Lagged Predictions of Next Day Alcohol Use for Adaptive and Personalized Continuing Care

Kendra Wyant

Gaylen E. Fronk

Jiachen Yu

John J. Curtin

2025-01-27

Abstract

We evaluated machine learning models predicting future alcohol lapses within 24-hour prediction windows that were systematically lagged further into the future (1 day, 3 days, 1 week, and 2 weeks). We engineered features from 4x daily ecological momentary assessment. Participants (N=151; 51% male; mean age=41; 87% White, 97% Non-Hispanic) in early recovery from alcohol use disorder provided data for up to three months. We used nested cross-validation to select and evaluate models. Median posterior probabilities for auROCs were high (0.84–0.90). Still, performance declined with increasing lags (probabilities > .98). Models also performed worse for disadvantaged groups (not White vs. non-Hispanic White, below poverty vs. above poverty, female vs. male; probabilities > .97). This study demonstrates the feasibility of predicting next-day alcohol lapses up to two weeks into the future. This advanced notice offers time to implement support options not immediately available. However, fairness concerns remain and are discussed further in the paper.

Source: [Article Notebook](https://jjcurtin.github.io/study_lag/index.qmd.html)

Source: [Article Notebook](https://jjcurtin.github.io/study_lag/index.qmd.html)

# Introduction

Alcohol and other substance use disorders (SUDs) are serious chronic conditions, characterized by high relapse rates(McLellan et al. 2000; Dennis and Scott 2007), substantial co-morbidity with other physical and mental health problems(Substance Abuse and Mental Health Services Administration n.d.; Dennis and Scott 2007), and an increased risk of mortality (Hedegaard et al. n.d.; Centers for Disease Control and Prevention (CDC) n.d.). Too few individuals receive medications or clinician-delivered interventions to help them initially achieve abstinence and/or reduce harms associated with their use (Substance Abuse and Mental Health Services Administration n.d.). Yet, this problem is even worse for subsequent continuing care during SUD recovery. Continuing care, including both risk monitoring and ongoing support, is the gold standard for managing chronic health conditions such as diabetes, asthma, and HIV. Yet, continuing care for SUDs is largely lacking despite ample evidence that SUDs are chronic, relapsing conditions (Substance Abuse and Mental Health Services Administration n.d.; Stanojlović and Davidson 2021; Socías, Volkow, and Wood 2016).

When available, an important focus of continuing care during SUD recovery is the prevention of lapses (i.e., single instances of goal-inconsistent substance use) and full relapse back to problematic use (Marlatt and Gordon 1985; Witkiewitz and Marlatt 2004). Critically, the risk factors that instigate lapses during recovery are individualized, numerous, dynamic, interactive, and non-linear (Witkiewitz and Marlatt 2007; Brandon, Vidrine, and Litvin 2007). Therefore, the optimal supports to address these risk factors and encourage continued, successful recovery vary both across individuals and within an individual over time. Given this, continuing care could benefit greatly from a precision mental health approach that seeks to provide the right support to the right individual at the right time, every time (Bickman, Lyon, and Wolpert 2016; DeRubeis 2019; Kranzler and McKay 2012). However, such monitoring and personalized support must also be highly scalable to address the substantial unmet need for SUD continuing care.

Recent advances in both smartphone sensing (Mohr, Zhang, and Schueller 2017) and machine learning (Hastie, Tibshirani, and Friedman 2009) hold promise as a scalable foundation for monitoring and personalized support during SUD recovery. Smartphone sensing approaches (e.g., ecological momentary assessment, geolocation sensing) can provide the frequent, longitudinal measurement of proximal risk factors that is necessary for prediction of future lapses with high temporal precision. Ecological momentary assessment (EMA) may be particularly well-suited for lapse prediction because it can provide privileged access to the subjective experiences (e.g., craving, affect, stress, motivation, self-efficacy) that are targets for change in evidence based approaches for relapse prevention (Marlatt and Gordon 1985; Witkiewitz and Marlatt 2004; Bowen et al. 2021). Furthermore, individuals with SUDs have found EMA to be acceptable for sustained measurement for up to a year with relatively high compliance (Wyant et al. 2023; Moshontz et al. 2021), suggesting that this method is feasible for long-term monitoring throughout SUD recovery.

Machine learning models are well-positioned to use EMAs as inputs to provide temporally precise prediction of the probability of future lapses with sufficiently high performance to support decisions about interventions and other supports for specific individuals. These models can handle the high dimensional feature sets that may result from feature engineering densely sampled raw EMA over time (Wyant et al. 2024). They can also accommodate non-linear and interactive relationships between features and lapse probability. And rapid advances in the tools for interpretable machine learning (e.g, Shapley values (Lundberg and Lee 2017)) now allow us to probe these models to understand which risk features contribute most strongly to a lapse prediction for a specific individual at a specific moment in time. Interventions, supports, and/or lifestyle adjustments can then be personalized to address these risks following from our understanding about relapse prevention.

