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

John J. Curtin

Author note

Correspondence concerning this article should be addressed to .

Introduction

About 1 in 10 adults in the United States met diagnostic criteria for alcohol use disorder (AUD) in 2022 [1]. While some individuals will experience natural recovery (i.e., improvement without intervention) [2], for others AUD will present as a chronic, relapsing disorder marked by periods of recovery interspersed with returns back to harmful use [3], [4]. For such individuals, continued monitoring may be beneficial in assisting with the maintenance of recovery goals and in identifying precipitants to lapses, or single instances of goal-inconsistent use that may lead to relapse [5]. 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 [6]. Because these devices are already ubiquitous within our day-to-day lives, one benefit of porting these data to clinical use is their proven ability to be collected unobtrusively and continuously. Importantly, these data do not require individuals to change their behavior or routines in any way. 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 both to 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 [7]. 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 [8]–[10]. This link with lapse risk has translated into the integration of coping skills that target substance-associated contexts in several treatment strategies like mindfulness-based relapse prevention [11]. 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 [12].

Within the substance use literature, geolocation data have historically been used to examine risky locations, such as the influence of neighborhood characteristics on use [13], [14] and individual physical proximity to locations of potential or past harmful use such as bars (either estimated using geofencing or user-defined) [15]–[19]. 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 instead more closely on factors relating to mood. Geolocation data have been used to estimate loneliness and isolation [20], to demonstrate increases in positive affect from seeking out novel environments [21], and to quantify depressive symptoms [22]. Moreover, these data have not only been harnessed to measure mood symptoms, but to also predict their emergence [23].

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).

Model Evaluation

Data selection, however, is only one component of the successful development of a continuous risk monitoring algorithm. Following development, 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 could be expected to generalize to new populations and is key when aiming to develop algorithms for real-world healthcare implementation. While standard performance metrics like model accuracy, for example, have been standard reporting practice for years, recent literature has begun to urge researchers to also include assessments of how *fair* a model is [24], [25]. A fair algorithm is one with no preference in performance with respect to inherent or acquired characteristics (e.g., gender, race, socioeconomic status; [26]). 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 health care 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 health care 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 [27], [28].

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 [29]. 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 particular 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.

The Current Study

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. To this end, 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 lapse 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 in a validation set, 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). 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).

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). 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 [30]. 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 [31]. Models were trained using high-throughput computing resources provided by the University of Wisconsin Center for High Throughput Computing [32].

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 be 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) 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 [33] 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 [34]. 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 [35]. 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. 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 [36]. 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 [37], 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

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.

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 [34]. 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 [38]. Figure 1 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).

Next, we performed model calibration in order to improve our trust in model predictions. Results of model calibration are displayed in Figure 3, 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, 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. 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 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; [38]], 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 (.5). This study also provided explanatory insights by way of quantifying feature importance as well as a crucial examination of fairness of model performance 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; [39]). 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 [40]. 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 than lapses). Instead, we see the opposite. This may be because our features may be biased in favor of 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 risk monitoring system. In the future development of such a system, we are not interested in communicating exact probabilities to individuals about their lapse risk (e.g., *"There is a 92% chance that you lapse back to use today"*). Rather, we are more interested in communicating *relative* levels of risk (e.g., *"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 [41], [42]. 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 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 (i.e., 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 four broad classes 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. It is likely that we are seeing effects of both lack of representation as well as historically entrenched biases in the literature in our sample. For instance, even collapsing across dichotomous categories for both race and ethnicity (i.e., white and non-white), 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 upsample cases of the minority class (in this case, non-white participants) such that the model has more data on which to base its predictions [43].

We also see divergent performance across men and women in our sample, a class that is well-balanced (*N* male = 74, *N* female = 72). Our features may not be as salient of predictors for lapse for women as they are for men, perhaps stemming from a historical literature that has been built primarily from studying the experiences of AUD in white men. 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 sample [44].

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 [45]–[47].

Limitations and future directions

- Baseline model?
- Add in more affective features
- Add in risk-terrain modeling features
- Add in other important features that could contribute to movement patterns like day of the week and weather
- Test final model
- Further calibration

- Break down three top performing features into their subcomponents (i.e., high risk locations, medium risk locations, and low risk locations; yes there is alcohol available here or no there is not alcohol available here) to obtain a more nuanced understanding of model performance.
- More information about feature importance can be gleaned from examining local Shapley values using a Sina plot.

