Dynamic lapse risk prediction in a national sample of individuals with opioid use disorder using personal sensing and machine learning

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 made available in this manuscripts supplemental materials. Finally, our data, 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 document changes made to the pre-registered report as follows:

1. Feature sets. 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 prior personal sensing work on alcohol use disorder by our group, technological constraints, and the need to balance feature diversity for capturing lapse complexity and minimizing participant burden. 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, due to Apple's strict restrictions on app access to communications, these data could only be collected from Android users. Thus, inclusion of these features would result in a model that could only be deployed within an android operating system. Daily video check-ins were only collected from the first handful of participants due to technical issues in the A-CHESS app. App usage data were sparse and unusable. This is likely due to us using a research version of the A-CHESS app not optimized for sustained engagement. Self-report measures can dramatically increase burden on the individual providing data. Monthly surveys were lengthy (20–30 minutes) and, based on prior work with alcohol use disorder, add no incremental predictive value beyond sensing data. However, individual differences in severity of use and stability of recovery at the time of intake have shown predictive signal [ndividual differences in severity of use and stability of recovery at the time of intake has]. We also retained key demographic variables known to influence treatment access and clinical outcomes.

2. Resampling method. Our 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 lower (N = 300). To maximize data use, we eliminated the held-out test set and relied on repeated cross-validation. We believe 6 repeats of cross-validation sufficiently mitigate optimization bias and provide generalizable performance estimates.

Participants

We nationally recruited 451 participants in early recovery from opioid use disorder receiving medication for opioid use disorder (MOUD) in 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 (sex, age, race, and ethnicity), and geographic location (urban, suburban, and rural residence). We required participants:

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

We did not exclude participants for comorbid substance use disorders or other psychopathologic conditions.

Procedure

All procedures were approved by the University of Wisconsin-Madison Institutional Review Board (Study #2019-0656). All participants provided written informed consent.

A-CHESS

Measures

[1]Adherence was defined as taking their monthly MOUD medication regularly or daily MOUD medication on most days or every day.

Individual Differences

Ecological Momentary Assessment

Geolocation Sensing

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 engineered from two types of data:

1. Static individual differences in demographic and OUD characteristics.
2. Dynamic EMA and geolocation sensing data.

- geolocation - within 50 meters of a known location

We imputed missing data using median imputation for numeric features and mode imputation for nominal features. We selected coarse median/mode methods for handling missing data due to the computational costs associated with more advanced forms of imputation (e.g., KNN imputation, multiple imputation). Importantly, our imputation calculations are done using only held-in data and can be applied to any new observation.

Prediction Windows and Labels

- Window starts 6am in participants own timezone (derived from gps)

Model Selection

Our performance metric for model selection and evaluation was area under the Receiver Operating Characteristic curve (auROC). .

The best model configuration was selected using 6 repeats of participant-grouped 5-fold cross-validation. Folds were stratified by a between-subject measure of our outcome (no lapse vs. any lapse).

Model configurations differed on:

1. Statistical algorithm (and hyperparameter tuning)
2. Resampling of minority outcome. All resampling was exclusively done with only held-in training data to prevent biasing performance estimates (Vandewiele et al., 2021).

We repeated the above process to select a best baseline model that was limited to features derived from individual differences in demographics and OUD characteristics.

Model Evaluation

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 performing model. We used weakly informative, data-dependent priors to regularize and reduce overfitting. Random intercepts were included for repeat and fold (nested within repeat).

- Model contrast

Using the same 30 held-out test sets, we calculated the median posterior probability and 95% Bayesian CI for auROC for each model separately by 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 (small town/rural vs. urban/suburban), 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

We calibrated our probabilities using Platt scaling. We calculated Brier scores to assess the accuracy of our raw and calibrated probabilities for our best model. Brier scores range from 0 (perfect accuracy) to 1 (perfect inaccuracy). We provide calibration plots for the raw and calibrated probabilities in the manuscript and histograms of predicted probabilities by true lapse outcome in the supplement.

Feature Importance

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, 10 participants were excluded due to evidence of careless responding on EMAs. Our final analysis

[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.

sample consisted of 300 participants. Participant demographic and OUD characteristics is presented in Table 1.