Preliminary research is now emerging that uses features derived from EMAs in machine learning models to predict the probability of future alcohol use Wyant et al. (2024). 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 people. And perhaps most importantly, Wyant et al. (2024) demonstrated that machine learning models using EMA can provide predictions with very high temporal precision at clinically implementable levels of performance. Specifically, they developed models that predict lapses in the immediate future (i.e., the next day and even the next hour) with area under the receiver operating characteristic curve of 0.91 and 0.93, respectively.

Wyant et al. (2024)’s next day lapse prediction model can provide personalized support recommendations to address immediate risks for possible lapses in that next day. Features derived from past EMAs can be updated in the early morning to yield the predicted lapse probability for an individual that day. Personalized supports that target the top features contributing to that prediction can be provided to assist them that day. For example, if predicted lapse probability is high due to recent frequent craving, they could be reminded about the benefits of urge surfing or distracting activities during brief periods when cravings arise. Conversely, guided relaxation techniques could be recommended if lapse probability was high due to recent past and anticipated stressors that day. Patients could also be assisted to implement any of these recommendations by videos or other tools within a digital therapeutic. Curtin and colleagues are currently evaluating outcomes associated with the use of this “smart” (machine learning guided) monitoring and personalized support system for patients in recovery from alcohol use disorder (Wyant and Curtin in prep).

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

In this study, we evaluated the performance of machine learning models that predict the probability of future lapses within 24-hour prediction windows that were systematically lagged further into the future. We considered several meaningful lags for these prediction windows: 1 day, 3 days, 1 week, and 2 weeks. We conducted pre-registered analyses of both the absolute performance of these lagged models and their relative performance compared to a baseline model that predicted lapse probability in the immediate next day (i.e., no lag). In addition to the aggregate performance of these models, we also evaluated algorithmic fairness by comparing model performance across important subgroups that have documented disparities in treatment access and/or outcomes. These include comparisons by race/ethnicity (Pinedo 2019; Kilaru et al. 2020), income (Olfson et al. 2022), and sex at birth (Greenfield et al. 2007; Kilaru et al. 2020). Finally, we calculated Shapley values for feature categories defined by EMA items to better understand how these models make their prediction and how these features can be used to recommend personalized supports.

# Methods

## Transparency and Openness

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

## Participants

We recruited participants in early recovery (1-8 weeks of abstinence) from moderate to severe alcohol use disorder in Madison, Wisconsin, US for a three month longitudinal study. One hundred fifty one participants were included in our analyses. We used data from all participants included in our previous study (see (Wyant et al. 2024) for enrollment and disposition information). 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 that participants:

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

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

## Procedure

Participants completed five study visits over approximately three months. After an initial phone screen, participants attended an in-person screening visit to determine eligibility, complete informed consent, and collect self-report measures. Eligible, consented participants returned approximately one week later for an intake visit. Three additional follow-up visits occurred about every 30 days that participants remained on study. Participants were expected to complete four daily EMAs 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). Participants could earn up to $150/month if they completed all study visits, had 10% or less missing EMA data and opted in to provide data for other personal sensing data streams.

## Measures

### Ecological Momentary Assessments

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

On all EMAs, participants reported dates/times of any unreported past alcohol use. Next, participants rated the maximum intensity of recent (i.e., since last EMA) experiences of craving, risky situations, stressful events, and pleasant events. Finally, participants rated their current affect on two bipolar scales: valence (Unpleasant/Unhappy to Pleasant/Happy) and arousal (Calm/Sleepy to Aroused/Alert).

On the first EMA each day, participants also rated the likelihood of encountering risky situations and stressful events in the next week and the likelihood that they would drink alcohol in the next week (i.e., abstinence self-efficacy).

### Individual Characteristics

We collected self-report information about demographics (age, sex, race, ethnicity, education, marital status, employment, and income) and AUD symptom count to characterize our sample. Demographic information was also included as features in our models and a subset (sex, race, ethnicity, and income) used for model fairness analyses.

As part of the aims of the parent project we collected many other trait and state measures throughout the study. A complete list of all measures can be found on our study’s OSF page.

## Data Analytic Strategy

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

### Predictions

A *prediction time point* is the hour at which our model calculated a predicted probability of a lapse. All available data up until, but not including, the prediction time point (i.e., the scoring epoch) was used to generate these predictions (see Feature Engineering section). Prediction time points were updated hourly ([Figure 1](#fig-methods), Panel A). The first prediction time point for each participant was 24 hours from midnight on their study start date. This ensured at least 24 hours of past EMAs for future lapse prediction at these first time points. Subsequent prediction time points for each participant repeatedly rolled hour-by-hour until the end of their study participation.