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

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

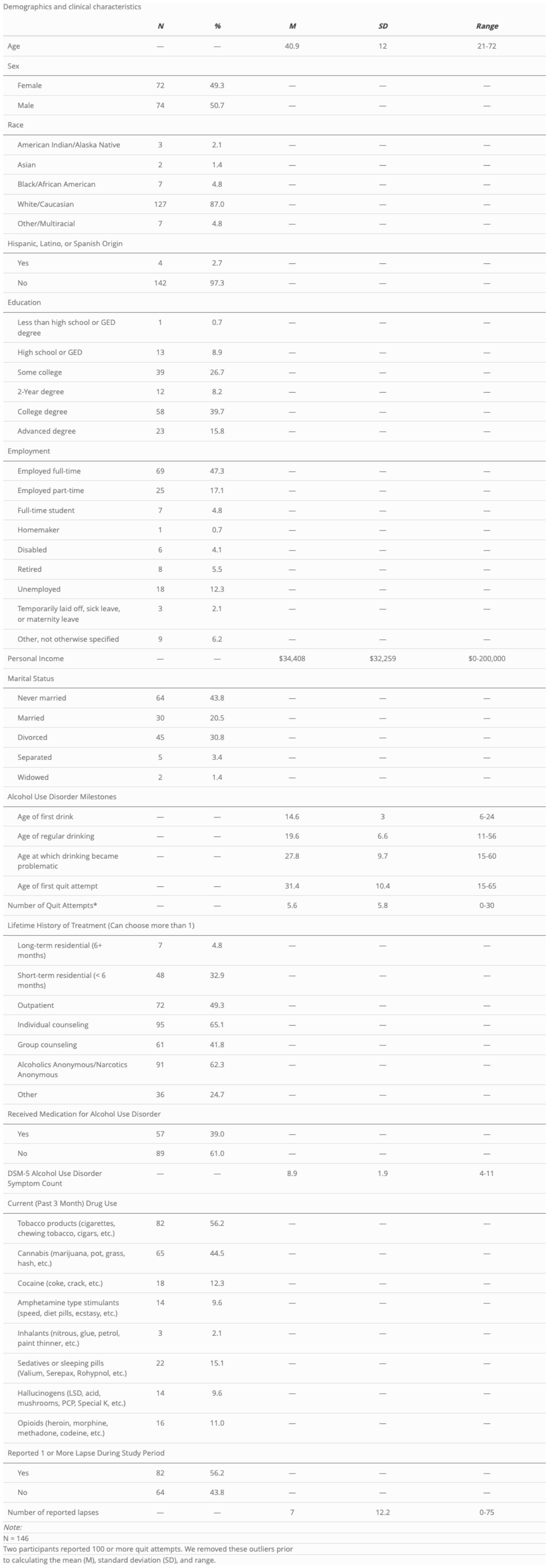
Table 2: Demographic and relevant alcohol use history variables sampled at screening visit.

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: Location information collected from frequently visited locations.

Question	Responses
Address	
Type of place	Work, School, Volunteer, Health care, 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 characteristics of the sample (N = 146).



Demographics and clinical characteristics

	N	*%*	*M*	*SD*	*Range*
Age	—	—	40.9	12	21-72
Sex					
Female	72	49.3	—	—	—
Male	74	50.7	—	—	—
Race					
American Indian/Alaska Native	3	2.1	—	—	—
Asian	2	1.4	—	—	—
Black/African American	7	4.8	—	—	—
White/Caucasian	127	87.0	—	—	—
Other/Multiracial	7	4.8	—	—	—
Hispanic, Latino, or Spanish Origin					
Yes	4	2.7	—	—	—
No	142	97.3	—	—	—
Education					
Less than high school or GED degree	1	0.7	—	—	—
High school or GED	13	8.9	—	—	—
Some college	39	26.7	—	—	—
2-Year degree	12	8.2	—	—	—
College degree	58	39.7	—	—	—
Advanced degree	23	15.8	—	—	—
Employment					
Employed full-time	69	47.3	—	—	—
Employed part-time	25	17.1	—	—	—
Full-time student	7	4.8	—	—	—
Homemaker	1	0.7	—	—	—
Disabled	6	4.1	—	—	—
Retired	8	5.5	—	—	—
Unemployed	18	12.3	—	—	—
Temporarily laid off, sick leave, or maternity leave	3	2.1	—	—	—
Other, not otherwise specified	9	6.2	—	—	—
Personal Income	—	—	$34,408	$32,259	$0-200,000
Marital Status					
Never married	64	43.8	—	—	—
Married	30	20.5	—	—	—
Divorced	45	30.8	—	—	—
Separated	5	3.4	—	—	—
Widowed	2	1.4	—	—	—
Alcohol Use Disorder Milestones					
Age of first drink	—	—	14.6	3	6-24
Age of regular drinking	—	—	19.6	6.6	11-56
Age at which drinking became problematic	—	—	27.8	9.7	15-60
Age of first quit attempt	—	—	31.4	10.4	15-65
Number of Quit Attempts*	—	—	5.6	5.8	0-30
Lifetime History of Treatment (Can choose more than 1)					
Long-term residential (6+ months)	7	4.8	—	—	—
Short-term residential (< 6 months)	48	32.9	—	—	—
Outpatient	72	49.3	—	—	—
Individual counseling	95	65.1	—	—	—
Group counseling	61	41.8	—	—	—
Alcoholics Anonymous/Narcotics Anonymous	91	62.3	—	—	—
Other	36	24.7	—	—	—
Received Medication for Alcohol Use Disorder					
Yes	57	39.0	—	—	—
No	89	61.0	—	—	—
DSM-5 Alcohol Use Disorder Symptom Count	—	—	8.9	1.9	4-11
Current (Past 3 Month) Drug Use					
Tobacco products (cigarettes, chewing tobacco, cigars, etc.)	82	56.2	—	—	—
Cannabis (marijuana, pot, grass, hash, etc.)	65	44.5	—	—	—
Cocaine (coke, crack, etc.)	18	12.3	—	—	—
Amphetamine type stimulants (speed, diet pills, ecstasy, etc.)	14	9.6	—	—	—
Inhalants (nitrous, glue, petrol, paint thinner, etc.)	3	2.1	—	—	—
Sedatives or sleeping pills (Valium, Serepax, Rohypnol, etc.)	22	15.1	—	—	—
Hallucinogens (LSD, acid, mushrooms, PCP, Special K, etc.)	14	9.6	—	—	—
Opioids (heroin, morphine, methadone, codeine, etc.)	16	11.0	—	—	—
Reported 1 or More Lapse During Study Period					
Yes	82	56.2	—	—	—
No	64	43.8	—	—	—
Number of reported lapses	—	—	7	12.2	0-75

Note:

N = 146

Two participants reported 100 or more quit attempts. We removed these outliers prior to calculating the mean (M), standard deviation (SD), and range.

