The Feasibility and Equity of Geolocation Data for Lapse Prediction in AUD

Claire Punturieri

John J. Curtin

2024-10-20

## Introduction

About 1 in 10 adults in the United States met diagnostic criteria for alcohol use disorder (AUD) in 2022 (“Highlights for the 2022 National Survey on Drug Use and Health,” n.d.). While some individuals will experience natural recovery (i.e., improvement without intervention) (Tucker, Chandler, and Witkiewitz 2020), for others AUD will present as a chronic, relapsing disorder marked by periods of recovery interspersed with returns back to harmful use (“The Science of Drug Use and Addiction: The Basics NIDA Archives” n.d.; Brandon, Vidrine, and Litvin 2007). For such individuals, continued monitoring may be beneficial in assisting with maintaining recovery goals and identifying precipitants to lapses, or single instances of goal-inconsistent use that may lead to relapse (Witkiewitz and Marlatt 2004). One sustainable and scalable way to provide this continuous monitoring to individuals who need it most is through developing algorithms to predict lapses using both personal sensing data and machine learning.

Personal sensing data are data derived via embedded sensors in technology such as smartphones, smartwatches, or wearables (Mohr, Zhang, and Schueller 2017). Because these devices are already ubiquitous within our day-to-day lives, these data can feasibly be collected unobtrusively and continuously for clinical purposes as well. Importantly, these data do not require individuals to change their behavior or routines in any way, making data collection both sustainable and ecologically valid. Moreover, when paired with machine learning models, statistical patterns connecting antecedents to lapse derived from these data (e.g., changes in mood, difficulty with close social connections, proximity to risky locations) to true lapse events can be uncovered. This is crucial for several reasons: 1) even when someone anticipates an oncoming lapse, it may be difficult to pinpoint the specific driving forces behind it; 2) these precipitating factors will have great variation both between- and within-people; and 3) uncovering these factors may help relieve some of the cognitive burden of recovery (i.e., constant monitoring of potential environmental risk factors).

### Geolocation Data for Risk Monitoring

Recovery and return to use are dynamic processes. Factors that contribute to both maintenance of recovery and return to use change from person-to-person and from moment-to-moment. A shift in social supports (e.g., a move, a break-up) may precede a lapse for one individual but not another. Time spent in locations where alcohol is available (e.g., bars, restaurants, concert venues) may precede a given lapse for another individual, but will not necessarily precede future lapses in that same individual. In order to best capture this fluidity, the ideal data type used within continuous risk monitoring systems should be able to provide a correspondingly appropriate level of granularity. Geolocation data are one such promising source.

Geolocation data consist of latitude and longitude coordinates and can be sampled at regular intervals using applications on smartphones with little to no input from the user beyond initial set-up. Many smartphones and smartwatches automatically collect these data by default. This fact, paired with increasing rates of smartphone ownership, suggest that there is high potential for these data to be feasibly harnessed for use in a risk-monitoring system (Areàn, Hoa Ly, and Andersson 2016). The importance of location, such as environmental cues or one’s perceived riskiness of a setting, has been shown to play an important role in lapse (Janak and Chaudhri 2010; Maureen A. Walton et al. 2003; M. A. Walton, Reischl, and Ramanthan 1995). This link with lapse risk has led to integrating coping skills that target substance-associated contexts in several treatment strategies like mindfulness-based relapse prevention (LeCocq et al. 2020). These findings underscore not only the potential wealth of information relating to lapse risk that an individual’s location can provide, but also demonstrate the proven integration of location information into treatment. Furthermore, geolocation data have been specifically identified as being of particular use in both understanding the precipitants to harmful substance use and its effective treatment (Stahler, Mennis, and Baron 2013).

Within the substance use literature, geolocation data have historically been used to examine risky locations, such as the influence of neighborhood characteristics on use (Epstein et al. 2014; Kwan et al. 2019) and individual physical proximity to locations of potential or past harmful use such as bars (either estimated using geofencing or user-defined) (Attwood et al. 2017; Carreiro et al. 2021; Gonzalez and Dulin 2015; Gustafson et al. 2014; Naughton et al. 2016). Several of the applications implemented in these studies enable real-time notifications about locations to their users (e.g., a pop-up message on a smartphone which reads *“You are entering a high-risk zone”*).

On the other hand, affective scientists have focused have focused more closely on the intersection of geolocation with factors relating to mood. Geolocation data have been used to estimate loneliness and isolation (Doryab et al. 2019), to demonstrate increases in positive affect from seeking out novel environments (Heller et al. 2020), and to quantify depressive symptoms (Raugh et al. 2020). Moreover, these data have not only been harnessed to measure mood symptoms, but to also predict their emergence (for review, see Shin and Bae 2023).

An integration across these subfields can be in part accomplished by enriching geolocation data with brief, intermittent surveys probing specific information about frequently visited locations. For example, some of the more nuanced facets captured within location are associations with others (or lack thereof, e.g., social isolation), associations with previous drinking behaviors (e.g., whether or not alcohol is present), and associations with affect (i.e., negative versus positive emotions tied to a given location). Using a combination of these insights in building a model for use within a continuous risk monitoring system will result in the ability to identify a wider variety of potential lapse precipitants and, therefore, theoretically perform better across heterogenous experiences of recovery.

### Model Evaluation

Data selection, however, is only one component of the successful development of a continuous risk monitoring algorithm. It is imperative that these models be rigorously evaluated using performance metrics and eventually tested using independent observations (i.e., using the model to predict outcomes for individuals whose data were not used in development). This workflow in machine learning is what enables researchers to anticipate how well a model can be expected to generalize to new populations and is key when aiming to develop algorithms for real-world healthcare implementation.

Although performance metrics like model accuracy, for example, have been standard reporting practice for years, recent literature has begun to urge researchers to add assessments of how *fair* a model is (Rajkomar et al. 2018; Wawira Gichoya et al. 2021). A fair algorithm is one with no preference in performance with respect to inherent or acquired characteristics (e.g., sex at birth, race, socioeconomic status; (X. Wang, Zhang, and Zhu 2022)). In the context of a continuous risk monitoring algorithm for AUD, this would mean that lapse predictions are reasonably accurate and do not favor or disadvantage any particular group solely due to group membership status.

