HHS Public Access
Author manuscript
Addiction. Author manuscript; available in PMC 2024 May 09.

Published in final edited form as:
Addiction. 2023 October ; 118(10): 2014–2025. doi:10.1111/add.16228.

Development and validation of a model for measuring alcohol consumption from transdermal alcohol content data among college students

Sina Kianersi[1,2], **Christina Ludema**[1], **Jon Agley**[3], **Yong-Yeol Ahn**[4], **Maria Parker**[1], **Sophie Ideker**[5], **Molly Rosenberg**[1]

[1] Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, IN, USA

[2] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[3] Prevention Insights, Department of Applied Health Science, Indiana University School of Public Health-Bloomington, Bloomington, IN, USA

[4] Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA

[5] Epidemiology Department, Columbia University's Mailman School of Public Health, New York City, NY, USA

Abstract

Background and aims: Transdermal alcohol content (TAC) data collected by wearable alcohol monitors could potentially contribute to alcohol research, but raw data from the devices are challenging to interpret. We aimed to develop and validate a model using TAC data to detect alcohol drinking.

Design: Model development and validation

Setting: Indiana, USA

Participants: In March-April 2021, we enrolled 84 college students who reported drinking at least once a week (Median age=20 years, 73% white, 70% female). We observed participants' alcohol drinking behavior for one week.

Measurements: Participants wore BACtrack® Skyn monitors (TAC data), provided self-reported drinking start times in real-time (smartphone app), and completed daily surveys about their prior day of drinking. We developed a model using signal filtering, peak detection algorithm, regression, and hyperparameter optimization. The input was TAC and outputs were alcohol drinking frequency, start time, and magnitude. We validated the model using daily surveys (internal validation) and data collected from college students in 2019 (external validation).

Corresponding author: Sina Kianersi (skianersi@bwh.harvard.edu).

Declaration of interests: none to declare

Findings: Participants (N=84) self-reported 213 drinking events. Monitors collected 10,915 hours of TAC. In internal validation, the model had a sensitivity of 70.9% (95% confidence interval 64.1%−77.0%) and a specificity of 73.9% (68.9%−78.5%) in detecting drinking events. The median absolute time difference between self-reported and model detected drinking start times was 59 minutes. Mean absolute error (MAE) for the reported and detected number of drinks was 2.8 drinks. In an exploratory external validation among five participants, number of drinking events, sensitivity, specificity, median time difference, and MAE were 15, 67%, 100%, 45 minutes, and 0.9 drinks, respectively. Our model's output was correlated with breath alcohol concentration data [Spearman's correlation (95% confidence interval): 0.88 (0.77, 0.94)].

Conclusion: This study, the largest of its kind to date, developed and validated a model for detecting alcohol drinking using transdermal alcohol content data collected with a new generation of alcohol monitors. The model and its source code are available as supplementary materials.

INTRODUCTION

Background

In young adults, alcohol use is a preventable risk factor for all-cause deaths (1). For years, compared to non-college-attending peers, college students report more frequently engaging in excessive alcohol drinking (2, 3). In 2017, about 54% of American full-time college students aged 18–22 years reported past-month drinking, 35% reported binge drinking, and 10% reported heavy drinking (4). Each of these rates are higher than those of non-college-attending-peers (4, 5). Alcohol use is linked to multiple problematic outcomes in the US, including 1,519 unintentional injury deaths (6), 696,000 assaults, and 97,000 sexual assaults (7) each year. Further, excessive alcohol use can cause many clinical disorders such as liver disease, diseases of the central nervous system, and cancer (8).

To study alcohol use among this high-risk population, it is important to have accurate data. However, traditional approaches of alcohol use data collection are imperfect, especially because they cannot easily facilitate continuous measurement. Common tools for collecting alcohol use data include self-report and breathalyzers (9). Both these measurement tools are considered reliable and are widely used in alcohol research. However, self-reports, while often cost-effective and useful for large scale prevalence studies, are also prone to recall bias, social desirability bias, and other types of measurement errors that might influence validity (10). Breathalyzers theoretically produce more objective data, but mouth alcohol and shallow breaths may cause inaccurate data (11). Further, breathalyzers cannot be used to produce continuous data due to their active data collection demands (11). A plausible alternative to both approaches is measurement of transdermal alcohol concentration (TAC) to collect objective and continuous alcohol use data passively and unobtrusively. However, devices to measure TAC remain nascent and need validation. In particular, they need to be validated against current commonly used measurement tools.