Table 1: Demographic and Clinical Characteristics

	N	%
Age		
18-21	2	0.7
22-25	7	2.3
26-35	108	36.0
36-45	106	35.3
46-55	60	20.0
56-65	13	4.3
Over 65	4	1.3
Gender		
Man	158	52.7
Woman	135	45.0
Non-binary	4	1.3
Not listed above	2	0.7
Orientation		
Straight, that is, not gay or lesbian	238	79.3
Lesbian or gay	16	5.3
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	79.0
Hispanic, Latino, or Spanish origin	25	8.3
Not listed above	5	1.7
Geographic Location		
small town/rural	62	20.7
urban/suburban	274	91.3
[illegible]teek	[illegible]	[illegible]
N = 300		

Adherence, Features, and Labels

Mean days on study across participants was 298 days (range 32-395 days). 83% of participants (250/300) 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, 313 daily geolocation points (range 20-825).

Our final feature set consisted of 640 features. The proportion of missing values across features was low (mean=.03, range = 0-.12). Across participants we generated a total of 88,973 day-level labels. Forty percent of participants (119/300) reported an opioid lapse while on study (mean=5.73, range 0-117). This resulted in 1.93% of the labels positive for lapse (1,720/88,973 labels).

Performance

The best 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 information gathered at intake (i.e., individual differences in demographic and OUD characteristics) to evaluate the incremental predictive value of adding dynamic features. The median posterior probability for the baseline model was 0.74 (95% CI [0.71, 0.77]). A Bayesian

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

model comparison revealed extremely strong evidence that the best model was more predictive than the baseline model (probability = 1.00).

Our fairness subgroup comparisons revealed no evidence that performance meaningfully differed by geographic location (probability = 0.68), weak evidence that performance differed by education (probability = 0.83), and strong evidence that performance differed by gender, income, and race and ethnicity (probabilities > 0.99). 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.

