, however, has fallen short of these thresholds. Moreover, Apple places strict restrictions on app access to communications, meaning inclusion of these features would result in a model that could only be deployed within an Android operating system.
2. We discontinued collecting daily video check-ins about 6 months into the 2.5 year data collection due to technical issues and App usage data (beyond the required study tasks) were generally sparse and inconsistent across participants.
3. Self-report measures can substantially increase data collection burden and expand the feature space. Therefore, we opted to not include all measures. Monthly surveys were lengthy (20–30 minutes) and, based on our work with alcohol use disorder, added no incremental predictive value beyond the dynamic sensing data. However, individual differences in severity of use and stability of recovery at the time of intake have demonstrated some predictive signal. Key demographic variables known to influence OUD treatment access and clinical outcomes are also important to reduce model bias by preventing these effects being encoded into the model indirectly by proxy variables. See Measures section for a comprehensive list of all retained self-report items.

Resampling Method

Our initial protocol proposed repeated cross-validation for model selection and a single held-out test set for evaluation. Although we nearly reached our recruitment goal (N = 451/480), the

number of participants with usable data was much lower (N = 299). To maximize data use, we eliminated the held-out test set and relied on repetitions in our cross-validation to mitigate optimization bias and provide generalizable performance estimates (i.e., 6 repeats of 5-fold cross-validation to generate 30 held-out test sets).

Participants

We recruited 451 participants in early recovery from opioid use disorder from across the United States. We recruited through national digital advertising and collaborations with treatment providers at MOUD clinics. Our recruitment strategy was designed to create a diverse sample with respect to demographics (gender, age, race, and ethnicity), and geographic location (urban and rural). We required participants:

- were age 18 or older,
- could read, write, and speak in English,
- were enrolled in and adherent with an MOUD program for at least 1 month but no longer than 12 months, and
- had an Android smartphone with an active cellular plan.

We did not exclude participants for comorbid substance use or other psychiatric disorders. Participants were compensated up to $75 per month for completing study tasks (i.e., EMAs, monthly surveys and sharing sensing data) and were paid $50 per month to offset the cost of maintaining a cellphone plan.

Procedure

Participants completed three video or phone visits over approximately 12 months. During the enrollment visit, study staff obtained written informed consent and collected demographic information. They then walked participants through how to download the Smart Technology for Addiction Recovery (STAR) study app, provided a set of video tutorials to learn how to use the app, and instructed participants to complete the intake survey within the app. The STAR app was developed with the UW Madison Center for Health Enhancement Systems Studies (CHESS) and used for all data collection. Within the app participants could control their data sharing options, monitor completed study tasks, receive reminder notifications about tasks, message staff, and access CHESS's suite of resources and tools for people in recovery from AUD. Enrolled participants met with study staff 1 week later to troubleshoot technical issues. At the end of the study enrollment period participants met briefly with study staff for a debriefing session. While on study, participants were expected to complete daily EMA, monthly surveys, and share geolocation sensing data. Other sensing data streams (i.e., daily video check-ins, cellular communications, and app usage data) were collected as part of the parent grant's aims (NIDA R01 DA047315). All procedures were approved by the University of Wisconsin-Madison Institutional Review Board (Study #2019-0656).

Measures

Individual Differences

We collected self-report information about demographics (age, gender, orientation, race, ethnicity, education, and income). Zip codes from participants' reported home addresses were linked to Rural–Urban Commuting Area (RUCA) codes to characterize the rural–urban status of their residence .

We also collected information about OUD history to characterize OUD severity and recovery stability, including self-reported Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition OUD symptoms (American Psychiatric Association, 2022), recovery satisfaction, motivation, and confidence, whether they intended to avoid using other drugs, perceived efficacy and likelihood of adhering to their MOUD medication (Morisky et al., 1986), past month opioid use, past month residential treatment for OUD, past month receipt of psychiatric medication, frequency of counseling sessions and self-help meetings in the past month, preferred opioid, preferred route of administration, and lifetime history of overdose. 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 in our registered report (Moshontz et al., 2021).

Ecological Momentary Assessment

Participants completed a brief (1-2 minute) EMA each day on study through the STAR app. The EMA became available in the app at 5:00 AM CST each morning and participants had 24 hours to complete it. Participants could enable push notifications for reminder prompts to complete the

assessment. Each EMA had 16-items that asked participants to report the date and time of any recent opioid use for nonmedical reasons not yet reported. These reports served as the primary outcome for the lapse risk prediction model. Participants also reported any other drugs that they had used in the past 24 hours (in a select all that apply format) and whether they took their MOUD as prescribed. Next, participants rated the maximum intensity of recent (i.e., since last EMA) experiences of pain, craving, risky situations, stressful events, and pleasant events. Next, participants rated their sleep and how depressed, angry, anxious, relaxed, and happy they have felt in the past 24 hours. Lastly, participants responded to 2 future-facing items that asked about participants' motivation and confidence to continue to avoid using opioids for nonmedical reasons over the next week. The full EMA questionnaire is available in the supplement.

