

**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 disorders Precision mental health

Introduction

Methods

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. 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/).

-TRIPOD+AI checklist (Collins et al., 2024)

Participants

Procedure

All procedures were approved by the University of Wisconsin-Madison Institutional Review Board. All participants provided written informed consent.

Measures

Individual Characteristics

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).

Model Selection

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. Feature set
3. Resampling of minority outcome

Features had missing values if the participant did not respond to the relevant EMA question

during the associated scoring epoch. The proportion of missing values across features was low .

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.

Model Evaluation

We used a Bayesian hierarchical generalized linear model to estimate the posterior distribution and 95% credible intervals (CI) 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).

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.

- following best practices for model reporting (Van Calster et al., 2025)

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 also provide calibration plots for the raw and calibrated probabilities.

Feature Importance

Results

Participants

Data were collected from April 2021 through December 2024. A total of 336 participants were eligible, consented, and remained on study for at least one month. Our final analysis sample consisted of 301 participants. Ten participants were excluded due to unusually low adherence, 13 participants were excluded due to insufficient context data for gps points, 4 participants were excluded due to frequent lapses suggesting they no longer had a goal of abstinence, 8 participants were excluded due to suspected fraud and careless responding. A summary of 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	35.9
36-45	107	35.5
46-55	60	19.9
56-65	13	4.3
Over 65	4	1.3
Gender		
Man	160	53.2
Woman	134	44.5
Non-binary	4	1.3
Not listed above	2	0.7
Orientation		
Straight, that is, not gay or lesbian	238	79.1
Lesbian or gay	16	5.3
Bisexual	40	13.3
Not sure	1	0.3
Not listed above	4	1.3
Race and Ethnicity (select all that apply)		
American Indian/Alaskan Native	17	5.6
Asian	4	1.3
Black/African American	46	15.3
Native Hawaiian/Other Pacific Islander	3	1.0
White/Caucasian	237	78.7
Hispanic, Latino, or Spanish origin	24	8.0
Not listed above	5	1.7
Education		
8th grade or less	6	2.0
Some high school, but did not graduate	25	8.3
High school graduate	200	66.8
Some college	51	16.9
College graduate	19	6.3
Postgraduate	0	0.0
N = 301		

Adherence and Labels

Median days on study across participants was 356 days (range 32-395 days). 83% of participants 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

Across participants we generated a total of 89,125 labels. Thirty-nine percent of participants reported an opioid lapse while on study (mean=5.34, range 0-76). This resulted in 1.80% of the labels positive for lapse (1,608/89,125 labels).

Performance

The best model configuration was a down-sampled xgboost statistical algorithm that used only the dynamic feature set (i.e., features from EMA and geolocation). 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 baseline (demographics and individual differences in OUD characteristics) to evaluate the 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 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.

