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 .

Specific Aims

Smart digital therapeutics ("smart DTx") refer to treatments delivered via technology such as smartphones that use machine learning algorithms to recommend engagement with particular in-app modules at the optimal time. In the context of alcohol use disorder (AUD), these algorithms could be used 1) to predict an oncoming lapse; and 2) to encourage use of specific tools to minimize risk of that lapse occurring based on factors specifically relevant to that individual (e.g., changes in mood, difficulty with close social connections, proximity to triggering locations). In other words, these algorithms have the potential to predict both *when* and *why* a lapse may occur. This is beneficial with respect to AUD where it may be difficult for an individual to know the exact precipitants to a lapse, even if the lapse itself is anticipated. These algorithms not only need to be developed outright, but also need to perform well (i.e., pinpoint lapses accurately) and identify meaningful, actionable lapse precursors. However, before accurate personalized recommendations can be offered, smart DTx must also demonstrate similar performance across sub-groups and provide a sustainable method of data collection for users (i.e., minimally burdensome).

A fair algorithm is one with no preference in performance with respect to inherent or acquired characteristics (e.g., gender, race, socioeconomic status; [1]). In the context of AUD, this would mean that lapse predictions are reasonably accurate and do not favor or disadvantage any particular group. The performance of an algorithm across sub-groups can be assessed using measures of *algorithmic fairness*, which quantify relative differences in predictions like accuracy.

Lapse prediction errors could result in both poorer outcomes and eroded trust of these tools. Assessing algorithmic fairness during the development of a smart DTx seeks to ensure that the same standard of care is provided to everyone utilizing it.

The inputs to these algorithms may also affect their fairness. For example, many DTx use self-report measures administered several times per day (ecological momentary assessment or "EMA"). However, EMA may be burdensome for individuals who do not have the flexibility of filling out repeated assessments every few hours. One solution is to utilize data collected via unobtrusive means, like geolocation data, which captures facets of activity patterns such as movement, frequency, duration, and time. Smartphones come pre-equipped with sensors to be able to capture this granular data continuously. Furthermore, geolocation requires little maintenance following initial set-up. Minimal additional contextual data can also be solicited from individuals about frequently visited locations to distill targets for intervention. A smart DTx could take these salient features and identify helpful in-app modules accordingly. Taken together, supplemented geolocation data can provide insight into how activity patterns and location semantics interface to increase (or decrease) the probability of lapse.

For my first-year project, I intend to use geolocation data collected from smartphones and corresponding self-reported contextual information for frequently visited locations to build a machine learning algorithm to predict next-day alcohol use lapse among participants with a diagnosis of AUD and a recovery goal of abstinence. I will also evaluate the algorithmic fairness of the best-performing model.

To accomplish these goals, I have identified the following specific aims:

Aim 1: Train several machine learning algorithms to predict alcohol lapse from geolocation data. Features will be derived from continuous geolocation data in combination with contextual information about frequently visited (>2 per one month period) locations. Several candidate classification algorithms will be trained and evaluated to predict next-day lapse.

Aim 2: Evaluate the best-performing algorithm. Area under the Receiver Operating Characteristic Curve (auROC) will be used as the primary performance metric to select and evaluate algorithm configurations. Secondary performance metrics (e.g., sensitivity, specificity) and interpretability metrics (i.e., SHAP values) will also be used to evaluate algorithms.

Aim 3: Evaluate the algorithmic fairness of the top performing algorithm across subgroups who experience known treatment disparities in the context of AUD. Algorithmic fairness will be examined across race/ethnicity, socioeconomic status, and sex. Fairness will be evaluated using metrics calculated across discrete subgroups which compare disadvantaged groups to the corresponding privileged group within said category (respectively: white, high socioeconomic status, and male individuals).