Geolocation Sensing

The STAR app passively recorded participants' time-stamped geolocations (i.e., latitude and longitude) every 1.5-15 minutes, depending on their movement. We augmented the geolocation data with self-reported subjective contexts. On each monthly survey we asked a set of 6 questions about frequently visited locations (i.e., spending more than 3 minutes at a location 2 or more times in a month) from the previous month. Participants were asked to describe the type of place, what they typically do there, the general frequency of pleasant and unpleasant experiences associated with the place, and the extent to which spending time there supports or undermines their recovery.

Data Analytic Strategy

Data preprocessing, modeling, and Bayesian analyses were done in R using the tidymodels ecosystem (Goodrich et al., 2023; Kuhn, 2022; Kuhn & Wickham, 2020). 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).

Feature Engineering

Features were calculated using only data collected before the start of each prediction window to ensure our models were making true future predictions. We caclulated a total of 793 features from three data sources:

1. *The prediction window.* We created dummy-coded features for day of the week for the start of the prediction window.
2. *Static individual differences in demographic and OUD characteristics.* For demographics, we created ordinal features for age, education, and income based on predefined response ranges (see Table 1 for ranges), an ordinal feature for RUCA code associated with home address (range 1-10), and dummy coded features for gender (male vs. not male) and race and ethnicity (non-Hispanic White vs. Hispanic and/or not White). For OUD characteristics we created ordinal features for recovery satisfaction, recovery motivation, recovery confidence, MOUD side effects experienced, perceived efficacy of MOUD medication, and likelihood of continuing MOUD (ranges 0-4), ordinal features for frequency of counseling sessions and self-help meetings attended in the past month (ranges 0-3), a quantitative feature for number

of self-reported DSM-5 OUD symptoms, an ordinal feature for lifetime history of overdose (range 0-4), and dummy coded features for past month opioid use (yes vs. no), past month detox or residential treatment (yes vs. no), past month psychiatric medication (yes vs. no), preferred opioid (fentanyl vs. heroin vs. prescription opioid not for opioid treatment vs. medication for opioid treatment), and preferred route of administration (injection vs. oral vs. smoke vs. sniff/snort vs. other).

3. *Dynamic EMA and geolocation sensing data.* For both sets of dynamic features we calculated two types of features: raw and difference features. Raw features represent the feature value calculated within a scoring epoch (e.g., the maximum urge rating reported on EMA during the 48 hours immediately preceding the start of the prediction window). Difference features capture participant-level changes from their baseline scores. Specifically, we subtracted each participant's mean score for each feature (using all available data prior to the prediction window) from the associated raw feature (e.g., the participant's average urge rating across all prior EMAs subtracted from the maximum urge rating in the preceding 48 hours).

We used three scoring epochs (48, 72, and 168 hours before the start of the prediction window) to create features from the daily EMA. We calculated raw and difference features for min, max, median, and most recent responses for the 13 5-point likert scale items (items 4-16; pain, urge, risky situation, stressful event, pleasant event, sleep, depressed, angry, anxious, relaxed, happy, abstinence motivation, and abstinence confidence) across all EMAs in each epoch for a given participant. We also calculated raw and difference rate features based on

counts of opioid lapses, other drug use, missed MOUD doses (items 1-3), and completed EMAs across all EMAs in each scoring epoch.

We used six scoring epochs (6, 12, 24, 48, 72, and 168 hours before the start of the prediction window) to create features from the densely sampled geolocation data. Raw geolocation points were cross-checked against known locations with reported subjective context. We used a threshold of 50 meters for matching context to geolocation points. We calulated raw and difference features for sum duration of time spent in transit (i.e., moving faster than 4 miles per hour) and time out in the evenings (i.e., not at their home between the hours of 7:00pm-4:00am). We calulated raw and difference features for sum duration of time spent at locations according to what they indicated they do at the location (spend time with friends, socialize with new people, religious activities, relax, spend time with family, volunteer, receive mental health care, receive physical health care, receive MOUD treatment, drink alcohol, take classes, work), how pleasant and unpleasant their experiences typically are at the location (Always, Most of the time, Sometimes, Rarely, Never), and how helpful and harmful the location is to their recovery (Extremely, Considerably, Moderately, Mildly, Not at all). We also calculated a raw and difference features of location variance (i.e., the extent to which a participant's location changes over a scoring epoch).

All features (sets 1-3 above) were included in our full model. We also fit a baseline model that excluded the dynamic EMA and geolocation features (i.e., only using feature sets 1 and 2 above) to assess the incremental predictive value of these sensing features.

Other feature engineering steps performed during cross-validation included imputing missing values (median imputation for numeric features and mode imputation for nominal features), dummy coding nominal features, normalizing features to have a mean of zero and standard deviation of 1, bringing outlying values (|z-score| > 5) to the fence, and removing zero and near-zero variance features as determined from held-in data. We selected coarse median/mode methods for handling missing data due to the low rates of missing values and computational costs associated with more advanced forms of imputation (e.g., KNN imputation, multiple imputation). A sample feature engineering script (i.e., tidymodels recipe) containing all feature engineering steps is available on our OSF study page.