The *prediction window* is the window of time in which a lapse might occur. The prediction window width for all models was 24 hours (i.e., models predicted the probability of a lapse occurring within a 24-hour window). Prediction windows rolled forward hour-by-hour with the prediction time point ([Figure 1](#fig-methods), Panel B).

Finally, there were five possible *lag times* between the prediction time point and start of the prediction window. A prediction window either started immediately after the prediction time point (no lag) or was lagged by 1 day, 3 days, 1 week, or 2 weeks ([Figure 1](#fig-methods), Panel B).

Therefore, our models provided hour-by-hour probabilities of an alcohol lapse in the next 24 hours pushed out up to 2 weeks into the future. For example, for a participant on their 30th day on study, the model would use all 30 days of data and make predictions about alcohol use starting at 5 different times (immediately, 1 day from now, 3 days from now, 1 week from now, and 2 weeks from now).

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

Source: [Make All Figures for Main Manuscript](https://jjcurtin.github.io/study_lag/notebooks\mak_figures-preview.html#cell-fig-methods)

### Labels

The start and end date/time of past drinking episodes were reported on the first EMA item. A prediction window was labeled *lapse* if the start date/hour of any drinking episode fell within that window. A window was labeled *no lapse* if no alcohol use occurred within that window +/- 24 hours. If no alcohol use occurred within the window but did occur within 24 hours of the start or end of the window, the window was excluded. [[1]](#footnote-36)

We ended up with a total of 274,179 labels for our baseline (no lag) model, 270,911 labels for our 1-day lagged model, 264,362 labels for our 3-day lagged model, 251,458 labels for our 1-week lagged model, and 228,420 labels for our 2-week lagged model.

### Feature Engineering

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

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

1. *Prediction window*: We dummy-coded features for day of the week for the start of the prediction window.
2. *Demographics*: We created quantitative features for age and personal income, and dummy-coded features for sex, race/ethnicity, marital status, education, and employment.
3. *Previous EMA responses*: We created raw EMA and change features for varying scoring epochs (i.e., 12, 24, 48, 72, and 168 hours) before the prediction time point for all EMA items. Raw features included min, max, and median scores for each EMA item across all EMAs in each epoch for that participant. We calculated change features by subtracting the participants’ overall mean score for each EMA item (using all EMAs collected before the start of the prediction window) from the associated raw feature. We also created raw and change features based on the most recent response for each EMA question and raw and change rate features from previously reported lapses and number of completed EMAs.

Other generic feature engineering steps included imputing missing data (median imputation for numeric features, mode imputation for nominal features) and removing zero and near-zero variance features as determined from held-in data (see Cross-validation section below).

### Model Training and Evaluation

#### Model Configurations

We trained and evaluated five separate classification models: one baseline (no lag) model and one model for 1 day, 3 day, 1 week, and 2 week lagged predictions. We considered four well-established statistical algorithms (elastic net, XGBoost, regularized discriminant analysis, and single layer neural networks) that vary across characteristics expected to affect model performance (e.g., flexibility, complexity, handling higher-order interactions natively) (Kuhn and Johnson 2018).

Candidate model configurations differed across sensible values for key hyperparameters. They also differed on outcome resampling method (i.e., no resampling and up-sampling and down-sampling of the outcome using majority/no lapse to minority/lapse ratios ranging from 1:1 to 5:1).

#### Cross-validation

We used participant-grouped, nested cross-validation for model training, selection, and evaluation with auROC. auROC indexes the probability that the model will predict a higher score for a randomly selected positive case (lapse) relative to a randomly selected negative case (no lapse). Grouped cross-validation assigns all data from a participant as either held-in or held-out to avoid bias introduced when predicting a participant’s data from their own data. We used 1 repeat of 10-fold cross-validation for the inner loops (i.e., *validation* sets) and 3 repeats of 10-fold cross-validation for the outer loop (i.e., *test* sets). Best model configurations were selected using median auROC across the 10 validation sets. Final performance evaluation of those best model configurations used median auROC across the 30 test sets.

#### Bayesian Model

We used a Bayesian hierarchical generalized linear model to estimate the posterior probability distributions and 95% Bayesian credible intervals (CIs) from the 30 held-out test sets for our five best models. Following recommendations from the rstanarm team and others (RStudio Team 2020; Gabry and Goodrich 2023), we used the rstanarm default autoscaled, weakly informative, data-dependent priors that take into account the order of magnitude of the variables to provide some regularization to stabilize computation and avoid over-fitting.[[2]](#footnote-41) We set two random intercepts to account for our resampling method: one for the repeat, and another for the fold nested within repeat. We specified two sets of pre-registered contrasts for model comparisons. The first set compared each lagged model to the baseline no lag model (no lag vs. 1-day lag, no lag vs. 3-day lag, no lag vs. 1-week lag, no lag vs. 2-week lag). The second set compared adjacently lagged models (1-day lag vs. 3-day lag, 3-day lag vs. 1-week lag, 1-week lag vs. 2-week lag). auROCs were transformed using the logit function and regressed as a function of model contrast.