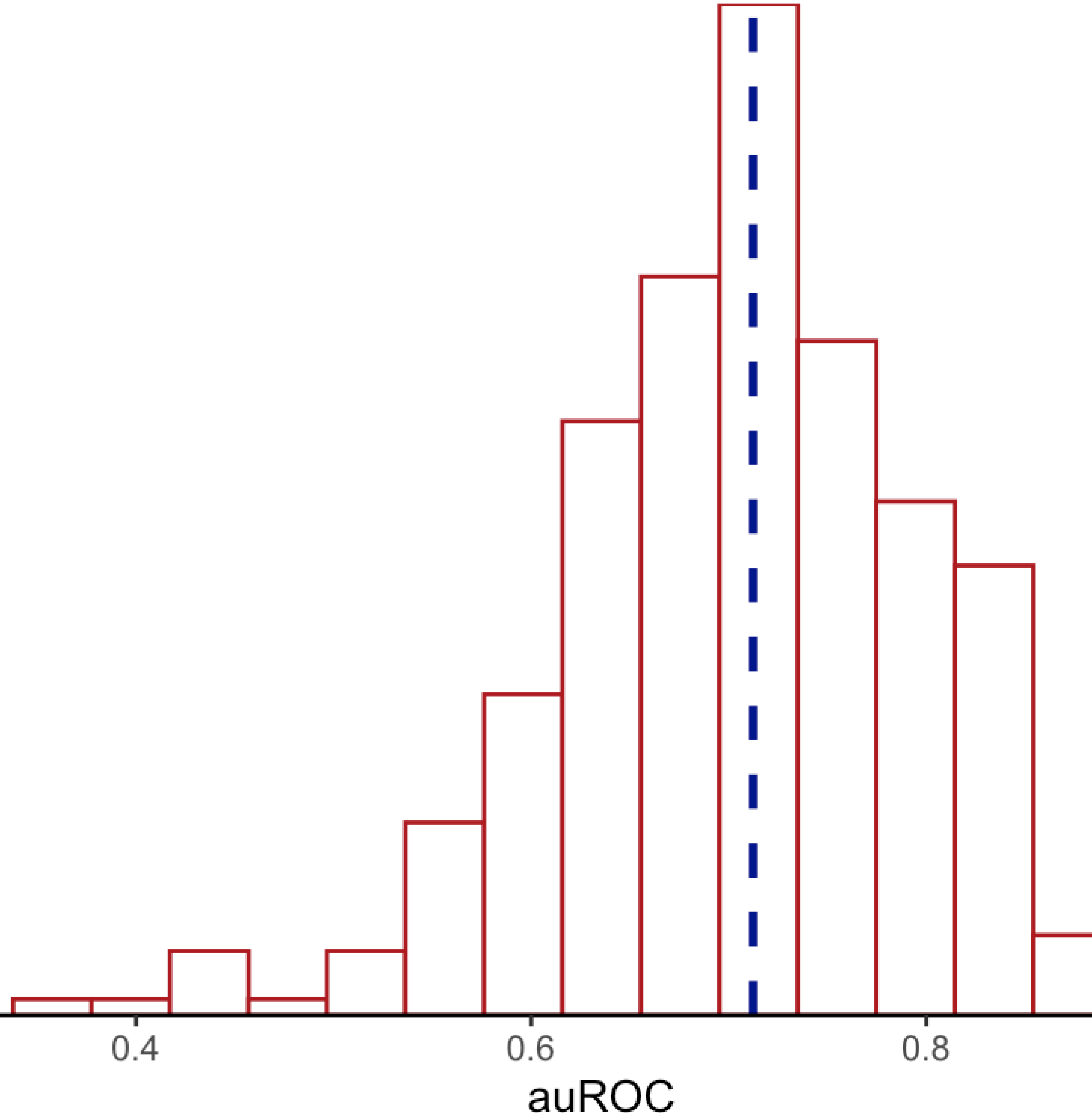


Figure 1: Area under the receiver operating characteristic (auROC) curves for each of 300 (10 x 30) cross validation splits. The dashed line represents the median auROC across all 300 splits.