The motivating factors behind this call to action are clear. In the broader context of health-related data, historical patterns of healthcare inequities will almost certainly and unavoidably be embedded within data used to train algorithms. These inequities may unintentionally be carried forward, and maybe exacerbated, by machine learning models if not critically examined. Without a careful eye towards these foreseeable consequences, monitoring algorithms run the risk of providing sub-optimal mental healthcare to individuals who already face disadvantages.

There are at least two clear pathways whereby algorithmic performance may diverge between subgroups. The first is non-representative sampling. During model training, an algorithm learns to associate patterns in observations with a given outcome. Model performance will therefore suffer if there is limited information from which to learn (e.g., few instances of a given demographic trait), particularly without the use of resampling techniques to amend these imbalances (Japkowicz 2000; A. Wang, Ramaswamy, and Russakovsky 2022).

The second is historical biases in the literature. This is of particular concern specifically in the context of AUD, where the literature has historically been built upon research developed with male, predominantly white, participants. Despite the call to action brought forth by the NIH through their *Guidelines on Inclusion of Women and Minorities in Research*, recent work has highlighted that seminal research in the field on medications for the treatment of AUD have failed to consistently report participant demographics (Schick, Spillane, and Hostetler 2020). This lack of reporting makes it difficult to assess how and if this lack of representation is being corrected. By the very nature of its historically limited participant pool, AUD research and its theory have been developed from a particular perspective using a limited group of individuals. This means that the variables that researchers decide are important to measure and input into models, informed by knowledge of AUD theory, will inherently be biased and may favor these groups. Therefore, researchers can also not assume that balanced classes are enough to compensate for biases brought on by the broader societal context.

Both of these facts motivate the reasoning behind examining algorithmic fairness in the context of developing continuous risk monitoring systems. In order for these continuous risk monitoring systems to be implemented in the real-world, these models must both be developed outright and rigorously evaluated on both standard performance metrics and for algorithmic fairness.

### The Current Study

This study utilized geolocation data collected from smartphones and corresponding self-reported contextual information for frequently visited locations to build a machine learning model to predict next-day alcohol use among individuals with a diagnosis of AUD and a recovery goal of abstinence. Model features were engineered from both raw geolocation data and change in these data over the previous 6, 12, 24, 48, 72, and 168 hours.

Here we present characterization of model performance for this prediction model across validation sets, including a discussion of feature importance and an evaluation of model fairness. This study constitutes a preliminary evaluation of a model designed to predict lapse back to alcohol use using minimally burdensome data that has the potential to be integrated within a continuous risk monitoring platform.

## Methods

### Participants

One hundred and forty six individuals in early-recovery (1-8 weeks of abstinence) for AUD were recruited from the Madison area to take part in a three-month study on how mobile health technology can provide recovery support between 2017 and 2019 (R01 AA024391). Recruitment approaches included social media platforms (e.g., Facebook), television and radio advertisements, and clinic referrals. Prospective participants completed a phone screen to assess match with eligibility criteria ([Table 1](#tbl-elig)). Participants were excluded if they exhibited severe symptoms of paranoia or psychosis (a score <= 2.24 on the SCL-90 psychosis scale or a score <= 2.82 on the SCL-90 paranoia scale administered at screening). Participants completed a baseline measure of demographics and other constructs relevant to lapse at the screening visit, which was used for fairness assessments ([Table 2](#tbl-demo-1)).

### Procedure

Participants enrolled in a three-month study consisting of five in-person visits, daily surveys, and continuous passive monitoring of geolocation data. Following screening and enrollment visits in which participants consented to participate, learned how to manage location sharing (i.e., turn off location sharing when desired), and reported frequently visited locations, participants completed three follow-up visits one month apart. At each visit, participants were asked questions about frequently visited (>2 times during the course of the previous month) locations ([Table 3](#tbl-context)). Participants were debriefed at the third and final follow-up visit. Participants were expected to provide continuous geolocation data while on study. Other personal sensing data streams (EMA, cellular communications, sleep quality, and audio check-ins) were collected as part of the parent grant’s aims (R01 AA024391).

### Geolocation and contextual data

To enable collection of geolocation data, participants downloaded either the Moves app or the FollowMee app during the intake visit. Moves was bought-out and subsequently deprecated while the study was ongoing (July 2018) and data collection continued using FollowMee until the end of the study. Both apps continuously tracked location via GPS and WiFi positioning technology.

After completion of the study, data were processed to filter out duplicated points, fast movement speeds (>100mph), sudden positional jumps, and periods of long duration suggesting sampling error issues (>24 hours with no movement or >2 hours with a positional jump of more than 0.31 miles or 500 meters). Data points were classified as “in transit” when spacing between individual positions suggested a movement speed of greater than 4mph per NIH health guidelines (“Physical Activity Guidelines for Americans, 2nd Edition,” n.d.). Participants were considered to be at a known contextual location if they were within 0.031 miles (50 meters) of a reported frequently visited location.

### Data analytic strategy

Data preprocessing, modeling, and Bayesian analyses were done in R using the tidymodels ecosystem (Kuhn and Wickham 2020). Models were trained using high-throughput computing resources provided by the University of Wisconsin Center for High Throughput Computing (Center for High Throughput Computing 2006).

### Lapses

Alcohol lapses were used as the outcome variable in this study and were used to provide labels for model training, for testing model performance, and for testing issues of algorithmic fairness. Future lapse occurrence (here conceptualized as next-day lapse) was predicted in 24-hour windows, beginning at 4:00am on a participant’s second day of participation to ensure one full day of data collection for the first window, and at every subsequent day on study thereafter. *Lapse* and *no lapse* occurrences were identified from the daily survey question, *“Have you drank any alcohol that you have not yet reported?”*. Participants who responded *yes* to this question were then asked to report the date and hour of the start and the end of the drinking episode. In this case, the prediction window was labeled *lapse*. Prediction windows were labeled *no lapse* if no alcohol use was reported within that window.