From the Bayesian model we obtained the posterior distribution (transformed back from logit) and Bayeisan CIs for all five models. To evaluate our models’ overall performance we report the median posterior probability for auROC and Bayesian CIs. This represents our best estimate for the magnitude of the auROC parameter for each model. If the confidence intervals do not contain .5 (chance performance), this suggests our model is capturing signal in the data.

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

#### Fairness Analyses

We calculated the median posterior probability and 95% Bayesian CI for auROC for each model separately by race and ethnicity (not White vs. non-Hispanic White), income (below poverty vs. above poverty[[3]](#footnote-43)), and sex at birth (female vs. male). We conducted Bayesian group comparisons to assess the likelihood that each model performs differently by group. We summarize the differences in posterior probabilities for auROC across models. Individual Bayesian fairness contrasts for all five models are available in the supplement.[[4]](#footnote-44)

#### Feature Importance

We calculated Shapley values in log-odds units for binary classification models from the 30 test sets to provide a description of the importance of categories of features across our five models (Lundberg and Lee 2017). We averaged the three Shapley values for each observation for each feature (i.e., across the three repeats) to increase their stability. An inherent property of Shapley values is their additivity, allowing us to combine features into feature categories. We created separate feature categories for each of the nine EMA questions, and the rate of past alcohol use. We calculated the local (i.e., for each observation) importance for each category of features by adding Shapley values across all features in a category, separately for each observation. We calculated global importance for each feature category by averaging the absolute value of the Shapley values of all features in the category across all observations. These local and global importance scores based on Shapley values allow us to contextualize relative feature importance for each model.

# Results

## Demographic and Lapse Characteristics

[Table 1](#tbl-demohtml) provides a detailed breakdown of the demographic and lapse characteristics of our sample (N = 151).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Demographic and Lapse Characteristics   | var | 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 |  |  |  | | DSM-5 AUD Symptom Count |  |  | 8.9 | 1.9 | 4-11 | | 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 | |

Source: [Make All Tables for Main Manuscript](https://jjcurtin.github.io/study_lag/notebooks\mak_tables-preview.html#cell-tbl-demohtml)

## Model Evaluation

Histograms of the full posterior probability distributions for auROC for each model are available in the supplement. The median auROCs from these posterior distributions were 0.90 (no lag), 0.88 (1-day lag), 0.87 (3-day hour lag), 0.86 (1-week lag), and 0.84 (2-week lag). These values represent our best estimates for the magnitude of the auROC parameter for each model. The 95% Bayesian CI for the auROCs for these models were relatively narrow and did not contain 0.5: no lag [0.88-0.92], 1-day lag [0.86-0.90], 3-day lag [0.85-0.89], 1-week lag [0.84-0.88], 2-week lag [0.81-0.86].

## Model Comparisons

[Table 2](#tbl-contrast) presents the median difference in auROC, 95% Bayesian CI, and posterior probability that that the auROC difference was greater than 0 for all baseline and adjacent lag contrasts. There was strong evidence (probabilities > .98) that the lagged models performed worse than the baseline (no lag) model, with average drops in auROC ranging from 0.02-0.06, and previous adjacent lagged model, with average drops in auROC ranging from 0.01-0.02.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Median difference in auROC, 95% Bayesian credible interval (CI), and posterior probability that that the auROC difference was greater than 0 for all baseline and adjacent lag contrasts.   | Contrast | Median | Bayesian CI | Probability | | --- | --- | --- | --- | | Baseline Contrasts |  |  |  | | No lag vs. 1 day | 0.02 | [0.013, 0.027] | 1 | | No lag vs. 3 days | 0.032 | [0.025, 0.04] | 1 | | No lag vs. 1 week | 0.043 | [0.035, 0.051] | 1 | | No lag vs. 2 weeks | 0.064 | [0.054, 0.074] | 1 | | Adjacent Contrasts |  |  |  | | 1 day vs. 3 days | 0.012 | [0.005, 0.02] | 0.998 | | 3 days vs. 1 week | 0.01 | [0.003, 0.018] | 0.986 | | 1 week vs. 2 weeks | 0.021 | [0.012, 0.03] | 1 | | Fairness Contrasts (No Lag) |  |  |  | | male vs. female | 0.042 | [0.027, 0.058] | 1 | | non-Hispanic White vs. not White | 0.215 | [0.057, 0.422] | 0.991 | | above poverty vs. below poverty | 0.042 | [0.007, 0.078] | 0.978 | | Fairness Contrasts (2-week Lag) |  |  |  | | male vs. female | 0.097 | [0.074, 0.12] | 1 | | non-Hispanic White vs. not White | 0.116 | [0.069, 0.169] | 1 | | above poverty vs. below poverty | 0.107 | [0.059, 0.157] | 0.999 | |