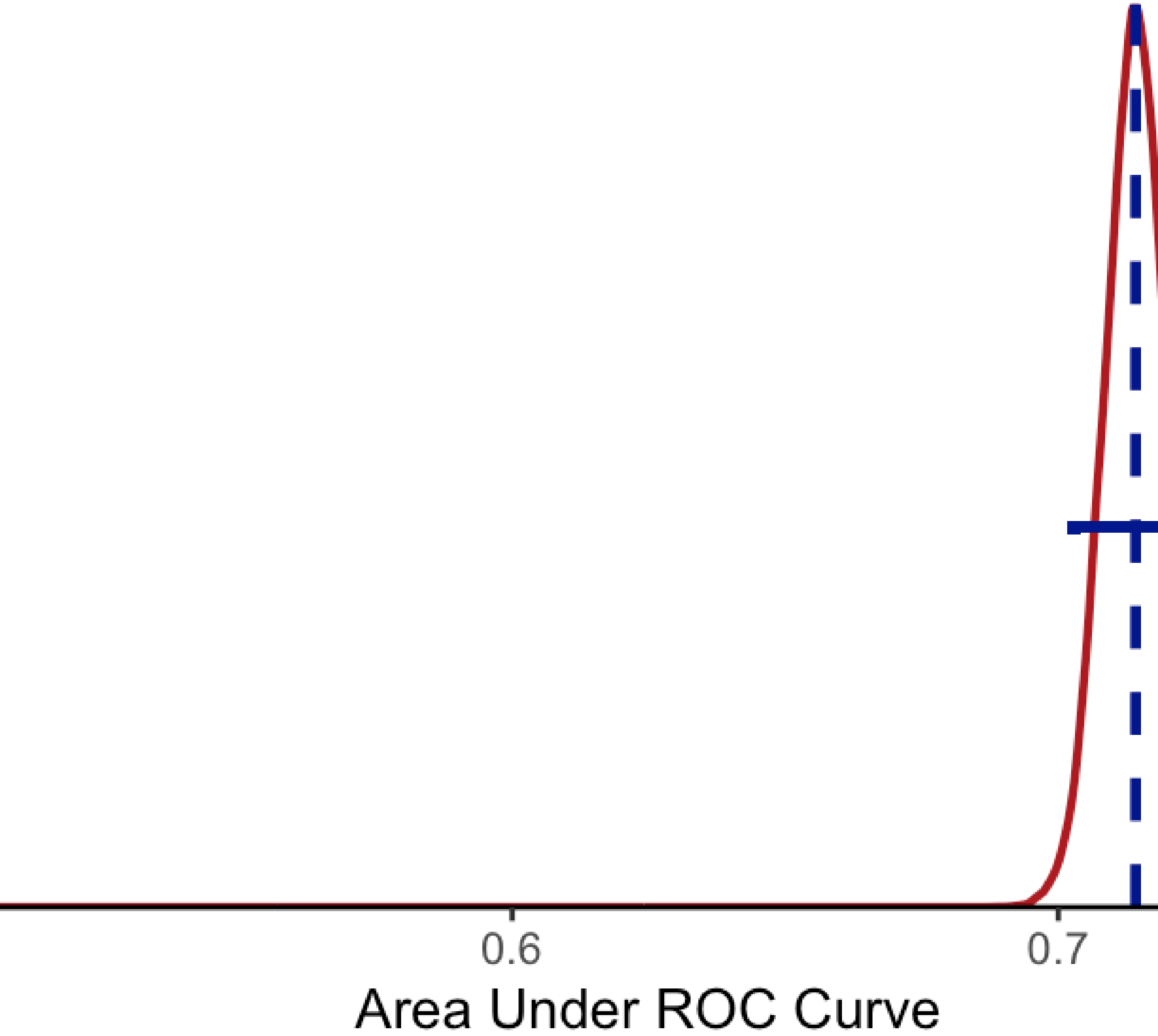


Figure 2: Posterior probability distribution of model performance with a 95% credible interval. The dashed line represents median auROC across the sampling distribution, while the dotted line represents chance performance (auROC = 0.50).

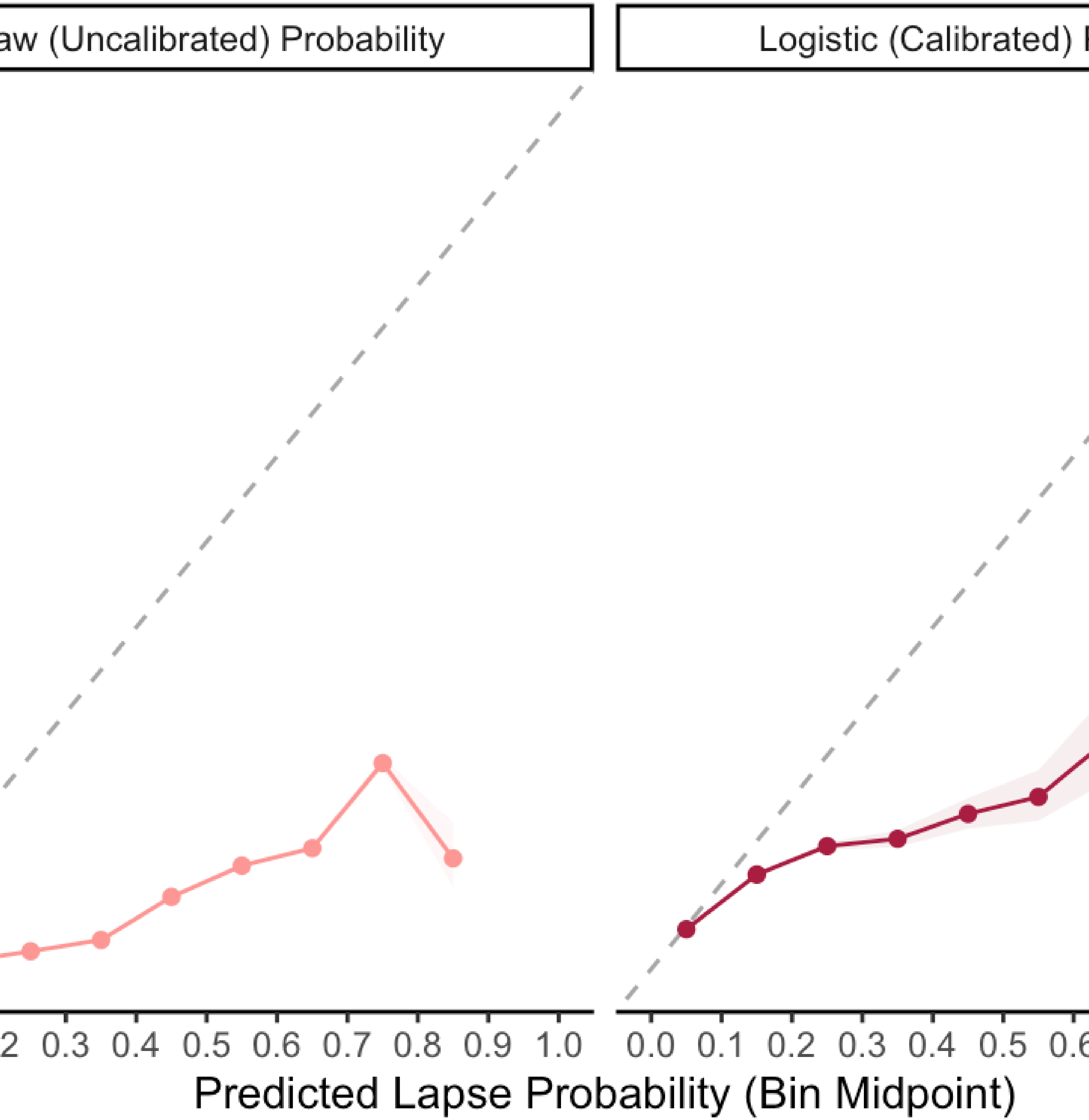


Figure 3: Comparison between raw (uncalibrated) and logistic (calibrated) probabilities.

