Dissertation Proposal

Kendra Wyant

2025-04-23

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

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

# 1. Introduction

## 1.1 Substance Use Disorders

In 2023, over 46 million U.S. adults had a substance use disorder in the past year (Substance Abuse and Mental Health Services Administration, n.d.). This is nearly 18% of the U.S. adult population.

Substance use disorders are associated with high rates of morbidity and mortality. Opioid overdose rates remain high and continue to increase each year (Friedman et al., 2022; *Increases in Drug and Opioid Overdose Deaths — United States, 2000–2014*, n.d.). Excessive alcohol use is a leading preventable cause of death, with the majority of these deaths caused by alcohol-attributed cancer, heart problems and stroke, and liver cirrhosis (*ARDI Alcohol-Attributable Deaths, US CDC*, n.d.). Additionally, alcohol-impaired driving accounts for over 30% of traffic fatalities each year (United States. Department of Transportation. National Highway Traffic Safety Administration. National Center for Statistics and Analysis, 2024).

The economic cost of substance use disorders is substantial. In 2016 the economic cost associated with substance use disorders was estimated to exceed $442 billion in lost productivity, health care expenses, law enforcement, and other criminal justice costs (Substance Abuse and Mental Health Services Administration (US) & Office of the Surgeon General (US), 2016). When also accounting for costs associated with loss of life and reduced quality of life, one research group estimated that in 2017 the cost of opioid use disorder alone exceeded $1 trillion (Florence et al., 2021).

## 1.2 Treatment

Substance use disorders are chronic conditions, characterized by high relapse rates. (Dennis & Scott, 2007; McLellan et al., 2000), substantial co-morbidity with other physical and mental health problems (Dennis & Scott, 2007; Substance Abuse and Mental Health Services Administration, n.d.), and an increased risk of mortality (Centers for Disease Control and Prevention (CDC), n.d.; Hedegaard et al., n.d.).

Initial treatments, including medication (e.g., methadone, buprenorphine, and naltrexone for opioid use disorder and acamprosate, disulfiram, and naltrexone for alcohol use disorder) and evidenced-based psychosocial treatments (e.g., relapse prevention (Marlatt & Donovan, 2007; Marlatt & Gordon, 1985), mindfulness-based relapse prevention (Bowen et al., 2021), and cognitive-behavioral therapy (Liese & Beck, 2022)), are efficacious for symptom stabilization and harm reduction when provided. Unfortunately, too often individuals do not receive treatment (Substance Abuse and Mental Health Services Administration, n.d.).

Equally concerning, is the lack of continuing support and care provided to individuals after completing initial treatment (Socías et al., 2016; Stanojlović & Davidson, 2021). Continuing care, including on-going monitoring, tailored adjustments to lifestyle and behaviors over time, and when necessary re-engagement with more intensive treatment, is the gold standard for treating chronic conditions, like diabetes, hypertension, and asthma. Similarly, substance use treatment has been shown to be most effective when care is prescribed over longer durations and involves active efforts to keep patients engaged in their recovery (McKay, 2021). Yet, for reasons including cost and insurance reimbursement issues, lack of collaborative provider teams, passive referral processes, geographic barriers to accessing services, patient dropout, and changes in the patient’s clinical needs over time (Dennis & Scott, 2007; McKay, 2021; Stanojlović & Davidson, 2021; Tai & Volkow, 2013), our current treatment system for substance use disorders does not not appear to have the capacity for long-term clinician-delivered care. This leaves individuals left to determine on their own “How can I best support my recovery today?”

This type of self-monitoring can be extremely difficult. The risk factors that precede lapse (i.e., single instances of goal-inconsistent substance use) and full relapse back to harmful use during recovery are individualized, numerous, dynamic and interactive (Brandon et al., 2007; K. Witkiewitz & Marlatt, 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.

## 1.3 Recovery Monitoring and Support System

An algorithm-guided recovery monitoring and support system could help patients monitor their current and future risk of lapse and make adjustments in their activities and supports to meet their recovery goals after initial treatment. Such a tool could offer personalized, adaptive recommendations aligned with evidenced-based care (i.e., relapse prevention model) and prompt individuals to engage with support at times of high risk. For example, individuals could receive daily messages about changes in their lapse risk and receive personalized recommendations based on top features contributing to their risk, like an urge surfing recommendation for someone with strong cravings. Moreover, it would provide a scalable option for long-term monitoring and support to address the substantial unmet need for continuing care for substance use disorders.

For such a system to exist, at least three pre-requisites must be met[[1]](#footnote-23). One, the system must be able to collect a rich and densely sampled source (or sources) of risk-relevant data. Two, the system must have access to a model that can predict substance use with high performance and temporal precision and have interpretable model inputs for support recommendations to be mapped onto. Three, the model must perform fairly. The accuracy of the predictions and usefulness and relevance of the recommendations should be similar for everyone. Advances in both smartphone sensing (Mohr et al., 2017) and machine learning (Hastie et al., 2009) now make this possible.

Smartphone sensing approaches (e.g., ecological momentary assessment [EMA], geolocation sensing) can provide the frequent, longitudinal measurement of proximal risk factors that is necessary for prediction of future lapses with high temporal precision.

Machine learning models can handle the high dimensional feature sets that may result from feature engineering densely sampled raw EMA over time. They can also accommodate non-linear and interactive relationships between features and lapse probability. And methods from interpretable machine learning can be used to understand which risk features contribute most strongly to a lapse prediction for a specific individual at a specific moment in time.

## 1.4 Aims of this proposal

This dissertation proposal describes a program of research that seeks to develop an algorithm to be used in a recovery monitoring and support system. Two studies (chapters 2 and 3) are completed and two studies (chapters 3 and 4) are planned and under development. These studies are explained more below:

**Chapter 2:** Machine learning models for temporally precise lapse prediction in alcohol use disorder

This study demonstrated that 4x daily EMA could be used to build temporally precise machine learning models to predict alcohol lapses. Specifically, we built three models to predict the probability of an alcohol lapse in the next week, day, and hour. Our models performed exceptionally well with median posterior probabilities for area under the ROC curve (auROC) of .90, .91, and .93 for predicting lapse probabilities into the next week, day, and hour, respectively.

In this study we also demonstrated it is possible to understand important features overall for a model (i.e., across all observations for all participants) and important features for an individual prediction (i.e., for a single individual at a specific moment). We found on average past use, craving, and abstinence self-efficacy were the most important features affecting predicted probability scores across the three models. We also saw a wide range of individual feature importance values across all people and all days, suggesting that even features with relative low overall importance were still clinically important for some people at some moments.

These findings suggest that it is feasible to identify periods of high lapse risk at varying levels of temporal precision and extract important risk features. If embedded in a recovery monitoring and support system, output from models that predict immediate lapses, like these, could be used to provide individuals critical information about their immediate risk of lapse. Daily (or more or less frequent) messages could be sent to individuals that relay information about changes in their risk and make supportive recommendations for immediate action based on the top features contributing to their risk. For example, recommending a coping with craving activity for someone experiencing strong cravings, recommending a guided relaxation video for someone reporting recent stressful events, or encouraging individuals to reflect on recent past successes or reasons for choosing abstinence or moderation when reported self-efficacy is low. Importantly, this assumes that the recommendations can be implemented immediately. Thus, the individual has already likely learned the skill and it is not contingent on scheduling or other people.

However, often the most clinically appropriate support recommendation takes time to set up. Scheduling positive or pleasant activities, increasing social engagement, or attending a peer-led recovery meeting may require days or weeks to implement. In these cases, patients would benefit from advanced warning about changes in their risk. This advanced warning could be obtained by lagging lapse probability predictions further into the future (e.g., predicting the probability of a lapse in a 24-hour window that begins two weeks in the future).

Finally, in this study, we failed to assess model fairness. In recent years, the machine learning field has begun to understand the critical importance of evaluating model fairness when algorithms are used to inform important decisions (e.g., healthcare services offered, eligibility for loans, early parole). Algorithms that perform favorably for only a single majority group could widen existing disparities in access to resources and important clinical outcomes (Veinot et al., 2018). Future studies included in this proposal will explicitly evaluate model performance parity across important subgroups.

**Chapter 3:** Lagged predictions of next day alcohol lapse risk for personalized and adaptive continuing care support

This study showed that the same 4x daily EMA could be used to predict alcohol use occurring within a 24-hour window up to two weeks into the future. We considered several meaningful lags between the prediction timepoint and start of the prediction window: 1 day, 3 days, 1 week, and 2 weeks. These models with lagged prediction windows can be used to provide personalized support recommendations with the added benefit of advanced warning. Thus, giving an individual extra time to implement the recommendation.

Model performance decreased as the start of the prediction window was lagged further from the prediction timepoint with median posterior probabilities for auROC of .88 (1-day lag), .87 (3-day lag), .86 (1-week lag), and .84 (2-week lag). This decrease in performance is not surprising given what we know about prediction and alcohol lapse. Many important lapse risk factors are fluctuating processes that can change day-by-day, if not more frequently. As lag times increase, features become less proximal to the start of the prediction window. Still, auROCs of .80 and higher are generally considered to indicate good performance and the benefit of advanced notice of lapse risk likely outweighs the cost to performance.

In this study, we also assessed model fairness by comparing model performance (posterior probability for auROC) across important subgroups with known disparities in substance use treatment access and/or outcomes - race/ethnicity (not White vs. non-Hispanic White), income (below poverty vs. above poverty), and sex at birth (female vs. male). There was strong evidence that our models performed better for individuals who were non-Hispanic White, male, and had a personal income above the Federal poverty line.

The relative ordering of important features remained somewhat consistent across the models. Past use, future efficacy, and craving were the top three features for all models. However on average these features became less important as model predictions lagged further into the future, consistent with their lower performance. As with the previous study, we found a wide range of individual importance values for features across the models.

These results are promising, however, several limitations exist. Most notably was a lack of fairness in the performance of our models among subgroups. The largest contributing factor is likely the lack of diversity in our training data. Even with our coarse dichotomous grouping of race we only had 20 participants in the not White group (compared to 131 in the White group). We also saw a similar pattern in our income groups (49 people in the below poverty group compared to 102 people in the above poverty group).

However, we had equal representation of men and women and still found our models performed better for men. We chose our 9 EMA items based on the extant relapse risk literature. Historically women were excluded from early substance use research due to their childbearing potential. It is possible that these constructs more precisely describe relapse risk factors for men than for women.

One solution, would be to add additional EMA items to the daily surveys in hopes to capture important risk factors for women. This however, would quickly increase the burden of our surveys. EMA acceptability studies suggest that longer surveys, but not more frequent prompts, promote increased perceived burden and compromised data quantity and quality (Eisele et al., 2020). Therefore, supplementing EMA with another lower burden sensing method may be preferred. Additionally, passive sensing data may be well-suited for data-driven (bottom-up) feature engineering approaches. Compared to traditional, theory-driven (top-down) methods, data-driven features can identify patterns and characteristics predictive of lapse in specific groups and reduce potential bias in features by minimizing researcher involvement.

A few other limitations were inherent in the data set. First, was the length of participation. As participants only provided data for up to three months, we were limited by how far into the future we could lag predictions. It is likely that even more advanced warning (e.g., 1 month) would be helpful for implementing more intensive supports (e.g., re-engagement with clinician-delivered care). Second, the results presented in chapters 2-3 focus on alcohol lapse prediction. Alcohol differs from other substances in several ways. It is legal and generally viewed to be socially acceptable (e.g., it is often integrated into celebrations and social gatherings). As a result, individuals who use other substances (e.g., opioids) may face more stigma and be less willing to report substance use and risk information. Therefore, it is not clear that similar prediction models will perform as well.

**Chapter 4:** Using sensing data to predict opioid lapse risk in a national sample of patients with opioid use disorder

In this study, we will take advantage of an existing dataset of personal sensing data (1X daily EMA and geolocation) and opioid lapse reports from a national sample of people with opioid use disorder to predict immediate (i.e., in the next 24 hours) lapses back to opioid use.

This study will allow us to generalize lapse prediction algorithms to other drugs beyond alcohol. Notably, a successful model will demonstrate that lapse prediction can be done with a drug where its use is illegal and people may be less willing to provide information about lapses and risk factors.

These data also offer more diversity with regard to race/ethnicity and income. This will allow us to determine if improving the quality of the training data with respect to diversity is sufficient to address issues of fairness. These data also offer diversity across geographic location (e.g., rural vs. urban vs. suburban), likely another important factor in evaluating fairness.

In this proposed study, model features will be derived from two complementary sensing methods: 1X daily EMA and continuous geolocation data. Geolocation sensing, a passive sensing method, could compliment EMA well. It could provide insight into information difficult to measure with self-report (e.g., the amount of time spent in risky locations, or changes in routine that could indicate life stressors) or that would add additional burden by increasing the number of questions on the EMA. Furthermore, by adding more data sources gives us more features and that could mean better personalization of predictions and recommendations for more people.

