Addictive Behaviors 83 (2018) 42–47

Contents lists available at ScienceDirect

Addictive Behaviors

journal homepage: www.elsevier.com/locate/addictbeh

Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions

Sangwon Bae[a], Tammy Chung[b], Denzil Ferreira[c], Anind K. Dey[a], Brian Suffoletto[d,*]

[a] *Human Computer Interaction Institute, Carnegie Mellon University, United States*
[b] *Department of Psychiatry, University of Pittsburgh, United States*
[c] *Center for Ubiquitous Computing, University of Oulu, Finland*
[d] *Department of Emergency Medicine, University of Pittsburgh, United States*

HIGHLIGHTS

- Mobile phone sensor data is useful in building accurate models to detect periods of drinking.
- Useful sensor features relate to activity/movement, phone use/calls, and keystrokes.
- Interventions could use phone sensor features to trigger remote support when it is most needed.

ARTICLE INFO

Keywords:
Alcohol
Smartphone sensors
Machine learning
AWARE app

ABSTRACT

Background: Real-time detection of drinking could improve timely delivery of interventions aimed at reducing alcohol consumption and alcohol-related injury, but existing detection methods are burdensome or impractical.
Objective: To evaluate whether phone sensor data and machine learning models are useful to detect alcohol use events, and to discuss implications of these results for just-in-time mobile interventions.
Methods: 38 non-treatment seeking young adult heavy drinkers downloaded AWARE app (which continuously collected mobile phone sensor data), and reported alcohol consumption (number of drinks, start/end time of prior day's drinking) for 28 days. We tested various machine learning models using the 20 most informative sensor features to classify time periods as non-drinking, low-risk (1 to 3/4 drinks per occasion for women/men), and high-risk drinking (> 4/5 drinks per occasion for women/men).
Results: Among 30 participants in the analyses, 207 non-drinking, 41 low-risk, and 45 high-risk drinking episodes were reported. A Random Forest model using 30-min windows with 1 day of historical data performed best for detecting high-risk drinking, correctly classifying high-risk drinking windows 90.9% of the time. The most informative sensor features were related to time (i.e., day of week, time of day), movement (e.g., change in activities), device usage (e.g., screen duration), and communication (e.g., call duration, typing speed).
Conclusions: Preliminary evidence suggests that sensor data captured from mobile phones of young adults is useful in building accurate models to detect periods of high-risk drinking. Interventions using mobile phone sensor features could trigger delivery of a range of interventions to potentially improve effectiveness.

1. Introduction

Binge drinking, defined as consuming > 4/5 drinks (women/men) per occasion, is a serious but preventable public health problem, with young adults disproportionately affected (Center for Behavioral Health Statistics and Quality, 2016). Digital interventions are a promising strategy to reduce excessive alcohol consumption, with most evidence for effectiveness in young adults (Carey, Scott-Sheldon, Elliott, Garey, & Carey, 2012; Fowler, Holt, & Joshi, 2016). Still, effects of digital interventions are typically small (Berman, Gajecki, Sinadinovic, & Andersson, 2016; Suffoletto et al., 2015), suggesting that designs are not optimized.

To improve longitudinal engagement and effects of digital interventions, the right support material should be delivered to the right

[*] Corresponding author at: Iroquois Building, Suite 400A, 3600 Forbes Avenue, Pittsburgh, PA 15261, United States.
E-mail address: suffbp@upmc.edu (B. Suffoletto).

https://doi.org/10.1016/j.addbeh.2017.11.039
Received 11 July 2017; Received in revised form 22 November 2017; Accepted 25 November 2017
Available online 27 November 2017

person at the right time (Nahum-Shani et al., 2016). Therefore, a digital intervention aimed at reducing binge drinking should deliver support “in the moment”, that is, in the context of a drinking episode to enhance motivation for setting and keeping drinking limits, and to reduce the likelihood of negative alcohol-related consequences (i.e., reinforce explicit intentions). To accomplish these goals, it is critical that a digital intervention be able to detect when the person is drinking.

Recent developments in sensor miniaturization provide the ability to collect multi-modal data continuously from mobile phones with minimal participant burden. Continuous smartphone sensing can capture time-stamped data elements that can be used to track a person's daily routine in line with a computer science-based “context aware” theoretical framework (Abowd et al., 1999). Phone sensor data has been shown to be useful in inferring other states such as mood (Mohr, Zhang, & Schueller, 2017). Still, it remains unknown whether phone sensors could be useful in detecting periods of drinking.