Lapse Labels

Prediction windows started at 6:00am in participants' local timezones and ended at 5:59am the next day. This window start time was selected to match a typical wake-sleep cycle as opposed to midnight-to-midnight calendar day. For each participant the first prediction window began at 6:00am on their second day of participation and rolled forward day-by-day until their participation ended (i.e., the last prediction window ended at 5:59am on the day of their last recorded EMA).

Participants reported lapses on the first EMA item. If participants responded "yes" to the question "Have you used any opioids for non-medical reasons that you have not yet reported?", they were prompted to select the day(s) and time(s) of the lapse(s). Times were reported in 6-hour increments (12:00am–5:59am, 6:00am–11:59am, 12:00pm–5:59pm, 6:00pm–11:59pm). These responses were used to label each prediction window as lapse or no lapse. For example, if a

participant reported a lapse on December 12 from 12:00am-5:59am the prediction window spanning 6:00am on December 11 through 5:59am on December 12 was labeled as a lapse.

Model Configurations

Candidate model configurations differed on statistical algorithm, hyperparameter values, resampling of outcome, and feature set. We considered five statistical algorithms that differed in terms of assumptions and bias-variance tradeoff: elastic net, random forest, XGBoost, a single-layer neural network, and a support vector machine. Configurations differed across sensible values for key hyperparameters. Configurations also differed on outcome resampling method (i.e., up-sampling and down-sampling of the outcome at ratios ranging from 5:1 to 1:1). All resampling was exclusively done in the held-in training data (i.e., test data was not resampled) to prevent biasing performance estimates (Vandewiele et al., 2021). Lastly, configurations differed on feature set (full vs. baseline). Our primary full model configurations used all available features (see Feature Engineering section). Our baseline model configurations used only features from the prediction window (day of week) and demographic and OUD characteristics (i.e., it excluded EMA and geolocation features).

The best full and baseline model configurations were selected using 6 repeats of participant-grouped 5-fold cross-validation. 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 by a between-subject measure of our outcome (no lapse vs. any lapse).

Model Evaluation

We evaluate the best full model's probability predictions across three domains: discrimination, calibration, and overall performance. We follow best recommendations for reporting measures and plots to characterize these performance domains (Van Calster et al., 2025). Classification and clinical utility are two other important domains for evaluating a model intended to make or inform a decision (e.g., whether to send a support message to an individual). In our scenario there is no decision to be made. Hypothetically, every day an individual would receive a message regardless of lapse risk, equivalent to a treat all condition. Therefore, we have constrained our evaluation to focus on the probability estimates for evaluating performance.

auROC and Model Comparisons. Our performance metric for model selection and evaluation was the area under the receiver operating characteristic curve (auROC). auROC represents the probability that the model will assign a higher predicted probability to a randomly selected positive case (lapse) compared to a randomly selected negative case (no lapse). auROC is an aggregate measure of discrimination across all possible decision thresholds. This is important because optimal thresholds depend on the setting, outcome prevalence, and the relative costs of misclassification, and should be addressed separately (e.g., with a decision curve analysis).

We used a Bayesian hierarchical generalized linear model to estimate the posterior distribution and 95% credible intervals (CI) for auROC for the 30 held-out test sets for our best full and baseline models. We used weakly informative, data-dependent priors to regularize and reduce overfitting. 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 the repeat. auROCs were transformed using the logit function and regressed as a function of model contrast (full vs. baseline). From the Bayesian model we obtained the posterior distribution for auROC for the full and baseline models. We reported the median posterior probability for auROC and 95% CIs for each model. We then conducted a Bayesian model comparison to determine the probability that the full and baseline models' performances differed systematically from each other.

We performed five dichotomous subgroup analyses to assess the fairness of our model's predictions. Using the same 30 held-out test sets and the same modeling procedure as above, we calculated the median posterior probability and 95% CI for auROC for each model separately by gender (not male vs. male), race/ethnicity (Hispanic and/or non-White vs. non-Hispanic White), income (below poverty line vs. above poverty line), geographic location (rural vs. urban)[1], and education (high school or less vs. some college). We conducted Bayesian group comparisons to assess the likelihood that each model performs differently by group.

Calibration and Overall Performance. Calibration is an indicator of how well a model's predicted probabilities correspond to the true observed outcomes. For example, a well-calibrated

[1]We followed guidelines from the United States Health Resources and Services Administration and define urban as an area where the primary commuting flow is within a metropolitan core of 50,000 or more people (RUCA code = 1) and rural as anything not urban (RUCA codes 2-10).

model that assigns a 30% lapse risk prediction should observe lapses in approximately 30% of such cases. We used Platt scaling to calibrate our full model's raw probabilities (Platt, 1999). We provided a calibration plot of these raw and calibrated probabilities. To characterize overall performance we reported Brier scores for the raw and calibrated probabilities. Brier scores are the mean squared difference between the predicted probabilities and observed outcome and range from 0 (perfect accuracy) to 1 (perfect inaccuracy). We also provided histograms of risk probability distributions by true lapse outcome.