Predicted lapse probability represents the predicted probabilities derived from the model, whereas

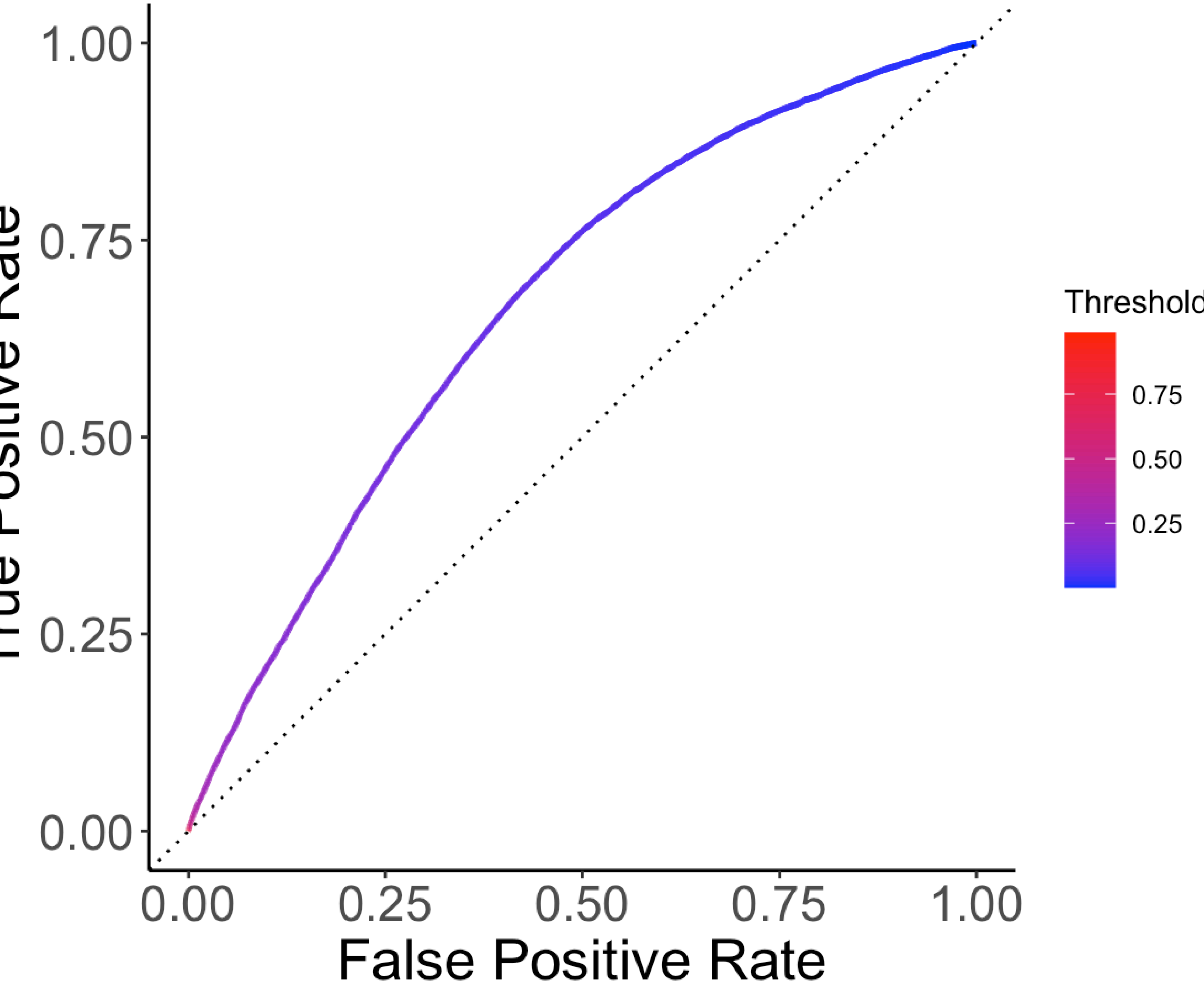


Figure 4: Area under the receiver operating characteristic (auROC) curve for overall validation set performance across all possible classification thresholds.