In previous work (Bae et al., 2017), we described the computer engineering methods involved in using phone sensors for detection of drinking periods. In this study, we expand upon this work by describing how sensor features differ between periods of high-risk (e.g., binge) drinking, low-risk drinking, and non-drinking. We hypothesized that phone sensor features related to time (Del Boca, Darkes, Greenbaum, & Goldman, 2004), movement patterns (Freisthler, Lipperman-Kreda, Bersamin, & Gruenewald, 2014; Gruenewald, Remer, & LaScala, 2014), communication (Cavazos-Rehg, Krauss, Sowles, & Bierut, 2015; Moewaka Barnes et al., 2016), and psychomotor impairment (Scholey, Benson, Neale, Owen, & Tiplady, 2012; Suffoletto, Gharani, Chung, & Karimi, 2017; Suffoletto, Goyal, Puyana, & Chung, 2017) would contribute to detection models. We also examined the time it takes for machine learning models to reach stability in accuracy, and differences in model performance on weekends versus weekdays. We discuss implications of our findings for delivery of just-in-time mobile interventions.

2. Methods

This prospective study recruited a convenience sample of young adults with hazardous drinking to provide phone sensor and self-reported measures of alcohol consumption for 28 consecutive days. All participants provided informed consent and were offered resources for alcohol treatment. This study was approved by the Institutional Review Boards at the University of Pittsburgh and Carnegie Mellon University.

2.1. Participants

Recruitment occurred through an Emergency Department (ED) and college campus, using similar methods. From the ED, 51 medically stable patients who were not seeking treatment for substance use, not intoxicated, and who were going to be discharged to home were screened for eligibility. At the college campus, 17 students who responded to study flyers or a Craigslist posting were screened for eligibility. At both sites, individuals who were between the ages of 21–28 years of age, reported recent hazardous alcohol consumption based on Alcohol Use Disorder Identification Test for Consumption (AUDIT-C) score of ≥ 3 for women or ≥ 4 for men (Bradley et al., 2007) and at least one high risk drinking occasion (> 4/5 drinks for women/men) on any day in the prior month were eligible for participation. We excluded those who did not own an iOS or Android phone. A total of 38 participants (21 ED patients, 17 students; see Table 1) met enrollment criteria and completed informed consent.

2.2. Procedures

Enrolled ED patients completed a brief questionnaire and downloaded the AWARE app (Ferreira, Kostakos, & Dey, 2015) to their phone in the ED. Enrolled college students presented to an on-campus office to complete the same questionnaire and download the AWARE app. All participants were instructed to keep the AWARE app open on their phone and to refrain from any non-drinking substance use (excluding cigarette use) during the study period. During enrollment, participants were provided with the definition of a standard drink (e.g., 12 oz. can of beer or 5 oz. glass of wine or 1.5 oz. 80-proof liquor) as well as an illustration of a typical standard drink for common beverage types: beer, wine, liquor. From the day after enrollment through 28 days, participants were sent a text-message (EMA) at 10 am: “Did you drink alcohol yesterday?” If they reported drinking, they received the following text queries: “Approximately what time did you start drinking?”, “Approximately what time did you stop drinking?”, and “How many standard drinks did you have during this period?” If there were multiple drinking episodes in a day, participants were instructed to report the episode when the largest number of drinks was consumed. All other potential drinking periods that day were coded as non-drinking. Participants received $20 for completing the baseline survey and $2 for each day they completed EMA.

Table 1
Sample characteristics.

Characteristics		ED patients (*n* = 21)	College students (*n* = 17)
Age, mean (SD)		23.1 (1.7)	23.9 (1.9)
Female sex, n (%)		7 (33.3)	8 (47.1)
Race	White	8 (38.1)	4 (23.5)
	Black	11 (52.3)	1 (5.9)
	Asian	1 (4.8)	12 (70.6)
	Other	1 (4.8)	0
Highest education	< High school	2 (9.5)	0
	High school grad.	5 (23.8)	1 (5.9)
	Some college	11 (52.4)	3 (17.7)
	College grad.	3 (14.3)	13 (76.4)
AUDIT-C score		6.0 (2.2)	6.2 (3.4)
Other drug use, last month	Daily or almost daily tobacco	2 (9.5)	2 (11.8)
	Any cannabis	12 (57.2)	6 (35.2)

2.3. AWARE app

When downloaded, AWARE app (Ferreira et al., 2015), for iOS and Android, places an icon on the phone screen which, when opened, automatically begins recording sensor data without requiring further participant interaction. When AWARE is opened for the first time, a unique IDwas randomly generated for research purposes. AWARE temporarily stored the sensor data on a participant's device and then synchronized this information to a university server over a secure connection via Wi-Fi every 30 min, when available. We configured AWARE to collect 56 sensor features related to time (e.g., day of week, time of day), movement patterns (e.g., accelerometry, rotation), communication (e.g., phone calls, texts), and psychomotor impairment (e.g., keystroke speed; available for Android phones only).