After each drink, around 1% of ingested alcohol is excreted through the skin and sweat glands (12). This can be measured as TAC and is hypothesized to correspond to blood alcohol content (BAC). While early TAC sensors were large and sometimes obtrusive (13), newer, small, wearable alcohol monitors have been designed to be worn on the wrist and can potentially collect reliable continuous data for research purposes and for personal feedback (13, 14). Wearable devices that detect TAC allow passive, objective, real-time, and continuous measurement of alcohol use (9, 15). Because the data are collected passively and objectively, measurement errors originating from participants (e.g., recall bias or missing data) are theoretically removed or minimized (16). Other potential advantages of such devices include compliance, comfort, high acceptability, and power efficiency (13). Yet, these wrist-worn alcohol monitors have a limited history of application in the real world, and even fewer instances of use among college student populations at high risk for extreme drinking events. Moreover, the TAC data they collect is in the form of a timeseries signal with "noise," by which we mean unwanted fluctuations and disturbances in TAC signal that do not correspond to a change in drinking status (Figure S1). Noise might be present in the signal, for instance, when individuals wear perfumes or use hand sanitizers. Hence, TAC data cannot readily be translated to commonly used alcohol use measures, such as alcohol use frequency and magnitude, using currently available approaches. Validation research and development of generalizable data processing procedures and models that can translate TAC data are needed.

Further hindering the transition to broader use of next-generation TAC monitors, nearly all TAC monitor validation studies have been conducted in laboratory settings using older monitors (i.e., SCRAM) which collect quasi-continuous data or suffer from high failure rates (17–20). TAC data produced by the new generation of wearable alcohol monitors is different in dimension from that produced by older devices. Models and rules developed for older devices are not applicable to the newer ones. While a small number of validation studies have been conducted using newer wrist-worn alcohol monitors, those studies were mainly conducted using early prototypes and in laboratory settings (13, 21), with fewer studies conducted in the field (22). Two recent studies have developed models for finding drinking start time (23) and breath alcohol concentration (BrAC) (24) based on TAC data produced by newer devices. Both models are yet to be validated for use with TAC data collected in the field and among general populations, such as college students, though one pilot field validation study found promising results for the latter model (22).

No model has been developed and validated to simultaneously capture drinking frequency, start time, and magnitude using TAC data collected with newer devices. Model development and validation studies on new wearable alcohol monitors in naturalistic drinking environments and among college students are required before researchers can reliably use these devices to better understand alcohol use in and develop alcohol interventions for college students (13, 21, 22), especially those who have high-risk drinking patterns (4, 6).

Objectives

We aimed to 1) develop a model that uses TAC data produced by the new generation of wearable alcohol monitors to detect alcohol drinking event start time, drinking frequency, and drinking magnitude, and to 2) evaluate the performance of this model relative to established alcohol use data collection tools.

METHODS

We followed relevant items from the STARD (25, 26) and TRIPOD (27, 28) guidelines when reporting our findings. The protocol for this study was approved by the Indiana University Bloomington (IUB) institutional review board (Protocol #2012949660). Participants provided electronic consent before participating in the study.

Study design

The study was conducted on the IUB campus, a large Big Ten state school in Monroe County, Indiana, with more than 33,000 undergraduate students. In this model development and validation study, we used a weeklong longitudinal prospective study design to collect alcohol use data.

Participants

Inclusion criteria were aged 18 years or older, enrollment in IUB courses in Spring 2021, living in Monroe County, IN, self-reported alcohol consumption at least once a week, and good general health. Additionally, because the study's alcohol sensors connected only to iPhones, non-iPhone users were excluded during recruitment.

Sampling: We used two sampling techniques; 1) one-stage random cluster sampling where clusters were IUB Spring 2021 classes and 2) a network sampling technique known as acquaintance sampling (sampling the friends of randomly selected individuals) (29). Potential participants were contacted via email. Owing to complexity, sampling is fully described in supplemental materials (see Sampling in Supplement Material).

Study procedures

In an online survey (REDCap), participants provided electronic informed consent, filled out a baseline survey about demographics and alcohol use history, and scheduled a baseline visit. Baseline visits took place on the IUB campus in March-April 2021. Participants used four tools to collect alcohol use data: wearable alcohol monitors, daily surveys, an ecological momentary assessment (EMA) methodology, and breathalyzers (Figure 1). Participants learned about study procedures in a baseline visit. We asked participants to wear their alcohol monitors most of the time and only take them off/turn them off when showering or charging the device. To efficiently use our alcohol monitor supply, we divided participants into three groups and collected data in three consecutive weeks, one week of data collection per group.