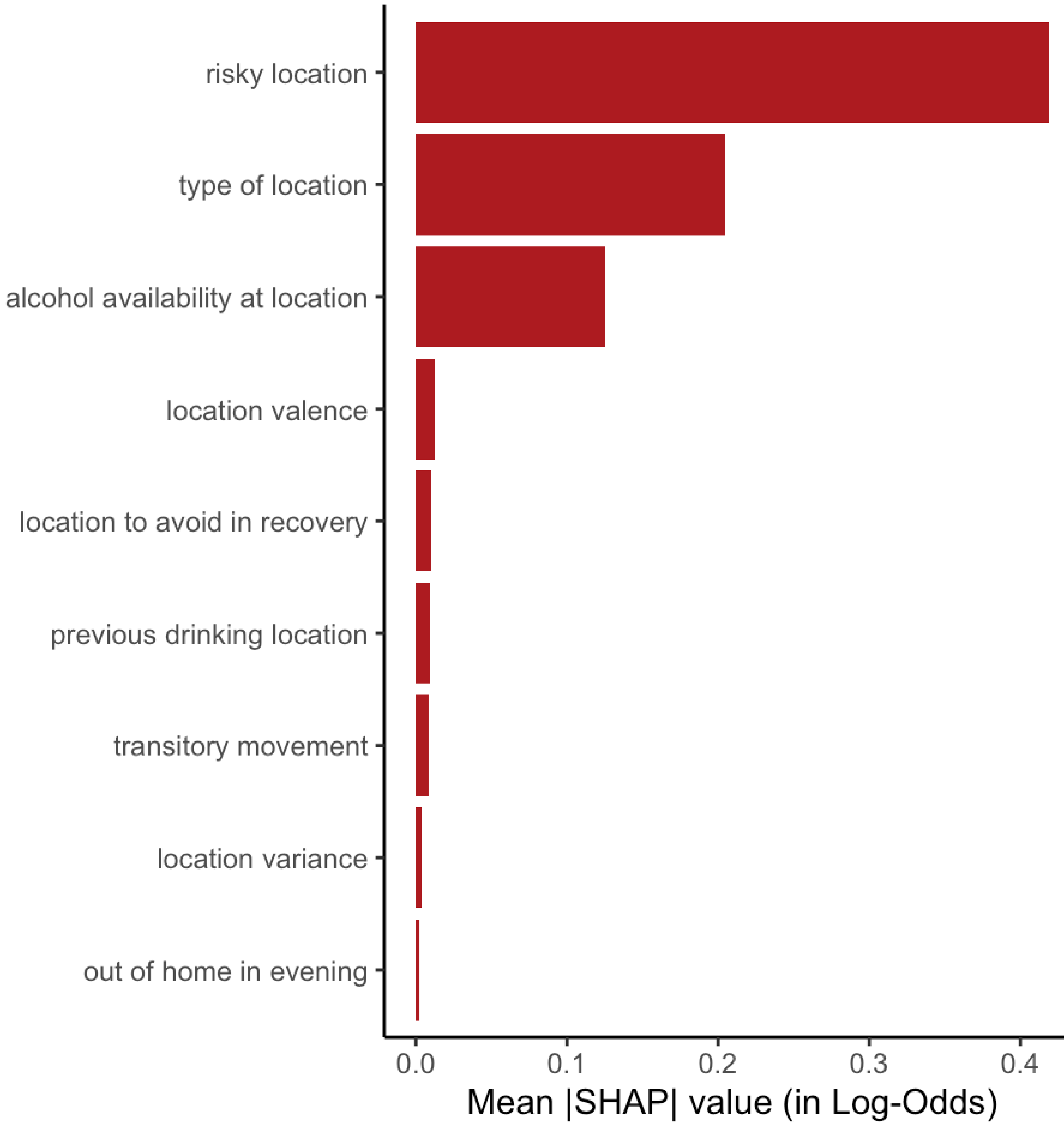


Figure 5: Grouped SHAP values displaying relative feature importance calculated using mean absolute values. Larger log-odds values indicate greater contribution to predictions in the model.

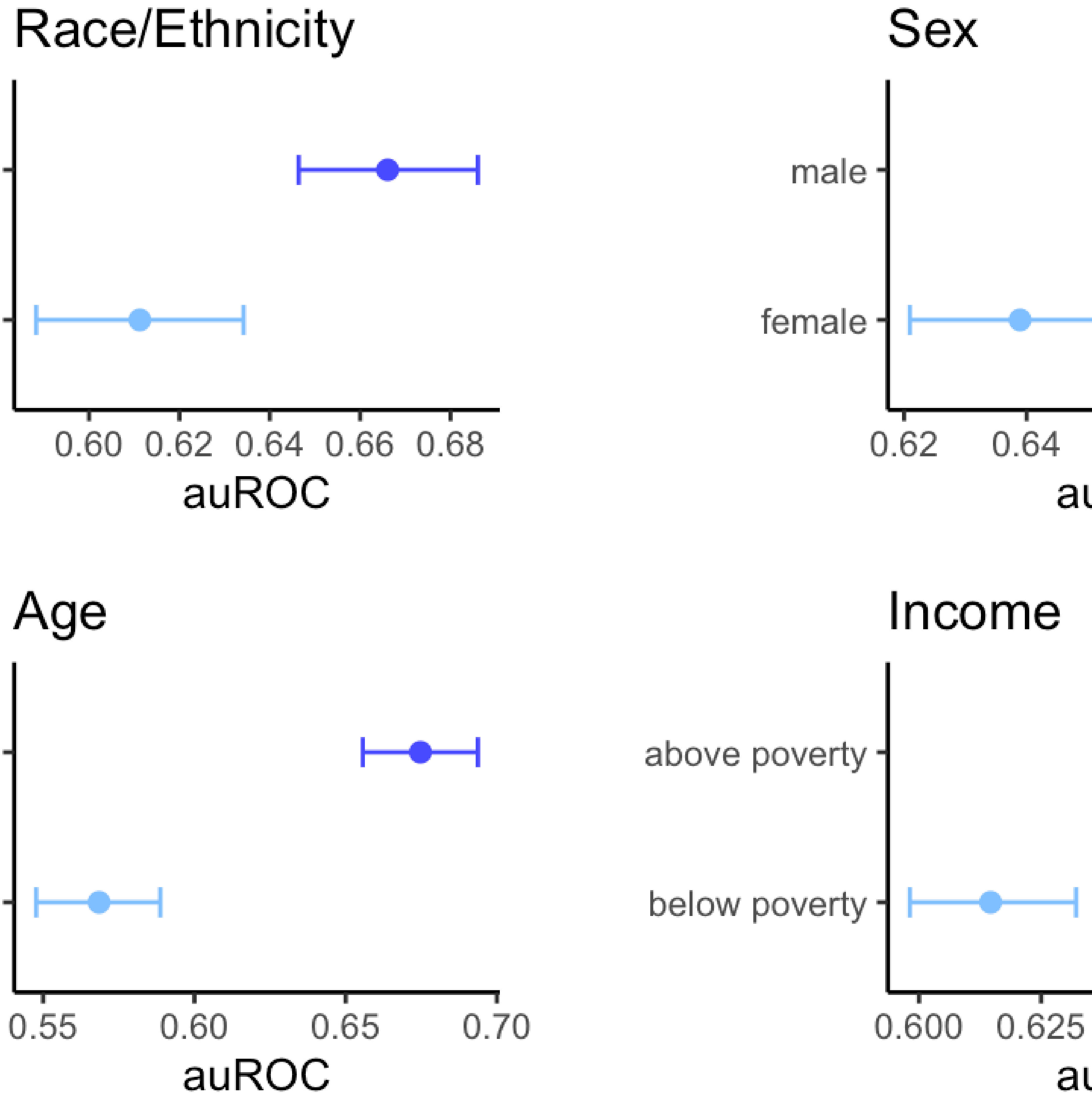


Figure 6: 95% credible intervals across posterior probability distributions by subgroup at differential levels of privilege.