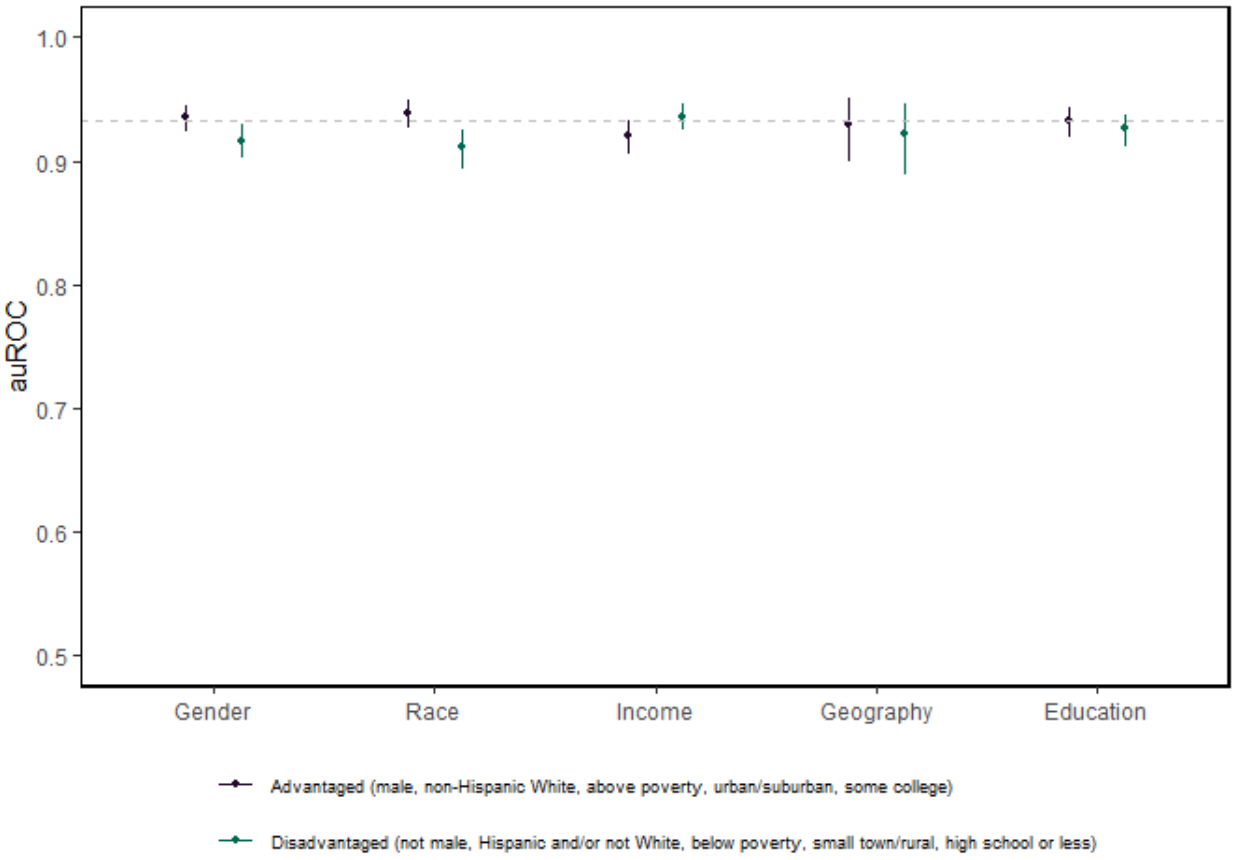

Figure 1: Posterior probabilities for area under the receiver operating curve (auROC) by demographic subgroup. auROC ranges from .5 (chance performance) to 1 (perfect performance). Subgroups advantaged in access to substance use treatment and outcomes (male, non-Hispanic White, above poverty, urban or surburban geographic 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, small town or rural geographic location, and high school education or less) are depicted in green. Overall model performance across groups is depicted as the dashed grey line.

Calibration

Our raw model predicted probabilities were generally well-calibrated (brier score = 0.015). We attempted to improve calibration with Platt scaling but results were comparable (brier score = 0.015). Calibration plots for the raw uncalibrated probabilities and calibrated probabilities with Platt scaline are presented in Figure 2. Histograms of predicted probabilities by true lapse outcome are available in the supplement.