Feature Importance

Feature importance values provide insight into the features that have the most influence on the model's predictions. For every prediction, we can extract feature importance values providing actionable insight into intervenable targets for lapse risk (i.e., for a specific individal at a specific moment). We calculated raw Shapley values for each observation in the held-out test sets (Lundberg & Lee, 2017). The magnitude of the raw Shapley value indicates how much the feature score for that observation adjusted the prediction (in log-odds units) 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. In other words, higher Shapley values suggest the feature increases lapse risk and lower values suggest the feature decreases lapse risk. Shapley values are inherently additive. For any observation, Shapley values can be summed to create a total adjustment score for the predicted value. We created feature categories by collapsing features that differed only by scoring epoch and/or dummy-coded level into a single feature category. We plotted raw Shapley

values and feature categories as partial dependence plots to illustrate these feature-risk relationships.

Feature importance values can also be aggregated across all participants and all observations to provide a relative rank ordering of the most important features. We calculated overall feature importance in two ways. First, we used a traditional approach in which we calculated the mean absolute Shapley value for each feature category across all observations. This approach summarizes overall feature importance by averaging the magnitude of each feature's contribution. However, it can be skewed toward features that exhibit infrequent but very large Shapley values, potentially overstating the importance of features that are strongly associated with the outcome, but only come up in a small subset of observations. The second way we calculated feature importance was by calculating the proportion of observations in which each feature category had the highest Shapley value. This approach summarizes how frequently a feature category is influential across observations (i.e., considering both magnitude and prevalence). We provided a plot of the relative ranking of feature categories by their overall feature importance using these two methods.

Results

Participants

We recruited 451 participants across 47 states in the United States from April 2021 through December 2024. A total of 336 participants were eligible, consented, and remained on study for at

least one month. Of these, 11 participants were excluded due to unusually low adherence[2], 13 participants were excluded due to insufficient context data for geolocation points[3], 1 participant was excluded due to geolocation data indicating they were not residing in the United States, 11 participants were excluded due to evidence of careless responding on EMAs and/or no longer having a goal of abstinence. Our final analysis sample consisted of 299 participants. Participant demographic and OUD characteristics is presented in Table 1.

[2]One participant completed only 3 EMA prompts over 88 days and 10 participants had fewer than 20 geolocation points per day on average.

[3]We required participants have at least two contextualized locations other than their home.

Table 1: Demographic and Clinical Characteristics

	N	%
Age		
18-21	2	0.7
22-25	7	2.3
26-35	107	35.8
36-45	106	35.5
46-55	60	20.1
56-65	13	4.3
Over 65	4	1.3
Gender		
Man	158	52.8
Woman	134	44.8
Non-binary	4	1.3
Not listed above	2	0.7
Orientation		
Straight, that is, not gay or lesbian	237	79.3
Lesbian or gay	16	5.4
Bisexual	39	13.0
Not sure	1	0.3
Not listed above	4	1.3
Race and Ethnicity (select all that apply)		
American Indian/Alaskan Native	17	5.7
Asian	4	1.3
Black/African American	44	14.7
Native Hawaiian/Other Pacific Islander	3	1.0
White/Caucasian	237	79.3
Hispanic, Latino, or Spanish origin	24	8.0
Not listed above	5	1.7
Geographic Location		
rural	62	20.7
urban	274	91.6
[illegible]teek	[illegible]	[illegible]
N = 299		

Adherence, Features, and Labels

Mean days on study across participants was 297 days (range 32-395 days). 83% of participants (249/299) remained on study for at least six months. EMA adherence was high. On average participants completed 73% of the daily EMA prompts (range 24-100%). Participants provided, on average, 311 daily geolocation points (range 20-825).

Our final feature set consisted of 674 features. The proportion of missing values across features was low (mean=.02, range = 0-.11). Across participants we generated a total of 88,607 day-level labels. Training sets (N=30) had, on average, 70886 labels (range 69025-73129) from 239 participants (range 239-240). Forty percent of participants (119/299) reported an opioid lapse while on study (mean=5.36, range 0-76). This resulted in 1.81% of the labels positive for lapse (1,603/88,607 labels). We stratified the data on a variable of whether someone lapsed on study to ensure our imbalanced outcome was evenly split over folds (mean=.018, range =.016-.020).

auROC and Model Comparisons

The best full model configuration used an xgboost statistical algorithm and up-sampled the minority class.[4] The median posterior probability for the best model was 0.93, with narrow 95% CI ([0.92, 0.94]).