Bibliography

[1] "Highlights for the 2022 National Survey on Drug Use and Health".

[2] J. A. Tucker, S. D. Chandler, and K. Witkiewitz, "Epidemiology of Recovery From Alcohol Use Disorder", *Alcohol Research : Current Reviews*, vol. 40, no. 3, p. 2, Nov. 2020, doi: 10.35946/arcr.v40.3.02.

[3] "The Science of Drug Use and Addiction: The Basics | NIDA Archives".

[4] T. H. Brandon, J. I. Vidrine, and E. B. Litvin, "Relapse and Relapse Prevention", *Annual Review of Clinical Psychology*, vol. 3, no. Volume3, 2007, pp. 257–284, Apr. 2007, doi: 10.1146/annurev.clinpsy.3.022806.091455.

[5] K. Witkiewitz and G. A. Marlatt, "Relapse Prevention for Alcohol and Drug Problems: That Was Zen, This Is Tao", *The American Psychologist*, vol. 59, no. 4, pp. 224–235, 2004, doi: 10.1037/0003-066X.59.4.224.

[6] D. C. Mohr, M. Zhang, and S. M. Schueller, "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning", *Annual Review of Clinical Psychology*, vol. 13, no. 1, pp. 23–47, May 2017, doi: 10.1146/annurev-clinpsy-032816-044949.

[7] P. A. Areàn, K. Hoa Ly, and G. Andersson, "Mobile Technology for Mental Health Assessment", *Dialogues in Clinical Neuroscience*, vol. 18, no. 2, pp. 163–169, Jun. 2016.

[8] P. H. Janak and N. Chaudhri, "The Potent Effect of Environmental Context on Relapse to Alcohol-Seeking After Extinction", *The Open Addiction Journal*, vol. 3, pp. 76–87, Jan. 2010, doi: 10.2174/1874941001003010076.

[9] M. A. Walton, F. C. Blow, C. R. Bingham, and S. T. Chermack, "Individual and Social/ Environmental Predictors of Alcohol and Drug Use 2 Years Following Substance Abuse

Treatment”, *Addictive Behaviors*, vol. 28, no. 4, pp. 627–642, Jun. 2003, doi: 10.1016/s0306-4603(01)00284-2.

[10] M. A. Walton, T. M. Reischl, and C. S. Ramanthan, “Social Settings and Addiction Relapse”, *Journal of Substance Abuse*, vol. 7, no. 2, pp. 223–233, 1995, doi: 10.1016/0899-3289(95)90006-3.

[11] M. R. LeCocq, P. A. Randall, J. Besheer, and N. Chaudhri, “Considering Drug-Associated Contexts in Substance Use Disorders and Treatment Development”, *Neurotherapeutics: The Journal of the American Society for Experimental NeuroTherapeutics*, vol. 17, no. 1, pp. 43–54, Jan. 2020, doi: 10.1007/s13311-019-00824-2.

[12] G. J. Stahler, J. Mennis, and D. A. Baron, “Geospatial Technology and the "Exposome": New Perspectives on Addiction”, *American Journal of Public Health*, vol. 103, no. 8, pp. 1354–1356, Aug. 2013, doi: 10.2105/AJPH.2013.301306.

[13] D. H. Epstein *et al.*, “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*, vol. 134, pp. 22–29, Jan. 2014, doi: 10.1016/j.drugalcdep.2013.09.007.

[14] M.-P. Kwan, J. Wang, M. Tyburski, D. H. Epstein, W. J. Kowalczyk, and K. L. Preston, “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*, vol. 33, no. 6, pp. 1176–1195, Jun. 2019, doi: 10.1080/13658816.2018.1503276.

[15] S. Attwood, H. Parke, J. Larsen, and K. L. Morton, “Using a Mobile Health Application to Reduce Alcohol Consumption: A Mixed-Methods Evaluation of the Drinkaware Track & Calculate Units Application”, *BMC public health*, vol. 17, no. 1, p. 394, May 2017, doi: 10.1186/s12889-017-4358-9.

[16] S. Carreiro, M. Taylor, S. Shrestha, M. Reinhardt, N. Gilbertson, and P. Indic, “Realize, Analyze, Engage (RAE): A Digital Tool to Support Recovery from Substance Use Disorder”, *Journal of Psychiatry and Brain Science*, vol. 6, p. e210002, 2021, doi: 10.20900/jpbs.20210002.

[17] V. M. Gonzalez and P. L. Dulin, “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*, vol. 83, no. 2, pp. 335–345, Apr. 2015, doi: 10.1037/a0038620.

[18] D. H. Gustafson *et al.*, “A Smartphone Application to Support Recovery From Alcoholism: A Randomized Clinical Trial”, *JAMA Psychiatry*, vol. 71, no. 5, p. 566, May 2014, doi: 10.1001/jamapsychiatry.2013.4642.

[19] F. Naughton *et al.*, “A Context-Sensing Mobile Phone App (Q Sense) for Smoking Cessation: A Mixed-Methods Study”, *JMIR mHealth and uHealth*, vol. 4, no. 3, p. e106, Sep. 2016, doi: 10.2196/mhealth.5787.