### Feature engineering

Feature engineering is the process of creating variables (or *“features”*) from unprocessed data and was used to transform raw data from geolocation data collected the prior day. Separate feature categories were created for the six contextual geolocation categories (presented in [Table 3](#tbl-context)) and for three movement-based categories: variability in location, time spent outside of the home in the evening, and time spent in transit. All features were calculated both as raw and change features based on previous geolocation data (i.e., change from past 6, 12, 24, 48, 72, and 168 hour periods) in order to capture individual variation.

Imputation of missing data and removal of zero-variance features are additional general processing steps that were completed during feature engineering.

### Algorithm development & performance

We trained and assessed several configurations of an XGBoost machine learning algorithm. The choice of using an XGBoost algorithm was motivated by two main reasons: 1) the calculation of Shapley values, used to understand the relative contributions of features in predictions, is optimized for XGBoost; and 2) previous work in our lab has made use of XGBoost algorithms in model development (Wyant et al. 2024) and the ability to eventually integrate features across models is of high priority. Configurations of the XGBoost algorithm varied across a relevant and appropriate range of model-specific hyperparameters (mtry, tree depth, learning rate) as well as resampling techniques (up-sampling of the positive class, lapse, and down-sampling of the negative class, no lapse ranging from 1:1 to 5:1) to account for the class imbalance in our outcome variable.

Models were trained and assessed using participant-grouped, nested *k*-fold cross-validation. Grouped cross-validation ensures that all data from a given participant are retained as either held-in or held-out. This prevents the introduction of bias from a participant’s data being used to predict 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 (Jonathan, Krzanowski, and McCarthy 2000). Results from the validations sets (i.e., inner loops) are presented here.

The primary performance metric for model selection and evaluation of the validation set was area under the Receiver Operating Characteristic (auROC) curve (Kuhn and 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 levels of class imbalance. The best model configuration was selected using median auROC across all validation sets.

Shapley values were computed in log-odd units as our interpretability metric. Shapley values measure the unique contribution of features in an algorithm’s predictions and therefore identify the relative importance of difference features (Lundberg and Lee 2017). Global feature importance for each broad feature category was calculated by averaging the absolute values of Shapley values across all observations per feature category. Highly important features represent relevant, actionable potential antecedents to lapse (and therefore points of intervention) that will be relevant in the future development of a continuous risk monitoring system. However, these are descriptive analyses because standard errors or other indices of uncertainty for importance scores are not available for Shapley values.

Finally, a Bayesian hierarchical generalized linear model was used to estimate the posterior probability distributions and 95% Bayesian credible intervals (CIs) for auROC.

### Algorithmic fairness

Subgroups were defined on the basis of personal individual characteristics divided such that groups reflected coarse dichotomies of groups which experience relatively increased and decreased societal privilege. This resulted in four broad classes: white versus non-white, younger than 55 versus equal to or older than 55, above or below the federal poverty line (MBA 2024), and sex at birth (male versus female). A Bayesian hierarchical generalized linear model was used to estimate the posterior probability distributions and 95% Bayesian CIs for auROC across these four classes. Model contrasts were used in order to identify the likelihood of differential performance of our model between subgroups within each class.

## Results

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

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

### Demographics

A total of 192 individuals were eligible to participate in the study, of which 191 consented to participate and 169 enrolled in the study. Fifteen participants were excluded prior to the first monthly follow-up visit. One participant was excluded for not maintaining a recovery goal of abstinence during their time on study. Two participants were excluded due to evidence of low compliance and careless responding. A further five individuals were excluded due to poor geolocation data quality as a result of insufficient data (resulting from software incompatibility), resulting in a final sample size of 146.

The average age of the final sample was 40.9 years (SD = 12 years, range = 21-72 years). There was an approximately equal number of men (n = 74, 50.7%) and women (n = 72, 49.3%). The majority of the sample was White/Caucasian (n = 127, 86.99%) and non-Hispanic (n = 142, n = 97%). The mean income of participants was $34,408 (SD = $32,259, range = $0-$200,000). On average, participants self-reported a mean number of 8.9 DSM-V symptoms of AUD (range = 4-11). A detailed breakdown of participant characteristics is presented in [Table 4](#tbl-demo-2).

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

### Model Evaluation

We selected and evaluated the best performing XGBoost model across all validation sets. This may result in a slight optimization bias in our model performance, though we believe this is largely offset through our use of 10 x 30 cross-validation (which averages model performance across 300 folds). Cross-validation maintains separation between data used to train the models, select the best models, and evaluate those best models, thereby minimizing optimization bias (Jonathan, Krzanowski, and McCarthy 2000). Evaluating the validation set was important to do because model development is still in progress, and as such it would not have been appropriate to examine independent test set performance at this stage.

The median auROC over all validation sets was 0.712. auROCs above .7 are typically considered as having “acceptable” performance and indicate that the model correctly assigns a higher probability of lapse to a positive case (rather than a negative case) 70% of the time (Mandrekar 2010). [Figure 1](#fig-auroc-histogram) displays a histogram of model performance distribution across all folds.

Posterior probability distributions for the auROCs for our best performing validation set model were then used to formally characterize model performance. The median auROC was 0.714 (95% CI [0.70-0.73]), indicating that there is a probability > .95 that our model is performing above chance (i.e., auROC > .5; [Figure 2](#fig-pp)).

Next, we performed model calibration in order to improve our trust in model predictions. Results of model calibration are displayed in [Figure 3](#fig-calibration), showing that this model *over* predicts lapse probability even after calibrating the model. In other words, our model is more likely to predict that an individual will lapse than the true observed rate of lapse in our sample.

Finally, a receiver operating characteristic curve is displayed in [Figure 4](#fig-auroc-plot), representing aggregate predicted lapse logistic (calibrated) probabilities across all validation sets.