Source: [Make All Tables for Main Manuscript](https://jjcurtin.github.io/study_lag/notebooks\mak_tables-preview.html#cell-tbl-contrast)

## Fairness Analyses

[Table 2](#tbl-contrast) presents the median difference in auROC, 95% Bayesian CI, and posterior probability that that the auROC difference was greater than 0 for our no lag and 2-week lag models for the three fairness contrasts: race/ethnicity (not White; *N* = 20 vs. Non-Hispanic White; *N* = 131), sex at birth (female; *N* = 74 vs. male; *N* = 77), and income (below poverty; *N* = 18 vs. above poverty; *N* = 133). Individual Bayesian fairness contrasts for all five models are available in the supplement. There was strong evidence (probabilities > .97) that our models performed better for the advantaged groups (White, male, above poverty) compared to the disadvantaged groups (not-White, female, below poverty). On average there was a median decrease in auROC of 0.15 (range 0.11-0.27) for participants who were not White compared to non-Hispanic White participants. On average there was a median decrease in auROC of 0.05 (range 0.04-0.10) for female participants compared to male participants. On average there was a median decrease in auROC of 0.06 (range 0.04-0.11) for participants below the federal poverty line compared to participants above the federal poverty line.

## Feature Importance

The top three globally important (i.e., highest mean |Shapley value|) feature categories for all models were past use, future efficacy, and craving. This was also consistent across demographic groups (plots of global feature importance by demographic group are availble for the no lag and two week lag models in the supplement). Panel A of [Figure 2](#fig-3) shows the relative ranking of feature categories for the no lag and 2-week lag models. Global feature importance plots for all 5 modes is included in the supplement. In the supplement we also provide global feature importance plots for the no lag and 2-week lag model separately by fairness contrast. Panel B of [Figure 2](#fig-3) shows the variation (min and max) in local feature importance for each EMA item for the no lag and 2-week lag models.

|  |
| --- |
| Figure 2: Panel A displays the global importance (mean |Shapley value|) for feature categories for the no lag and 2-week lag models. Feature categories are ordered by their aggregate global importance. The importance of each feature category for each model is displayed separately by color. Panels B displays the variation in local feature importance for the no lag and 2-week lag models. Lines start at minimum Shapley value and end at maximum Shapley value. |

Source: [Make All Figures for Main Manuscript](https://jjcurtin.github.io/study_lag/notebooks\mak_figures-preview.html#cell-fig-3)

## Discussion

## Model Performance

Our models performed exceptionally well with median posterior probabilities for auROCs of .84 - .90. This suggests we can achieve clinically meaningful performance up to two weeks out. Our rigorous resampling methods (grouped, nested, k-fold cross-validation) make us confident that these are valid estimates of how our models would perform with new individuals.

Nevertheless, model performance did decrease as models predicted further into the future. This is unsurprising given what we know about prediction and substance use. Many important relapse risk factors are fluctuating processes that can change day-by-day, if not more frequently. As lag time increases, features become less proximal to the prediction time point. Still, we wish to emphasize that our lowest auROC (.84) is still excellent, and the benefit of advanced notice likely outweighs the cost to performance.

The relative ordering of important features remained somewhat consistent across models. Past use, future efficacy, and craving were the top three features for all models. However the magnitude of their importance varried somewhat by lag time. Additionally, for the two-week model future risky situations emerged as an important feature, whereas with the no lag model past stressful events were more important.

## Model Fairness

All models performed worse for people who were not White, and for people who had an income below the poverty line. The largest contributing factor is likely the lack of diversity in our training data. For example, even with our coarse combination of race/ethnicity, the not White group was largely underrepresented relative to the non-Hispanic White group. Similarly, our below poverty group was underrepresented relative to the above poverty group.

One obvious potential solution to this problem is to recruit a more representative sample. In a separate project, we recruited a national sample of participants with opioid use disorder (Moshontz et al. 2021). In addition to achieving better representation in income and race/ethnicity, we also ensured diversity across geographic location (e.g., rural vs. urban) as this is likely another important factor in evaluating fairness.

Computational solutions to mitigate these issues in the current data may also exist. We could explore upsampling disadvantaged group representation in the data (e.g., using synthetic minority oversampling technique). We also could adjust the penalty weights so that prediction errors for disadvantaged groups are weighted more heavily than prediction errors for majority groups. We could also consider using personalized modeling approaches that consider the characteristics and behaviors important to an individual rather than generalizing across a population. For example, state space models inherently capture time series data and allow for the modeling of how an individual’s risk evolves over time from observable and latent states.