Wearable alcohol monitor

We used BACtrack® Skyn (firmware version 2.0.8) to collect TAC data (30). Skyn was the first-prize winner of National Institute on Alcohol Abuse and Alcoholism (NIAAA) Wearable Alcohol Biosensor Challenge in 2016 (31). Skyn is an alcohol monitor with sensors that collect TAC data in near real-time. The TAC data are indexed in time order and are stored in secure HIPAA-compliant servers. The device is worn on the wrist with the sensor on the palm side and connects to an iOS companion app on the user's iPhone via Bluetooth. This allows for the wireless transfer of the data from the monitor to the servers. Skyn collects one TAC [unit= μg/L(air)] data point every 20 seconds. The data can be downloaded from the server as a CSV file. We sent prompts to participants to sync and transfer the TAC data every two days and charge the devices whenever necessary.

Reference standard tests

We evaluated the accuracy of TAC data collected with Skyn relative to the data collected with three separate tools, some of which are arguably the reference standard tests (i.e., the best available method) in the field (26): daily surveys, an EMA methodology developed specifically for the current study, and breathalyzers.

Daily Survey—Every day at noon, participants completed an online survey about their previous day's drinking, drinking start time, and number of standard drinks consumed (Figure S2). A standard drink was defined as a drink that contains 14 grams (0.6 fluid ounces) of alcohol (32). A picture of common standard drinks was included in the daily survey. To improve the response rate, we sent SMS reminders to non-responders throughout the day. The validity and reliability of alcohol use data collected in daily surveys have been established (33, 34), with caveats (as described in the Background) that are mitigated by the triangulation of reference data collection.

EMA methodology—The EMA methodology was a timesheet survey created on REDCap (35, 36) to collect real-time drinking start times (37). Each participant worked with the research team to set up the survey so that the bookmark looked like an app on participants' phone Home screen. Opening the app revealed a "Now" button that recorded the current timestamp when pressed. Just before drinking every new alcoholic beverage, participants opened the app and clicked on the "Now" button to capture the drinking start time for that beverage. In an internal validation study, this EMA approach performed well relative to both retrospective self-reports and BrAC data (37). Every timestamp in the EMA app dataset indicated consumption of one standard drink. Timestamps within 5 hours of each other were coded as one drinking event and the earliest one of these timestamps was coded as the drinking event start time.

Breathalyzers—To assess the recall bias of self-reports in our study, we collected objective BrAC data from a random subsample of 25 participants. This random sample was selected using the Pandas package in Python; the breathalyzer was given to the next participant if the randomly selected individual missed their baseline visit appointment. We used smart, portable BACtrack® C6™ keychain breathalyzers that estimate BrAC level through exhaled breath and used an iPhone app to store the BrAC readings on the user's phone via Bluetooth. We asked participants to record their BrAC readings four times following their last drink and/or meal, once every 20–30 minutes. The maximum value among these readings was coded as the event BrAC level.

Covariates

In the baseline survey, we collected data on sex at birth (female/male), age (18 to ≥21 years), race (white/other), residence (off-campus/on-campus), year in school (1st to 4th and 5th), Greek membership (yes/no), annual income (≥$25,000 vs. <$25,000), and alcohol use history (frequency, magnitude, estimated BrAC level after a typical night of drinking). We used the validated self-report version of USAUDIT questionnaire to measure high-risk alcohol drinking at baseline (38–40). Higher USAUDIT scores show higher risk. We used a USAUDIT score of 7 for female and 8 for male participants as the cut-off point for high-risk drinking. Body mass index (BMI) was measured at baseline visit using a scale and tape measure (<18.5, 18.5–24.9, 25–29.9, and ≥30).

Model development (index test)

We developed a model that uses TAC data collected by Skyn to capture alcohol drinking frequency, start time, and magnitude. Our model consists of three consecutive procedures, 1) TAC data processing, 2) peak detection algorithm, and 3) regression analysis (Figure S3). We used the EMA app data as a reference standard test when developing the model.

In procedure one, we recoded negative TAC values as zero, and implemented median filter and moving average consecutively on recoded TAC data to remove the signal noise. Each peak in the processed TAC signal potentially represents a drinking event. However, a peak does not always correspond to an alcohol drinking event and could be due to environmental alcohol exposure, such as cleaning products or even alcohol drink spills. These naturally occurring environmental exposures provide an important justification for validating these procedures outside of a lab setting. We expected that the shapes of drinking event peaks would be different from those of other peaks. Thus, in procedure two, we used a peak detection algorithm to detect drinking events in the processed TAC data based on peak properties (e.g., prominence and width) (41). For each detected peak, the algorithm returned three time points: 1) Left base (which we defined as drinking start time), peak maximum (point with the maximum TAC value), and right base (last point in the peak timeseries). Lastly, for each detected peak, we calculated the area under the curve (AUC) from left base to right base of the detected peak using the composite Simpson’s rule.