### Feature Importance

Global importance (mean absolute Shapley values) for feature categories is shown in [Figure 5](#fig-shaps-group). Three aggregated feature categories were identified as being particularly important in contributing to model predictions: time spent at risky locations, time spent at different types of location, and time spent at locations with varying levels of alcohol availability. Other aggregated feature groups, both context-supplemented and independent, did not appear to be strong, unique global contributors to model predictions.

### Algorithmic Fairness

[Figure 6](#fig-fairness-subgroups) shows differences in model performance across race (*N* white = 127, *N* non-white = 19), sex (*N* male = 74, *N* female = 72), age (*N* younger than 55 = 126, *N* older than or equal to 55 = 20), and income (*N* below federal poverty line = 48, *N* above federal poverty line = 98). All group comparisons were reliably different (probability > .95) across models, such that identities with higher assumed privilege were associated with improved model performance. White, non-Hispanic participants demonstrated 0.055 greater model performance than Hispanic and/or non-white participants (range=0.027-0.084, probability=1.000). Male participants demonstrated 0.037 greater model performance than female participants (range=0.013-0.060, probability=0.998). Younger participants demonstrated 0.107 greater model performance than older participants (range=0.079-0.133, probability=1.000). Finally, participants above the poverty line demonstrated 0.056 greater model performance than those below the poverty line (range=0.033-0.078, probability=1.000).

## Discussion

### Model Performance

Our day-level model of lapse prediction using geolocation data performs at an “acceptable” threshold [auROC between .7 and .8; Mandrekar (2010)], suggesting that, while there is still a considerable amount of improvement to be made in model performance, geolocation data can predict future alcohol lapse in the next day with fair sensitivity and specificity. Bayesian model comparisons corroborated that this model performed unilaterally better than chance (auROC = .5). This study also provided explanatory insights by way of quantifying feature importance as well as a crucial examination of model fairness across subgroups, detailed in the following section.

Model calibration is the process of fine-tuning model predictions to more closely align with the true likelihood of a given outcome and was carried out in order to improve our trust in model predictions (in this case, to better align model predictions against the observed lapse rate in our sample; Dormann (2020)). XGBoost is not a probabilistic model and as such it is expected that probabilities would need to be calibrated. Yet, even following calibration, our model overpredicts occurrence of lapses in our sample. Identifying this in the validation stage enables us to make further changes to our algorithm, such as refitting the model, prior to moving onto the final evaluation stage (Van Calster et al. 2019). Two potential solutions would be to explore other calibration methods outside of logistic calibration, such as beta calibration, and to examine the distribution of our feature set. Typically class imbalances result in an overprediction of the majority class and not the minority class (here, we would expect an overprediction of no lapses relative to lapses). Instead, we see the opposite. This may be because our features favor the minority class. For example, time spent at risky locations, our most predictive feature, is a unipolar scale that is focused on the *riskiness* and not the *protectiveness* of a location. However, it should be noted that this oversensitivity may not be an issue depending on what information we hope to relay to individuals using a continuous risk monitoring system. In the future development of such a system, we are not necessarily interested in communicating exact probabilities to individuals about their lapse risk (such as, *“There is a 92% chance that you lapse back to use today”*). Rather, we are more interested in communicating *relative* levels of risk (such as, *“You are at a low risk level of lapse today,”* or, *“Your risk of lapse is higher this week compared to last week”* where *low risk* corresponds to a designated probability threshold).

We used Shapley values to quantify global feature importance. The top performing Shapley values were time spent at risky locations, time spent at different location types (e.g., home, bars, work), and time spent at locations with varying levels of alcohol availability. Time spent at risky locations was associated with a 2x *greater* log-odds change in lapse risk as compared to the next highest performing feature of time spent at different location types. These results are well-aligned with the extant AUD literature, notably the focus of high-risk situations as an immediate determinant to relapse within the relapse prevention model (Larimer, Palmer, and Marlatt 1999; Marlatt and Gordon 1985). These three features were all generated utilizing additional context supplied by participants after a given location was identified as frequently visited (> 2x in the previous month). However, it should be noted that these features may be able to be generated without user feedback. For example, location types could be classified using public map data and consumer data could be used to identify establishments that sell alcohol. This could further reduce the burden on an individual using a continuous risk monitoring system by not requiring individual input. On the other hand, self-classifying locations as risky might be encoding nuance that could not be feasibly obtained using public data. For instance, a location might be labeled as risky from user input because it is a person-specific triggering location (e.g., scene of a traumatic event).

Interestingly, location valence, or the emotion tied to a given location, is the fourth-highest Shapley value, yet appears to be minimally contributing to model predictions. This may be because participants were asked retrospectively about these locations at one month follow-up visits, and so our measures of emotional quality of a location may be too distal to be meaningful. While other measures of global feature importance were low, it will be important to examine local feature importance to better understand their potential utility. For example, a given feature may have low global importance (low mean absolute Shapley value), but may have a unidirectional relationship with a particular class (e.g., always associated with lapse predictions).

### Model Fairness

All models exhibited differential performance across subgroups of race/ethnicity, sex at birth, age, and income, such that model performance was worse for non-white, female, and older participants, as well as those below the poverty line. One of the driving forces behind these disparities is a lack of representation in these data. For instance, even collapsing across dichotomous categories for both race and ethnicity (i.e., white and non-Hispanic and non-white and/or Hispanic), our non-white and/or Hispanic sample only reflects 13% of the total sample. Outside of recruiting a more diverse sample in future studies, one potential solution is to synthetically up-sample cases of the minority class (in this case, non-white participants) such that the model has more data on which to base its predictions (Kabir et al. 2024).

Yet, we also see divergent performance across men and women in our sample, a class that *is* well-balanced (*N* male = 74, *N* female = 72). This may be due, as predicted, to how constructs of AUD have been conceptualized. The historical literature has been built primarily on studying the experiences of AUD in men. Our features may not be as salient of predictors for lapse for women as they are for men, resulting in the differences in auROCs that we see here. This underscores the importance of *procedural* fairness (i.e., considering the fairness of feature inputs themselves) in addition to outcome fairness (Grgić-Hlača et al. 2018). Using this *means* and *ends* approach may result in a more refined understanding of different facets impacting performance across groups.