We compared our best model's performance to a baseline model that only used day of week and demographic and OUD characteristic features to evaluate the incremental predictive value of adding the dynamic EMA and geolocation sensing features. The median posterior

[4]The best model configuration used 1:1 upsampling of the minority class and the following hyperparameter values: learning rate = .01, tree depth = 5, mtry = 50.

probability for the baseline model was 0.74 (95% CI [0.71, 0.77]). A Bayesian model comparison revealed extremely strong evidence that the best model was more predictive than the baseline model (probability = 1.00).

Our fairness subgroup comparisons for the full model revealed no evidence that performance meaningfully differed by education (probability = 0.55) and strong evidence that performance differed by gender, income, race and ethnicity, and geographic location (probabilities > 0.90). Notably, our model performed better for individuals with an annual income below the federal poverty line compared to individuals above the federal poverty line, thus favoring the disadvantaged group. While differences in performance estimates exist across subgroups, they are not likely clinically meaningful as all of our subgroups yielded median auROCs between 0.91 - 0.94 (Figure 1). A table of fairness subgroup comparisons is available in the supplement.

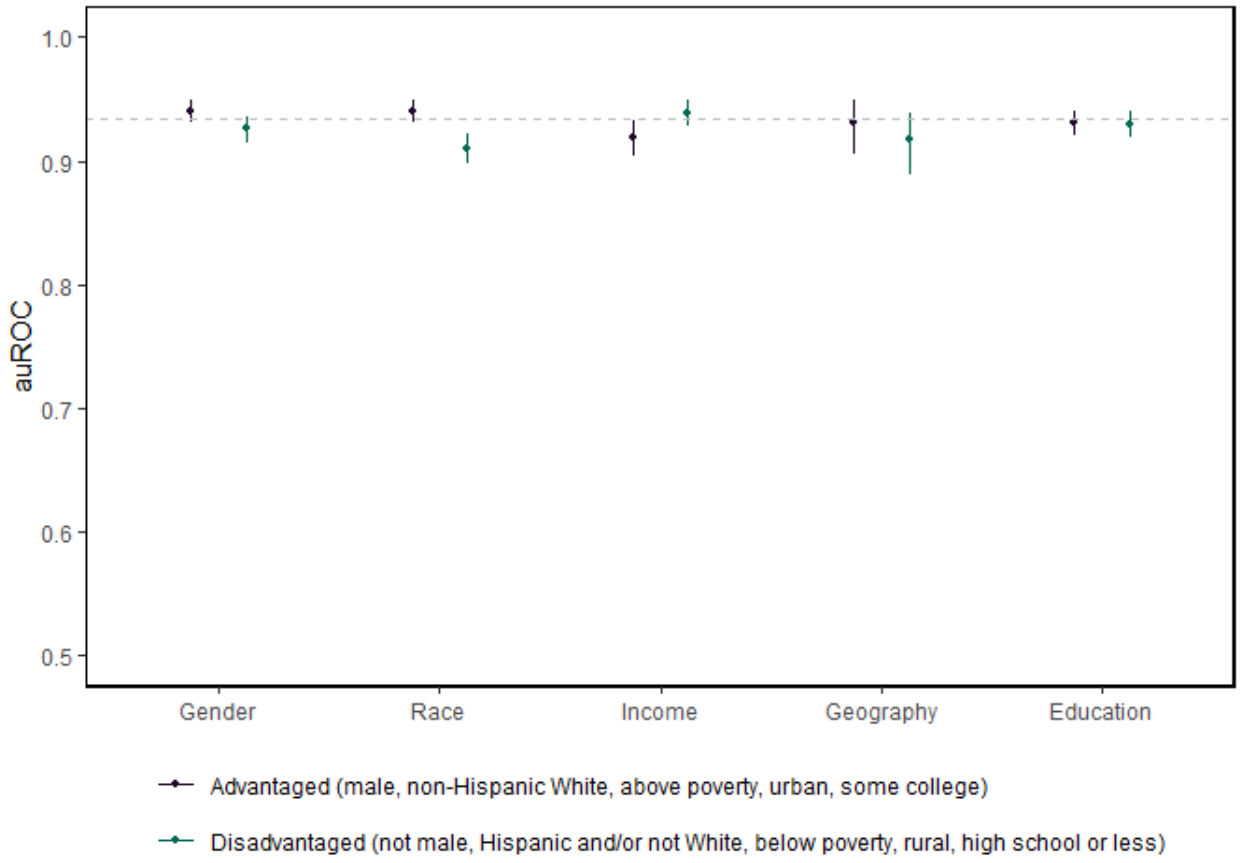

Figure 1: Posterior probabilities for area under the receiver operating curve (auROC) by demographic subgroup. auROC ranges from .5 (chance performance) to 1 (perfect performance). Subgroups advantaged in access to substance use treatment and outcomes (male, non-Hispanic White, above poverty, urban location, and some college education) are depicted in dark purple. Subgroups disadvantaged in access to substance use treatment and outcomes (not male, Hispanic and/or not White, below poverty, rural location, and high school education or less) are depicted in green. Overall model performance across groups is depicted as the dashed grey line.