[20] A. Doryab *et al.*, “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*, vol. 7, no. 7, p. e13209, Jul. 2019, doi: 10.2196/13209.

[21] A. S. Heller *et al.*, “Association between Real-World Experiential Diversity and Positive Affect Relates to Hippocampal-Striatal Functional Connectivity”, *Nature Neuroscience*, vol. 23, no. 7, pp. 800–804, Jul. 2020, doi: 10.1038/s41593-020-0636-4.

[22] I. M. Raugh *et al.*, “Geolocation as a Digital Phenotyping Measure of Negative Symptoms and Functional Outcome”, *Schizophrenia Bulletin*, vol. 46, no. 6, pp. 1596–1607, Dec. 2020, doi: 10.1093/schbul/sbaa121.

[23] J. Shin and S. M. Bae, “A Systematic Review of Location Data for Depression Prediction”, *International Journal of Environmental Research and Public Health*, vol. 20, no. 11, p. 5984, May 2023, doi: 10.3390/ijerph20115984.

[24] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, “Ensuring Fairness in Machine Learning to Advance Health Equity”, *Annals of Internal Medicine*, vol. 169, no. 12, pp. 866–872, Dec. 2018, doi: 10.7326/M18-1990.

[25] J. Wawira Gichoya, L. G. McCoy, L. A. Celi, and M. Ghassemi, “Equity in Essence: A Call for Operationalising Fairness in Machine Learning for Healthcare”, *BMJ health & care informatics*, vol. 28, no. 1, p. e100289, Apr. 2021, doi: 10.1136/bmjhci-2020-100289.

[26] X. Wang, Y. Zhang, and R. Zhu, “A Brief Review on Algorithmic Fairness”, *Management System Engineering*, vol. 1, no. 1, p. 7, Nov. 2022, doi: 10.1007/s44176-022-00006-z.

[27] N. Japkowicz, “The Class Imbalance Problem: Significance and Strategies”, in *Proc. of the Int'l Conf. on Artificial Intelligence*, 2000, pp. 111–117.

[28] A. Wang, V. V. Ramaswamy, and O. Russakovsky, “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, in FAccT '22. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 336–349. doi: 10.1145/3531146.3533101.

[29] M. R. Schick, N. S. Spillane, and K. L. Hostetler, “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*, vol. 44, no. 10, pp. 1933–1951, 2020, doi: 10.1111/acer.14440.

[30] “Physical Activity Guidelines for Americans, 2nd Edition”.

[31] M. Kuhn and H. Wickham, “Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles”. 2020.

[32] Center for High Throughput Computing, “Center for High Throughput Computing”. Center for High Throughput Computing, 2006. doi: 10.21231/GNT1-HW21.

[33] K. Wyant, S. J. K. Sant'Ana, G. Fronk, and J. J. Curtin, “Machine Learning Models for Temporally Precise Lapse Prediction in Alcohol Use Disorder”, *Psychopathology and Clinical Science*, 2024, doi: 10.31234/osf.io/cgsf7.

[34] P. Jonathan, W. J. Krzanowski, and W. V. McCarthy, “On the Use of Cross-Validation to Assess Performance in Multivariate Prediction”, *Statistics and Computing*, vol. 10, no. 3, pp. 209–229, Jul. 2000, doi: 10.1023/A:1008987426876.

[35] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, 1st ed. 2013, Corr. 2nd printing 2018 edition. New York: Springer, 2018. doi: 10.1007/978-1-4614-6849-3.

[36] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions”, in *Proceedings of the 31st International Conference on Neural Information Processing*

Systems, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.

[37] T. B. MBA DVM, “2024 Federal Poverty Rates Published: Why That Matters for Your Student Loans”. Jan. 2024.

[38] J. N. Mandrekar, “Receiver Operating Characteristic Curve in Diagnostic Test Assessment”, *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, vol. 5, no. 9, pp. 1315–1316, Sep. 2010, doi: 10.1097/JTO.0b013e3181ec173d.

[39] C. F. Dormann, “Calibration of Probability Predictions from Machine-Learning and Statistical Models”, *Global Ecology and Biogeography*, vol. 29, no. 4, pp. 760–765, 2020, doi: 10.1111/geb.13070.

[40] B. Van Calster *et al.*, “Calibration: The Achilles Heel of Predictive Analytics”, *BMC Medicine*, vol. 17, no. 1, p. 230, Dec. 2019, doi: 10.1186/s12916-019-1466-7.

[41] M. E. Larimer, R. S. Palmer, and G. A. Marlatt, “Relapse Prevention: An Overview of Marlatt's Cognitive-Behavioral Model”, *Alcohol Research & Health*, vol. 23, no. 2, p. 151, 1999.

[42] G. A. Marlatt and J. R. Gordon, Eds., *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors*, First edition. New York: The Guilford Press, 1985.

[43] M. A. Kabir, M. U. Ahmed, S. Begum, S. Barua, and M. R. Islam, “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*, in ICMLT '24. New York, NY, USA: Association for Computing Machinery, Sep. 2024, pp. 21–29. doi: 10.1145/3674029.3674034.

[44] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, “Algorithmic Fairness: Choices, Assumptions, and Definitions”, *Annual Review of Statistics and Its Application*, vol. 8, no. Volume8, 2021, pp. 141–163, Mar. 2021, doi: 10.1146/annurev-statistics-042720-125902.

[45] B. Green, “Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness”, *Philosophy & Technology*, vol. 35, no. 4, p. 90, Oct. 2022, doi: 10.1007/s13347-022-00584-6.

[46] B. Green and L. Hu, “The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning”.

[47] R. Ochigame, “The Long History of Algorithmic Fairness”. Jan. 2020.