2.4. Measures

2.4.1. Baseline questionnaire

Demographics. Participants reported age, sex, race, ethnicity, and education.

Drug use. NM-ASSIST (Humeniuk et al., 2008) assessed frequency of past month drug use (e.g., tobacco, cannabis, opiates).

Alcohol Consumption. AUDIT-C (Bradley et al., 2007), includes 3 items on drinking quantity and frequency in the past 3 months. AUDIT-C score > 4 for men, and > 3 for women is considered positive (Rubinsky, Dawson, Williams, Kivlahan, & Bradley, 2013).

2.4.2. Event-level alcohol use

We used daily text message reports of alcohol use to label time periods (i.e. windows) as non-drinking, low-risk (1 to 3/4 drinks per episode for women/men), and high-risk (> 4/5 drinks per episode for women/men).For example, if a female reported consuming 5 drinks on the prior day, starting at 5 pm and ending at 8 pm, the time between 5 pm–8 pm was labeled "high-risk drinking", whereas all other times that day were labeled "non-drinking" (Supplemental Fig. 1). We labeled the time after drinking offset as "non-drinking" although blood alcohol concentration may still be elevated, given that activities likely differ between active drinking and end of drinking.

2.5. Analyses

2.5.1. Protocol adherence and data preparation

We measured two main components of protocol adherence: 1) completion of daily text queries, and 2) time running the AWARE app on their phone. We excluded individuals who did not provide > 1 report of a day with no drinking and > 1 report of a day on which alcohol was consumed over the 28 days ($n = 2$); or manually disabled sensor plug-ins, closed the AWARE app, or turned off the smartphone for > 80% of days ($n = 6$). If there were brief time periods (< 1 h) when sensor data was not captured, we interpolated average values based on neighboring data. The non-drinking, low-risk drinking, and high-risk drinking episodes with sensor data were first divided into non-overlapping 5 minute segments which were used to extract sensor features. The analyses examined 30-min, 1-h, and 2-h "windows", which aggregated (e.g., averaged for numerical values) sensor data over the relevant 5-min segments. As an example outcome of interest, using the 30-min window, there were 12,442 total segments across participants, which were labeled "non-drinking" ($n = 11{,}798$), "low-risk drinking" ($n = 243$), or "high-risk drinking" ($n = 401$).

We created three non-overlapping datasets using all of the coded segments across participants (e.g., $N = 12{,}442$ coded segments using 30-min window size). The "training" dataset (60% of coded segments) was used to select features and build the initial model, "cross-validation" dataset (20%) optimized the initial model (e.g., modified feature weighting to optimize performance), and "test" data (20%) evaluated the optimized model's performance. To reduce imbalances that can bias model building due to under-represented events (i.e., low- and high-risk drinking), in the training dataset, we used Synthetic Minority Over-Sampling Technique (SMOTE) (Chawla, Bowyer, Hal, & Kegelmeyer, 2002). For example, prior to SMOTE, the training dataset (30-min window) included 7078 non-drinking, 145 low-risk drinking, and 240 high-risk drinking segments. After SMOTE, non-drinking segments did not change ($n = 7078$), whereas low-risk increased to 1160, and high-risk segments increased to 960, reducing imbalance.

2.5.2. Model building

For both raw sensor data (e.g., number of calls) and processed sensor data (e.g., min/max, standard deviation), we ran correlation and Information Gain (Xu & Chen, 2010) analyses to identify the 20 most informative features in the "training" dataset. We used the "cross-validation" dataset to select the top 20 features and to optimize the model built with the training dataset. Using only the top-20 features identified in training and cross-validation datasets, we evaluated the following machine learning classifiers on the "test" dataset: C4.5 decision tree, Bayesian Network (BN) and Random Forest (RF) for each window size (30-min, 1-h, 2-h). Multiple methods evaluated model performance (Powers, 2011): accuracy, Kappa, F-score, and Receiver Operating Characteristic (ROC) (Bae et al., 2017). We report Kappa and ROC as they provide good representation of results across methods.