The models also performed more poorly for women compared to men, despite the fact that they were well represented. This finding suggests representation in our data is not the only factor affecting model fairness. We chose our EMA items based on domain expertise and years of relapse risk research. It is possible that these constructs more precisely describe relapse risk factors for men than for women. This could mean that more research is needed to identify relapse risk factors for women (and other groups underrepresented in the literature more broadly). Additionally, data driven (bottom-up) approaches to creating features could be one way to remove some of the bias in domain driven (top-down) approaches. For example, using natural language processing on text message content could allow for new categories of features to emerge.

## Additional Limitations and Future Directions

We believe lapse prediction models will be most effective when embedded in a recovery monitoring and support system designed to deliver adaptive and personalized continuing care. This system could send daily, weekly, or less frequent messages to patients with personalized feedback about their risk of lapse and provide support recommendations tailored to their current recovery needs. As described earlier, we previously built day- and hour-level models to predict the probability of an immediate lapse (i.e., within 24 hours, within 1 hour). We can use these models with high temporal precision to guide individuals to take actionable steps to maintain their recovery goals and support them in implementing these steps (e.g., pointing them to a specific module in an app).

This study demonstrated lagged models can be used to shift the 24-hour prediction window up to weeks out. This lag provides individuals with advanced warning of their lapse risk. These models are well-suited to support recovery needs that cannot be addressed within an app, such as scheduling an appointment or attending a support group. To be clear, we do not believe an app alone is sufficient to deliver continuing care. We expect individuals will require additional support throughout their recovery from a mental health provider (e.g., motivational enhancement, crisis management, skill building), a peer (e.g., sponsor, support group), or family member. Importantly, these types of supports take time to set up; highlighting the value of the lagged week model.

Despite building successful prediction models, it is still unclear the best way to provide risk and support information to people. For a recovery monitoring and support system to be successful, it is important that participants trust the system, engage with the system and find the system beneficial. In an ongoing grant, our group is working to optimize the delivery of daily support messages by examining whether the inclusion or exclusion of risk-relevant message components (e.g., lapse probability, lapse probability change, important features, and a risk-relevant recommendation) increase engagement in recovery tools and supports, trust in the machine learning model, and improve clinical outcomes (Wyant and Curtin in prep).

For a system using lagged models, we can imagine that even longer lags (i.e., more advanced warning) would be better still. In the present study, we were limited by how much time we could lag predictions. Participants only provided EMA for up to three months. Therefore, a lag time of two weeks between the prediction time point and start of the prediction window means data from 2 out of the 12 possible weeks is not being used. This loss of data could be one reason we saw a decrease in model performance with increased lag times. In a separate NIH protocol underway, participants are providing EMA and other sensed data for up to 12 months (Moshontz et al. 2021). By comparing models built from these two datasets, we will better be able to evaluate whether this loss of data impacted model performance and if we can sustain similar performance with even longer lags in these data.

A recovery monitoring and support system will require new data to update model predictions. A model only using EMA could raise measurement burden concerns. Research suggests people can comply with effortful sensing methods (e.g., 4x daily EMA) while using substances (Wyant et al. 2023; Jones et al. 2019). However, it is likely that frequent daily surveys will eventually become too burdensome when considering long-term monitoring. We have begun to address this by building models with fewer EMAs (1x daily) and have found comparable performance. Additionally, reinforcement learning could potentially be used for adaptive EMA sampling. For example, each day the algorithm could make a decision to send out an EMA or not based on inferred latent states of the individual based on previous EMA responses and predicted probability of lapse.

Additionally, we have begun to explore how we can supplement our models with data from other lower burden sensing methods. Geolocation is a passive sensing method that could compliment EMA well. First, it could provide insight into information not easily captured by self-report. For example, the amount of time spent in risky locations, or changes in routine that could indicate life stressors. Second, the near-continuous sampling of geolocation could offer risk-relevant information that would otherwise be missed in between the discrete sampling periods of EMA. Ultimately, passive sensing offers the opportunity to capture additional risk features that would be difficult to measure with self-report or would add additional burden by increasing the number of questions on the EMA.

## Conclusion

This study suggests it is possible to predict next day alcohol lapses up to two weeks into the future. This advanced notice could allow patients to implement support options not immediately available. Important steps are still needed to make these models clinically implementable. Most notably, is the increased fairness in model performance. However, we remain optimistic as we have already begun to take several steps in addressing these barriers.

## References

Bae, Sangwon, Tammy Chung, Denzil Ferreira, Anind K. Dey, and Brian Suffoletto. 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 (August): 42–47. <https://doi.org/10.1016/j.addbeh.2017.11.039>.