In procedure three, we used peak maximum and peak AUC to predict number of standard drinks consumed in each drinking event. This procedure was only conducted on the true positive drinking events detected with the peak detection algorithm (i.e., peaks that were validated with self-report).

Hyperparameter optimization

A parameter that is used for improving the performance of an algorithm is called a hyperparameter. The process of identifying a good value for hyperparameters is known as hyperparameter optimization (42, 43). The optimization is mainly done with machine learning techniques based on pre-determined performance scores. The best hyperparameters are the ones that result in the best performance score.

Our model had multiple hyperparameters. We performed hyperparameter optimization using random grid search (also known as randomized parameter optimization) and finetuning (42, 43). In random grid search, we used group 5-fold cross-validation, an internal validation technique [GroupKFold (44)], which ensured the same participants were not included in both training and test sets. Hyperparameter optimization of procedures I and II was conducted simultaneously. Here, we used the balanced accuracy score, which is the arithmetic mean of sensitivity and specificity of the model in detecting drinking events. Balanced accuracy score prevents inflated performance in imbalanced datasets (45), such as that in our study. For procedure III, we conducted a hyperparameter optimization to select the best regression technique and its best hyperparameters out of four different commonly used regression techniques (regression technique was itself a hyperparameter in procedure III). To quantify performance of procedure III in predicting number of standard drinks in a drinking event we used mean absolute error (MAE) for the paired measures of EMA app recorded and model predicted number of standard drinks consumed in a drinking event. Annotated source code and details on model development are available as a Jupyter Notebook as well as an HTML file in supplemental materials (see Model Development Source Code.ipynb).

We used Python (version 3.9.1, Python Software Foundation, Beaverton, OR, US) when developing our model (46). SciPy was used in peak detection (47). We developed our estimator class for conducting procedures I and II to be compatible with scikit-learn (48). All machine learning procedures, including hyperparameter optimization, were conducted in scikit-learn (49). The final model is available as Python code in supplemental materials (Final model.ipynb).

Sample size

In pre-hoc power analyses, we estimated the minimum required number of participants to be 64 in validation analyses and 118 in correlation analyses (see Sample Size Calculation in Supplemental Material).

Model validation and statistical analysis

In internal validation analyses, we calculated the model performance relative to daily survey data in 1) detecting drinking events, 2) drinking event start times, and 3) drinking magnitude (i.e., number of standard drinks consumed in a drinking event).

To quantify model performance in detecting drinking events we reported sensitivity and specificity measures along with balanced accuracy. Each model-detected peak was counted as one drinking event in the index dataset. For a detected peak to be considered a true positive, its left base (start of the peak) needed to be within 5 hours of the self-reported drinking start time on the daily survey. Participants could report one drinking event start time in daily surveys. However, more than one peak might form in TAC data when participants drink intermittently throughout the day. Therefore, in cases where more than one peak was detected in a drinking day, all peaks in that day were counted as one drinking event (this occurred for 35 out of 146 accurately detected drinking days). We calculated sensitivity/specificity and the 95% exact CIs using SAS software, Version 9.4 of the SAS System for Windows 10 (Cary, NC, USA).

We calculated the absolute time difference between model-detected drinking event start time and the start time reported in the daily surveys to evaluate model performance in detecting drinking event start time. This comparison variable could range from 0 to 300 minutes (5 hours), with a value of 0 indicating that the start time in both data collection tools (Skyn alcohol monitors and daily survey) matched exactly with less than 1 minute of variability.

To quantify model performance in detecting drinking magnitude we calculated mean absolute error (MAE) for the paired measures of daily survey self-reported number of consumed standard drinks in a drinking event and model-predicted number of standard drinks consumed in a drinking event. We estimated the Spearman's correlation coefficient (50) for the continuous measures (number of drinks in daily surveys, peak maximum and AUC, and BrAC). We used complete case analysis; we included days where ≥1 TAC data points were collected, and the corresponding daily surveys were completed.

Exploratory external validation: In 2019, our team had collected alcohol use data from five IUB students, selected with convenience sampling, using study procedures similar to that of the current study (51). These five students wore earlier prototypes of Skyn wearable alcohol monitors and simultaneously reported their alcohol use with daily surveys for five consecutive days (EMA app and BrAC data were not collected). In a sensitivity analysis, using this small dataset collected in our 2019 study, we explored our model's external validity on TAC data.