Using the "test" dataset, we also determined whether phone sensor data collected before drinking onset (1-, 2-, 3-days; "historical sensor data"; Supplemental Fig. 1) improved model performance, hypothesizing that there may be changes in patterns of activity that routinely precede alcohol use. We then evaluated model performance for detecting drinking on weekends (Friday to Sunday) versus weekdays (Monday to Thursday), based on research showing that drinking on these days involve different processes (Lau-Barraco, Braitman, Linden-Carmichael, & Stamates, 2016). Finally, we compared model performance using (1) only time of day and day of week, and (2) all top-20 sensor features against a naïve model (ZeroR) that predicts the most frequent (largest N) class. We also explored the amount of time in days for the best performing model to stabilize (i.e., accuracy does not change beyond a small threshold, by adding more additional training data). This is a measure of how many days of data must be collected to build a detection model for a new individual. We also examined whether single sensor features (e.g., number of calls), in addition to time of day and day of week, improved accuracy of the best performing model.

3. Results

3.1. Protocol adherence: daily query and sensor data

3.1.1. Adherence rates

To the 38 enrolled, we sent out a total of 1064 daily surveys, among which 764 (71.8%) were completed. Completion rates decreased from 87.9% on day 1 to 53.9% by day 28, with the greatest decline in the last week (Fig. 1). Sensor data was captured on 57% of the 28 days. Across the 30 participants in the analyzed sample, who did not differ significantly in baseline attributes from the 38 enrolled, there were 207 days on which no drinking was reported, 41 days on which a low-risk drinking episode was reported and 45 days on which a high-risk drinking episode was reported where both sensor and drinking data were available. Participants reported an average of 6.9 (SD = 6.7) non-drinking, 1.4 (SD = 1.9) low-risk drinking, and 1.5 (SD = 1.6) high-risk drinking episodes. 22 participants provided at least 1-week of data, 20 provided at least 2-weeks of data, and 10 provided at least 3 weeks of data.

3.1.2. Drinking episode characteristics

Mean number of drinks consumed during an entire low-risk drinking episode was 2.2 (SD = 1.0), and during a high-risk drinking

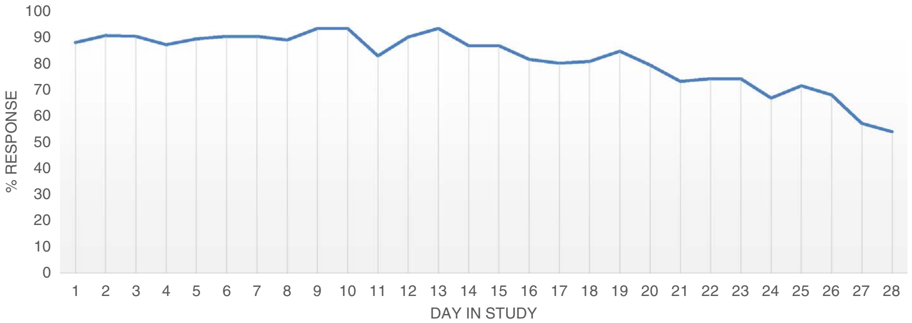

Fig. 1. Daily query response rates over 28 study days.

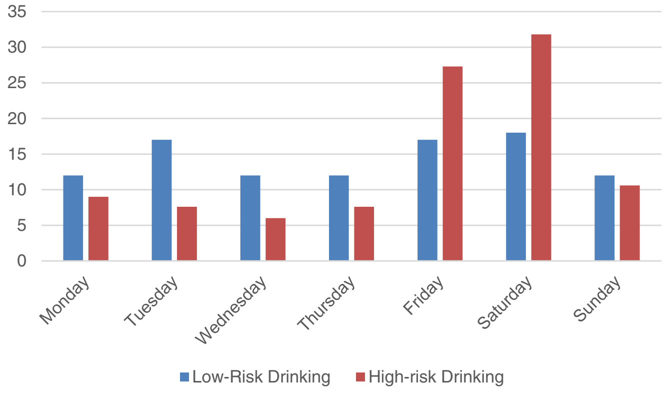

Fig. 2. Drinking episodes by day of week.

episode was 7.6 (SD = 2.7), with a maximum of 15 drinks. Low- and high-risk drinking episodes were distributed across participants (Supplementary Fig. 2). Regarding drinking onset, 41.2% of drinking episodes commenced between 8 and 9 pm and 40.7% stopped after midnight. Almost half (47%) of low-risk drinking episodes occurred on weekends whereas 69.7% of high-risk drinking occurred on weekends (Fig. 2).

3.2. *Correlation of sensor features with drinking*

Using 1-h windows, drinking categories were significantly correlated with time of day ($r = 0.11$) and day of week ($r = 0.06$), in addition to 9 other sensor-based features measuring phone usage (e.g., screen interaction, $r = 0.07$) and movement features (e.g., transitions between walking and in vehicle; $r = 0.03$). Many of these correlations increased when adding historical data of 23-, 47- and 71-h prior to that categorized hour.