Significance

Lapse as intervention target

Chronic, recurrent patterns of relapse back to harmful alcohol use are, for many individuals diagnosed with AUD, a common occurrence during the recovery process [2], [3]. As a result, many researchers have identified relapse's potential utility as an intervention target [4]. However, the term *relapse* is poorly defined in the literature, with studies using divergent definitions (e.g.,

any use at all versus any use after remission) or not providing a definition at all [5]. Alternatively, lapses, or single instances of goal-inconsistent use that may lead to relapse [6] are clearly defined, easy to observe, and by their very definition always precede relapse. Moreover, lapses can be tied to clinically meaningful events, such as the abstinence violation effect [7], in which one occurrence of goal-inconsistent use can lead to relapse. In sum, lapses are a strong candidate for early intervention and are therefore utilized as the primary intervention target in this study.

Despite its clear definition, factors which precipitate lapse change both person-by-person and moment-by-moment, highlighting the importance of understanding not only when a lapse may happen, but also why, in order to facilitate optimally effective intervention. Because of its dynamic and continuous nature, it is imperative that measures which seek to quantify risk of lapse can provide the appropriate level of granularity.

The role of location in lapse

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.

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). Moreover, these different factors may play differential roles in lapse risk both within and across individuals.

A key benefit to using location data to predict lapse is its inherent ability to take in a wide range of information relating to both physical place, as well as mood states and behavioral patterns associated with a given location. Because precipitants to lapse are heterogeneous, utilizing such a far-reaching data source may enable our model to better encapsulate these diverse factors. Not only is there potential for this to then improve our predictions, but it may also help us better elucidate what *specific* features *explain* those predictions. Therefore, there may be benefits both to our prediction as well as explanatory goals in this project.

Geolocation data for lapse prediction

One way that location information can be assessed is using geolocation data. Also referred to as global positioning system (GPS) data, geolocation data quantify spatial positioning and movement across time. Many forms of technology that are now ubiquitous in daily life, such as smartphones and smartwatches, automatically and continually collect these data. This fact, paired with increasing rates of smartphone ownership, suggest that there is high potential for these data to be feasibly harnessed for mental health applications [12]. In the context of substance use, geolocation data has been specifically identified as being of particular use in both understanding the precipitants to harmful use and its effective treatment [13].

While previous mental health research utilizing geolocation data has shown promise, the features used and outcomes explored have largely been siloed across psychopathology subfields. For example, within the substance use literature, geolocation data have historically been used to examine risky locations, such as the influence of neighborhood characteristics on use [14], [15] and individual physical proximity to locations of potential or past harmful use such as bars (either estimated using geofencing or user-defined) [16]–[20]. 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 and behavior. Geolocation data have been used to estimate loneliness and isolation [21], to demonstrate increases in positive affect from seeking out novel environments [22], and to quantify depressive symptoms [23]. Moreover, these data have not only been harnessed to measure mood symptoms, but to also predict their emergence [24].

Latent characteristics of location – such as isolation, mobility, novelty seeking, and emotional valence – are notably lacking in the substance use geolocation literature, but have historically been captured in other ways (for example, using ecological momentary assessment, *EMA*, data). In substance use research, EMA surveys often ask information about mood, social experiences, and daily activities. However, geolocation data provides greater temporal sensitivity and specificity compared to its EMA counterpart, which is collected at static intervals several times per day compared to the continuous sampling of geolocation data. While there is no doubt in the predictive utility of EMA measures in the context of lapse [25], the affective science

literature suggests that it may be possible to quantify some of these features utilizing location data. In the context of a smart DTx, this would mean reduced patient burden (i.e., fewer surveys to fill out) and a potentially more equitable system (i.e., easier for people who can't fill out multiple surveys a day to engage with and, therefore, benefit from).

My first-year project will address a crucial gap in the literature between clinical science subdisciplines by applying feature engineering techniques from both the substance use and affective science literatures, therefore providing a novel, dynamic lens to the use of geolocation data in predicting lapse. Moreover, it will utilize a lower burden, finer-grained temporal data source compared to other techniques like EMA. Taking the limitations of previous work into account, my first-year project will develop a sensing system that can predict lapse risk day-to-day utilizing a broad range of low-burden features which capture facets of both movement and mood.