Participants provided data for up to 12 months. This extended window of recovery (12 months vs. 3 months) is critical for evaluating the value of an algorithm intended for ongoing continuing care support and understanding how lapse risk evolves as people progress in their recovery (i.e., past 3 months). Unfortunately, our ability to address explanatory questions about the time course of lapse risk, and how individuals might cluster on different recovery trajectories is limited with traditional machine learning methods. These methods do not capture the repeated nature of sensing data. Each lapse prediction is treated as a new independent observation. We can account for the repeated observations in our sensing data by engineering features that capture individual changes over time to produce unbiased and precise estimates of predictive performance. However, more traditional time series models are better suited for understanding the temporal dynamics of lapse risk over long periods of recovery.

The longer duration of participation also provides the opportunity to experiment with lagged prediction windows further than two weeks into the future (i.e., 1 month). However, in a machine learning framework, different models must be built to predict lapses at varying times in the future. For example, a recovery monitoring support system that detects both immediate lapse risk (i.e., in the next 24 hours) and future lapse risk (i.e, in the next 2 weeks and in the next month) would need three models. This approach is time consuming, cumbersome, and still only provides coarse understanding of time-course. Therefore, this study will only predict immediate lapse risk.

**Chapter 5:** State-space models for idiographic risk monitoring and recovery support recommendations

In this study, we propose Hierarchical Bayesian state-space models as an alternative approach for prediction models. State-space models model measured inputs (e.g., ema responses, time spent in risky locations, time spent at home) and outputs (i.e., lapse or no lapse) from time series data with latent states. The hierarchical nature, will allow us to better use and understand time-varying information. State-space models explicitly model how the latent state of an individual’s lapse risk evolves over time.

Given the heterogeneity in lapse risk and the complex interactions between environment and individual differences, a time series model that use an individual’s own data to make future predictions may perform better than models trained at the group-level. Therefore, we will build an individual model for each participant using their own data. Although our immediate lapse risk models have been performing quite well, individual models could improve performance for our lagged prediction models. Individual models also may help mitigate issues of unfairness, as the model will weigh the individual’s own data more heavily than group level estimates.

Additionally, time series models could potentially improve the efficiency and performance of lagged prediction models. A single model can be used to predict a lapse at any point in the future, eliminating the need for multiple models for predicting immediate and future lapse risk.

Therefore, in this study, we will evaluate the performance and fairness of a state-space model approach for opioid lapse risk prediction using the EMA and geolocation data set introduced in Chapter 4. We will evaluate both immediate (i.e., in the next 24 hours) and future (i.e., next 2 weeks and next 1 month) lapse risk.

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

## 2.1 Introduction

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

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

Digital therapeutics may help to overcome these barriers associated with in-person, clinician-delivered treatments. Digital therapeutics provide evidence-based interventions and other supports via smartphones to prevent, treat, or manage a medical disorder, either independently or in conjunction with traditional treatments (Jacobson et al., 2022). They offer highly scalable, on-demand therapeutic support that is accessible whenever and wherever it is needed most. Several large, randomized controlled trials have confirmed that digital therapeutics for AUD improve clinical outcomes (Campbell et al., 2014; Gustafson et al., 2014; Jacobson et al., 2022). Additionally, US adults (including patients with AUD (Wyant et al., 2023)) display high rates of smartphone ownership (over 85% in 2021), with minimal variation across race, ethnicity, socioeconomic status, and geographic settings. Therefore, digital therapeutics may not only mitigate in-person treatment barriers but also combat associated disparities (Jacobson et al., 2022).

### 2.1.1 Improving Digital Therapeutics via Personal Sensing

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

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

Successful, continuous monitoring of risk for relapse and its contributing factors would enable patients to adapt their lifestyle, behaviors, and supports to their changing needs. Successful monitoring could also direct patients to engage with the most appropriate digital therapeutic modules, addressing the unique risks present at any given moment throughout their recovery. Such continuous monitoring is now feasible via personal sensing (i.e., in-situ data collection via sensors embedded in individuals’ daily lives) (Bae et al., 2018; Chih et al., 2014; Epstein et al., 2020; Moshontz et al., 2021; Soyster et al., 2022; Wyant et al., 2023).

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

### 2.1.2 Promising Preliminary Research

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

Despite this initial promise, several important limitations exist. Some prediction models have been trained using convenience samples (e.g., college students) (Bae et al., 2018; Soyster et al., 2022). Other models have been developed to predict hazardous alcohol use in non-treatment-seeking populations (Walters et al., 2021). In both these instances, features that predict planned or otherwise intentional alcohol use among individuals not motivated to change their behavior may not generalize to people in AUD recovery. Moreover, individuals who have not yet begun to contemplate and/or commit to behavior change regarding their alcohol use are unlikely to use digital therapeutics designed for AUD recovery (Prochaska et al., 1992).

A handful of other models have been trained to predict putative precursors of substance use, such as craving (Burgess-Hull et al., 2022; Dumortier et al., 2016) and stress (Epstein et al., 2020). Although craving and stress may be associated with substance use, their relationships with relapse are complex, inconsistent, and not always very strong (Fronk et al., 2020; Sayette, 2016). For these reasons, we believe that explicit substance use may be a better target for prediction.

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

An early alcohol lapse prediction model developed by Gustafson and colleagues (Chih et al., 2014) provided the foundation on which our current project builds. Participants completed EMAs once per week for 8 months while using a digital therapeutic after discharge from an inpatient treatment program for AUD. These EMAs were used as features in a machine learning model to predict lapses. However, the temporal precision for both the features and outcome was coarse. Model predictions were updated only once per week at most, and lapse onsets could occur anytime within the next two weeks. This coarseness restricts the model from being used to implement *just-in-time* interventions (e.g., guided mindfulness or other stress reduction techniques, urge surfing) that are well-suited to digital therapeutics.

### 2.1.3 The Current Study

The current study addresses these limitations of previously developed prediction models. We trained our models using participants in early recovery from moderate to severe AUD who reported a goal of alcohol abstinence. We developed three separate models that provide hour-by-hour probabilities of a future lapse back to alcohol use with increasing temporal precision: lapses in the next week, next day, and next hour. Model features were engineered from raw scores and longitudinal change in responses to 4X daily EMAs. These features were derived to measure theoretically-implicated risk factors and contexts that have considerable support as predictors of lapses including past use, craving, past pleasant events, past and future risky situations, past and future stressful events, emotional valence and arousal, and self-efficacy (Fronk et al., 2020; for reviews, see Marlatt & Gordon, 1985; K. Witkiewitz & Marlatt, 2007).

In this study, we characterize the performance of these three prediction models in held-out data (i.e., for observations from participants who were not used to train the models). We also evaluated the relative feature importance of key relapse prevention constructs in the models as part of the model validation process and to contribute to the relapse prevention literature. This research represents an important step toward the development of a “smart” (machine learning guided) sensing and prediction system that can be embedded within a digital therapeutic both to identify periods of peak lapse risk and to recommend specific supports to address factors contributing to this risk.

## 2.2 Method

### 2.2.1 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. Finally, our data, analysis scripts, annotated results, questionnaires, and other study materials are publicly available (<https://osf.io/w5h9y/>).

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

### 2.2.2 Participants

We recruited 151 participants in early recovery (1-8 weeks of abstinence) from AUD in Madison, Wisconsin, US. This sample size was determined based on traditional power analysis methods for logistic regression (Hsieh, 1989) because comparable approaches for machine learning models have not yet been validated. Participants were recruited through print and targeted digital advertisements and partnerships with treatment centers. We required participants:

1. were age 18 or older,
2. could write and read in English,
3. had at least moderate AUD (>= 4 self-reported DSM-5 symptoms),
4. were abstinent from alcohol for 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.

### 2.2.3 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).

### 2.2.4 Measures

#### 2.2.4.1 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 intensity of four recent experiences:

* craving [“How intense was your greatest urge to drink?”],
* risky situations [“Did you encounter any risky situations (people, places, or things)? If yes, rate the intensity of the situation.”],
* stressful events [“Has a hassle or stressful event occurred? If yes, rate the intensity of the event.”],
* pleasant events [Has a pleasant or positive event occurred? If yes, rate the intensity of the event.”].

For each of these experiences, participants rated the maximum intensity since their last EMA on a 12-point ordinal scale (mid- and end-point anchors of “Mild”, “Moderate”, and “Strong”). If they did not experience an event since their last EMA, participants selected “No” to indicate that no experience occurred for that respective question.

Next, participants rated their current affect using 11-point bipolar scales measuring valence (end-point anchors of “Unpleasant/Unhappy” to “Pleasant/Happy”) and arousal (end-point anchors of “Calm/Sleepy” to “Aroused/Alert”).

On the first EMA each day, participants used an 11-point bipoloar scale (end-point anchors of “Very Unlikely” to “Very Likely”) to rate the likelihood of:

* future risky situations [“How likely are you to encounter risky situations (people, places, or things) within the next week?”],
* future stressful events [“How likely are you to encounter a stressful event within the next week?”],
* abstinence efficacy [“How likely are you to drink any alcohol within the next week?”].

#### 2.2.4.2 Individual Differences

We collected self-report information about demographics (age, sex, race, ethnicity, education, marital status, employment, and income) and clinical characteristics (AUD milestones, number of quit attempts, lifetime AUD treatment history, lifetime receipt of AUD medication, DSM-5 AUD symptom count, and current drug use (WHO ASSIST Working Group, 2002)). This information was collected primarily to characterize the sample and to evaluate the diversity of the training data. We also included demographic features in our models to quantify the importance of relapse prevention constructs beyond these static characteristics, given known disparities in AUD and other health outcomes (Jacobson et al., 2022)[[2]](#footnote-36).

### 2.2.5 Data Analytic Strategy

Data preprocessing, modeling, and Bayesian analyses were done in R using the tidymodels ecosystem (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).

#### 2.2.5.1 Lapse Labels

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

For each participant, the first prediction window for all three widths began at midnight on their second day of participation and ended one week, one day, or one hour later. This ensured at least 24 hours of past EMAs for future lapse prediction in these first windows. Subsequent windows for each participant were created by repeatedly rolling the window start/end forward one hour until the end of their study participation (i.e., each participant’s last prediction window started one week, one day, or one hour before their last recorded EMA).

We labeled each prediction window as *lapse* or *no lapse* using participants’ reports from the EMA question “Have you drank any alcohol that you have not yet reported?”. If participants answered yes to this question, they entered the date and hour of the start and end of the drinking episode. During monthly follow-up sessions, participants could review and correct their lapses reported by EMA and report to staff any additional lapses.

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. 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.

#### 2.2.5.2 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 created features for both baseline and full models. The baseline models were developed to determine how well we could predict lapses using a simple model based only on the participants’ histories of previous lapses. The full models used all EMA responses combined with demographic and day/time features.

The baseline models had only one dummy-coded feature: lapse frequency (high vs. low). The median number of lapses across participants during the study period was 1. Therefore, the lapse frequency feature was coded low when the participant had a history of 1 or fewer lapses before that prediction window. This feature was coded high when the participant had more than 1 lapse before that window.

Features for the full model were derived from three sources: 1) common demographic characteristics, 2) day of the week and hour of the day at prediction window onset, and 3) previous EMA responses. We created a quantitative feature for age, and dummy-coded features for sex (male vs. female), race/ethnicity (White/Non-Hispanic vs. other), marital status (never married vs. married vs. other), and education (high school or less vs. some college vs. 4-year degree or more). We created dummy-coded features to indicate time of day (5pm - midnight vs. any other time) and day of week that the prediction window began.

We created raw EMA features for varying scoring epochs before the start of the prediction window for all EMA items excluding the alcohol use question. For the six EMA questions that appeared on all four daily EMAs, we used five scoring epochs of 12, 24, 48, 72, and 168 hours. For the three EMA questions that only appeared on the morning EMA, we used three scoring epochs of 48, 72, and 168 hours. Raw features included min, max, and median scores for each EMA question across all EMAs in each epoch for that participant. We calculated change features by subtracting the participant’s mean score for each EMA question (using all EMAs collected before the start of the prediction window) from the associated raw feature. These change features allowed us to capture within-subject effects by comparing recent EMA responses relative to an individual’s own baseline. For both raw and change features, the feature was set to missing (and later imputed; see below) if no responses to the specific EMA question were provided by the participant within the associated scoring epoch.

We also created raw and change features based on the most recent response for each EMA question (excluding the alcohol use question). This generated two features for each EMA question: 1) raw value of the most recent previous response, and 2) difference between that raw value and the mean response to that EMA question over all EMAs collected before that prediction window.

We also calculated raw and change rate features from previously reported lapses. We calculated lapse rate features using the same five scoring epochs described earlier. Raw lapse rate features were generated by dividing the total number of previously observed lapses within a scoring epoch by the duration of that epoch. For change rate features, we subtracted the rate of previous lapses for that participant (i.e., total number of lapses while on-study divided by total hours on-study before the prediction window) from their associated raw lapse rate. We employed a similar approach to calculate raw and change rate of missing EMAs (i.e., number of full EMA surveys that were requested but not completed in a scoring epoch / duration of epoch).