Calibration and Overall Performance

The full model produced generally well-calibrated raw probabilities (brier score = 0.015). We attempted to improve calibration with Platt scaling and results were comparable (brier score =

0.015). A Calibration plot revealed our raw model predictions tended to systematically overpredict risk probabilities (Figure 2). Platt scaling appeared to produce probabilities closer to the ideal line were predicted probabilities perfectly match observed rates. Histograms of raw risk probability distributions separately by true lapse outcome are presented in Figure 2.

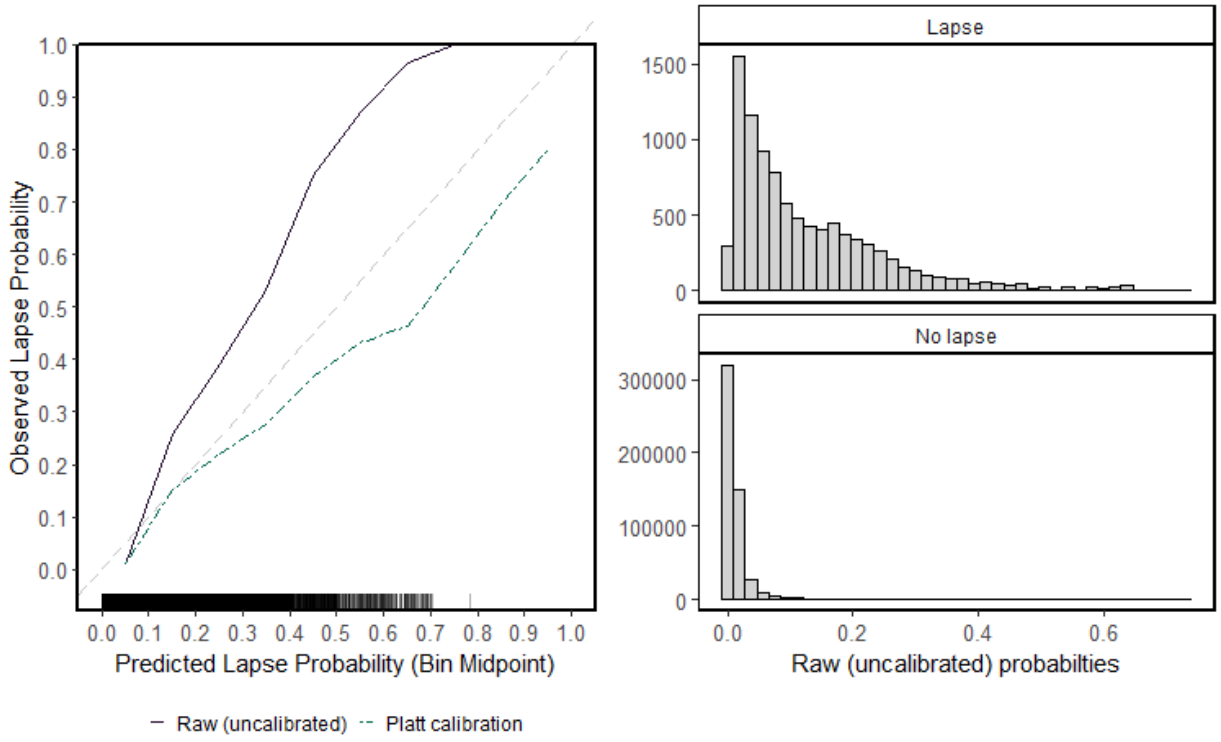

Figure 2: The left panel presents a calibration plot of raw and Platt scaled risk probabilities. Predicted probabilities (x-axis) are binned into deciles. Observed lapse probability (y-axis) represents the proportion of actual lapses observed in each bin. The dashed diagonal represents perfect calibration. Points below the line indicate overestimation and points above the line indicate underestimation. Raw probabilities are depicted as the dark purple line. Platt calibrated probabilities are depicted as the green dashed line. The rug plot along the x-axis depicts observation frequency in each bin. The right panel presents histograms of raw (uncalibrated) risk probability distributions separately by true lapse outcome.

Feature Importance

For brevity, we are only displaying the top 30 features, as defined by the 30 largest absolute mean shapley values. In the supplement, we provide full figures of all possible feature categories.