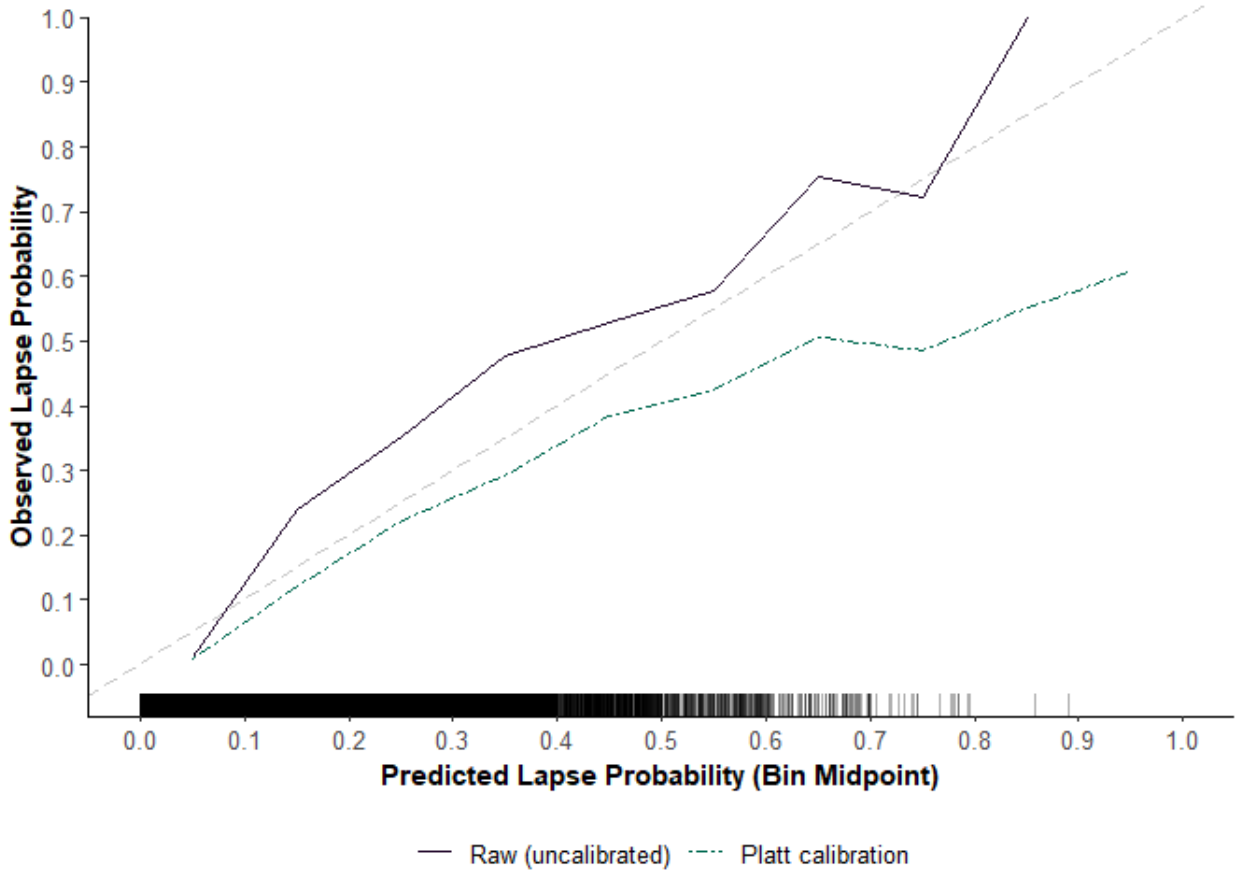

Figure 2: Calibration plots of raw and calibrated lapse probabilities. Predicted probabilities (x-axis) are binned into deciles. Observed lapse probability (y-axis) represents the proportion of actual lapses observed in each bin. The dashed diagonal represents perfect calibration. Points below the line indicate overestimation and points above the line indicate underestimation. Raw probabilities are depicted as dark purple lines Platt calibrated probabilities are depicted as green dashed lines.

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.

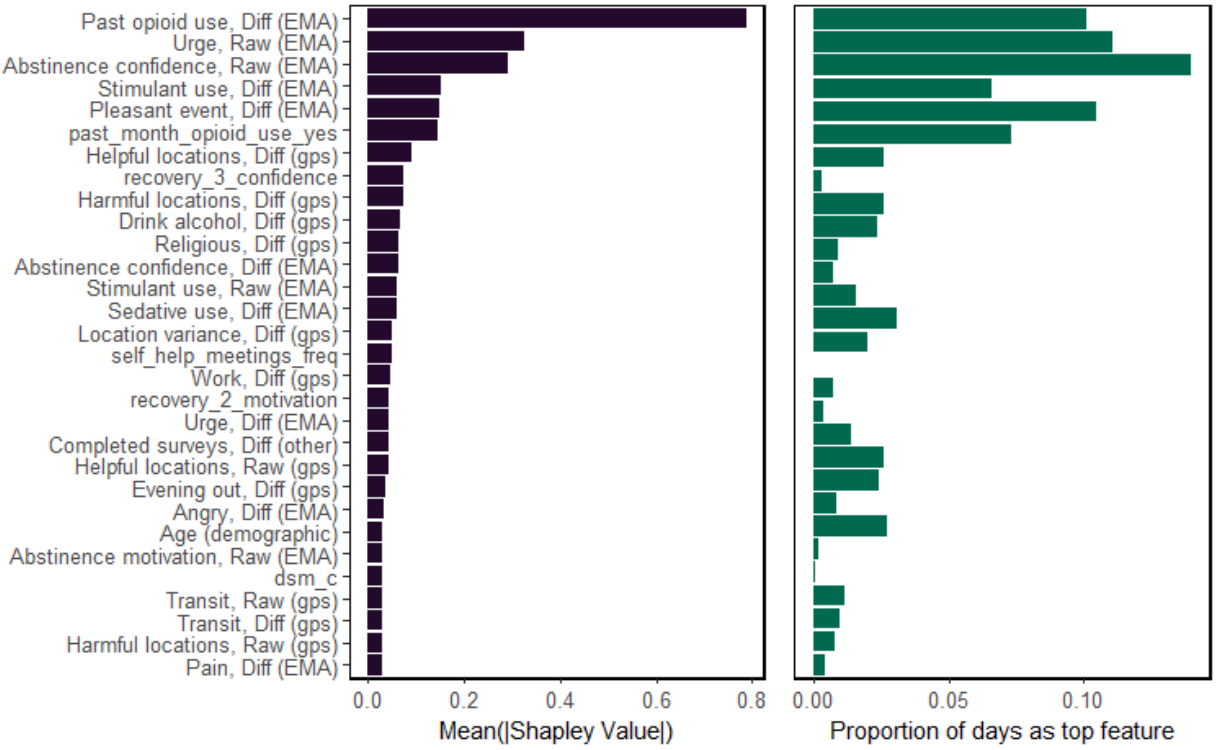
Past opioid use, Diff (EMA)
Urge, Raw (EMA)
Abstinence confidence, Raw (EMA)
Stimulant use, Diff (EMA)
Pleasant event, Diff (EMA)
past_month_opioid_use_yes
Helpful locations, Diff (gps)
recovery_3_confidence
Harmful locations, Diff (gps)
Drink alcohol, Diff (gps)
Religious, Diff (gps)
Abstinence confidence, Diff (EMA)
Stimulant use, Raw (EMA)
Sedative use, Diff (EMA)
Location variance, Diff (gps)
self_help_meetings_freq
Work, Diff (gps)
recovery_2_motivation
Urge, Diff (EMA)
Completed surveys, Diff (other)
Helpful locations, Raw (gps)
Evening out, Diff (gps)
Angry, Diff (EMA)
Age (demographic)
Abstinence motivation, Raw (EMA)
dsm_c
Transit, Raw (gps)
Transit, Diff (gps)
Harmful locations, Raw (gps)
Pain, Diff (EMA)
0.0
0.2
0.4
0.6
0.8
Mean(|Shapley Value|)
0.00
0.05
0.10
Proportion of days as top feature