Other generic feature engineering steps included: 1) imputing missing data (median imputation for numeric features, mode imputation for nominal features); 2) dummy coding for nominal features; and 3) removing zero-variance features. Medians/modes for missing data imputation and identification of zero variance features were derived from held-in (training) data and applied to held-out (validation and test) data (see Cross-validation section below). We recognize that median/mode imputation is a coarse method for handling missing data; however, computational costs of more sophisticated methods (e.g., KNN imputation, multiple imputation) were not practical for this study. A sample feature engineering script (i.e., tidymodels recipe) containing all feature engineering steps is available on our OSF study page.

#### 2.2.5.3 Model Training and Evaluation

##### 2.2.5.3.1 Statistical Algorithm and Hyperparameters.

We trained and evaluated six separate classification models: one baseline and one full model for each prediction window (week, day, and hour). We initially considered four well-established statistical algorithms (XGBoost, Random Forest, K-Nearest Neighbors, and Elastic Net) that vary across characteristics expected to affect model performance (e.g., flexibility, complexity, handling higher-order interactions natively) (Kuhn & Johnson, 2018). However, preliminary exploratory analyses suggested that XGBoost consistently outperformed the other three algorithms[[3]](#footnote-41). Furthermore, the Shapley Additive Explanations (SHAP) method, which we planned to use for explanatory analyses of feature importance in our full models, is optimized for XGBoost. Consequently, we focused our primary model training and evaluation on the XGBoost algorithm only.

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

Model training and evaluation used all participants (N = 151), regardless if they had any positive labels (i.e., lapses) because XGBoost itself does not use grouping of observations within participants. This grouping is handled instead by a participant-grouped cross-validation procedure (below).

##### 2.2.5.3.2 Performance Metric.

Our primary performance metric for model selection and evaluation was area under the Receiver Operating Characteristic Curve (auROC) (Kuhn & Johnson, 2018). auROC indexes the probability that the model will predict a higher score for a randomly selected positive case (lapse) relative to a randomly selected negative case (no lapse). This metric was selected because it 1) combines sensitivity and specificity, which are both important characteristics for clinical implementation; 2) is an aggregate metric across all decision thresholds, which is important because optimal decision thresholds may differ across settings and goals; and 3) is unaffected by class imbalance, which is important for comparing models with differing prediction window widths and levels of class imbalance.

##### 2.2.5.3.3 Cross-validation.

We used participant-grouped, nested cross-validation for model training, selection, and evaluation with auROC. Grouped cross-validation assigns all data from a participant as either held-in or held-out to avoid bias introduced when predicting a participant’s data from their own data. Nested cross-validation uses two nested loops for dividing and holding out folds: an outer loop, where held-out folds serve as *test sets* for model evaluation; and inner loops, where held-out folds serve as *validation sets* for model selection. Importantly, these sets are independent, maintaining separation between data used to train the models, select the best models, and evaluate those best models. Therefore, nested cross-validation removes optimization bias from the evaluation of model performance in the test sets and can yield lower variance performance estimates than single test set approaches (Jonathan et al., 2000).

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

#### 2.2.5.4 Bayesian Estimation of auROC and Model Comparisons

We used a Bayesian hierarchical generalized linear model to estimate the posterior probability distributions and 95% Bayesian confidence intervals (CIs) for auROC for the six best models. Bayesian analyses were accomplished using the tidyposterior (Kuhn, 2022) and rstanarm (Goodrich et al., 2023) packages in R. Following recommendations from the rstanarm team and others (Gabry & Goodrich, 2023; RStudio Team, 2020), 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. Specifically, the 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).

To estimate the probability that the full model outperformed the baseline model, we regressed the auROCs (logit transformed) from the 30 test sets for each model as a function of model type (baseline vs. full). To determine the probability that full models’ performances differed systematically from each other, we regressed the auROCs (logit transformed) from the 30 test sets for each full model as a function of prediction window width (week vs. day vs. hour). Following recommendations from the tidymodels team (Kuhn, 2022), we set two random intercepts: one for the repeat, and another for the fold within repeat (folds are nested within repeats for 3x10-fold cross-validation). We report the 95% (equal-tailed) Bayesian CIs from the posterior probability distributions for our models’ auROCs. We also report 95% (equal-tailed) Bayesian CIs for the differences in performance associated with the Bayesian comparisons.

#### 2.2.5.5 Shapley Additive Explanations for Feature Importance

We computed Shapley Values (Lundberg & Lee, 2017) to provide a consistent, objective explanation of the importance of categories of features (based on EMA questions) across our three full models. Shapley values possess several useful properties including: Additivity (Shapley values for each feature can be computed independently and summed); Efficiency (the sum of Shapley values across features must add up to the difference between predicted and observed outcomes for each observation); Symmetry (Shapley values for two features should be equal if the two features contribute equally to all possible coalitions); and Dummy (a feature that does not change the predicted value in any coalition will have a Shapley value of 0).

We calculated Shapley values from the 30 test sets using the SHAPforxgboost package that provides Shapley values in log-odds units for binary classification models. We averaged the three Shapley values for each observation for each feature across the three repeats to increase their stability. The additivity property of Shapley values allowed us to create 18 feature categories from the 286 separate features. We created separate feature categories for each of the nine EMA questions (excluding the alcohol use question), the rates of past alcohol use and missing surveys, the time of day and day of the week of the start of the prediction window, and the five demographic variables included in the models. For the EMA questions and rates of past alcohol use and missing surveys, these categories included all individual raw and change features across the three to five scoring epochs (see Feature Engineering above) and the most recent response. To calculate the local (i.e., for each observation) importance for each category of features, we added Shapley values across all features in a category, separately for each observation. To calculate global importance for each feature category, we averaged 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 answer questions of relative feature importance. However, these are descriptive analyses because standard errors or other indices of uncertainty for importance scores are not available for Shapley values.

## 2.3 Results

### 2.3.1 Demographic and Clinical Characteristics

One hundred ninety-two participants were eligible. Of these, 191 consented to participate, and 169 subsequently enrolled in the study. Fifteen participants discontinued before the first monthly follow-up visit. We excluded data from one participant who did not maintain a goal of abstinence during their participation. We also excluded data from two participants due to evidence of careless responding and unusually low compliance. Our final sample consisted of 151 participants.

The final sample included approximately equal numbers of men (N = 77, 51.0%) and women (N = 74, 49.0%) who ranged in age from 21 to 72 years old. The sample was majority White (N = 131, 86.8%) and non-Hispanic (N = 147, 97.4%). Participants self-reported a median of 9.0 DSM-5 symptoms of AUD (M = 8.9, SD = 1.9, range = 4.0–11.0) and a median of 4.0 previous quit attempts (M = 5.5, SD = 5.8, range = 0.0–30.0). Most participants (N = 84, 55.6%) reported one or more lapses during participation. The median number of lapses per participant while on study was 1.0 (M = 6.8, SD = 12.0, range = 0.0–75.0). [Table 1](#tbl-dem) provides more detail on demographic and clinical characteristics of the sample.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/tables-preview.html#cell-tbl-dem)

### 2.3.2 EMA Compliance, Features, and Prediction Window Labels

Participants on average completed 3.1 (SD=0.6) of the four EMAs each day (78.4% compliance overall). Participants completed at least one EMA on 95.0% of days. Across individual weeks on-study, EMA compliance percentages ranged from 75.3%-86.8% completion for all of the 4x daily EMAs and from 91.7%-99.1% for at least one daily EMA completed.

Using these EMA reports, we created data sets with 270,081, 274,179, and 267,287 future prediction windows for the week, day, and hour window widths, respectively. Each data set contained 286 features and an outcome labeled as lapse or no lapse. These data sets were unbalanced with respect to the outcome such that lapses were observed in 68,467 (25.4%) week windows, 21,107 (7.7%) day windows, and 1,017 (0.4%) hour windows.

Features had missing values if the participant did not respond to the relevant EMA question during the associated scoring epoch. The median proportions of missing values across features were relatively low: 0.020 (range = 0 - 0.121), 0.022 (range = 0 - 0.125), and 0.023 (range = 0 - 0.127) for the week, day, and hour prediction windows. There were no missing values for demographic features, the hour and day of the start of the prediction window, or lapse rate and missing survey rate features.

### 2.3.3 Model Performance

#### 2.3.3.1 auROC for Baseline Models

We selected the best *baseline model* (previous lapse frequency feature only) configurations using auROCs from the *validation sets*. We evaluated these best baseline model configurations using *test set* performance to remove the optimization bias present in performance metrics from validation sets. The median auROC across the 30 test sets was moderate for the week (Mdn = 0.792, IQR = 0.079, range = 0.671–0.915), day (Mdn = 0.784, IQR = 0.070, range = 0.687–0.890), and hour (Mdn = 0.779, IQR = 0.077, range = 0.675–0.884) prediction windows.

We used the 30 test set auROCs to estimate the posterior probability distribution for the auROC of these baseline models. The median auROCs from these posterior distributions were 0.798 (week), 0.785 (day), and 0.780 (hour). These values represent our best estimates for the magnitude of the auROC parameter for each model. The 95% Bayesian CIs for the auROCs for these models were relatively narrow and did not contain 0.5 (chance performance) for any window width: week [0.770–0.822], day [0.757–0.810], and hour [0.752–0.806].

#### 2.3.3.2 auROCs for Full Models

We next selected the best *full model* (which included all features) configurations using auROCs from the *validation sets*. We evaluated these best full model configurations using *test set* performance. The median auROC across the 30 test sets was high for the week (Mdn = 0.891, IQR = 0.043, range = 0.785– 0.963), day (Mdn = 0.899, IQR = 0.05, range = 0.788–0.969), and hour (Mdn = 0.929, IQR = 0.045, range = 0.847–0.972) prediction windows. [Figure 1](#fig-1-ema) (left panel) displays the ROC curves by prediction window derived by aggregating predicted lapse probabilities across all test sets.

The right panel of [Figure 1](#fig-1-ema) displays posterior probability distributions for the auROC for the full models by prediction window. The median auROCs from these posterior distributions were 0.895 (week), 0.905 (day), and 0.930 (hour). These values represent our best estimates for the magnitude of the auROC parameter for each model. The 95% Bayesian CIs for the auROCs for these models were relatively narrow and did not contain 0.5 (chance performance) for any window width: week [0.876–0.910], day [0.888–0.919], and hour [0.916–0.940].

|  |
| --- |
| Figure 1: ROC curves and posterior probabilities for auROCs by prediction window. The left panel depicts the aggregate receiver operating characteristic (ROC) curve for each model, derived by concatenating predicted lapse probabilities across all test sets. The dotted line represents the expected ROC curve for a random classifier. The histograms on the right depict the posterior probability distribution for the areas under the receiver operating characteristic curves (auROCs) for each model. The vertical lines represent the median posterior probability and the horizontal line represents the boundaries 95% CI. |

Source: [Make All Figures for Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/figures-preview.html#cell-fig-1-ema)

#### 2.3.3.3 Bayesian Comparisons of Baseline vs. Full Models

We used the posterior probability distributions for the auROCs to formally compare the baseline versus full models (matched for prediction window). The median increase in auROC for the full versus baseline week model was 0.097 (95% CI [0.081–0.114]), yielding a probability of 1.000 that the full week model had superior performance. The median increase in auROC for the full versus baseline day model was 0.120 (95% CI [0.102–0.138]), yielding a probability of 1.000 that the full-day model had superior performance. The median increase in auROC for the full versus baseline hour model was 0.149 (95% CI [0.131–0.170]), yielding a probability of 1.000 that the full hour model had superior performance.

#### 2.3.3.4 Bayesian Comparisons of Full Models by Prediction Window

We also used the posterior probability distributions for the auROCs for the three full models to formally compare the differences in performance by prediction window width. The median increase in auROC for the hour versus the day model was 0.025 (95% CI [0.017–0.034]), yielding a probability of 1.000 that the hour (vs. day) model had superior performance. The median increase in auROC for the hour versus the week model was 0.035 (95% CI [0.026–0.045]), yielding a probability of 1.000 that the hour model (vs. week) had superior performance. The median increase in auROC for the day versus the week model was 0.010 (95% CI [0.001–0.020]), yielding a probability of 0.982 that the day (vs. week) model had superior performance.