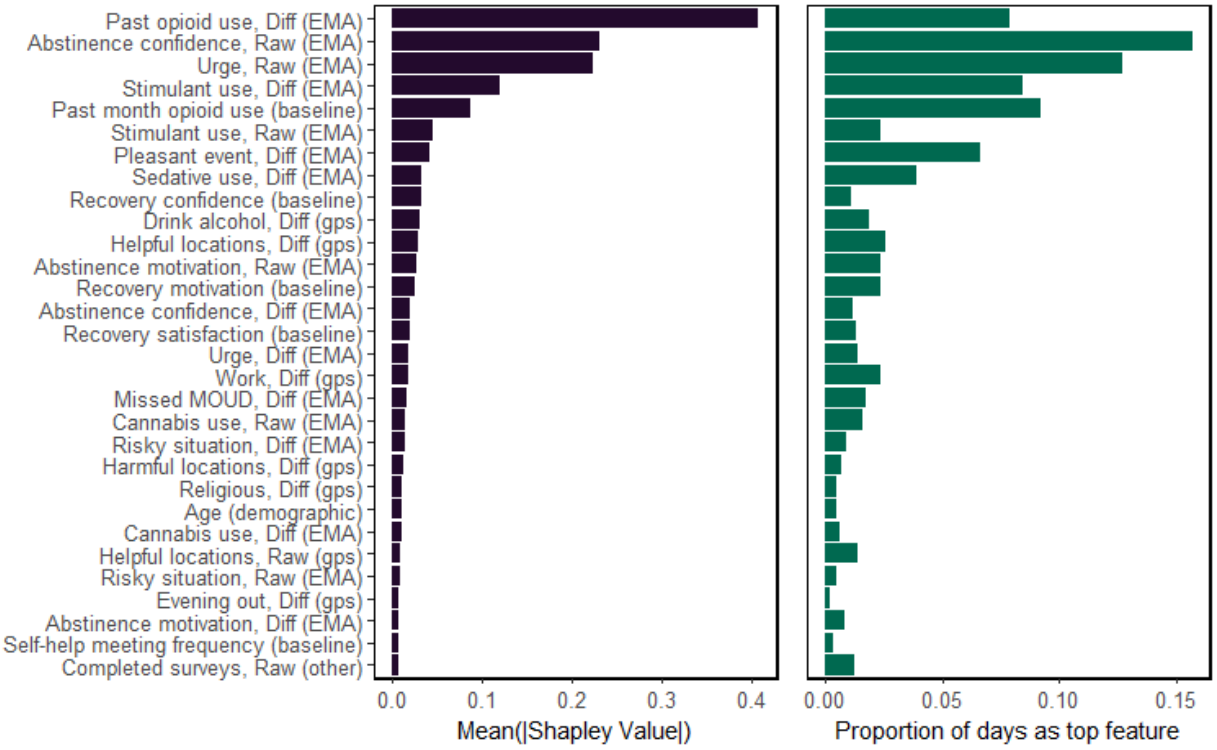
Past opioid use, Diff (EMA)
Abstinence confidence, Raw (EMA)
Urge, Raw (EMA)
Stimulant use, Diff (EMA)
Past month opioid use (baseline)
Stimulant use, Raw (EMA)
Pleasant event, Diff (EMA)
Sedative use, Diff (EMA)
Recovery confidence (baseline)
Drink alcohol, Diff (gps)
Helpful locations, Diff (gps)
Abstinence motivation, Raw (EMA)
Recovery motivation (baseline)
Abstinence confidence, Diff (EMA)
Recovery satisfaction (baseline)
Urge, Diff (EMA)
Work, Diff (gps)
Missed MOUD, Diff (EMA)
Cannabis use, Raw (EMA)
Risky situation, Diff (EMA)
Harmful locations, Diff (gps)
Religious, Diff (gps)
Age (demographic)
Cannabis use, Diff (EMA)
Helpful locations, Raw (gps)
Risky situation, Raw (EMA)
Evening out, Diff (gps)
Abstinence motivation, Diff (EMA)
Self-help meeting frequency (baseline)
Completed surveys, Raw (other)
0.0
0.1
0.2
0.3
0.4
Mean(|Shapley Value|)
0.00
0.05
0.10
0.15
Proportion of days as top feature