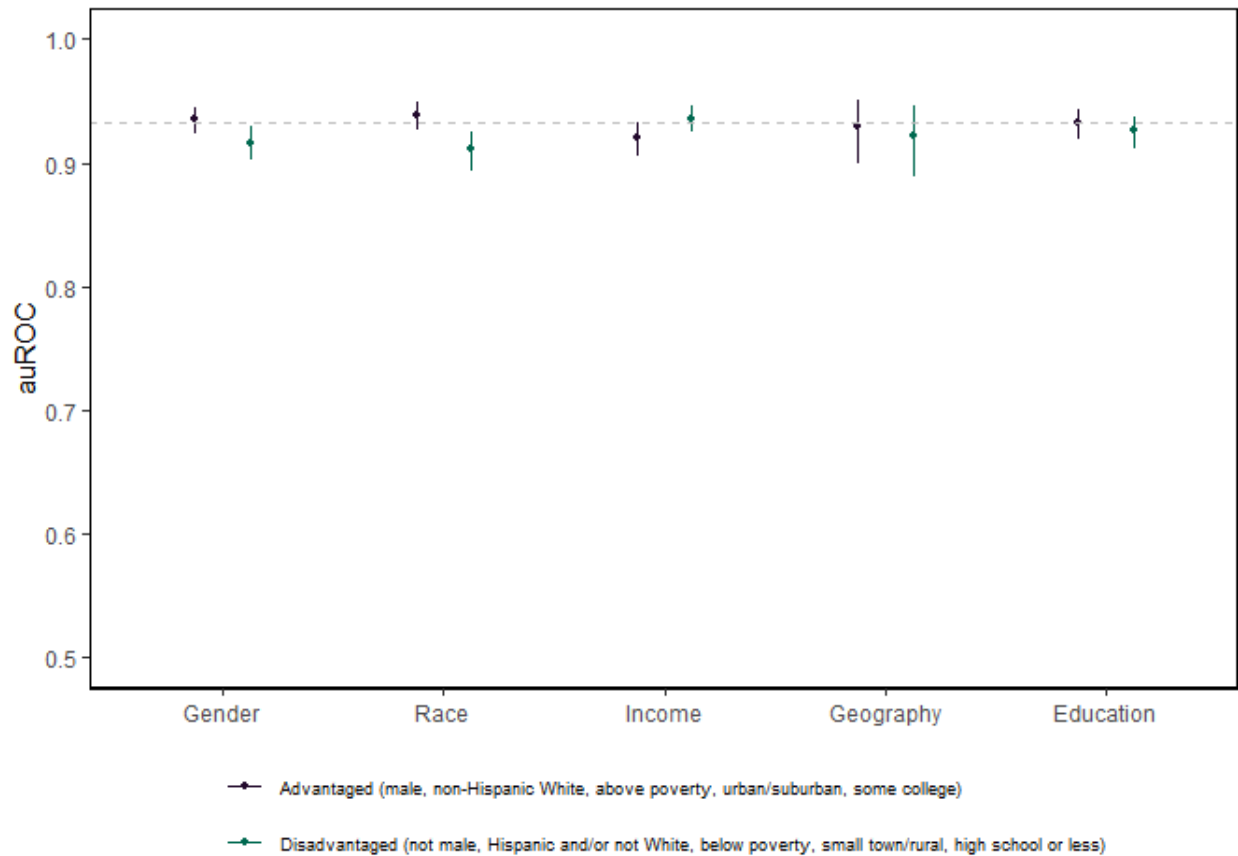


Figure 1: Posterior probabilities for area under the receiver operating curve (auROC) by demographic subgroup. auROC ranges from .5 (chance performance) to 1 (perfect performance). Subgroups advantaged in access to substance use treatment and outcomes (male, non-Hispanic White, above poverty, urban or suburban 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 **fig-cal**. Histograms of predicted probabilities by true lapse outcome are available in the supplement.