#### 2.3.3.5 Other Performance Metrics for the Full Models

We evaluated the sensitivity, specificity, balanced accuracy, PPV, and NPV when the predicted lapse probabilities were used for binary classification (*lapse* vs. *no lapse*) with decision thresholds identified by Youden’s Index. All three full models had high sensitivity, specificity, balanced accuracy, and NPV ([Table 2](#tbl-perf-ema)). PPV, however, notably declined as the prediction window width decreased.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Performance Metrics for Full models by Prediction Window   | Metric | Week | Day | Hour | | --- | --- | --- | --- | | auROC | 0.8906310 | 0.8989930 | 0.9293432 | | sensitivity | 0.8232301 | 0.8275224 | 0.8636420 | | specificity | 0.8192581 | 0.8454889 | 0.8813554 | | balanced accuracy | 0.8275871 | 0.8345976 | 0.8536644 | | positive predictive value | 0.6301088 | 0.2997223 | 0.0254053 | | negative predictive value | 0.9435759 | 0.9877148 | 0.9994361 | |

Source: [Make All Tables for Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/tables-preview.html#cell-tbl-perf-ema)

PPV can be increased by increasing the decision threshold; however, increasing the decision threshold will also lower the model’s sensitivity. To evaluate the trade-off between PPV (i.e., precision) and sensitivity (i.e., recall) across decision thresholds, we created Precision-Recall curves by concatenating predicted lapse probabilities across the 30 test sets ([Figure 2](#fig-2-ema)). For example, the dotted lines in [Figure 2](#fig-2-ema) depict the sensitivities (0.718, 0.473, and 0.327 for week, day, and hour models, respectively) associated with decision thresholds that yield 0.700 PPV for each model.

|  |
| --- |
| Figure 2: Precision-recall Curves by Prediction Window for the Full Models. The plot depicts the aggregate precision-recall curves for each full model, derived by concatenating predicted lapse probabilities across all test sets. The dotted lines depict the sensitivities (0.718, 0.473, and 0.327 for week, day, and hour models, respectively) associated with decision thresholds that yield 0.700 positive predictive value for each of those models. |

Source: [Make All Figures for Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/figures-preview.html#cell-fig-2-ema)

### 2.3.4 Feature Importance for Full Models

Global importance (mean |Shapley value|) for feature categories for each full model appears in Panel A of [Figure 3](#fig-3-ema). Past use was the most important feature category for lapse prediction across prediction window widths. Future abstinence efficacy was also globally important across window widths. Time-varying constructs (craving, time of day) appear to have more impact in lapse prediction for the hour model compared to the day and week models.

Sina plots of local Shapley values (i.e., the influence of feature categories on individual observations) for each model show that some feature categories (e.g., past pleasant events, future stressful events) impact lapse probability for specific individuals at specific times even if they are not globally important across all observations ([Figure 3](#fig-3-ema), Panels B-D).

|  |
| --- |
| Figure 3: Feature importance (Shapley values) for Full Models by Prediction Window. Panel A displays the global importance (mean |Shapley value|) for feature categories for each full model. Raw EMA features are grouped into categories by the original question from the EMA. Features based on the rates of previous lapses and previous missing surveys, as well as demographics, and the time of day and day of the week for the start of the prediction window are also included. Feature categories are ordered by their aggregate global importance (i.e., total bar length) across the three models. The importance of each feature category for specific models is displayed separately by color. Panels B-D display local Shapley values that quantify the influence of feature categories on individual observations (i.e., a single prediction window for a specific participant) for each model. |

Source: [Make All Figures for Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/figures-preview.html#cell-fig-3-ema)

## 2.4 Discussion

### 2.4.1 Model Performance

All baseline models, which used only past frequency of lapses to predict future lapses, performed moderately well with auROCs in the upper .70s. These results confirm what we would expect: past behavior is a relatively good predictor of future behavior. However, there was still substantial room for increased predictive performance. Furthermore, these baseline models do not identify specific risk factors contributing to lapse predictions at any moment in time for each participant.

All three full models performed exceptionally well, yielding auROCs of 0.89, 0.90, and 0.93 for week, day, and hour level models, respectively. auROCs above .9 are generally described as having “excellent” performance; the model will correctly assign a higher probability to a positive case (e.g., lapse) than a negative case 90% of the time (Mandrekar, 2010). Bayesian comparisons indicated that these full models performed better than the baseline models for the same prediction window. This confirms that EMA can predict future alcohol lapses in the next week, next day, and next hour with high sensitivity and specificity for new individuals. And, as we describe later, using features that map onto important relapse prevention risk constructs may illuminate momentary contributors to predicted lapses.

This study addressed several important limitations of previous research to advance toward robust sensing and prediction models that can be embedded within digital therapeutics. First, our models were trained on a relatively large, treatment-seeking sample of adults in early recovery from AUD that closely matches the individuals most likely to benefit from such models within a digital therapeutic. Second, we explicitly predicted episodes of goal-inconsistent alcohol use (i.e., lapses) because features that predict goal-inconsistent use likely differ from those that predict other types of alcohol use. Third, we measured EMA features and alcohol use with sufficient frequency and granularity to train well-performing models with high temporal resolution - specifically, hour-by-hour predicted probabilities for lapses in the next week, day, and hour. Fourth, we collected features and outcomes over three months during a high risk period (initial remission (Hagman et al., 2022) from AUD). Fifth, we used cutting-edge resampling methods (grouped, nested, k-fold cross-validation) to provide valid estimates of how our models would perform with new individuals. Finally, we used interpretable machine learning methods (SHAP (Lundberg & Lee, 2017; Molnar, 2022)) to better understand how our models made predictions globally and locally for specific participants at discrete moments in time.

### 2.4.2 Understanding & Contextualizing Model Performance

We used SHAP to describe the relative importance of key relapse prevention model constructs (represented by categories of features) to predicted lapses in our three full models. Some constructs consistently emerged as globally important across week, day, and hour level models. Unsurprisingly, the largest contribution to lapse prediction was past use. This is consistent with decades of research on relapse precipitants and our understanding of human behavior more generally (i.e., past behavior predicts future behavior) (Marlatt & Gordon, 1985). Decreases in abstinence self-efficacy were also strongly associated with increased probability of future lapses across windows.

The relative importance of some constructs descriptively differed by window width. Punctate, time-varying constructs (e.g., craving, arousal, recent risky situation) had greater impact on predicted lapse probabilities in the hour model compared to day or week models. The time of day feature was relatively important (top four) in the hour model, such that lapses were more likely for hour-level prediction windows that began in the evenings. The day of week feature made a small contribution to the hour and day models given that lapses were more likely on weekends. The time and day features were not useful in the week model because its associated prediction window (a full week) spanned all days and times, making the time and day that the window began irrelevant. The increased global importance for all these punctate features/constructs to immediate lapse risk likely contributed to the hour model outperforming the day and week models. These important global differences in next hour lapse risk also highlight the need for just-in-time interventions that can address these imminent but short-lived risks.

The individual, local Shapley values also shed light on the multidimensional and heterogeneous nature of lapse risk in our sample. Sina plots of local Shapley values ([Figure 3](#fig-3-ema)) display meaningful ranges of scores for most feature categories. This means that even feature categories with lower global importance (e.g., past pleasant events, future stressful events) still consequentially impacted predictions for some individuals at specific times. This variability in locally important features highlights the potential benefits of recommending optimal interventions and other supports that are personalized for that person at that moment in time.

Our demographic features did not display high global or local importance. Despite the diversity in sex, age, education, and marital status in our sample, these features did not meaningfully contribute to lapse prediction. Although this does not preclude these features’ predictive utility, it does suggest that other EMA feature categories may be more relevant for lapse prediction than these characteristics. Race/ethnicity also did not emerge as globally or locally important features. However, the limited representation of participants of color in our sample warrants caution in drawing conclusions about the predictive utility of race and ethnicity at this time.

### 2.4.3 Considerations for Clinical Implementation

#### 2.4.3.1 Smart Digital Therapeutics

We believe these full models may be most effective when embedded in a “smart” digital therapeutic that guides patients toward optimal, adaptive engagement to address their ongoing and momentary risks. These models can provide the patient’s predicted future lapse probability and the features that meaningfully contribute to that probability. We consciously selected EMA items that map onto well-known risk factors from the relapse prevention literature. Consequently, these outputs can be used to recommend specific intervention and support modules that are risk-relevant for each patient - much like a clinician would do if they were available in-the-moment. For example, during sensed periods of high stress, stress reduction techniques (e.g., guided mindfulness) could be recommended. If increased time with risky people or locations is driving lapse risk, the digital therapeutic can support patients to attend support meetings, or encourage participation in the in-app discussion board.

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

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

A smart digital therapeutic can potentially improve clinical outcomes in multiple ways. First, feedback from the prediction model could improve patient insight and self-monitoring by connecting their daily experiences to changing risk. Second, it can remove patient uncertainty by guiding selection from the substantial content available. Third, a smart digital therapeutic could encourage risk-relevant engagement. Rather than trying to increase overall time using the digital therapeutic, patients could be guided to use the supports that specifically target their personal risk factors at that moment in time. Thus, smart digital therapeutics are well-positioned to pursue the precision mental health goal to “provide the right treatment to the right patient at the right time, every time” (Kaiser, 2015).

#### 2.4.3.2 Categorical Lapse Predictions

Our models natively provide quantitative predictions of lapse probabilities. These lapse probabilities can also be used to make categorical predictions (lapse vs. no-lapse) by applying a decision threshold to the quantitative predicted lapse probabilities (i.e., predict lapse when the probability exceeds the decision threshold).

We observed high sensitivity and specificity for these categorical predictions at a decision threshold selected to balance these two performance metrics. However, the PPV (proportion of predicted lapses that were true lapses) of these categorical predictions in our full models was moderate to very low at this threshold (ranging from .630 down to .025 across window widths). For this reason, categorical predictions should be provided to patients with extreme caution, if at all. Instead, we favor the quantitative lapse probabilities as risk indicators to guide intervention and support recommendations.

If categorical predictions are necessary, PPV can be improved by raising the decision threshold, but this comes at the cost of reduced sensitivity. We explored this trade-off in the precision-recall curves displayed in [Figure 2](#fig-2-ema). From these curves, it is clear decision thresholds that yield higher PPV (e.g., .700) exist for all three full models, but the associated sensitivity will be lower (e.g., 0.718, 0.473, and 0.327 for the week, day, and hour models, respectively, at this threshold). Clinical implementation of categorical predictions will require selecting an optimal decision threshold after weighing the cost of missing true lapses (low sensitivity) vs. predicting lapses that subsequently do not occur (low PPV). Different thresholds could be used depending on the purpose, context, available resources, or even patient preference.

### 2.4.4 Additional Limitations and Future Directions

Successful clinical implementation of our models will require several important steps to address limitations in our work to-date. First, we need to enrich the training data to include diversity across race, ethnicity, and geographic region. Our current prediction models may not work well for people of color or people from rural communities. Prediction models must use diverse training samples to avoid exacerbating rather than mitigating existing disparities. We must also collect data from individuals in later stages of recovery beyond initial remission; features that predict lapses may differ in these later periods. We are intentionally addressing these issues in a current NIH protocol that recruits nationally for demographic and geographic diversity and follows participants for up to 1.5 years into their recovery (Moshontz et al., 2021).

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

Our current models predict probability of imminent lapses. The hour and day full models are well-positioned to identify and recommend just-in-time interventions to address these immediate risks. However, the week model may not have sufficient temporal specificity to recommend immediate patient action. Instead, its clinical utility may improve if we shift this coarser window width into the future. For example, we could train a model to predict the probability of lapse at any point during a week window that begins two weeks in the future. This “time-lagged” model could provide patients with increased lead time to implement supports that might not be immediately available to them (e.g., schedule therapy appointment, request support from an AA sponsor).

Finally, XGBoost does not take advantage of grouping observations within participants or systematic variation unique to individual participants[[4]](#footnote-84). Independence of observations is not necessary for statistically valid prediction. When observations are grouped/repeated within participants, linear mixed effects models or other statistical models that can estimate both population-level (fixed) effects and participant-level (random) effects may predict better for the participants on which they were trained than would XGBoost. However, we are not interested in making predictions for participants in our training set. We want to know how well our models will work with new individuals like those that will use smart digital therapeutics in the future.

In some domains, there has been increasing interest in idiographic approaches where models are trained and then implemented for the same individual(Fisher et al., 2019; Wright & Zimmermann, 2019). Such approaches may also yield superior predictive performance but are not possible to implement for outcomes like alcohol use lapse. A person-specific lapse prediction model requires a sufficient number of positive labels (i.e., lapses) for that individual. It may be too late to prevent relapse if we must wait until an individual has lapsed multiple (perhaps many) times to offer help. We believe the most promising approaches may involve first developing population-based models and updating these models with person-specific information as the patient uses the system (Zhou et al., 2018). We are pursuing these cutting-edge models as a near-term future direction.

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

# 3. Lagged predictions of next day alcohol lapse risk for personalized and adaptive continuing care support

## 3.1 Introduction

Alcohol and other substance use disorders (SUDs) are serious chronic conditions, characterized by high relapse rates(Dennis & Scott, 2007; McLellan et al., 2000), substantial co-morbidity with other physical and mental health problems(Dennis & Scott, 2007; Substance Abuse and Mental Health Services Administration, n.d.), and an increased risk of mortality (Centers for Disease Control and Prevention (CDC), n.d.; Hedegaard et al., 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 (Socías et al., 2016; Stanojlović & Davidson, 2021; Substance Abuse and Mental Health Services Administration, n.d.).