If maintaining a top-down approach to feature selection from the literature is a primary goal, more research will need to be done to create a better understanding of AUD experiences and recovery for specific groups who have historically been under-studied in this area. There have been several recent calls to action with respect to studying AUD across sex suggesting that factors which precipitate lapse and contribute to recovery maintenance may be different for women as compared to men (McCaul et al. 2019; McCrady, Epstein, and Fokas 2020). Alternatively, bottom-up approaches, such as allowing an algorithm to identify patterns in the data that are most predictive of lapse, may get around issues of bringing in biased features from the literature. This may be at the expense of generating features that are not easily clinically translatable or particularly meaningful. Together, these results suggest that we are seeing the aftereffects of both statistical bias (i.e., inadequate sampling) and societal bias (i.e., constructs which are of limited value to certain groups) in our model (Mitchell et al. 2021).

While it is important to consider ways to remedy this in our model so that it may be able to be equitably used in the real-world, it is also important to note that the goal of this work is not to suggest that quantitative definitions of fairness are sufficient to fix deeply rooted issues of societal injustice (Green 2022; Green and Hu, n.d.; Ochigame 2020).

### Limitations and future directions

This study is an important step forward in building a model using geolocation data to predict next-day lapse back to drinking in individuals with a diagnosis of AUD and a recovery goal of abstinence. However, there are several limitations to this work.

First, our model has only a fair level of performance. A simple strategy to improve this is to explore additional features. Risk-relevant features that can be generated without additional participant-supplied contextual information include geographical spatial risk indicators and circadian rhythm data. Risk-terrain modeling is a spatial analysis technique which enables the estimation of environmental risk factors on a given outcome, such as contact with high liquor outlet density or high crime areas preceding lapse (Gajos et al. n.d.). Circadian rhythm data, on the other hand, can be extracted using a Lomb-Scargle periodogram to derive the spectrum of geolocation data, therefore quantifying the (ir)regularity of daily activity (VanderPlas 2018). Circadian rhythm patterns derived from geolocation data have been widely used in the affective science literature for predicting mood (Chikersal et al. 2021; Saeb, Zhang, Karr, et al. 2015; Saeb, Zhang, Kwasny, et al. 2015). Moreover, encouraging the maintenance of a daily routine may be helpful in the context of AUD recovery specifically (Hühne, Hoch, and Landgraf 2021; Tamura et al. 2021). These features are not only clinically meaningful, but are also clinically *intervenable*. In the context of a continuous risk monitoring system, individuals could receive notifications relating to increasing amounts of time spent in riskier areas (beyond self-identified risky locations) or irregularity of movement patterns suggesting circadian disruption. Potential interventions in these examples would be to suggest relocating to a “safer” area or encouraging following a more regular schedule. We might also be interested in features that, while not being clinically-actionable, may add more predictive value to our model by increasing precision. Examples of such features are day of the week and weather, two factors that contribute to daily behavior and thus geolocation data (Heller et al. 2020).

Secondly, we assessed auROCs from our validation sets against change performance (auROC = .5). While this is an important standard to meet, it may be more useful to instead design a baseline model against which to compare performance. For example, we could compare the predictive power of our model against a model which predicts future lapse from past lapse behavior (as in Wyant et al. 2024). Alternatively, we might be interested in comparing our model against another model using more simplistic features that we think might be related to both geolocation patterns and drinking behavior, such as a model which uses day of the week to predict lapse. A more rigorous model contrast will set a higher standard with which to evaluate our model against prior to implementation in a real-world context.

Next, we have examined the feature importance and have identified three geolocation features that appear to be strongly predictive of next-day lapse: time spent at risky locations, time spent across various types of places, and time spent at locations where alcohol is available. These features are collapsed across particular response-types and across time periods. A first next step will be to break down these top performing features into their subcomponents, such as looking at the contributions of time spent at high, medium, and low risk locations to lapse predictions separately. It may also be of value to look at particular time points. Duration features in this model were calculated at intervals starting from 6 hours prior up until 168 hours prior. In the context of geolocation data, it may be interesting to more closely exmaine more proximal as opposed to distal time points. These two avenues are both still at the global (i.e., across all subjects) level of feature importance. We can also examine how these features perform at an individual level for a given prediction to get a sense of how stable they are over time, within and between people.

Finally, following final evaluation using an independent test set, a future direction of this work is to integrate the most-predictive features derived from this geolocation-based model with features from an ecological momentary assessment (EMA)-based model designed within our lab (Wyant et al. 2024). Geolocation features may provide unique insight into lapse risk that survey-based assessments do not (or cannot) capture, and may also enable the use of less-frequent questionnaires. The final model built from these two separate endeavors will be used to optimize the development of a continuous risk monitoring system for AUD in a recently funded grant (1R01AA031762-01).

### Conclusion

This study demonstrates that it is feasible to predict lapse with a fair level of accuracy using geolocation data, suggesting that geolocation data is a viable supplement for risk prediction monitoring systems. However, our model demonstrates differential performance across vulnerable subgroups. Moving forward, additional risk-relevant features will be added to the model in an effort to improve prediction and the final model will be evaluated.

## References

Areàn, Patricia A., Kien Hoa Ly, and Gerhard Andersson. 2016. “[Mobile Technology for Mental Health Assessment](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4969703).” *Dialogues in Clinical Neuroscience* 18 (2): 163–69.

Attwood, Sophie, Hannah Parke, John Larsen, and Katie L. Morton. 2017. “Using a Mobile Health Application to Reduce Alcohol Consumption: A Mixed-Methods Evaluation of the Drinkaware Track & Calculate Units Application.” *BMC Public Health* 17 (1): 394. <https://doi.org/10.1186/s12889-017-4358-9>.

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

Carreiro, Stephanie, Melissa Taylor, Sloke Shrestha, Megan Reinhardt, Nicole Gilbertson, and Premananda Indic. 2021. “Realize, Analyze, Engage (RAE): A Digital Tool to Support Recovery from Substance Use Disorder.” *Journal of Psychiatry and Brain Science* 6: e210002. <https://doi.org/10.20900/jpbs.20210002>.