Figure 3

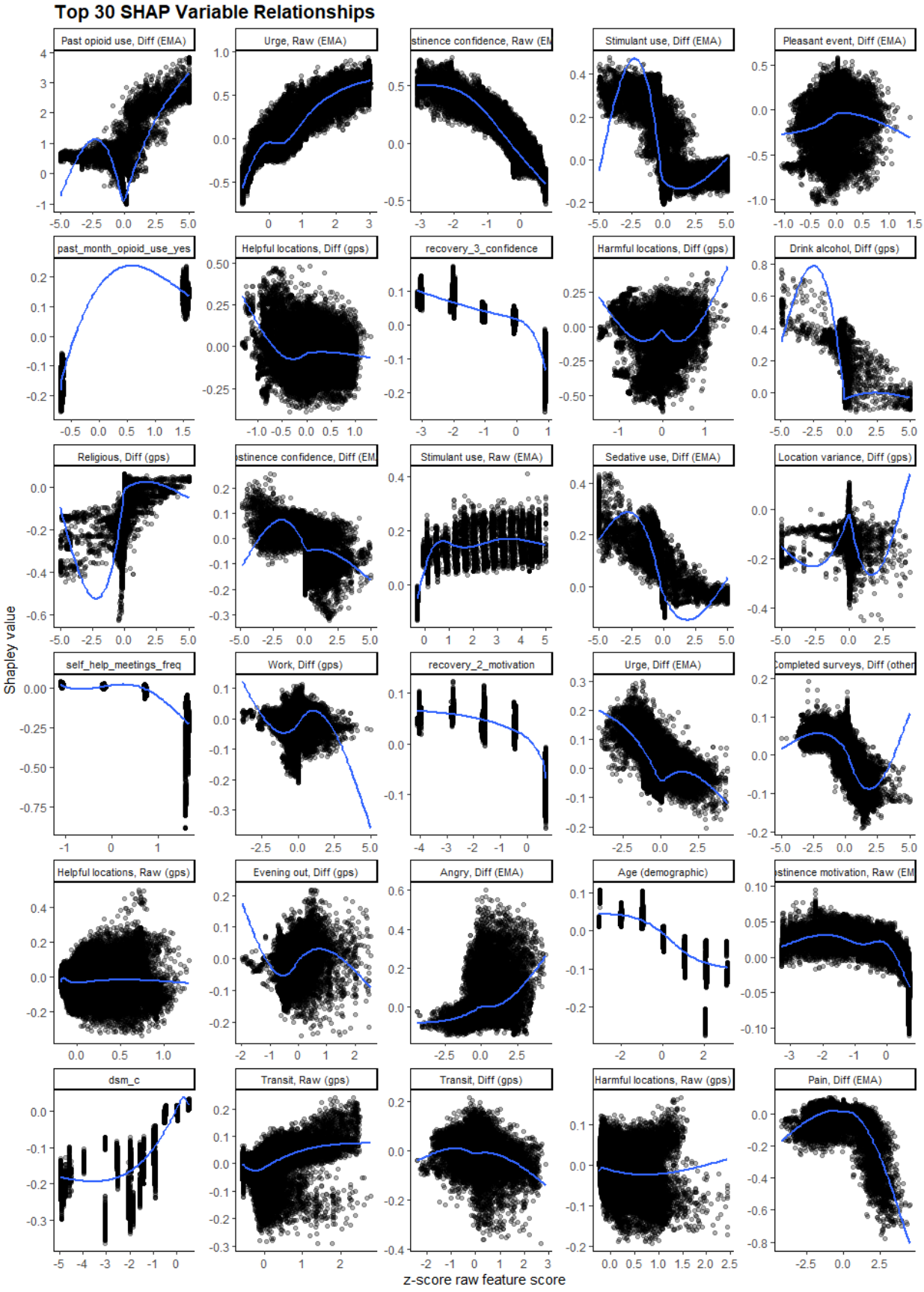

Figure 4: Feature importance partial dependence plots.

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

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*.

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

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