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 harmful use (Marlatt & Gordon, 1985; K. Witkiewitz & Marlatt, 2004). Critically, the risk factors that instigate lapses during recovery are individualized, numerous, dynamic, interactive, and non-linear (Brandon et al., 2007; K. Witkiewitz & Marlatt, 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 et al., 2016; DeRubeis, 2019; Kranzler & 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 et al., 2017) and machine learning (Hastie et al., 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 (Bowen et al., 2021; Marlatt & Gordon, 1985; K. Witkiewitz & Marlatt, 2004). Furthermore, individuals with SUDs have found EMA to be acceptable for sustained measurement for up to a year with relatively high compliance (Moshontz et al., 2021; Wyant et al., 2023), 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 that are likely necessary for accurate prediction of lapse probability. And rapid advances in the tools for interpretable machine learning (e.g, Shapley values (Lundberg & 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 (Bae et al., 2018; Soyster et al., 2022; Walters et al., 2021; 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 et al., 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 to assist the patient to acquire these new skills. 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. A 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 (Kilaru et al., 2020; Pinedo, 2019), 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.

## 3.2 Methods

### 3.2.1 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. 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/>).

### 3.2.2 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. 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 moderately severe alcohol use disorder (>= 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.[[5]](#footnote-92)

One hundred ninety-two participants were eligible. Of these, 191 consented to participate in the study at the screening visit, and 169 subsequently enrolled in the study at the enrollment visit, which occurred approximately one week later. Fifteen participants discontinued before the first monthly follow-up visit. We excluded data from one participant who did not maintain a goal of abstinence during their participation. We also excluded data from two participants due to evidence of careless responding and unusually low compliance. Our final sample consisted of 151 participants. 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.

### 3.2.3 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.

### 3.2.4 Measures

#### 3.2.4.1 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 previously 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).

#### 3.2.4.2 Individual Characteristics

We collected self-report information about demographics (age, sex at birth, race, ethnicity, education, marital status, employment, and income) and clinical characteristics (AUD milestones, number of quit attempts, lifetime AUD treatment history, lifetime receipt of AUD medication, DSM-5 AUD symptom count, current drug use (WHO ASSIST Working Group, 2002), and presence of psychological symptoms (Derogatis, L.R., n.d.)) to characterize our sample. DSM-5 AUD symptom count and presence of psychological symptoms were also used to determine eligibility. Demographic information was included as features in our models. A subset of these variables (sex at birth, race, ethnicity, and income) were used for model fairness analyses, as they have documented disparities in treatment access and outcomes.

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.

### 3.2.5 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).

#### 3.2.5.1 Predictions

A *prediction timepoint* ([Figure 4](#fig-methods-lag), Panel A) is the hour at which our model calculates a predicted probability of a lapse within a future 24-hour prediction window for any specific individual. We calculated the features used to make predictions at each prediction timepoint within a feature scoring epoch that included all available EMAs up until, but not including, the prediction timepoint. The first prediction timepoint for each participant was 24 hours from midnight on their study start date. This ensured at least 24 hours of past EMAs were available in the feature scoring epoch. Subsequent prediction timepoints for each participant repeatedly rolled forward hour-by-hour until the end of their study participation.

The *prediction window* ([Figure 4](#fig-methods-lag), Panel B) spans a period 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 specific 24-hour period). Prediction windows rolled forward hour-by-hour with the prediction timepoint. However, there were five possible *lag times* between the prediction timepoint and start of the associated 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 into the future.

Given this structure, our models provided hour-by-hour predicted probabilities of an alcohol lapse in a future 24 hour period. Depending on the model, that future period (the prediction window) might start immediately after the prediction timepoint or up to 2 weeks into the future. For example, at midnight on the 30th day of participation, the feature scoring epoch would include the past 30 days of EMAs. Separate models would predict the probability of lapse for 24 hour periods staring at midnight that day, or similar 24 hour periods starting 1 day, 3 days, 1 week or 2 weeks after midnight on day 30.

|  |
| --- |
| Figure 4: 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 of time in which a lapse might occur) rolls forward hour-by-hour with the prediction time point with the prediction time point. The prediction window width for all our models was one week (i.e., models predicted the probability of a lapse occurring within a one week window). 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 (0 lag) or was lagged by 24, 72, 168, or 336 hours. |

Source: [Make All Figures for Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/figures-preview.html#cell-fig-methods-lag)

#### 3.2.5.2 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. [[6]](#footnote-104)

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.

#### 3.2.5.3 Feature Engineering

Features were calculated using only data collected in feature scoring epochs before each prediction timepoint to ensure our models were making true future predictions. For our no lag models the prediction timepoint 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 timepoint 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 (in years) and personal income (in dollars), and dummy-coded features for sex at birth (male vs. female), race/ethnicity (non-Hispanic White vs. not White), marital status (married vs. not married vs. other), education (high school or less vs. some college vs. college degree), and employment (employed vs. unemployed).
3. *Previous EMA responses*: We created raw and change features using EMAs in varying feature scoring epochs (i.e., 12, 24, 48, 72, and 168 hours) before the prediction timepoint 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 each participant’s baseline mean score for each EMA item from their raw feature. These baseline mean scores were calculated using all of their EMAs collected from the start of their participation until the prediction timepoint. 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).

#### 3.2.5.4 Model Training and Evaluation

##### 3.2.5.4.1 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 & 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).

##### 3.2.5.4.2 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.

##### 3.2.5.4.3 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 (Gabry & Goodrich, 2023; RStudio Team, 2020), 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.[[7]](#footnote-109) 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 auROCs 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 credible intervals do not contain .5 (chance performance), this provides strong evidence (> .95 probability) that 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.

##### 3.2.5.4.4 Fairness Analyses

We calculated the median posterior probability and 95% Bayesian CI for auROC for each model separately by race/ethnicity (not White vs. non-Hispanic White), income (below poverty vs. above poverty[[8]](#footnote-111)), 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.[[9]](#footnote-112)

##### 3.2.5.4.5 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 & 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.

## 3.3 Results

### 3.3.1 Demographic and Lapse Characteristics

[Table 1](#tbl-dem) in Chapter 2 provides a detailed breakdown of the demographic and clinical characteristics of our sample (N = 151).

## 3.4 Model Evaluation

[Figure 5](#fig-2-lag) presents the full posterior probability distributions for auROC for each model (no lag, 1-day, 3-day, 1-week, and 2-week lag). The median auROCs from these posterior distributions were 0.90 (no lag), 0.88 (1-day lag), 0.87 (3-day 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].

|  |
| --- |
| Figure 5: Posterior probability distributions for area under ROC curve (auROC) for each model (no lag, 1-day, 3-day, 1-week, and 2-week lag). Horizonatal lines depict 95% Bayesian credible intervals (CI) and vertical solid lines depict median posterior probability for auROC. Vertical dashed line represents expected performance from a random classifier (.5 auROC). |

Source: [Make All Figures for Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/figures-preview.html#cell-fig-2-lag)

## 3.5 Model Comparisons

[Table 3](#tbl-contrast-lag) 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. Median auROC differences greater than 0 indicate the more immediate model, on average, out-performed the more lagged model (e.g., no lag - 1-day lag, 1-day lag - 3-day lag). 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 the previous adjacent lagged model, with average drops in auROC ranging from 0.01-0.02.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: 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, 0.052] | 1 | | No lag vs. 2 weeks | 0.063 | [0.053, 0.073] | 1 | | Adjacent Contrasts |  |  |  | | 1 day vs. 3 days | 0.012 | [0.005, 0.02] | 0.999 | | 3 days vs. 1 week | 0.011 | [0.003, 0.018] | 0.989 | | 1 week vs. 2 weeks | 0.02 | [0.011, 0.029] | 1 | |

Source: [Make All Tables for Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/tables-preview.html#cell-tbl-contrast-lag)

## 3.6 Fairness Analyses

[Table 4](#tbl-fairness) presents the median difference in auROC, 95% Bayesian CI, and posterior probability that the auROC difference was greater than 0 for the three fairness contrasts: race/ethnicity (Non-Hispanic White; *N* = 131 vs. not White; *N* = 20), sex at birth (male; *N* = 77 vs. female; *N* = 74), and income (above poverty; *N* = 102 vs. below poverty; *N* = 49). Median auROC differences greater than 0 indicate the model, on average, performed better for the advantaged group (male, non-Hispanic White, and above poverty) compared to the non-advantaged group (female, not White, below poverty). In [Table 4](#tbl-fairness) we present fairness analyses for our baseline model (no lag) and for our longest lagged model (2-week lag), as this is likely the most clinically useful lagged model for providing advanced warning of lapse risk. There was strong evidence (probabilities > .96) that our models performed better for the advantaged groups compared to the non-advantaged groups. On average, across all five models, 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, across all five models, there was a median decrease in auROC of 0.05 (range 0.04-0.10) for female participants compared to male participants. On average, across all five models, there was a median decrease in auROC of 0.04 (range 0.03-0.06) for participants below the federal poverty line compared to participants above the federal poverty line.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: 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 | | --- | --- | --- | --- | | Fairness Contrasts (No Lag) |  |  |  | | male vs. female | 0.042 | [0.026, 0.058] | 1 | | non-Hispanic White vs. not White | 0.211 | [0.052, 0.421] | 0.99 | | above poverty vs. below poverty | 0.026 | [0.002, 0.051] | 0.964 | | Fairness Contrasts (2-week Lag) |  |  |  | | male vs. female | 0.098 | [0.077, 0.12] | 1 | | non-Hispanic White vs. not White | 0.109 | [0.062, 0.159] | 1 | | above poverty vs. below poverty | 0.058 | [0.025, 0.09] | 0.997 | |

Source: [Make All Tables for Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/tables-preview.html#cell-tbl-fairness)

## 3.7 Feature Importance

Global feature importance is an indicator of how important a feature category was to the model’s predictions, on average (i.e., across all participants and all observations). The top three globally important feature categories (i.e., highest mean |Shapley value|) for all models were past use, future efficacy, and craving. This was also consistent across demographic groups. Panel A of [Figure 6](#fig-3-lag) shows the relative ranking of feature categories for the no lag and 2-week lag models.

Local feature importance is an indicator of how important a feature category is at a specific prediction timepoint (i.e., for a single individual on a specific day). Local importance can be used to map feature categories onto clinical interventions and recommendations (e.g., What does this individual need right now?). Panel B of [Figure 6](#fig-3-lag) shows the range of local feature importance (minimum Shapley value and maximum Shapley value) for each EMA feature category for the no lag and 2-week lag models. This plot suggests that even feature categories with low global importance (e.g., past pleasant event) have a wide range of local importance values, suggesting that for some people at some moments in time these features are clinically important.

|  |
| --- |
| Figure 6: 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 Dissertation](https://KendraPaquette.github.io/study_dissertation/notebooks/figures-preview.html#cell-fig-3-lag)

## 3.8 Discussion

## 3.9 Model Performance

Our models performed exceptionally well. Our no lag models had a .90 median posterior probability for auROC. This model predicts the probability of an immediate (i.e., within 24 hours) lapse back to alcohol use. Our 2-week lagged model, our most lagged model, had a .84 median posterior probability for auROC, suggesting lagged models can be used to shift a 24-hour prediction window up to weeks out.

Across all models (no lag, 1 day, 3 days, 1 week, and 2 weeks) we saw model performance decrease as models predicted further into the future. All lagged models had lower performance compared to the no lag baseline model and to the preceding adjacent lag model. 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 start of the prediction window. Still, we wish to emphasize that our lowest auROC (.84) is still quite good, and the benefit of advanced notice (i.e., 2 weeks) likely outweighs the cost to performance.

Collectively, these results suggest 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.

When looking at global feature importance, or how important a feature category was to the model’s predictions, on average across all participants and all observations, the relative ordering of important features remained somewhat consistent across both our no lag and 2-week lag models. Past use, future efficacy, and craving were the top three features for both models. However the magnitude of their importance varied by lag time. For the 2-week lagged models, these top features were overall less important for the model’s predictions compared to the no lag model. This is consistent with the 2-week lagged model’s lower performance, compared to the no lag model.

When looking at local feature importance, or how important a feature category is for an individual prediction timepoint for a single individual on a specific day, we saw a wide variation in possible values for both the no lag and 2-week lag models. A wide range of possible values suggests that even feature categories with low global importance (e.g., past pleasant event) are important risk-relevant factors for some people on some days.

We believe our 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. This study provides initial support that immediate and lagged prediction models can be built with high accuracy using EMA for recovery monitoring. Furthermore, the high variance in importance of features for individual predictions is well suited for making tailored recovery support recommendations.

Our no lag models can be used to guide individuals to take immediate 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). For example, recommending an urge surfing activity when someone’s immediate risk is driven by strong craving, recommending a guided relaxation video when someone is reporting recent stressful events, or encouraging individuals to reflect on recent past successes or reasons for choosing abstinence or moderation when self-efficacy is low.

The 2-week lagged model 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 positive or pleasant activities, increasing social engagement, or attending a peer-led recovery meeting. 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 this lagged week model.