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

Chikersal, Prerna, Afsaneh Doryab, Michael Tumminia, Daniella K. Villalba, Janine M. Dutcher, Xinwen Liu, Sheldon Cohen, et al. 2021. “Detecting Depression and Predicting Its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection.” *ACM Trans. Comput.-Hum. Interact.* 28 (1): 3:1–41. <https://doi.org/10.1145/3422821>.

Dormann, Carsten F. 2020. “Calibration of Probability Predictions from Machine-Learning and Statistical Models.” *Global Ecology and Biogeography* 29 (4): 760–65. <https://doi.org/10.1111/geb.13070>.

Doryab, Afsaneh, Daniella K. Villalba, Prerna Chikersal, Janine M. Dutcher, Michael Tumminia, Xinwen Liu, Sheldon Cohen, et al. 2019. “Identifying Behavioral Phenotypes of Loneliness and Social Isolation with Passive Sensing: Statistical Analysis, Data Mining and Machine Learning of Smartphone and Fitbit Data.” *JMIR mHealth and uHealth* 7 (7): e13209. <https://doi.org/10.2196/13209>.

Epstein, David H., Matthew Tyburski, Ian M. Craig, Karran A. Phillips, Michelle L. Jobes, Massoud Vahabzadeh, Mustapha Mezghanni, Jia-Ling Lin, C. Debra M. Furr-Holden, and Kenzie L. Preston. 2014. “Real-Time Tracking of Neighborhood Surroundings and Mood in Urban Drug Misusers: Application of a New Method to Study Behavior in Its Geographical Context.” *Drug and Alcohol Dependence* 134 (January): 22–29. <https://doi.org/10.1016/j.drugalcdep.2013.09.007>.

Gajos, Jamie M., Alejandro Gimenez-Santana, Jeffery T. Walker, Karen L. Cropsey, Scott T. Walters, and Michael S. Businelle. n.d. “Using Risk Terrain Modeling and Geographically-Explicit Ecological Momentary Assessments to Examine Alcohol Use in Adults Experiencing Homelessness.” *Journal of Crime and Justice* 0 (0): 1–14. Accessed October 19, 2024. <https://doi.org/10.1080/0735648X.2024.2396406>.

Gonzalez, Vivian M., and Patrick L. Dulin. 2015. “Comparison of a Smartphone App for Alcohol Use Disorders with an Internet-based Intervention Plus Bibliotherapy: A Pilot Study.” *Journal of Consulting and Clinical Psychology* 83 (2): 335–45. <https://doi.org/10.1037/a0038620>.

Green, Ben. 2022. “Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness.” *Philosophy & Technology* 35 (4): 90. <https://doi.org/10.1007/s13347-022-00584-6>.

Green, Ben, and Lily Hu. n.d. “The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning.”

Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. “Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning.” *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1). <https://doi.org/10.1609/aaai.v32i1.11296>.

Gustafson, David H., Fiona M. McTavish, Ming-Yuan Chih, Amy K. Atwood, Roberta A. Johnson, Michael G. Boyle, Michael S. Levy, et al. 2014. “A Smartphone Application to Support Recovery From Alcoholism: A Randomized Clinical Trial.” *JAMA Psychiatry* 71 (5): 566. <https://doi.org/10.1001/jamapsychiatry.2013.4642>.

Heller, Aaron S., Tracey C. Shi, C. E. Chiemeka Ezie, Travis R. Reneau, Lara M. Baez, Conor J. Gibbons, and Catherine A. Hartley. 2020. “Association Between Real-World Experiential Diversity and Positive Affect Relates to Hippocampal-Striatal Functional Connectivity.” *Nature Neuroscience* 23 (7): 800–804. <https://doi.org/10.1038/s41593-020-0636-4>.

“Highlights for the 2022 National Survey on Drug Use and Health.” n.d.

Hühne, Anisja, Eva Hoch, and Dominic Landgraf. 2021. “DAILY—A Personalized Circadian Zeitgeber Therapy as an Adjunctive Treatment for Alcohol Use Disorder Patients: Study Protocol for a Randomized Controlled Trial.” *Frontiers in Psychiatry* 11 (January). <https://doi.org/10.3389/fpsyt.2020.569864>.

Janak, Patricia H., and Nadia Chaudhri. 2010. “The Potent Effect of Environmental Context on Relapse to Alcohol-Seeking After Extinction.” *The Open Addiction Journal* 3 (January): 76–87. <https://doi.org/10.2174/1874941001003010076>.

Japkowicz, Nathalie. 2000. “The Class Imbalance Problem: Significance and Strategies.” In *Proc. Of the Int’l Conf. On Artificial Intelligence*, 56:111–17.

Jonathan, P., W. J. Krzanowski, and W. V. McCarthy. 2000. “On the Use of Cross-Validation to Assess Performance in Multivariate Prediction.” *Statistics and Computing* 10 (3): 209–29. <https://doi.org/10.1023/A:1008987426876>.

Kabir, Md Alamgir, Mobyen Uddin Ahmed, Shahina Begum, Shaibal Barua, and Md Rakibul Islam. 2024. “Balancing Fairness: Unveiling the Potential of SMOTE-Driven Oversampling in AI Model Enhancement.” In *Proceedings of the 2024 9th International Conference on Machine Learning Technologies*, 21–29. ICMLT ’24. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3674029.3674034>.

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

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

Kwan, Mei-Po, Jue Wang, Matthew Tyburski, David H. Epstein, William J. Kowalczyk, and Kenzie L. Preston. 2019. “Uncertainties in the Geographic Context of Health Behaviors: A Study of Substance Users’ Exposure to Psychosocial Stress Using GPS Data.” *International Journal of Geographical Information Science* 33 (6): 1176–95. <https://doi.org/10.1080/13658816.2018.1503276>.