3.3. *Information gain and descriptive statistics for select sensor features*

Information Gain, using 1-h windows, applied to the "training" dataset identified the top 20 sensor features associated with drinking cateogires (Bae et al., 2017). During high-risk drinking windows, there were higher activity levels, and higher change in activity levels, yet lower distance traveled compared to low-risk and non-drinking (Table 2). There were also longer average periods of screen duration and lower frequency of phone unlocking during high-risk drinking windows compared to low-risk and non-drinking windows. Regarding communication features, there were longer and more missed calls during high-risk drinking windows, compared to low-risk drinking windows. For typing features, there was greater time between keypresses, more text deletions and insertions, and greater use of happy emoticons during high- versus low-risk drinking, and non-drinking windows.

Table 2
Mean values per 1-hour window of select sensor features by drinking classification.

Features	Non-drinking	Low-risk drinking	High-risk drinking
Movement features			
Activity level	4.98	3.67	5.75
Count of changes in activities	5.71	5.72	8.24
Distance traveled (meters)	43.3	56.27	28.35
Phone usage features			
Screen duration (secs)	23.19	16.32	57.27
Phone screen unlocks (per min)	13.41	3.24	1.48
Communication features			
Duration of outgoing calls (secs)	29.02	5.13	11.93
Time between keypress (msec)	513.63	502.16	743.35
Letter deletions	14.75	11.58	16.30

Note: Information gain analyses using 1-h window.

3.4. *Machine learning model performance*

3.4.1. *Best performing model and inclusion of "historical" sensor data*

Among the machine learning models tested, Random Forest (RF) generally performed best. The following results are for RF applied to "test" dataset. The best performing model to detect drinking overall used RF, 30-min windows and 3-days of historical data, which had a Kappa of 0.804 and ROC of 0.961, correctly classifying non-drinking 98.5% of the time, low-risk drinking 70.2% of the time, and high-risk drinking 84.4% of the time.

The best performing model to detect high-risk drinking, however, used RF, 30-min windows and 1-day of historical data. In this model, the "1-day" (including the current 30-min window being classified) corresponds to 23.5 h of historical data in 30-min windows, for a total of 47 windows, plus the 30 min window being evaluated. The total number of features in this model was 48 windows ∗ 20 top-features = 960 total features. In this model, Kappa increased to 0.842 and ROC to 0.976 (relative to RF using 30-min windows and 3-days of historical data), correctly classifying non-drinking 97.9% of the time, low-risk drinking 68.3% of the time, and high-risk drinking 90.9% of the time. In this optimal model to detect high-risk drinking, the 9.09% of actual high-risk drinking segments that were misclassified ($n = 7$), and the 31.7% of actual low-risk drinking segments that were misclassified ($n = 13$), were all incorrectly classified as non-drinking (Supplemental Table).

3.4.2. *RF model performance in detecting weekday versus weekend drinking*

Using RF to detect both weekend and weekday drinking using 30 min windows and 3-days of historical data in the "test" dataset, performance was slightly lower for detecting weekday (Kappa = 0.728, ROC = 0.951) compared to weekend drinking (Kappa = 0.832, ROC = 0.991). Based on this RF model, high-risk drinking was correctly classified 81.8% of weekdays versus 80.4% of weekends, whereas low-risk drinking was correctly classified 35.3% of weekdays versus 87.1% of weekends.

3.4.3. *RF models using time of day/day of week and single sensor feature to detect drinking*

Comparing the performance of RF models (30-min window, 3-day historical data) in the "test" dataset using only time of day and day of week, and all top-20 sensor features against a naïve model (ZeroR), the RF model using all top-20 features showed a relative improvement in accuracy of around 6% compared to a RF model using only time of day and day of week (Fig. 3). It took approximately 10 days for the RF model (30-min window, 3-day historical data) to reach a stable accuracy (around 96%) regardless of what subset we chose for training. Finally, when examining whether individual sensor features, in addition to time of day and day of week, improved model performance, for the RF model, 30-min window without any historical data, only time between keypress improved performance, albeit slightly. The two sensor features that independently improved performance in a RF model, 30-min window with 1-day of historical data, were number of incoming calls and screen interaction duration.