## 3.10 Model Fairness

In our previous study (Wyant et al., 2024), we failed to assess model fairness. In recent years, the machine learning field has begun to understand the critical importance of evaluating model fairness when algorithms are used to inform important decisions (e.g., healthcare services offered, eligibility for loans, early parole). Algorithms that perform favorably for only a single majority group could widen existing disparities in access to resources and important clinical outcomes (Veinot et al., 2018). Therefore, we are committed to continuing to improve our understanding of and methods for evaluating model fairness. In this study, we assessed model fairness by comparing model performance across important subgroups with known disparities in substance use treatment access and/or outcomes - race/ethnicity (not White vs. non-Hispanic White), income (below poverty vs. above poverty), and sex at birth (female vs. male).

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. Our sample was majority non-Hispanic White (*N*=131, 87%) with an average income of $34,000. As a result, even with our coarse combination of race/ethnicity, the not White group was severely 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. These data were collected from 2015-2017. Our group has since committed to making changes in recruitment strategies. We are currently working to collect data from a more diverse sample of individuals with alcohol use disorder to improve these models (Wyant et al., in prep). 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 non-advantaged group representation in the data (e.g., using synthetic minority oversampling technique). We also could adjust the penalty weights so that prediction errors for non-advantaged 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 can capitalize on time series data by building individual models for each participant using their own data. Individual models may help mitigate issues of unfairness, as the model will weigh the individual’s own data more heavily than group level estimates. In other words, how does lapse risk evolve over time for a specific individual.

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 decades of relapse risk research. However, prior to 1993 National Institute of Health Revitalization Act (Studies et al., 1994) that mandated the inclusion of minorities and women in research, women were mostly excluded from most substance abuse treatment research due to their childbearing potential (Vannicelli & Nash, 1984). As a result, 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 (e.g., interpersonal relationship problems (Walitzer & Dearing, 2006), hormonal changes (McHugh et al., 2018)), and other groups underrepresented in the literature more broadly.

An alternative approach could be to use a data-driven (bottom-up) approach to identify patterns and characteristics predictive of lapse in specific groups. Compared to traditional, theory-driven (top-down) methods, data-driven features reduce potential bias in features by minimizing researcher involvement. For example, applying natural language processing to text message content could allow new categories of features to emerge. These categories may or may not align with existing theoretical concepts of lapse and relapse, but because they are generated from participants’ own words, they may serve as reliable indicators of lapse risk for certain individuals.

It is also true that historically marginalized groups that have experienced systemic racism, exclusion, or other stigma around substance use (e.g., societal expectations for women regarding attractiveness, cleanliness and motherhood (Meyers et al., 2021)) may feel less trusting in disclosing substance use (Marwick & Boyd, 2018). These experiences could result in individuals under-reporting lapses and/or risk factors, contributing to the gap in model performance. We saw a comparable, or slightly higher, percentage of lapse reports for disadvantaged compared to advantaged groups: race/ethnicity (6%, not White vs. 8%, non-Hispanic White), income (12%, below poverty vs. 7%, above poverty), sex at birth (9%, female vs. 7%, male). However, comparable lapse labels (i.e., reported lapses) does not necessarily equate comparable reporting (i.e., proportion of true lapses reported).

## 3.11 Additional Limitations and Future Directions

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 et al., 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 (Jones et al., 2019; Wyant et al., 2023). 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 (Pulick et al., under review). 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.

## 3.12 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.

# 4. Using sensing data to predict opioid lapse risk in a national sample of patients with opioid use disorder

## 4.1 Introduction

Studies show high agreement between recent (i.e., 1-4 days) self-report and biological markers (i.e., urine, saliva, hair) of drug use (Bharat et al., 2023). This suggests people may be willing to report illicit substance use behaviors.

However, it is unclear if people in recovery from substance use disorders, other than alcohol, can sustain long-term adherence (e.g., one year or more) needed for a recovery monitoring support system that uses self-report data. Previous studies examining adherence to frequent self-report prompts among people with illicit substance use disorders have typically only prompted for 7-30 days (Jones et al., 2019) (except Kennedy et al. (2013) prompted for 175 days and found 75% adherence).

It is also unclear whether people in recovery from illicit substance use disorders can, or are willing to, accurately report other risk information needed to develop accurate prediction models. For example, people with substance use disorders may experience greater instability in their day-to-day lives (e.g., stigma or legal consequences may make access to healthcare, stable housing, or supportive relationships more difficult). This instability could make it difficult to recall and report recent behaviors and events promptly or accurately. It may also skew their baseline perception of what constitutes a risky or stressful experience.

Supplementing self-report data with passively sensed data (e.g., geolocation) could make up for imprecise reports of risk factors or be used for lapse prediction during periods of non-adherence to self-report surveys. Additionally, more data will produce more features that could allow for better personalization of support recommendations.

This study will use daily surveys and sensed geolocation for up to one year from a national sample of people with opioid use disorder to predict immediate (i.e., in the next 24 hours) lapses back to opioid use.

## 4.2 Specific Aims

In this study, we will expand our previous modeling procedure to opioid use disorder and evaluate performance, fairness, and top features for predicting opioid lapse risk.

Specifically, we will:

**1. Evaluate the performance (auROC) of a machine learning model that predicts opioid lapse risk from geolocation and daily surveys.** This aim will allow us to determine whether lapse prediction models can be generalized to other drugs beyond alcohol. Notably, a successful model will demonstrate that lapse prediction can be done with a drug where its use is illegal and people may be less willing to provide information about lapses and risk factors. It will also show the feasibility of using self-report data over long periods of reocvery (i.e., 12 months).

**2. Assess model fairness in performance across important subgroups with known disparities in substance use treatment access and/or outcomes - race/ethnicity (not White vs. non-Hispanic White), income (below $25,000 vs. above $25,000), sex at birth (female vs. male), and geographic location (rural vs. other).** These data offer more diversity with regard to race/ethnicity, income, and geographic location. This aim will allow us to determine if improving the quality of the training data with respect to diversity is sufficient to address issues of fairness.

**3. Describe the relative importance of features on model performance.** Model features will be derived from two complementary sensing methods: daily surveys and continuous geolocation data. Geolocation sensing, a passive sensing method, could compliment daily surveys well. It could provide insight into information difficult to measure with self-report (e.g., the amount of time spent in risky locations, or changes in routine that could indicate life stressors) or that would add additional burden by increasing the number of questions on the daily surveys. Furthermore, by adding more data sources gives us more features and that could mean better personalization of predictions and recommendations for more people. This aim will help determine whether a sufficient number of unique important features emerge from these data.

## 4.3 Methods

### 4.3.1 Participants

We recruited participants in early recovery from opioid use disorder. Participants were recruited through print and targeted digital advertisements (craigslist, reddit, Facebook) and partnerships with MOUD treatment centers. We required that participants:

1. were age 18 or older,
2. could write, speak, and read in English,
3. enrolled in an MOUD treatment program (for at least one month but not longer than 12 months) and adherent (taken daily medication every day or nearly every day in past month) or enrolled in or recently completed an intensive outpatient treatment program for opioid use disorder,
4. had a goal of abstinence from opioids,
5. had an android smartphone that they were willing to use as their single phone for duration of the study, and
6. had active cellular plan that they were willing to maintain for duration of the study.

Participants were considered enrolled in the study if they were eligible, consented, and provided data for at least one month. A total of 336 participants enrolled in the study. We excluded data from one participant whose geolocation data showed they did not reside in the US. We also excluded 11 participants due to careless responding on the daily surveys and/or lapses reported nearly every day on study, suggesting they did not have a goal of abstinence. Our final sample consisted of 324 participants from 38 states across the US.

The table below presents the demographic and clinical characteristics of the 324 participants in our analysis sample.

|  | N | % | M | SD | Range |
| --- | --- | --- | --- | --- | --- |
| Age |  |  |  |  |  |
| 18-21 | 2 | 0.6 |  |  |  |
| 22-25 | 7 | 2.2 |  |  |  |
| 26-35 | 120 | 37.0 |  |  |  |
| 36-45 | 115 | 35.5 |  |  |  |
| 46-55 | 62 | 19.1 |  |  |  |
| 56-65 | 14 | 4.3 |  |  |  |
| Over 65 | 4 | 1.2 |  |  |  |
| Gender |  |  |  |  |  |
| Man | 175 | 54.0 |  |  |  |
| Non-binary | 4 | 1.2 |  |  |  |
| Not listed above | 2 | 0.6 |  |  |  |
| Prefer not to say | 1 | 0.3 |  |  |  |
| Woman | 142 | 43.8 |  |  |  |
| Orientation |  |  |  |  |  |
| Bisexual | 42 | 13.0 |  |  |  |
| Lesbian or gay | 17 | 5.2 |  |  |  |
| Not listed above | 4 | 1.2 |  |  |  |
| Not sure | 1 | 0.3 |  |  |  |
| Straight, that is, not gay or lesbian | 258 | 79.6 |  |  |  |
| Race or Origin (Select all that apply) |  |  |  |  |  |
| American Indian or Alaska Native | 19 | 5.9 |  |  |  |
| Asian | 5 | 1.5 |  |  |  |
| Black or African American | 53 | 16.4 |  |  |  |
| Hispanic, Latino, or Spanish Origin | 29 | 9.0 |  |  |  |
| Native Hawaiian or Other Pacific Islander | 3 | 0.9 |  |  |  |
| Not Listed Above | 5 | 1.5 |  |  |  |
| White or Caucasian | 251 | 77.5 |  |  |  |
| Education |  |  |  |  |  |
| 8th grade or less | 6 | 1.9 |  |  |  |
| Some high school, but did not graduate | 28 | 8.6 |  |  |  |
| High school graduate or GED | 102 | 31.5 |  |  |  |
| Some college or 2-year degree | 140 | 43.2 |  |  |  |
| 4-year college graduate | 33 | 10.2 |  |  |  |
| More than 4-year or advanced degree | 13 | 4.0 |  |  |  |
| Employment |  |  |  |  |  |
| Disabled, not able to work | 36 | 11.1 |  |  |  |
| Employed, working 1-39 hours per week | 74 | 22.8 |  |  |  |
| Employed, working 40 or more hours per week | 60 | 18.5 |  |  |  |
| Not employed, NOT looking for work | 33 | 10.2 |  |  |  |
| Not employed, looking for work | 116 | 35.8 |  |  |  |
| Retired | 3 | 0.9 |  |  |  |
| Annual Household Income |  |  |  |  |  |
| Less than $25,000 | 180 | 55.6 |  |  |  |
| $25,000 - $34,999 | 45 | 13.9 |  |  |  |
| $35,000 - $49,999 | 39 | 12.0 |  |  |  |
| $50,000 - $74,999 | 33 | 10.2 |  |  |  |
| $75,000 - $99,999 | 14 | 4.3 |  |  |  |
| $100,000 - $149,999 | 6 | 1.9 |  |  |  |
| $150,000 - $199,999 | 4 | 1.2 |  |  |  |
| $200,000 or more | 1 | 0.3 |  |  |  |
| In a Committed Romantic Relationship |  |  |  |  |  |
| No | 141 | 43.5 |  |  |  |
| Yes | 181 | 55.9 |  |  |  |
| DSM-5 OUD Symptom Count |  |  | 9.9 | 2.2 | 0 - 11 |
| OUD Medication |  |  |  |  |  |
| I am not taking any medications for opioid use | 31 | 9.6 |  |  |  |
| Medication taken daily (e.g., Suboxone, Methadone) and medication taken monthly (e.g., Vivitrol) | 40 | 12.3 |  |  |  |
| Medications taken daily (e.g., Suboxone, Methadone) | 230 | 71.0 |  |  |  |
| Medications taken monthly (e.g., Vivitrol) | 22 | 6.8 |  |  |  |
| Number of Overdoses in Lifetime |  |  |  |  |  |
| 0 | 135 | 41.7 |  |  |  |
| 1 time | 46 | 14.2 |  |  |  |
| 2-3 times | 76 | 23.5 |  |  |  |
| 4-5 times | 27 | 8.3 |  |  |  |
| More than 5 times | 38 | 11.7 |  |  |  |
| Reported 1 or More Lapse During Study Period |  |  |  |  |  |
| Yes | 137 | 42.3 |  |  |  |
| No | 187 | 57.7 |  |  |  |
| Number of reported lapses |  |  | 9.1 | 25.0 | 0 - 164 |

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

### 4.3.2 Procedure

Participants who screened eligible were consented and onboarded over video or phone call. All participants were instructed to download the STARR study app (a version of CHESS). After this meeting participants were instructed to watch a set of video tutorials for learning how to use the app. One week later they participated in a check-in video or phone call with study staff to answer any questions about the app and troubleshoot any technical issues. At this time, study staff also mailed onboarding materials to participants, including a payment card. Participants were expected to complete monthly surveys to remain on study. At the end of the study participants met with study staff for a final debriefing video or phone call.

### 4.3.3 Measures

#### 4.3.3.1 Individual Characteristics

We collected self-report information about demographics (age, gender, orientation, race/ethnicity, education, employment, income, relationship status, location) and clinical characteristics (DSM-5 OUD symptom count, MOUD medication, number of lifetime overdoses) to characterize our sample. Demographic information will be included as features in our models. A subset of these variables (gender, race/ethnicity, income, and location) will be used for model fairness analyses, as they have documented disparities in treatment access and outcomes.