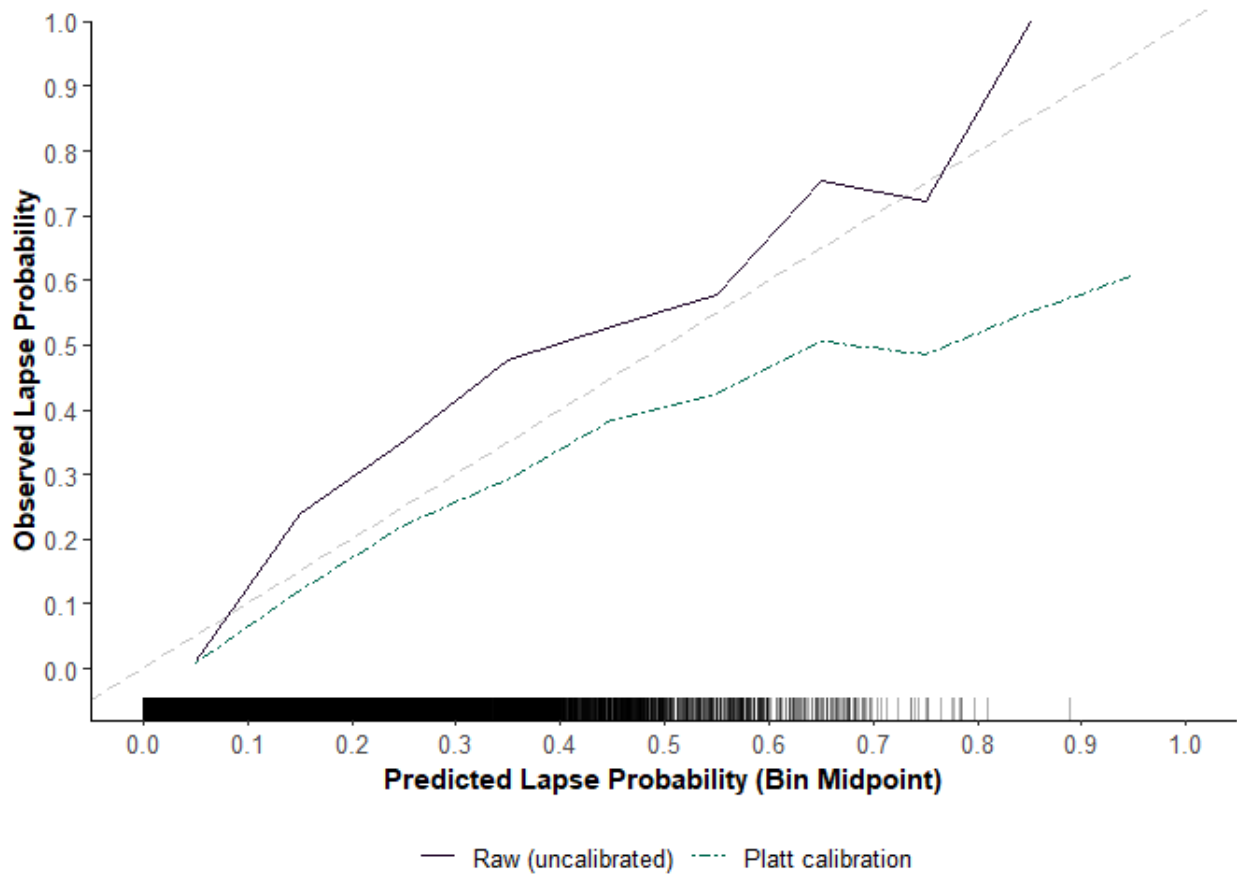


Figure 2: Calibration plots of raw and calibrated lapse probabilities. Predicted probabilities (x-axis) are binned into deciles. Observed lapse probability (y-axis) represents the proportion of actual lapses observed in each bin. The dashed diagonal represents perfect calibration. Points below the line 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