Algorithmic fairness

While a great strength of machine learning algorithms is their ability to elucidate latent statistical patterns of data, their greatest weakness lies in the necessary reliance on these data. 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 in perpetuity by machine learning models if not critically examined. Without examining algorithmic fairness prior to deployment in the real-world, smart DTx run the risk of providing sub-optimal mental health care to individuals who already face disadvantages.

Though the nature of data collection in this study avoids certain inequities (e.g., by virtue of not using data collected in a health care setting), there are particular concerns specific to this work which motivate interrogating the fairness of models developed during this project. Firstly, while the data set for this study has a range of age and sex representation (21-72 years of age, 49% women), it is primarily white (86.8%) and non-Hispanic (97.4%). Having a limited number of observations within underrepresented groups means that our models will not have as wide a range of individuals to learn from for making predictions of lapse as compared to white, non-Hispanic participants. Performance of these models for racialized minority individuals may therefore be less accurate as a result, particularly without the use of resampling techniques to amend these imbalances [26], [27].

Second, the AUD 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 [28]. 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 we decide are important to measure and input into our models in this study, informed by our knowledge of AUD theory, will inherently be biased and may favor men. Therefore, despite the fact that our models will have equivalent representation across sex, we should not anticipate that

this is enough to compensate for biases brought on by the broader societal context. Indeed, this fact should also be kept in mind when exploring model performance across other "well-balanced" categories such as age in our models. Both of these facts motivate the reasoning behind examining algorithmic fairness in this current project.

Conversations around machine learning and fairness in healthcare have peppered the literature for several years [29], [30], yet no concrete standards of reporting of algorithmic fairness have been delineated. This is of critical importance in considering that the ultimate goal of many of these algorithms is to be implemented in patient contexts. One potential legal marker of algorithmic fairness may be derived from the Equal Employment Opportunity Commission (EEOC) 80% rule, which states that hiring rates of individuals from protected groups (e.g., people of color, women) should be at least 80% that of white men. Applied in the context of algorithmic fairness, this would be something like model performance being at least 80% of the best-performing group. While this idea has begun to take shape at the intersection of law and machine learning [31], [32], it should be noted that this threshold is not a very rigorous standard, particularly in the context of health data. We should strive to do better.

Despite the fact that standards of reporting do not exist, there are many established metrics for statistically assessing algorithmic fairness. Common metrics of algorithmic fairness are predicated on the assignment of a "privileged" and "non-privileged" group. Once this has been established, measures such as accuracy can be examined between those two groups [33]. This can be repeated over many combinations of privileged and non-privileged groups to assess model performance across various identities.

At broad level, algorithmic fairness is paramount to consider in context of AUD because risk, access to and engagement with treatment, and negative consequences from drinking vary across socioeconomic status, race, and sex in AUD [34]–[36]. Location plays an important role in contributing to mental health inequities in general [37], and in the context of substance use, exposure to substances, neighborhood disadvantage, and barriers to treatment stratified by location all contribute to the initiation and maintenance of substance use disorders [38]. Furthermore, these inequities are not shared equally across the population: for example, alcohol outlet density is greater in predominantly Black census tracts [39]. Counterintuitively, this may lead algorithms to perform *better* in these contexts – for example, if alcohol outlet density is closely tied to lapse prediction, then an algorithm may perform better for Black individuals.

My first-year project will address several gaps in the mental health and algorithmic fairness literature. Firstly, it will advocate for the use of fairness metrics alongside other measures of algorithm performance. Secondly, it will apply fairness measures to geolocation data, which is known to encode disparities in the context of substance use, in an effort to critically examine my best-performing model.

Current study

My proposed first-year project will apply machine learning to predict lapse in AUD utilizing geolocation data with the aim of advancing both prediction and explanatory efforts in AUD. First, I will build (**Aim 1**) and evaluate (**Aim 2**) a model to predict oncoming lapse with an almost entirely passive data stream for individuals with a recovery goal of abstinence. **Aim 2** will also

include an examination of what geolocation features are most predictive of an oncoming lapse, thereby providing explanatory insight into which of these features may be most theoretically tied to lapse in AUD. Next (**Aim 3**), I will critically examine the fairness of the best performing model.