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 will be made available on our study’s OSF page.

#### 4.3.3.2 Daily Surveys

Participants completed one brief (16 questions) daily survey. Daily surveys became available in star app each morning at 6 am. The survey remind open until 5:59 am the next day. Push notifications were also sent to participants to remind them that they had a new task that had not been completed yet.

On each survey, participants reported dates/times of any previously unreported past opioid use. They also reported any other drugs or alcohol used in the past 24 hours 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. Finally, participants rated how motivated they were to completely avoid using opioids for non-medical reasons and how confident they were in their ability to completely avoid using opioids for non-medical reasons.

Participants were withdrawn early from the study if they did not complete at least 20 daily surveys in a four week period.

#### 4.3.3.3 Monthly Surveys

Monthly surveys consisted of several clinical scales as part of the parent project’s aims. The monthly survey was also personalized to ask a series of questions about participants’ frequent locations (identified by geolocation sensing - see Sensed Geolocation section below). For locations that participants visited twice in a month, they were asked to identify the type of location, what they do there, how pleasant and unpleasant their experience is there, and how much this place helps and harms their recovery from opioids.

Participants were withdrawn early from the study if they missed three monthly surveys.

#### 4.3.3.4 Sensed Geolocation

Continuous sensed geolocation was collected through the STARR app. Geolocation was contextualized by asking questions about frequently visited locations in each monthly survey (see Monthly Surveys section above).

Participants were shown how to temporarily turn off sharing geolocation with us. However, participants were expected to share their location with the STARR app and were withdrawn from the study if they did consistently provide geolocation data (i.e., disabling location sharing for more than 12 hours in a four-week period).

### 4.3.4 Planned Data Analyses

#### 4.3.4.1 Labels

Prediction windows are 24 hours in width. The 24-hour windows roll day-by-day starting at 6 am in the participant’s own time zone. The start and end date/time of past opioid use were reported on the first daily survey item. Each prediction window was labeled as a lapse if opioid use was reported as occurring between 6 am that day and 5:59 am the next morning. Windows with no reported opioid use were labeled as no lapse.

We ended up with a total of 93376 labels, with 2% labeled as lapses.

#### 4.3.4.2 Feature Engineering

Features will be derived from four sources:

1. Prediction window: We will dummy-code features for day of the week for the start of the prediction window.
2. Demographics: We will create dummy-coded features for age (18-25 years, 26-35 years, 36-45 years, 46-55 years, 56 years or older), personal income (less than $25,000, more than $25,000), sex at birth (male, female), race/ethnicity (non-Hispanic White vs. not White), geographic location (urban, suburban, rural), relationship status (in committed relationship, not in committed relationsip), education (high school or less, some college, college degree), and employment (employed, not employed).
3. Previous daily survey responses: We will create raw and change features using daily surveys in varying feature scoring epochs (i.e., 48, 72, and 168 hours) before the start of the prediction window for all daily survey items. Raw features will include min, max, and median scores for each daily survey item across all daily surveys in each epoch for that participant. We will also calculate change features by subtracting each participant’s baseline mean score for each daily survey item from their raw feature. These baseline mean scores will be calculated using all of their daily surveys collected from the start of their participation until the start of the prediction window. We also will create raw and change features based on the most recent response for each daily survey question and raw and change rate features from previously reported lapses and number of completed daily surveys.
4. Geolocation data: We will calculate raw and change features for time spent in locations harmful to recovery, time spent in locations helpful for recovery, time spent in pleasant locations, time spent in unpleasant locations, location variance, type of location, and activity done at location over varying feature scoring epochs (i.e., 6, 12, 24, 48, 72, and 168 hours).

Other generic feature engineering steps will include 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).

#### 4.3.4.3 Model Training and Evaluation

##### 4.3.4.3.1 Cross Validation

We will consider candidate Xgboost model configurations that differ across sensible values for key hyperparameters and 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).

We will use 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 will use 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 will be selected using median auROC across the 10 validation sets. Final performance of these best model configurations will be evaluated using median auROC across the 30 test sets.

##### 4.3.4.3.2 Bayesian Model

We will use 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 best model. We will use 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. We will set two random intercepts to account for our resampling method: one for the repeat, and another for the fold nested within repeat.

From the Bayesian model we will obtain the posterior distribution and Bayeisan CI for auROCs our best model. To evaluate our models’ overall performance we will report the median posterior probability for auROC and Bayesian CI. This represents our best estimate for the magnitude of the auROC parameter. If the credible interval does not contain .5 (chance performance), this provides strong evidence (> .95 probability) that our model is capturing signal in the data.

##### 4.3.4.3.3 Fairness Analyses

We will calculate the median posterior probability and 95% Bayesian CI for auROC for our best model separately by race/ethnicity (not White vs. non-Hispanic White), income (below $25,000 vs. above $25,000), sex at birth (female vs. male), and location (rural vs. other). We will conduct Bayesian group comparisons to assess the likelihood that each model performs differently by group. We will report the precise posterior probability for the difference in auROCs and the 95% Bayesian CIs for each model comparison.

##### 4.3.4.3.4 Feature Importance

We will calculate 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 best model. We will average 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 will create separate feature categories for each of the 15 daily survey questions, past opioid use, missing daily surveys, time spent in risky locations, and time spent at known locations (separate by type of location). We will calculate 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 will calculate 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 our model.

# 5. Chapter 5: State-space models for idiographic risk monitoring and recovery support recommendations

## 5.1 Introduction

Lapse risk is multidimensional. The extant relapse prevention literature suggests relapse is preceeded by a complex interplay of factors, including include emotional or cognitive states, environmental contingencies, and physiological states (K. A. Witkiewitz & Marlatt, 2007; K. Witkiewitz & Marlatt, 2007). For reasons related to burden and cost, researchers typically rely on a handful of course categories of risk-relevant features (e.g., self-reported craving, self-reported self-efficacy, geolocation-sensed time spent in risky locations) to model an unknown hidden state of an individual (i.e., their true lapse risk).

Moreover, lapse risk factors differ between individuals and within an individual over time. Personalized models (i.e., a model built for a specific individual using their own data) that make use of time-varying information in repeated measures may be able to better understand between and within subject differences in lapse risk trajectories, leading to more accurate predictions.

State-space models may be one approach for modeling the multidimensional, heterogeneous, and time-varying construct of lapse risk. State-space models are time series models that describe the relationship between the observed measured inputs and the unknown latent state while accounting for how this latent state evolves over time. They personalize lapse risk prediction by using an individual’s own data to make future predictions about lapse risk for that single individual. Furthermore, predictions can be made at any point into the future at a single prediction timepoint from a single model. These individual models that integrate time-varying information could improve performance for our lagged prediction models. They may also help mitigate issues of unfairness, as the model will weigh the individual’s own data more heavily than group level estimates.

## 5.2 Specific Aims

In this study, we will evaluate the performance and fairness of a hierarchichal Gaussian linear state-space model (also known as a dynamic linear model) for opioid lapse risk prediction using the same daily survey and geolocation data introduced in study 3 (Chapter 4).

Specifically, we will:

**1. Evaluate the performance (auROC) of personalized state-space models that predict immediate (next day) opioid lapse risk from geolocation and daily surveys.** In this aim we will compare the performance of state-space models to our traditional machine learning model in study 3 (chapter 4) to determine if individualized models that account for time are superior for lapse risk prediction.

**2. Use the same models from Aim 1 to evaluate their performance for predicting future lapse risk (i.e., lapse risk in the next two weeks and the next month).** A benefit of time series models, like state-space models, is that they could potentially improve the efficiency and performance of lagged prediction models (compared to traditional machine learning approaches). A single model can be used to predict a lapse in the next day or at any point in the future. We will evaluate how well our individual models predict lapse risk two weeks and one month into the future.

**3. Assess model fairness in model performance for immediate lapses across the same important subgroups assessed in study 3 - race/ethnicity (not White vs. non-Hispanic White), income (below $25,000 vs. above $25,000), sex at birth (female vs. male), and geographic location (rural vs. other).** Individual models may help mitigate issues of unfairness, seen in our previous group-level machine learning models, as the model will weigh the individual’s own data more heavily than group level estimates. Therefore, we will compare state-space model performance across different groups.

## 5.3 Methods

Refer to Chapter 4 [Section 4.3](#sec-methods) for a complete description of the data set and study procedures.

### 5.3.1 Planned Data Analyses

#### 5.3.1.1 State-space Model

State-space models consist of 1. a transition (or state) equation that describes how the latent state evolves over time based on its previous state and observed inputs (e.g., daily surveys, geolocation risk score), and 2. an observation equation that describes the functional relationship between the unobserved, latent states and observed inputs.

For both equations, we will use a linear formula with Gaussian noise.

**Transition equation:** , where represents the hidden state at the next timepoint, is the state transition matrix, is the hidden state at the current timepoint, and is zero mean Gaussian noise.

**Observation equation:** , where is the observed survey and Geolocation inputs, is the observation matrix, is the hidden state at the current timepoint and is zero mean Gaussian noise.

We will also assign prior distributions to all individual- and population-level model parameters, consistent with a Bayesian approach. We will use a Bayseian fitting approach called maximum a posteriori (MAP) estimation to establish prior distributions for the model parameters: transition matrix (), observation matrix (), and noise (,). These priors will be established from held-in data. Parameters and noise variance will be estimated for each individual in the held out data. Thus, model priors will be combined with observed data from a new individual to fit an idiographic model for that individual. When few data for the individual are available, the fit will rely more heavily on the prior distributions for the model parameters. As more data become available the model fit will primarily the participant’s data and the influence of the prior distributions diminishes.

State-space models use the observed data to estimate the unknown latent states, not to directly predict the outcome label. Therefore, state-space models can handle missing data without the need for imputation.

#### 5.3.1.2 Predictions

Prediction windows are 24 hours in width. The 24-hour windows roll day-by-day starting at 6 am in the participant’s own time zone. The start and end date/time of past opioid use were reported on the first daily survey item. Each prediction window was labeled as a lapse if opioid use was reported as occurring between 6 am that day and 5:59 am the next morning. Windows with no reported opioid use were labeled as no lapse.

The first label for each participant will be two weeks after their study start date. This will ensure we have at least two weeks of daily surveys and geolocation data.

All available data up until the start of the prediction window will be used. The predictors, or model inputs, will be the raw daily survey responses and a single geolocation risk score (calculated from the previous 24 hours of geolocation data). The formula for calculating the geolocation risk score will be informed by the top geolocation predictors that emerge in study 3 (Chapter 4).

In the first month models will be fit every 2 weeks. After month 1 the model will update (i.e., be refit with the additional data) each month.

#### 5.3.1.3 Model Evaluation

We will use participant-grouped 3 repeats of 10-fold cross-validation to assess model performance using area under the ROC curve (auROC). Grouped cross-validation assigns assigns all data from a participant into a single fold. Data from individuals in the held-in folds will be used to fit prior distributions for the model parameters.

Individual models will then be fit to each individual in the held-out fold. We will attach the group-level Bayesian priors generated from the held-in participants to help prevent over-fitting and improve stability of the models’ performance. Each individual model will output a predicted probability of lapse in the next day, in a 24-hour window two weeks from the prediction timepoint, and in a 24-hour window one month from the prediction timepoint.

We will use 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 each lag. We will use the rstanarm default autoscaled, weakly informative, data-dependent priors and set two random intercepts to account for our resampling method: one for the repeat, and another for the fold nested within repeat. We will report the median posterior probability for auROC and Bayesian CI for each lag. We will specify two sets of contrasts for model comparisons. The first contrast will compare the immediate next day state-space lapse prediction model to our next day traditional machine learning model (from study 3). The second set of contrasts will compare our three state-space models (immediate/no lag, 2-week lag, and 1-month lag). auROCs will be transformed using the logit function and regressed as a function of model contrast.

From the Bayesian model we will obtain the posterior distribution (transformed back from logit) and Bayeisan CIs for auROCs for the three state-space models and comparative traditional machine learning model. To evaluate our models’ overall performance we will report the median posterior probability for auROC and Bayesian CIs. We will then conducted Bayesian model comparisons using our two sets of contrasts. For both model comparisons, we will determine the probability that the models’ performances differed systematically from each other and report the precise posterior probability for the difference in auROCs and the 95% Bayesian CIs.

#### 5.3.1.4 Model Fairness

We will calculate the median posterior probability and 95% Bayesian CI for auROC for our immediate model separately by race/ethnicity (not White vs. non-Hispanic White), income (below $25,000 vs. above $25,000), sex at birth (female vs. male), and location (rural vs. other). We will conduct Bayesian group comparisons to assess the likelihood that each model performs differently by group. We will report the precise posterior probability for the difference in auROCs and the 95% Bayesian CIs for each model comparison.

*ARDI Alcohol-Attributable Deaths, US CDC*. (n.d.). https://nccd.cdc.gov/DPH\_ARDI/Default/Report.aspx?T=AAM&P=F1F85724-AEC5-4421-BC88-3E8899866842&R=EACE3036-77C9-4893-9F93-17A5E1FEBE01&M=7F40785C-D481-440A-970F-50EFBD21B35B&F=&D=.