Larimer, Mary E., Rebekka S. Palmer, and G. Alan Marlatt. 1999. “[Relapse Prevention: An Overview of Marlatt’s Cognitive-Behavioral Model](https://www.ncbi.nlm.nih.gov/pubmed/10890810).” *Alcohol Research & Health* 23 (2): 151.

LeCocq, Mandy Rita, Patrick A. Randall, Joyce Besheer, and Nadia Chaudhri. 2020. “Considering Drug-Associated Contexts in Substance Use Disorders and Treatment Development.” *Neurotherapeutics: The Journal of the American Society for Experimental NeuroTherapeutics* 17 (1): 43–54. <https://doi.org/10.1007/s13311-019-00824-2>.

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

Mandrekar, Jayawant 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): 1315–16. <https://doi.org/10.1097/JTO.0b013e3181ec173d>.

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

MBA, Tony Bartels, DVM. 2024. “2024 Federal Poverty Rates Published: Why That Matters for Your Student Loans.” *VIN Foundation*. https://vinfoundation.org/2024-federal-poverty-rates-published-why-that-matters-for-your-student-loans/.

McCaul, M. E., D. Roach, D. S. Hasin, C. Weisner, G. Chang, and R. Sinha. 2019. “Alcohol and Women: A Brief Overview.” *Alcoholism, Clinical and Experimental Research* 43 (5): 774. <https://doi.org/10.1111/acer.13985>.

McCrady, Barbara S., Elizabeth E. Epstein, and Kathryn F. Fokas. 2020. “Treatment Interventions for Women With Alcohol Use Disorder.” *Alcohol Research : Current Reviews* 40 (2): 08. <https://doi.org/10.35946/arcr.v40.2.08>.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D’Amour, and Kristian Lum. 2021. “Algorithmic Fairness: Choices, Assumptions, and Definitions.” *Annual Review of Statistics and Its Application* 8 (Volume 8, 2021): 141–63. <https://doi.org/10.1146/annurev-statistics-042720-125902>.

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

Naughton, Felix, Sarah Hopewell, Neal Lathia, Rik Schalbroeck, Chloë Brown, Cecilia Mascolo, Andy McEwen, and Stephen Sutton. 2016. “A Context-Sensing Mobile Phone App (Q Sense) for Smoking Cessation: A Mixed-Methods Study.” *JMIR mHealth and uHealth* 4 (3): e106. <https://doi.org/10.2196/mhealth.5787>.

Ochigame, Rodrigo. 2020. “The Long History of Algorithmic Fairness.” *Phenomenal World*.

“Physical Activity Guidelines for Americans, 2nd Edition.” n.d.

Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. “Ensuring Fairness in Machine Learning to Advance Health Equity.” *Annals of Internal Medicine* 169 (12): 866–72. <https://doi.org/10.7326/M18-1990>.

Raugh, Ian M., Sydney H. James, Cristina M. Gonzalez, Hannah C. Chapman, Alex S. Cohen, Brian Kirkpatrick, and Gregory P. Strauss. 2020. “Geolocation as a Digital Phenotyping Measure of Negative Symptoms and Functional Outcome.” *Schizophrenia Bulletin* 46 (6): 1596–1607. <https://doi.org/10.1093/schbul/sbaa121>.

Saeb, Sohrab, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. 2015. “Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study.” *Journal of Medical Internet Research* 17 (7): e4273. <https://doi.org/10.2196/jmir.4273>.

Saeb, Sohrab, Mi Zhang, Mary Kwasny, Christopher J. Karr, Konrad Kording, and David C. Mohr. 2015. “The Relationship Between Clinical, Momentary, and Sensor-Based Assessment of Depression.” In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 229–32. <https://doi.org/10.4108/icst.pervasivehealth.2015.259034>.

Schick, Melissa R., Nichea S. Spillane, and Katherine L. Hostetler. 2020. “A Call to Action: A Systematic Review Examining the Failure to Include Females and Members of Minoritized Racial/Ethnic Groups in Clinical Trials of Pharmacological Treatments for Alcohol Use Disorder.” *Alcoholism: Clinical and Experimental Research* 44 (10): 1933–51. <https://doi.org/10.1111/acer.14440>.

Shin, Jaeeun, and Sung Man Bae. 2023. “A Systematic Review of Location Data for Depression Prediction.” *International Journal of Environmental Research and Public Health* 20 (11): 5984. <https://doi.org/10.3390/ijerph20115984>.

Stahler, Gerald J., Jeremy Mennis, and David A. Baron. 2013. “Geospatial Technology and the "Exposome": New Perspectives on Addiction.” *American Journal of Public Health* 103 (8): 1354–56. <https://doi.org/10.2105/AJPH.2013.301306>.

Tamura, Eduardo K., Kallyane S. Oliveira-Silva, Felipe A. Ferreira-Moraes, Eduardo A. V. Marinho, and Natalí N. Guerrero-Vargas. 2021. “Circadian Rhythms and Substance Use Disorders: A Bidirectional Relationship.” *Pharmacology Biochemistry and Behavior* 201 (February): 173105. <https://doi.org/10.1016/j.pbb.2021.173105>.

“The Science of Drug Use and Addiction: The Basics NIDA Archives.” n.d. https://archives.nida.nih.gov/publications/media-guide/science-drug-use-addiction-basics. Accessed October 7, 2024.

Tucker, Jalie A., Susan D. Chandler, and Katie Witkiewitz. 2020. “Epidemiology of Recovery From Alcohol Use Disorder.” *Alcohol Research : Current Reviews* 40 (3): 02. <https://doi.org/10.35946/arcr.v40.3.02>.

Van Calster, Ben, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, et al. 2019. “Calibration: The Achilles Heel of Predictive Analytics.” *BMC Medicine* 17 (1): 230. <https://doi.org/10.1186/s12916-019-1466-7>.

VanderPlas, Jacob T. 2018. “Understanding the Lomb–Scargle Periodogram.” *The Astrophysical Journal Supplement Series* 236 (1): 16. <https://doi.org/10.3847/1538-4365/aab766>.