Approach

Overview

For my first-year project, I will conduct analyses using a subset of data collected between 2017 and 2019 under a larger grant funded though the National Institute of Alcohol Abuse and Alcoholism (R01 AA024391). As such, the following approach section represents aspects relevant only to my first-year project, though other data were collected under this grant.

Participants

One hundred and fifty one 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. 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).

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 1: Eligibility criteria for study enrollment. **Personal or study-provided.

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. 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).

Measures

Geolocation 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.

Contextual information

Contextual information for frequently visited locations (>2 times in the previous month) was obtained during an interview at each follow-up visit (at month 1, 2, and 3; Table 2).

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 2: Location information collected from frequently visited locations.

Participant characteristics

Participants completed a baseline measure of demographics and other constructs relevant to lapse at the screening visit, which will be used both for model building and for fairness assessments (Table 3).

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

Definition of subgroups for fairness assessments

Subgroups will be defined on the basis of personal individual characteristics (in the machine learning fairness literature, “sensitive attributes”) that are specifically associated with treatment disparities in AUD. Characteristics will be assessed using this method to ensure the development of a model that does not further exacerbate existing treatment disparities.

Lapses

Alcohol lapses will be used as our outcome variable in this study and will be used to provide labels for model training, for testing model performance, and for testing issues of algorithmic fairness across our predefined subgroups. Future lapse occurrence (here conceptualized as next-day lapse) will be predicted in 24-hour rolling windows, beginning at midnight on a participant’s second day of participation to ensure one full day of data collection for the first window, and at every subsequent hour (i.e., future lapse in the next 24-hours predicted each hour). This rolling

window approach will take advantage of the granular, continuous nature of geolocation data, enabling hourly updating of predictions. *Lapse* and *no lapse* occurrences will be 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 will be labeled *lapse*. Prediction windows will be labeled *no lapse* if no alcohol use was reported within that window. Windows will be excluded if no alcohol use occurred within the window but *did* occur within six hours of the start or end of the window for label fidelity.

Feature engineering

Feature engineering is the process of creating variables (or *"features"*) from unprocessed data and will be used in this project to transform raw data from three sources: 1) demographic characteristics relevant to treatment disparities in AUD collected at intake; 2) day at prediction window onset; and 3) geolocation data collected the prior day. Features will be created using all GPS features combined with demographic and day/time features.

We will generate a quantitative feature for age and dummy-coded features for race/ethnicity, marital status, education, and sex. Dummy-coded features indicating time of day (5pm - midnight versus any other time) and day of the week will also be generated. As a comparison to the geolocation model, a baseline model will also be developed utilizing only the time-of-day dummy-coded features to predict lapse. Features from geolocation data will be generated that

both utilize contextual information collected from monthly surveys (i.e., location valence and perceived riskiness) as well as features that are independent of further individual input.

Features will be derived using examples from previous literature and will also be generated using contextual features. For example, geo-fencing techniques to identify bars and liquor stores [18], [19] have been used in the AUD literature. More broadly, features such as isolation [21], [23] and seeking of novelty locations [22] have been used in the affective science literature as mood proxies. In addition to the wide range of possible features from the literature, this data set affords the ability to create unique features using person-specific contextual information in tandem with geolocation data, such as places where an individual has drank in the past and the emotional valence associated with that location. We will also calculate raw and change features based on previous geolocation data in order to capture individual variation.

Imputation of missing data and removal of zero-variance features are additional general processing steps that will also be undertaken during feature engineering.

Algorithm development & performance

The first aim of my project is to develop a machine learning algorithm which can predict lapse from contextualized geolocation data. To accomplish this, fseveral classification candidate machine learning algorithms will first be considered and will be trained and assessed using participant-grouped, nested *k*-fold cross-validation. Grouped cross-validation assigns all data from a participant as either held-in or held-out to avoid bias introduced when predicting a participant's data from their own data. Nested cross-validation uses two nested loops for dividing

and holding out folds: an outer loop, where held-out folds serve as test sets for model evaluation; and inner loops, where held-out folds serve as validation sets for model selection. Importantly, these sets are independent, maintaining separation between data used to train the models, select the best models, and evaluate those best models. Therefore, nested cross-validation removes optimization bias from the evaluation of model performance in the test sets and can yield lower variance performance estimates than single test set approaches [40].