4. Discussion

Preliminary evidence suggests that sensor data captured from mobile phones of young adults are useful in building accurate models to detect periods of high-risk drinking. Our study design has several noteworthy strengths. First, we recruited a diverse sample of young adults

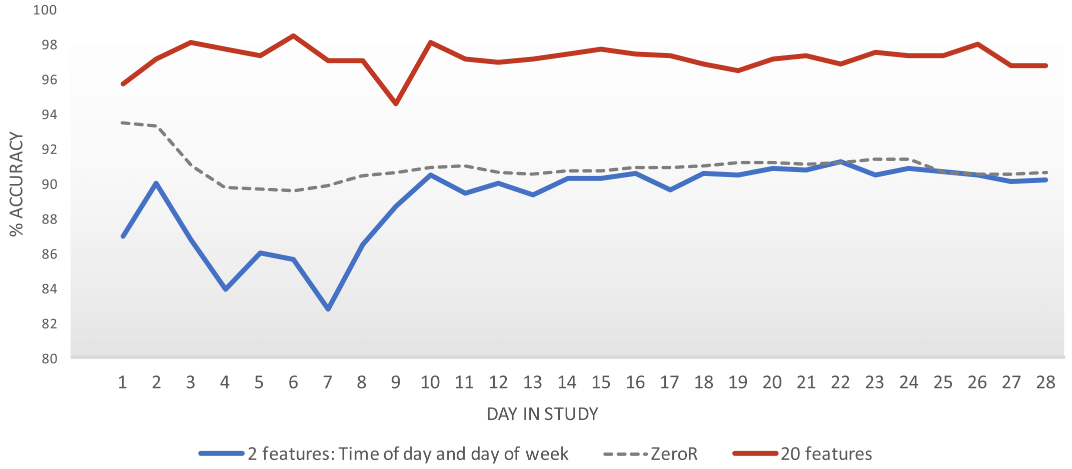

Fig. 3. Cumulative accuracy of random forest models (30-min window, 3-day historical data) over 28 days.
Legend: X-axis represents number of days (maximum 28 days), Y-axis represents accuracy of classifying non-drinking, low-risk, and high-risk drinking. The Red and Blue lines represent results from Random Forest (30-min window, 3-days of historical data) model, since it had the best overall performance in classifying non-drinking, low-risk, and high-risk drinking. Red line: only top-20 features were used for classification. Blue line: only 2 features, time of day and day of week, were used for classification. ZeroR model (dashed line) is a naïve model that just predicts the most frequent 'N' class. The graph depicts cumulative accuracy up to a given day, and not accuracy per day. Cumulative accuracy was determined by incrementally training models on successively larger sets of data. The figure shows higher classification accuracy (i.e., non-drinking, low-risk drinking, high-risk drinking) when using the top-20 features (red line) compared to the model using only time of day and day of week (blue line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

from two different settings that are both key points of contact for providing brief alcohol interventions (Carey et al., 2012; Suffoletto et al., 2015). We longitudinally captured onset and offset of prior day's drinking using daily recall, minimizing potential reporting biases (Shiffman, 2009). Our outcomes classified high- and low-risk drinking windows separately, since type of behavioral support provided for these events differ. We tested several machine learning models, varying time windows for detection and amounts of historical data to detect low- and high-risk drinking. Finally, we examined differences in accuracy of models for detecting weekend versus weekday drinking.

We found that time of day and day of week alone resulted in accuracy of 90% in detecting low- and high-risk drinking, which is consistent with research examining timing of young adult drinking (Del Boca et al., 2004). Adding other phone sensor features (e.g., change in activity) resulted in a 6% increase in accuracy, which represents 1 out of 16 (on average) times when we improve classification. In this regard, important phone sensor features to detect drinking included movement (activity change), device usage (e.g., screen duration), communication (e.g., call duration) and typing (speed/errors). There seemed to be more change in activities (e.g., walking, in a vehicle) associated with high-risk drinking. Also, prior to high-risk drinking, there were more incoming calls and longer screen interactions, which fit with the social-ecological model of alcohol use (Freisthler et al., 2014; Gruenewald et al., 2014), in which drinking and making plans to drink are social activities.

Within high-risk drinking windows, we found increased time between keystrokes, which is in line with a pharmacological model of the effects of alcohol on psychomotor functioning (Scholey et al., 2012; Suffoletto, Goyal et al., 2017). Alternatively, these sensor-based features could simply reflect, for example, distractions in social settings, or some combination of the acute effects of alcohol on psychomotor functioning and environmental context (Tiplady, Oshinowo, Thomson, & Drummond, 2009).