Walton, M. A., T. M. Reischl, and C. S. Ramanthan. 1995. “Social Settings and Addiction Relapse.” *Journal of Substance Abuse* 7 (2): 223–33. <https://doi.org/10.1016/0899-3289(95)90006-3>.

Walton, Maureen A., Frederic C. Blow, C. Raymond Bingham, and Stephen T. Chermack. 2003. “Individual and Social/Environmental Predictors of Alcohol and Drug Use 2 Years Following Substance Abuse Treatment.” *Addictive Behaviors* 28 (4): 627–42. <https://doi.org/10.1016/s0306-4603(01)00284-2>.

Wang, Angelina, Vikram V Ramaswamy, and Olga Russakovsky. 2022. “Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation.” In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 336–49. FAccT ’22. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3531146.3533101>.

Wang, Xiaomeng, Yishi Zhang, and Ruilin Zhu. 2022. “A Brief Review on Algorithmic Fairness.” *Management System Engineering* 1 (1): 7. <https://doi.org/10.1007/s44176-022-00006-z>.

Wawira Gichoya, Judy, Liam G. McCoy, Leo Anthony Celi, and Marzyeh Ghassemi. 2021. “Equity in Essence: A Call for Operationalising Fairness in Machine Learning for Healthcare.” *BMJ Health & Care Informatics* 28 (1): e100289. <https://doi.org/10.1136/bmjhci-2020-100289>.

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

Wyant, Kendra, Sarah June Kittleson Sant’Ana, Gaylen Fronk, and John J. Curtin. 2024. “Machine Learning Models for Temporally Precise Lapse Prediction in Alcohol Use Disorder.” *Psychopathology and Clinical Science*. <https://doi.org/10.31234/osf.io/cgsf7>.

## Tables and Figures

### Table 1: Eligibility Criteria

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 1: Eligibility criteria for study enrollment. \*\*Personal or study-provided.   | Eligibility Criteria | | --- | | >= 18 years of age | | Ability to read and write in English | | Diagnosis of moderate AUD (>= 4 self-reported DSM-5 symptoms) | | Abstinent from alcohol for 1-8 weeks | | Willing to use only one smartphone\*\* while on study | |

### Table 2: Collected demographic information

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Demographic and relevant alcohol use history variables sampled at screening visit.   | Variable | Measure | | --- | --- | | Demographics | Age | |  | Sex | |  | Race | |  | Ethnicity | |  | Employment | |  | Income | |  | Marital Status | | Alcohol | Alcohol Use History | |  | DSM-5 Checklist for AUD | |  | Young Adult Alcohol Problems Test | |  | WHO-The Alcohol, Smoking and Substance Involvement Screening Test | |

### Table 3: Contextual geolocation information

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Location information collected from frequently visited locations.   | Question | Responses | | --- | --- | | Address |  | | Type of place | Work, School, Volunteer, healthcare, Home of a friend, Home of a family member, Liquor store, Errands (e.g., grocery store, post office), Coffee shop or cafe, Restaurant, Park, Bar, Gym or fitness center, AA or recovery meeting, Religious location (e.g., church, mosque, temple), Other | | Have you drank alcohol here before? | No, Yes | | Is alcohol available here? | No, Yes | | How would you describe your experiences here? | Pleasant, Unpleasant, Mixed, Neutral | | Does being at this location put you at any risk to begin drinking? | No risk, Low risk, Medium risk, High risk | | Did the participant identify this place as a risky location they are trying to avoid now that they are sober? | No, Yes | |

### Table 4: Demographic responses

|  |
| --- |
| Table 4: Demographic characteristics of the sample (N = 146). |

### Figure 1: auROC histogram

|  |
| --- |
| Figure 1: Area under the receiver operating characteristic (auROC) curves for each of 300 (10 x 30) cross validation splits. The dashed line represents the median auROC across all 300 splits. |

Source: [Validation auROC histogram and posterior probability distribution](https://jjcurtin.github.io/study_gps/notebooks/auROC_distribution_posterior-preview.html#cell-fig-auroc-histogram)

### Figure 2: auROC posterior probability distribution

|  |
| --- |
| Figure 2: Posterior probability distribution of model performance with a 95% credible interval. The dashed line represents median auROC across the sampling distribution, while the dotted line represents chance performance (auROC = 0.50). |

Source: [Validation auROC histogram and posterior probability distribution](https://jjcurtin.github.io/study_gps/notebooks/auROC_distribution_posterior-preview.html#cell-fig-pp)

### Figure 3: Calibration plot

|  |
| --- |
| Figure 3: Comparison between raw (uncalibrated) and logistic (calibrated) probabilities. Predicted lapse probability represents the predicted probabilities derived from the model, whereas observed lapse rate reflects the true rate of lapses in the data. The dashed y = x line represents perfect performance, where predicted probabilities reflect true probabilities. Each point represents the midpoint of a given bin, which increase by 10% (i.e., 5% represents the midpoint from 0-10%). |

Source: [Validation set calibration](https://jjcurtin.github.io/study_gps/notebooks/calibration_plot-preview.html#cell-fig-calibration)

### Figure 4: Calibrated auROC plot

|  |
| --- |
| Figure 4: Area under the receiver operating characteristic (auROC) curve for overall validation set performance across all possible classification thresholds. |

Source: [Calibrated auROC](https://jjcurtin.github.io/study_gps/notebooks/auROC_plot-preview.html#cell-fig-auroc-plot)

### Figure 5: Global Feature Importance

|  |
| --- |
| Figure 5: Grouped SHAP values displaying relative feature importance calculated using mean absolute values. Larger log-odds values indicate greater contribution to predictions in the model. |

Source: [SHAPs](https://jjcurtin.github.io/study_gps/notebooks/shaps-preview.html#cell-fig-shaps-group)

### Figure 6: Posterior probability distribution 95% credible intervals by subgroup

|  |
| --- |
| Figure 6: 95% credible intervals across posterior probability distributions by subgroup at differential levels of privilege. |

Source: [Subgroup Analysis](https://jjcurtin.github.io/study_gps/notebooks/fairness-preview.html#cell-fig-fairness-subgroups)