The primary performance metric for model selection and evaluation will be area under the Receiver Operating Characteristic Curve (auROC) [41]. 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 will be selected using median auROC across all validation sets. Several secondary performance metrics including sensitivity, specificity, balanced accuracy, positive predictive value (PPV), and negative predictive value (NPV) will also be assessed.

SHAP (SHapley Additive exPlanations) values will be computed as our interpretability metric to identify the relative importance of different features in each final algorithm. SHAP values measure the unique contribution of features in an algorithm's predictions [42]. SHAP

values possess several useful properties including: *Additivity* (SHAP values for each feature can be computed independently and summed); *Efficiency* (the sum of SHAP values across features must add up to the difference between predicted and observed outcomes for each observation); *Symmetry* (SHAP values for two features should be equal if the two features contribute equally to all possible coalitions); and *Dummy* (a feature that does not change the predicted value in any coalition will have a SHAP value of 0). 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 smart DTx. However, these will be descriptive analyses because standard errors or other indices of uncertainty for importance scores are not available for SHAP values.

Finally, a Bayesian hierarchical generalized linear model will be used to estimate the posterior probability distributions and 95% Bayesian credible intervals (CIs) for auROC between the baseline (time-of-day) and geolocation models.

Algorithmic fairness

The second aim of my project will be to assess the algorithmic fairness of the best-performing model. Several fairness metrics will be assessed across demographic characteristics implicated in treatment disparities, such as the true positive rate (TPR), true negative rate (TNR), and accuracy. These metrics can be calculated using a number of open source R packages [33]. Differences in classification are calculated by first splitting the sample by group, where one is the privileged group and the other is the unprivileged group, and then examining assigned outcomes (in this case, *lapse* or *no lapse*). A Bayesian hierarchical generalized linear model will be used to estimate

the posterior probability distributions and 95% Bayesian CIs for auROC across different subgroups.

Contribution to Theory

If successful, this study will contribute both to our ability to both predict and explain lapse in AUD as well as how we approach designing equitable machine learning models. First, it will identify the most predictive features which contribute to lapse from a minimally burdensome and continuously collected data source, geolocation data. These features will be created leveraging modeling techniques from both the substance use and affective science subdisciplines, resulting in broadly relevant transdiagnostic constructs. Furthermore, this study will also provide a critical evaluation of model performance beyond AUC by including examination of algorithmic fairness. Though many fairness metrics have been used in simulated data or in publicly available data sets (e.g., *COMPAS* recidivism data set; [43]), few researchers in applied mental health settings have made use of these tools to critically examine their own models [44]. Thus, this study will also advocate for transparency in reporting fairness metrics alongside standard measures of model performance.

References

Bibliography

[1] 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.

[2] J. R. McKay and S. Hiller-Sturmhofel, “Treating Alcoholism as a Chronic Disease: Approaches to Long-Term Continuing Care”, *Alcohol Research & Health: The Journal of the National Institute on Alcohol Abuse and Alcoholism*, vol. 33, no. 4, pp. 356–370, 2011.

[3] R. H. Moos and B. S. Moos, “Rates and Predictors of Relapse after Natural and Treated Remission from Alcohol Use Disorders”, *Addiction (Abingdon, England)*, vol. 101, no. 2, pp. 212–222, Feb. 2006, doi: 10.1111/j.1360-0443.2006.01310.x.

[4] M. A. Stillman and J. Sutcliff, “Predictors of Relapse in Alcohol Use Disorder: Identifying Individuals Most Vulnerable to Relapse”, *Addiction and Substance Abuse*, vol. 1, no. 1, pp. 3–8, 2020, Accessed: Apr. 14, 2024. [Online]. Available: https://probiologists.com/Uploads/Articles/29_637447991624587373.pdf

[5] W. Sliedrecht, H. Roozen, p. u. family=Waart given=Ranne, G. Dom, and K. Witkiewitz, “Variety in Alcohol Use Disorder Relapse Definitions: Should the Term “Relapse” Be Abandoned?”, *Journal of Studies on Alcohol and Drugs*, vol. 83, no. 2, pp. 248–259, Mar. 2022, doi: 10.15288/jsad.2022.83.248.