Based on the accuracy of our optimized model, we can detect high-risk drinking periods with 90% accuracy, within 30 min after drinking onset, using data collected over a relatively short time (~10 days). Thus, for 9 out of 10 high-risk drinking periods, we theoretically have the potential to intervene early during a drinking episode (e.g., within 30-min after drinking onset) to enhance motivation toward setting drinking limits and/or intervene later to reduce the likelihood of negative alcohol-related consequences. Example interventions include delivering supportive messaging, or contacting supportive friends or family.

Despite high accuracy, the model also generated false classifications. As one day is composed of forty-eight 30-min windows, a 2.1% false positive rate means around one "30-min window" where an individual was not drinking was estimated as drinking per day. A 9.1% false negative rate means that around four "30-min windows" per day when an individual is in a high-risk drinking episode would be misclassified as non-drinking. To minimize intervening during non-drinking windows, without missing any potential drinking occasion, a program could send an electronic (e.g. text message) query to individuals to verify at that moment whether or not they are drinking. Missing high-risk drinking events completely is unlikely given that consecutive misclassification over two 30-min windows is probabilistically rare.

This pilot study is limited by the small sample size and by the amount of missing sensor data. Although we did not systematically record reasons for missing sensor data, feedback from participants who stopped the AWARE app involved concerns about privacy, data usage, or perception of battery drain. Generalizability of results may be limited, since participants who did not provide adequate data for analysis were excluded, and these excluded individuals would not be eligible for interventions that rely on the detection model. In addition, we used self-report of alcohol use, which has demonstrated reliability and validity (Simons, Wills, Emery, & Marks, 2015), but may be subject to bias. We only coded the episode of heaviest drinking on a given day if there were multiple drinking episodes, which could affect model accuracy (i.e., result in more false negatives). Future work could use some form of alcohol sensor (e.g., WrisTAS) to validate findings, and to examine number and pacing of drinks consumed. Machine learning models, which are data-driven, do not provide explanations regarding why a feature is related to an outcome, but may have potential to inform more dynamic theories of behavior (Riley et al., 2011). Finally, although we used training, cross-validation, and testing data sets, we did not externally validate our model.

5. Conclusions

Phone sensors can provide useful data for use in machine learning models to accurately detect high-risk drinking in young adults. Although these results need to be replicated in a larger sample, next generation mobile interventions could consider using phone sensor features analyzed in real-time by machine learning algorithms to trigger just-in-time behavioral support.

Role of funding sources

There was no direct funding for this project. BS and TC both receive funding from NIAAA for related work: Grant K23 AA023284 (to BS) and R01 AA023650 (to BS&TC).

Contributors

Authors BS, TC, and AD designed the study and wrote the protocol. Authors DF and SB conducted literature searches for background and discussion and provided summaries of previous research studies. Authors BS and SB conducted the statistical analysis. Author BS wrote the first draft of the manuscript and all authors contributed to and have approved the final manuscript.

Conflict of interest

The authors have no conflicts of interest to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.addbeh.2017.11.039.

References

Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., & Steggles, P. (1999). Towards a better understanding of context and context-awareness. In H. W. Gellersen (Ed.). *Handheld and ubiquitous computing*. Heidelberg: Springer-Berlin.

Bae, S., Ferreira, D., Suffoletto, B., Puyana, J. C., Kurtz, R., Chung, T., & Dey, A. K. (2017). *Detecting drinking episodes in young adults using smartphone-based sensors.* PACM Interactive Mobile Wearable Ubiquitous Technology (IMWUT)1.

Berman, A. H., Gajecki, M., Sinadinovic, K., & Andersson, C. (2016). Mobile interventions targeting risky drinking among university students: A review. *Current Addiction Reports, 3*, 166–174.

Bradley, K. A., DeBenedetti, A. F., Volk, R. J., Williams, E. C., Frank, D., & Kivlahan, D. R. (2007). AUDIT-C as a brief screen for alcohol misuse in primary care. *Alcoholism: Clinical and Experimental Research, 31*, 1208–1217.

Carey, K. B., Scott-Sheldon, L. A., Elliott, J. C., Garey, L., & Carey, M. P. (2012). Face-to-face versus computer-delivered alcohol interventions for college drinkers: A meta-analytic review, 1998 to 2010. *Clinical Psychology Review, 32*, 690–703.

Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S. J., & Bierut, L. J. (2015). "Hey everyone, I'm drunk" An evaluation of drinking-related twitter chatter. *Journal of Studies on Alcohol and Drugs, 76*, 635–643.

Center for Behavioral Health Statistics and Quality (2016). *Key substance use and mental health indicators in the United States: Results from the 2015 National Survey on Drug Use and Health.* HHS Publication. No. SMA 16–4984, NSDUH Series H-51). Available from: http://www.samhsa.gov/data/.