Figure 3

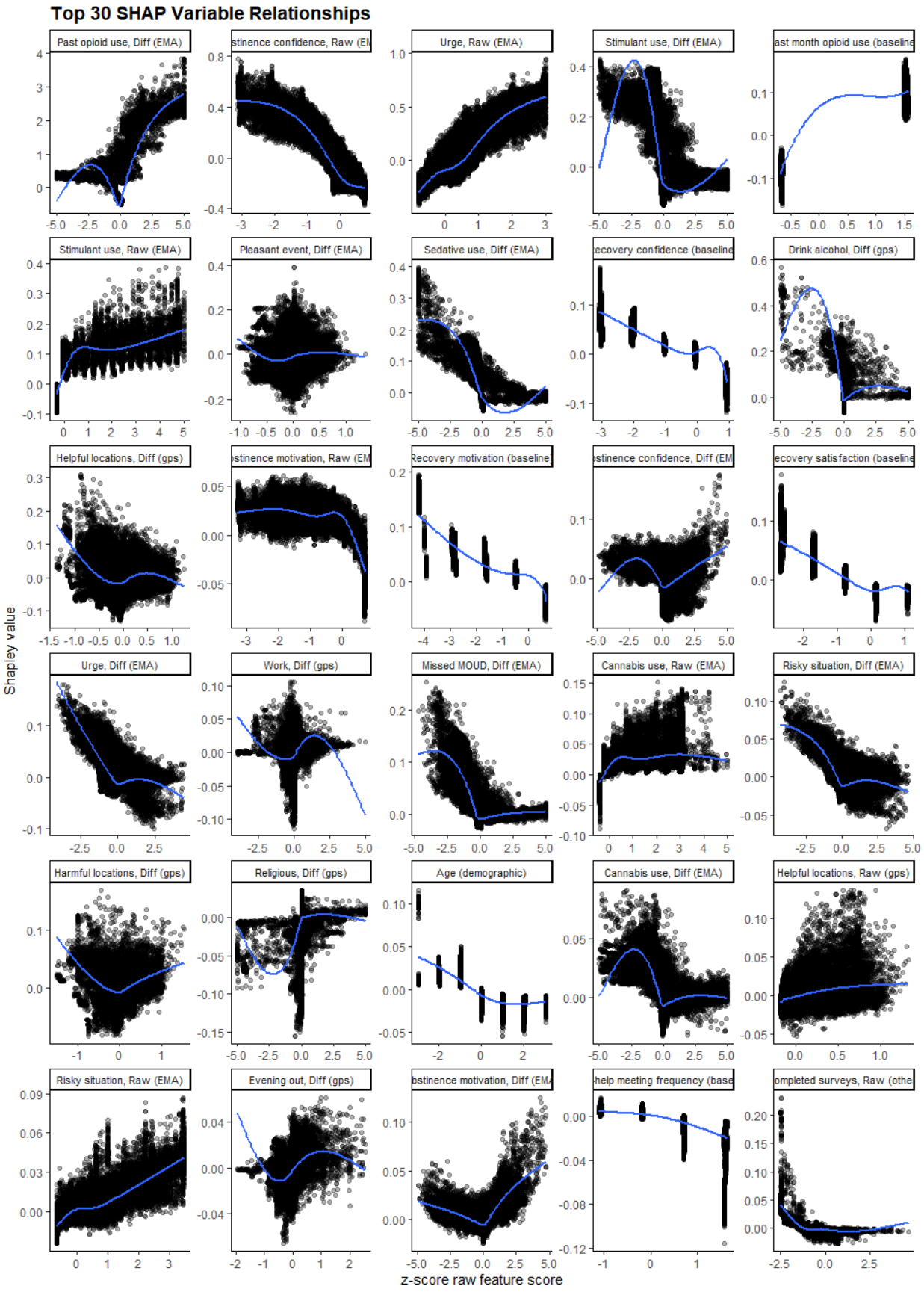

Figure 4: Feature importance partial dependence plots.

Discussion

- Discuss contrast between top global features and “actual” important features (i.e., what comes up day to day as being important). What does this mean for how we traditionally conceptualize feature importance?
- When discussing fairness - we don’t look at age. People over 65 is an important group to look at but they were not in our sample
- Decision not include misclassification cost analyses: acknowledge that discrimination and calibration don’t equate to clinical utilility. Decision analysis tools, such as net benefit curves could be used to quantify clinical benefit at relevant probability thresholds. Our model is designed to not be used to make decisions about whether or not to treat but to provide model feedback to individuals each day to help them monitor their risk for lapse. Therefore the decision has already been made to always treat. Should these models be used to inform when to deploy more cost-intensive interventions (e.g., communicating risk level to a therapist to initiate contact) these analyses will be important next steps.

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders : DSM-5-TR*. 5th edition, text revision. Washington, DC : American Psychiatric Association Publishing, 2022.

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

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2023). *Rstanarm: Bayesian Applied Regression Modeling via Stan*.

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

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

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.

Morisky, D. E., Green, L. W., & Levine, D. M. (1986). Concurrent and predictive validity of a self-reported measure of medication adherence. *Medical Care*, *24*(1), 67–74. https://doi.org/10.1097/00005650-198601000-00007

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

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers* (pp. 61–74). MIT Press.

Van Calster, B., Collins, G. S., Vickers, A. J., Wynants, L., Kerr, K. F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K. G., Hernandez-Boussard, T., Timmerman, D., McLernon, D. J., Van Smeden, M., & Steyerberg, E. W. (2025). Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance. *The Lancet Digital Health*, 100916. https://doi.org/10.1016/j.landig.2025.100916

Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J., Van Hoecke, S., & Demeester, T. (2021). Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine*, *111*, 101987. https://doi.org/10.1016/j.artmed.2020.101987

Wyant, K., Fronk, G. E., Yu, C., Punturieri, C. E., & Curtin, J. J. (under review). *Forecasting Risk of Alcohol Lapse up to Two Weeks in Advance using Time-lagged Machine Learning Models*.

Wyant, K., Sant'Ana, S. J. K., Fronk, G., & Curtin, J. J. (2024). Machine learning models for temporally precise lapse prediction in alcohol use disorder. *Psychopathology and Clinical Science*. https://doi.org/10.31234/osf.io/cgsf7