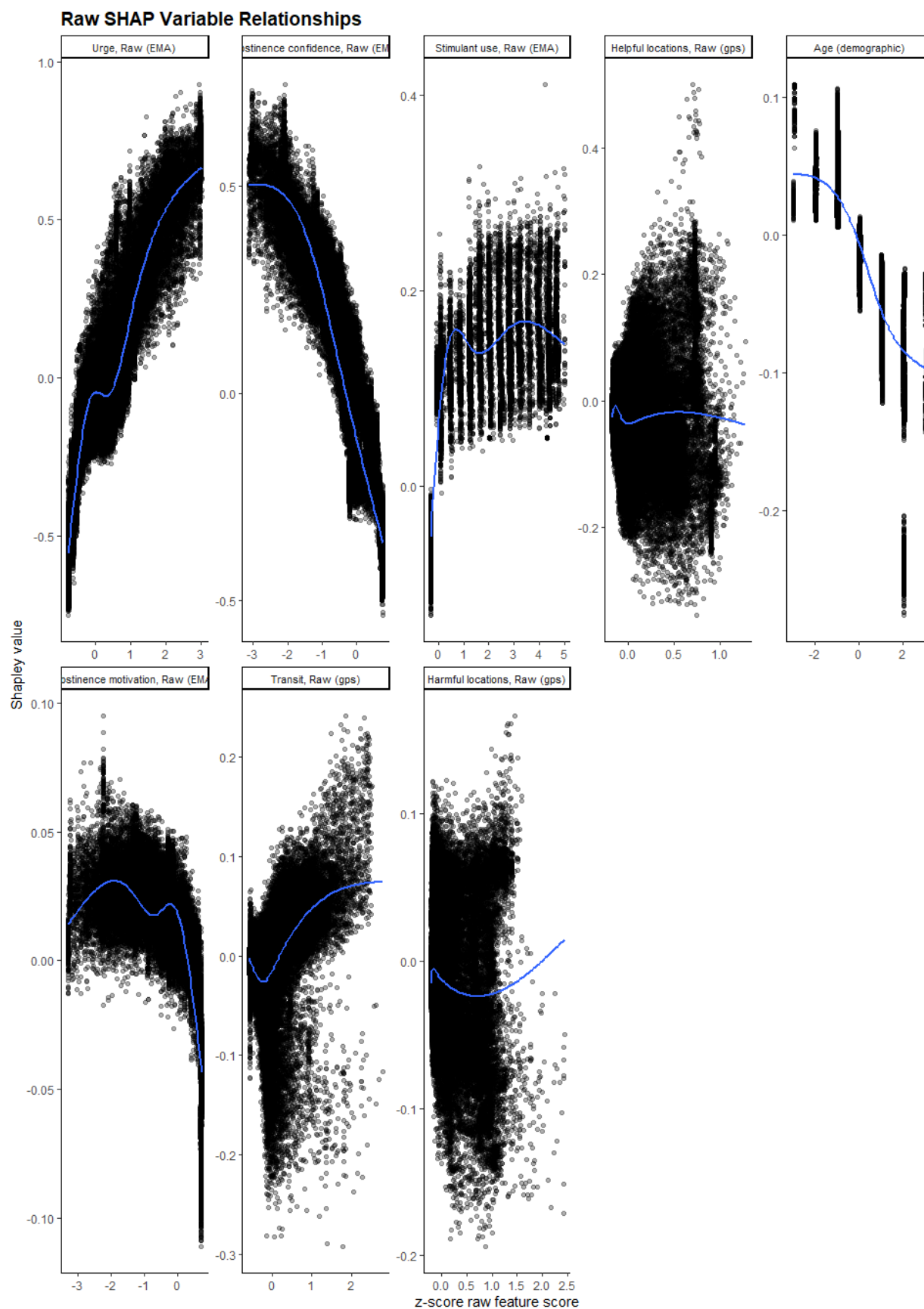


Figure 3: Raw feature importance partial dependence plots.

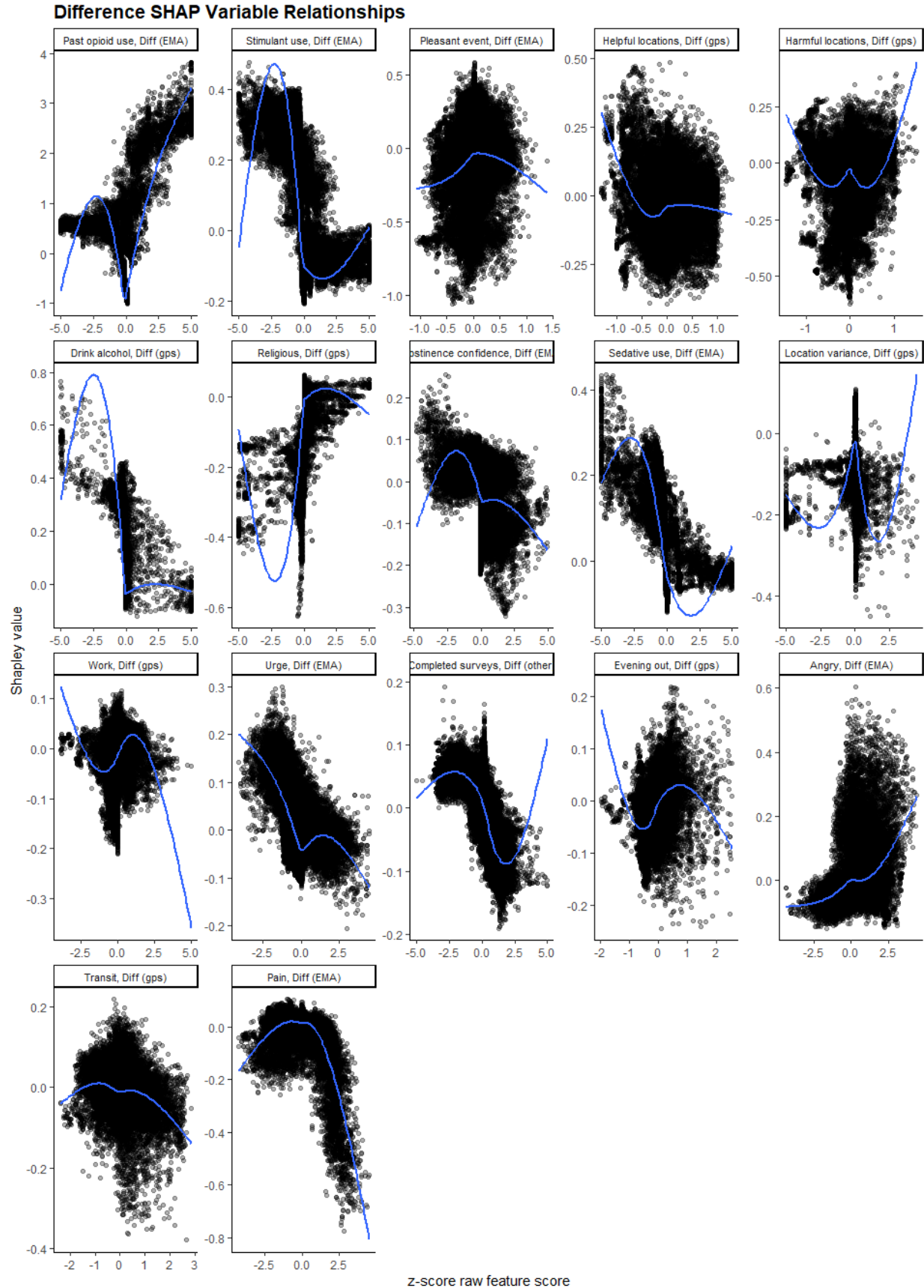


Figure 4: Difference feature importance partial dependence plots.

Discussion

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*.
- 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>