Chawla, N. V., Bowyer, K. W., Hal, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Del Boca, F. K., Darkes, J., Greenbaum, P. E., & Goldman, M. S. (2004). Up close and personal: Temporal variability in the drinking of individual college students during their first year. *Journal of Consulting and Clinical Psychology, 72*, 155–164.

Ferreira, D., Kostakos, V., & Dey, A. K. (2015). AWARE: Mobile context instrumentation framework. *Frontiers in ICT, 2*.

Fowler, L. A., Holt, S. L., & Joshi, D. (2016). Mobile technology-based interventions for adult users of alcohol: A systematic review of the literature. *Addictive Behaviors, 62*, 25–34.

Freisthler, B., Lipperman-Kreda, S., Bersamin, M., & Gruenewald, P. J. (2014). Tracking the when, where, and with whom of alcohol use: Integrating ecological momentary assessment and geospatial data to examine risk for alcohol-related problems. *Alcohol Research: Current Reviews, 36*, 29–38.

Gruenewald, P. J., Remer, L. G., & LaScala, E. A. (2014). Testing a social ecological model of alcohol use: The California 50-city study. *Addiction, 109*, 736–745.

Humeniuk, R., Ali, R., Babor, T. F., Farrell, M., Formigoni, M. L., Jittiwutikarn, J., ... Simon, S. (2008). Validation of the alcohol, smoking and substance involvement screening test (ASSIST). *Addiction, 103*, 1039–1047.

Lau-Barraco, C., Braitman, A. L., Linden-Carmichael, A. N., & Stamates, A. L. (2016). Differences in weekday versus weekend drinking among nonstudent emerging adults. *Experimental and Clinical Psychopharmacology, 24*, 100–109.

Moewaka Barnes, H., McCreanor, T., Goodwin, I., Lyons, A., Griffin, C., & Hutton, F. (2016). Alcohol and social media: Drinking and drunkenness while online. *Critical Public Health, 26*, 62–76.

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology, 13*, 23–47.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2016). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* (epub ahead of print).

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology, 2*, 37–63.

Riley, W. T., Rivera, D. E., Atienza, A. A., Nilsen, W., Allison, S. M., & Mermelstein, R. (2011). Health behavior models in the age of mobile interventions: Are our theories up to the task? *Translational Behavioral Medicine, 1*, 53–71.

Rubinsky, A. D., Dawson, D. A., Williams, E. C., Kivlahan, D. R., & Bradley, K. A. (2013). AUDIT-C scores as a scaled marker of mean daily drinking, alcohol use disorder severity, and probability of alcohol dependence in a U.S. general population sample of drinkers. *Alcoholism: Clinical and Experimental Research, 37*, 1380–1390.

Scholey, A. B., Benson, S., Neale, C., Owen, L., & Tiplady, B. (2012). Neurocognitive and mood effects of alcohol in a naturalistic setting. *Human Psychopharmacology, 27*, 514–516.

Shiffman, S. (2009). Ecological momentary assessment (EMA) in studies of substance use. *Psychological Assessment, 21*, 486–497.

Simons, J. S., Wills, T. A., Emery, N. N., & Marks, R. M. (2015). Quantifying alcohol consumption: Self-report, transdermal assessment, and prediction of dependence symptoms. *Addictive Behaviors, 50*, 205–212.

Suffoletto, B., Gharani, P., Chung, T., & Karimi, H. (2017). Using phone sensors and an artificial neural network to detect gait changes during natural drinking episodes. *Gait and Posture* (in press).

Suffoletto, B., Goyal, A., Puyana, J. C., & Chung, T. (2017). Can an app help identify psychomotor function impairments during drinking occasions in the real world? A mixed-method pilot study. Oct-Dec *Substance Abuse, 38*(4), 438–449. http://dx.doi.org/10.1080/08897077.2017.

Suffoletto, B., Kristan, J., Chung, T., Jeong, K., Fabio, A., Monti, P., & Clark, D. B. (2015). An interactive text message intervention to reduce binge drinking in young adults: A randomized controlled trial with 9-month outcomes. *PLoS One, 10*, e0142877.

Tiplady, B., Oshinowo, B., Thomson, J., & Drummond, G. B. (2009). Alcohol and cognitive function: Assessment in everyday life and laboratory settings using mobile phones. *Alcoholism: Clinical and Experimental Research, 33*, 2094–2102.

Xu, Y., & Chen, L. (2010). Term-frequency based feature selection methods for text categorization. *IEEE genetic and evolutionary computing (ICGEC), 2010 Fourth International Conference* (pp. 280–283). .