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

Bharat, C., Webb, P., Wilkinson, Z., McKetin, R., Grebely, J., Farrell, M., Holland, A., Hickman, M., Tran, L. T., Clark, B., Peacock, A., Darke, S., Li, J.-H., & Degenhardt, L. (2023). Agreement between self-reported illicit drug use and biological samples: A systematic review and meta-analysis. *Addiction*, *118*(9), 1624–1648. <https://doi.org/10.1111/add.16200>

Bickman, L., Lyon, A. R., & Wolpert, M. (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), 271–276. <https://doi.org/10.1007/s10488-016-0718-5>

Bowen, S., Chawla, N., Grow, J., & Marlatt, G. A. (2021). *Mindfulness-Based Relapse Prevention for Addictive Behaviors: A Clinician’s Guide* (Second edition). The Guilford Press.

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

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

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

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

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

Centers for Disease Control and Prevention (CDC). (n.d.). Annual Average for United States 2011–2015 Alcohol-Attributable Deaths Due to Excessive Alcohol Use, All Ages. In *2022 Alcohol Related Disease Impact (ARDI) Application Website*. https://nccd.cdc.gov/DPH\_ARDI/Default/Default.aspx.

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

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

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

Derogatis, L.R. (n.d.). *Brief Symptom Inventory 18 - Administration, scoring, and procedures manual*. NCS Pearson.

DeRubeis, R. J. (2019). The history, current status, and possible future of precision mental health. *Behaviour Research and Therapy*, *123*, 103506. <https://doi.org/10.1016/j.brat.2019.103506>

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

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

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

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

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). *The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population*.

Epstein, D. H., Tyburski, M., Kowalczyk, W. J., Burgess-Hull, A. J., Phillips, K. A., Curtis, B. L., & Preston, K. L. (2020). Prediction of stress and drug craving ninety minutes in the future with passively collected GPS data. *Npj Digital Medicine*, *3*(1), 26. <https://doi.org/10/ghqvcw>

Fisher, A. J., Bosley, H. G., Fernandez, K. C., Reeves, J. W., Soyster, P. D., Diamond, A. E., & Barkin, J. (2019). Open trial of a personalized modular treatment for mood and anxiety. *Behaviour Research and Therapy*, *116*, 69–79. <https://doi.org/10.1016/j.brat.2019.01.010>

Florence, C., Luo, F., & Rice, K. (2021). The economic burden of opioid use disorder and fatal opioid overdose in the United States, 2017. *Drug and Alcohol Dependence*, *218*, 108350. <https://doi.org/10.1016/j.drugalcdep.2020.108350>

Friedman, J., Godvin, M., Shover, C. L., Gone, J. P., Hansen, H., & Schriger, D. L. (2022). Trends in Drug Overdose Deaths Among US Adolescents, January 2010 to June 2021. *JAMA*, *327*(14), 1398–1400. <https://doi.org/10.1001/jama.2022.2847>

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

Gabry, J., & Goodrich, B. (2023). Prior Distributions for rstanarm Models. In *CRAN R-Project*. https://cran.r-project.org/web/packages/rstanarm/vignettes/priors.html.

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

Greenfield, S. F., Brooks, A. J., Gordon, S. M., Green, C. A., Kropp, F., McHugh, R. K., Lincoln, M., Hien, D., & Miele, G. M. (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>

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

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

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). Springer.

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

Hedegaard, H., Miniño, A. M., Spencer, M. R., & Warner, M. (n.d.). *Drug overdose deaths in the United States, 1999–2020*.

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

*Increases in Drug and Opioid Overdose Deaths — United States, 2000–2014*. (n.d.). https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm.

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

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

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction (Abingdon, England)*, *114*(4), 609–619. <https://doi.org/10/gfsjzg>

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

Kennedy, A. P., Epstein, D. H., Phillips, K. A., & Preston, K. L. (2013). Sex differences in cocaine/heroin users: Drug-use triggers and craving in daily life. *Drug and Alcohol Dependence*, *132*(1-2), 29–37. <https://doi.org/10.1016/j.drugalcdep.2012.12.025>

Kilaru, A. S., Xiong, A., Lowenstein, M., Meisel, Z. F., Perrone, J., Khatri, U., Mitra, N., & Delgado, M. K. (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, H. R., & McKay, J. R. (2012). Personalized Treatment of Alcohol Dependence. *Current Psychiatry Reports*, *14*(5), 486–493. <https://doi.org/10.1007/s11920-012-0296-5>

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

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

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

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

Liese, B. S., & Beck, A. T. (2022). *Cognitive-Behavioral Therapy of Addictive Disorders* (First edition). The Guilford Press.

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

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

Marlatt, G. A., & Donovan, D. M. (Eds.). (2007). *Relapse Prevention, Second Edition: Maintenance Strategies in the Treatment of Addictive Behaviors* (2nd edition). The Guilford Press.

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

Marwick, A. E., & Boyd, D. (2018). Privacy at the Margins Understanding Privacy at the Margins—Introduction. *International Journal of Communication*, *12*(0), 9.

McHugh, R. K., Votaw, V. R., Sugarman, D. E., & Greenfield, S. F. (2018). Sex and gender differences in substance use disorders. *Clinical Psychology Review*, *66*, 12–23. <https://doi.org/10.1016/j.cpr.2017.10.012>

McKay, J. R. (2021). Impact of Continuing Care on Recovery From Substance Use Disorder. *Alcohol Research : Current Reviews*, *41*(1), 01. <https://doi.org/10.35946/arcr.v41.1.01>

McLellan, A. T., Lewis, D. C., O’Brien, C. P., & Kleber, H. D. (2000). Drug dependence, a chronic medical illness: Implications for treatment, insurance, and outcomes evaluation. *JAMA*, *284*(13), 1689–1695. <https://doi.org/10.1001/jama.284.13.1689>

Meyers, S. A., Earnshaw, V. A., D’Ambrosio, B., Courchesne, N., Werb, D., & Smith, L. R. (2021). The intersection of gender and drug use-related stigma: A mixed methods systematic review and synthesis of the literature. *Drug and Alcohol Dependence*, *223*, 108706. <https://doi.org/10.1016/j.drugalcdep.2021.108706>

Mohr, D. C., Zhang, M., & Schueller, S. M. (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>

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

Moshontz, H., Colmenares, A. J., Fronk, G. E., Sant’Ana, S. J., Wyant, K., Wanta, S. E., Maus, A., Jr, D. H. G., Shah, D., & Curtin, J. J. (2021). Prospective Prediction of Lapses in Opioid Use Disorder: Protocol for a Personal Sensing Study. *JMIR Research Protocols*, *10*(12), e29563. <https://doi.org/10.2196/29563>

Office of the Surgeon General (US), Center for Mental Health Services (US), & National Institute of Mental Health (US). (2001). [*Mental Health: Culture, Race, and Ethnicity*](https://www.ncbi.nlm.nih.gov/pubmed/20669516). Substance Abuse and Mental Health Services Administration (US).

Olfson, M., Mauro, C., Wall, M. M., Choi, C. J., Barry, C. L., & Mojtabai, R. (2022). Healthcare coverage and service access for low-income adults with substance use disorders. *Journal of Substance Abuse Treatment*, *137*, 108710. <https://doi.org/10.1016/j.jsat.2021.108710>

Pinedo, M. (2019). A current re-examination of racial/ethnic disparities in the use of substance abuse treatment: Do disparities persist? *Drug and Alcohol Dependence*, *202*, 162–167. <https://doi.org/10.1016/j.drugalcdep.2019.05.017>

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

Pulick, E., Curtin, J. J., & Mintz, Y. (under review). *Idiographic lapse prediction with state space modeling: Algorithm development and validation*.

RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, Inc.

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

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

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

Socías, M. E., Volkow, N., & Wood, E. (2016). Adopting the “cascade of care” framework: An opportunity to close the implementation gap in addiction care? *Addiction*, *111*(12), 2079–2081. <https://doi.org/10.1111/add.13479>

Soyster, P. D., Ashlock, L., & Fisher, A. J. (2022). Pooled and person-specific machine learning models for predicting future alcohol consumption, craving, and wanting to drink: A demonstration of parallel utility. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *36*(3), 296–306. <https://doi.org/10.1037/adb0000666>

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

Studies, I. of M. (US). C. on E. and L. I. R. to the I. of W. in C., Mastroianni, A. C., Faden, R., & Federman, D. (1994). NIH Revitalization Act of 1993 Public Law 103-43. In *Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies: Volume I*. National Academies Press (US).

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.

Substance Abuse and Mental Health Services Administration (US), & Office of the Surgeon General (US). (2016). [*Facing Addiction in America*](https://www.ncbi.nlm.nih.gov/pubmed/28252892). US Department of Health and Human Services.

Tai, B., & Volkow, N. D. (2013). Treatment for Substance Use Disorder: Opportunities and Challenges under the Affordable Care Act. *Social Work in Public Health*, *28*(3-4), 165–174. <https://doi.org/10.1080/19371918.2013.758975>

United States. Department of Transportation. National Highway Traffic Safety Administration. National Center for Statistics and Analysis. (2024). *Traffic Safety Facts 2022 Data: Alcohol-Impaired Driving* (DOT HS 813 578).

Vannicelli, M., & Nash, L. (1984). Effect of Sex Bias on Women’s Studies on Alcoholism. *Alcoholism: Clinical and Experimental Research*, *8*(3), 334–336. <https://doi.org/10.1111/j.1530-0277.1984.tb05523.x>

Veinot, T. C., Mitchell, H., & Ancker, J. S. (2018). Good intentions are not enough: How informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association: JAMIA*, *25*(8), 1080–1088. <https://doi.org/10.1093/jamia/ocy052>

Walitzer, K. S., & Dearing, R. L. (2006). Gender differences in alcohol and substance use relapse. *Clinical Psychology Review*, *26*(2), 128–148. <https://doi.org/10.1016/j.cpr.2005.11.003>

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

WHO ASSIST Working Group. (2002). [The Alcohol, Smoking and Substance Involvement Screening Test (ASSIST): Development, reliability and feasibility](https://www.ncbi.nlm.nih.gov/pubmed/12199834). *Addiction (Abingdon, England)*, *97*(9), 1183–1194.

Witkiewitz, K. A., & Marlatt, G. A. (Eds.). (2007). *Therapist’s Guide to Evidence-Based Relapse Prevention* (1st edition). Academic Press.

Witkiewitz, K., & Marlatt, G. A. (2004). Relapse prevention for alcohol and drug problems: That was zen, this is tao. *American Psychologist*, *59*(4), 224–235. <https://doi.org/10.1037/0003-066X.59.4.224>

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

Wright, A. G. C., & Zimmermann, J. (2019). Applied Ambulatory Assessment: Integrating Idiographic and Nomothetic Principles of Measurement. *Psychological Assessment*, *31*(12), 1467–1480. <https://doi.org/10.1037/pas0000685>

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

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>

Wyant, K., Sant’Ana, S., Punturieri, C., Yu, J., Fronk, G., Kornfield, R., Wanta, S., Maggard, C., Herrmann, M., & Curtin, J. (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*.

Zhou, M., Fukuoka, Y., Mintz, Y., Goldberg, K., Kaminsky, P., Flowers, E., & Aswani, A. (2018). Evaluating Machine Learning–Based Automated Personalized Daily Step Goals Delivered Through a Mobile Phone App: Randomized Controlled Trial. *JMIR mHealth and uHealth*, *6*(1), e9117. <https://doi.org/10.2196/mhealth.9117>

1. Of course, these are not the only things needed for a successful recovery monitoring support system. For example, people must be willing and able to provide sensing data and the system must able to provide risk-relevant feedback to the individual in a useful and clinically helpful way. While important, these questions are outside the scope of the current proposal (see Wyant et al. (2023) and Wyant et al. (in prep)). [↑](#footnote-ref-23)
2. Features for income and employment were inadvertently excluded from all models. [↑](#footnote-ref-36)
3. In early exploratory analyses, we evaluated auROCs of all four algorithms using grouped k-fold cross-validation for models based on preliminary feature engineering using the EMAs. XGBoost models consistently outperformed other algorithms such that we focused all further development on XGBoost to reduce the substantial computational time associated with model training and evaluation. [↑](#footnote-ref-41)
4. Although XGBoost ignores participant-level information, we do leverage this information to some degree by including change features that anchor participants’ EMA responses to their own previous responses. [↑](#footnote-ref-84)
5. Defined as scores >2.2 or 2.8, respectively, on the psychosis or paranoia scales of the Symptom Checklist–90 (Derogatis, L.R., n.d.) [↑](#footnote-ref-92)
6. 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-104)
7. 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-109)
8. The poverty cutoff was defined from the 2024 federal poverty line for the 48 contiguous United States. Participants at or below $15,060 annual income were categorized as below poverty. [↑](#footnote-ref-111)
9. 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 non-advantaged group being absent in any single fold. [↑](#footnote-ref-112)