[6] 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.

[7] 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. [Online]. Available: https://psycnet.apa.org/record/2005-08721-000

[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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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”. Accessed: Mar. 26, 2024. [Online]. Available: https://osf.io/cgsf7

[26] N. Japkowicz, “The Class Imbalance Problem: Significance and Strategies”, in *Proc. of the Int’l Conf. on Artificial Intelligence*, 2000, pp. 111–117. Accessed: Apr. 14, 2024. [Online]. Available: https://site.uottawa.ca/~nat/Papers/ic-ai-2000.ps

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] S. Barocas and A. D. Selbst, “Big Data's Disparate Impact”, *SSRN Electronic Journal*, 2016, doi: 10.2139/ssrn.2477899.

[32] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, “Fairness Constraints: Mechanisms for Fair Classification”, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR, Apr. 2017, pp. 962–970. Accessed: Apr. 13, 2024. [Online]. Available: https://proceedings.mlr.press/v54/zafar17a.html

[33] J. Wiśniewski and P. Biecek, “Fairmodels: A Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models”, *The R Journal*, vol. 14, no. 1, pp. 227–243, Jun. 2022, doi: 10.32614/RJ-2022-019.

[34] S. Calling, H. Ohlsson, J. Sundquist, K. Sundquist, and K. S. Kendler, “Socioeconomic Status and Alcohol Use Disorders across the Lifespan: A Co-Relative Control Study”, *PLOS ONE*, vol. 14, no. 10, p. e224127, Oct. 2019, doi: 10.1371/journal.pone.0224127.

[35] P. A. C. Vaeth, M. Wang-Schweig, and R. Caetano, “Drinking, Alcohol Use Disorder, and Treatment Access and Utilization Among U.S. Racial/Ethnic Groups”, *Alcoholism: Clinical and Experimental Research*, vol. 41, no. 1, pp. 6–19, Jan. 2017, doi: 10.1111/acer.13285.

[36] J. Witbrodt, N. Mulia, S. E. Zemore, and W. C. Kerr, “Racial/Ethnic Disparities in Alcohol-Related Problems: Differences by Gender and Level of Heavy Drinking”, *Alcoholism: Clinical and Experimental Research*, vol. 38, no. 6, pp. 1662–1670, Jun. 2014, doi: 10.1111/acer.12398.

[37] J. Vallée, E. Cadot, C. Roustit, I. Parizot, and P. Chauvin, “The Role of Daily Mobility in Mental Health Inequalities: The Interactive Influence of Activity Space and Neighbourhood of Residence on Depression”, *Social Science & Medicine (1982)*, vol. 73, no. 8, pp. 1133–1144, Oct. 2011, doi: 10.1016/j.socscimed.2011.08.009.

[38] J. Mennis, G. J. Stahler, and M. J. Mason, “Risky Substance Use Environments and Addiction: A New Frontier for Environmental Justice Research”, *International Journal of Environmental Research and Public Health*, vol. 13, no. 6, p. 607, Jun. 2016, doi: 10.3390/ijerph13060607.

[39] J. Scott *et al.*, “Structural Racism in the Built Environment: Segregation and the Overconcentration of Alcohol Outlets”, *Health & Place*, vol. 64, p. 102385, Jul. 2020, doi: 10.1016/j.healthplace.2020.102385.

[40] 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.

[41] 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.

[42] 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.

[43] J. Dressel and H. Farid, “The Accuracy, Fairness, and Limits of Predicting Recidivism”, *Science Advances*, vol. 4, no. 1, p. eaao5580, Jan. 2018, doi: 10.1126/sciadv.aao5580.

[44] A. C. Timmons *et al.*, “A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health”, *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, vol. 18, no. 5, pp. 1062–1096, Sep. 2023, doi: 10.1177/17456916221134490.