Bickman, Leonard, Aaron R. Lyon, and Miranda Wolpert. 2016. “Achieving Precision Mental Health Through Effective Assessment, Monitoring, and Feedback Processes.” *Administration and Policy in Mental Health and Mental Health Services Research* 43 (3): 271–76. <https://doi.org/10.1007/s10488-016-0718-5>.

Bowen, Sarah, Neha Chawla, Joel Grow, and G. Alan Marlatt. 2021. *Mindfulness-Based Relapse Prevention for Addictive Behaviors: A Clinician’s Guide*. Second edition. New York: The Guilford Press.

Brandon, Thomas H., Jennifer Irvin Vidrine, and Erika B. Litvin. 2007. “Relapse and Relapse Prevention.” *Annual Review of Clinical Psychology* 3 (1): 257–84. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091455>.

Center for High Throughput Computing. 2006. “Center for High Throughput Computing.” Center for High Throughput Computing. <https://doi.org/10.21231/GNT1-HW21>.

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.” *2022 Alcohol Related Disease Impact (ARDI) Application Website*. https://nccd.cdc.gov/DPH\_ARDI/Default/Default.aspx. Accessed December 14, 2022.

Chih, Ming-Yuan, Timothy Patton, Fiona M. McTavish, Andrew J. Isham, Chris L. Judkins-Fisher, Amy K. Atwood, and David H. Gustafson. 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>.

Dennis, Michael, and Christy K Scott. 2007. “[Managing Addiction as a Chronic Condition](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2797101).” *Addiction Science & Clinical Practice* 4 (1): 45–55.

DeRubeis, Robert J. 2019. “The History, Current Status, and Possible Future of Precision Mental Health.” *Behaviour Research and Therapy* 123 (December): 103506. <https://doi.org/10.1016/j.brat.2019.103506>.

Gabry, Jonah, and Ben Goodrich. 2023. “Prior Distributions for Rstanarm Models.” *CRAN R-Project*. https://cran.r-project.org/web/packages/rstanarm/vignettes/priors.html.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2023. “Rstanarm: Bayesian Applied Regression Modeling via Stan.”

Greenfield, Shelly F., Audrey J. Brooks, Susan M. Gordon, Carla A. Green, Frankie Kropp, R. Kathryn McHugh, Melissa Lincoln, Denise Hien, and Gloria M. Miele. 2007. “Substance Abuse Treatment Entry, Retention, and Outcome in Women: A Review of the Literature.” *Drug and Alcohol Dependence* 86 (1): 1–21. <https://doi.org/10.1016/j.drugalcdep.2006.05.012>.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY: Springer.

Hedegaard, Holly, Arialdi M. Miniño, Merianne Rose Spencer, and Margaret Warner. n.d. “Drug Overdose Deaths in the United States, 1999–2020.” Accessed January 13, 2025.

Hsieh, F. 1989. “Sample Size Tables for Logistic Regression.” *Statistics in Medicine* 8: 795–802.

Jones, Andrew, Danielle Remmerswaal, Ilse Verveer, Eric Robinson, Ingmar H. A. Franken, Cheng K. Fred Wen, and Matt Field. 2019. “Compliance with Ecological Momentary Assessment Protocols in Substance Users: A Meta-Analysis.” *Addiction (Abingdon, England)* 114 (4): 609–19. <https://doi.org/10/gfsjzg>.

Kilaru, Austin S., Aria Xiong, Margaret Lowenstein, Zachary F. Meisel, Jeanmarie Perrone, Utsha Khatri, Nandita Mitra, and M. Kit Delgado. 2020. “Incidence of Treatment for Opioid Use Disorder Following Nonfatal Overdose in Commercially Insured Patients.” *JAMA Network Open* 3 (5): e205852. <https://doi.org/10.1001/jamanetworkopen.2020.5852>.

Kranzler, Henry R., and James R. McKay. 2012. “Personalized Treatment of Alcohol Dependence.” *Current Psychiatry Reports* 14 (5): 486–93. <https://doi.org/10.1007/s11920-012-0296-5>.

Kuhn, Max. 2022. “Tidyposterior: Bayesian Analysis to Compare Models Using Resampling Statistics.”

Kuhn, Max, and Kjell Johnson. 2018. *Applied Predictive Modeling*. 1st ed. 2013, Corr. 2nd printing 2018 edition. New York: Springer. <https://doi.org/10.1007/978-1-4614-6849-3>.

Kuhn, Max, and Hadley Wickham. 2020. “Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.”

Lundberg, Scott M., and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–77. NIPS’17. Red Hook, NY, USA: Curran Associates Inc.

Marlatt, G. Alan, and Judith R. Gordon, eds. 1985. *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors*. First edition. New York: The Guilford Press.

McLellan, A. T., D. C. Lewis, C. P. O’Brien, and H. D. Kleber. 2000. “Drug Dependence, a Chronic Medical Illness: Implications for Treatment, Insurance, and Outcomes Evaluation.” *JAMA* 284 (13): 1689–95. <https://doi.org/10.1001/jama.284.13.1689>.

Mohr, David C., Mi Zhang, and Stephen M. Schueller. 2017. “Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning.” *Annual Review of Clinical Psychology* 13 (1): 23–47. <https://doi.org/10.1146/annurev-clinpsy-032816-044949>.

Moshontz, Hannah, Alejandra J. Colmenares, Gaylen E. Fronk, Sarah J. Sant’Ana, Kendra Wyant, Susan E. Wanta, Adam Maus, David H. Gustafson Jr, Dhavan Shah, and John J. Curtin. 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>.

Olfson, Mark, Christine Mauro, Melanie M. Wall, C. Jean Choi, Colleen L. Barry, and Ramin Mojtabai. 2022. “Healthcare Coverage and Service Access for Low-Income Adults with Substance Use Disorders.” *Journal of Substance Abuse Treatment* 137 (June): 108710. <https://doi.org/10.1016/j.jsat.2021.108710>.

Pinedo, Miguel. 2019. “A Current Re-Examination of Racial/Ethnic Disparities in the Use of Substance Abuse Treatment: Do Disparities Persist?” *Drug and Alcohol Dependence* 202 (September): 162–67. <https://doi.org/10.1016/j.drugalcdep.2019.05.017>.

RStudio Team. 2020. “RStudio: Integrated Development for R.” Boston, MA: RStudio, Inc.

Socías, M. Eugenia, Nora Volkow, and Evan Wood. 2016. “Adopting the ‘Cascade of Care’ Framework: An Opportunity to Close the Implementation Gap in Addiction Care?” *Addiction* 111 (12): 2079–81. <https://doi.org/10.1111/add.13479>.

Soyster, Peter D., Leighann Ashlock, and Aaron J. Fisher. 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>.

Stanojlović, Milena, and Larry Davidson. 2021. “Targeting the Barriers in the Substance Use Disorder Continuum of Care With Peer Recovery Support.” *Substance Abuse: Research and Treatment* 15 (January): 1178221820976988. <https://doi.org/10.1177/1178221820976988>.

Substance Abuse and Mental Health Services Administration. n.d. “2023 NSDUH Detailed Tables CBHSQ Data.” https://www.samhsa.gov/data/report/2023-nsduh-detailed-tables. Accessed January 13, 2025.

Walters, Scott T., Michael S. Businelle, Robert Suchting, Xiaoyin Li, Emily T. Hébert, and Eun-Young Mun. 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 (August): 108417. <https://doi.org/10.1016/j.jsat.2021.108417>.

Witkiewitz, Katie, and G. Alan Marlatt. 2004. “Relapse Prevention for Alcohol and Drug Problems: That Was Zen, This Is Tao.” *American Psychologist* 59 (4): 224–35. <https://doi.org/10.1037/0003-066X.59.4.224>.

———. 2007. “Modeling the Complexity of Post-Treatment Drinking: It’s a Rocky Road to Relapse.” *Clinical Psychology Review*, New Approaches to the Study of Change in Cognitive Behavioral Therapies, 27 (6): 724–38. <https://doi.org/10.1016/j.cpr.2007.01.002>.

Wyant, Kendra, and John Curtin. in prep. “Optimizing Message Components of a Recovery Monitoring Support System for Engagement and Clinical Outcomes for Alcohol Use Disorder: Protocol for an Optimization Study,” in prep.

Wyant, Kendra, Hannah Moshontz, Stephanie B. Ward, Gaylen E. Fronk, and John J. Curtin. 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>.

Wyant, Kendra, Sarah June Kittleson Sant’Ana, Gaylen Fronk, and John J. Curtin. 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>.

1. We used this conservative 24-hour fence for labeling windows as no lapse (vs. excluded) to increase the fidelity of these labels. Given that most windows were labeled no lapse, and the outcome was highly unbalanced, it was not problematic to exclude some no lapse events to further increase confidence in those labels. [↑](#footnote-ref-36)
2. 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). [↑](#footnote-ref-41)
3. The poverty cutoff was defined from the 2024 federal poverty line for the 48 continguous United States. Participants at or below $1560 annual income were categorized as below poverty. [↑](#footnote-ref-43)
4. For our fairness analyses, we altered our outer loop resampling method from 3 x 10 cross-validation to 6 x 5 cross-validation. This method still gave us 30 held out tests sets, but by splitting the data across fewer folds (i.e., 5 vs. 10) we were able to reduce the likelihood of the disadvantaged group being absent in any single fold. [↑](#footnote-ref-44)