RESULTS

Participants

Overall, N=84 students participated in our study, n=46 from the random cluster sample (Figure S5) and n=38 from the Friends sample (Figure S6). Participants were ages 18 to 22 with a median age of 20 years (IQR=2 years). Participants were mostly white (73%), female (70%), off-campus residents (70%), first year students (32%), non-Greek affiliated (70%), normal weight (60%), and had an income of less than $25,000 annually (88%). Most demographics and alcohol use patterns were similar to the general undergraduate population and to other larger studies among IUB undergraduate students (Table 1) (52).

Descriptive results

Skyn: Alcohol monitors collected 1,964,713 TAC data points (10,915 hours), out of a maximum possible 2,492,640 (13,848 hours) TAC data points that could have been collected if the devices were never turned off. Participants wore the monitors for 79% of the data collection week. On average, each participant provided 23,389 TAC data points (~130 hours) [Median: 25,553 (142 hours); IQR: 7,602 (105 hours)]. The mean TAC value was 11.17 μg/L(air) (Median: −0.43, IQR: 5.63).

Daily surveys: Out of the 84 participants who completed the baseline visit, three participants completed their baseline visits on the third day of their data collection week. One participant opted out of the study on day four of their data collection week, contributing their data until day three. Out of the 577 daily surveys that we sent out to the participants, 568 (response rate = 98.4%) were completed. Participants self-reported 213 drinking events. On average, participants self-reported 2.5 drinking events in the data collection week (Median: 2.5, IQR: 1). Five participants self-reported no drinking event and two had 6 drinking events. The mean value for the total consumed standard drinks by each participant at the end of data collection week was 13.2 (Median: 9.3, IQR: 12.1). Further, on average, participants self-reported consumption of 5.2 (Median: 4, IQR: 5) standard drinks in each of the 213 drinking events.

EMA app: Six participants did not record any drinking event using the EMA app. Overall, 78 participants recorded 206 drinking events. On average, participants recorded 2.6 drinking events in the data collection week (Median: 3, IQR: 1). Six participants had more than one drinking event in a day (drinks more than 5 hours apart). The mean value for the total consumed standard drinks recorded in the EMA app by each participant at the end of data collection week was 11.0 (Median: 8.0, IQR: 10.5). The mean value for the number of standard drinks recorded in each of the 206 drinking events was 4.2 (Median: 3, IQR: 3).

Breathalyzer: A total of 142 BrAC readings were recorded by 25 participants in 52 drinking events. On average, breathalyzers recorded a maximum BrAC level of 0.066% (Median: 0.042%, IQR: 0.082%). The mean for the maximum recorded BrAC levels in each drinking event was 0.090% (Minimum=0.008%, Median: 0.080%, IQR: 0.099%, Maximum: 0.237%).

Correlation analyses

Spearman’s correlation coefficient between number of standard alcohol drinks consumed in a drinking event self-reported in daily surveys and AUC for the peaks detected with our model was moderate and significant [r_s (95% CI): 0.57 (0.45, 0.67)]. Maximum BAC level in a drinking event was recorded for 32 of the detected drinking events. Among these 32 drinking events, Spearman’s correlation coefficient between the maximum BAC level recorded using breathalyzers and AUC for the drinking events detected with the model was strong and significant [r_s (95% CI): 0.88 (0.77, 0.94)] (Figure 2).

Model performance

Model apparent performance: Relative to EMA app data, the best-balanced accuracy score in the model development phase was 84% for procedures I and II (true negatives=2064, true positives=148, false positive=170, and false negative=47). The best MAE score in procedure III was 2.2 standard drinks.

Model performance relative to daily surveys: Participants wore Skyn monitors and collected one or more TAC data points on 620 days. Each participant completed at least one daily survey. Overall, both daily survey and TAC data (≥1 TAC readings) were available for 543 days. Under the assumption that the self-reported data accurately represented real drinking events, there were 146 true positives, 249 true negatives, 88 false positives, and 60 false negatives detected by the alcohol monitors (Figure 3). Relative to daily surveys, the sensitivity of our model in detecting drinking events in TAC data collected by Skyn alcohol monitors was 70.9% (64.1%−77.0%). Specificity was slightly higher, 73.9% (68.9% −78.5%), which equals a balanced accuracy score of 72.4%.

Model performance varied at the individual level. Both sensitivity and specificity were 100% for 16 out of 84 (19%) participants. Both sensitivity and specificity were ≥80% for n=21 (25%) participants. Sensitivity or specificity was <80% for n=60 (71%) participants. Sensitivity was zero for 10 (12.0%) participants.

Drinking start times: This analysis was conducted only on observations with a drinking event self-reported and detected by the alcohol monitor. The average absolute time difference between daily survey self-reports and model-detected drinking event start times was 79 minutes (Median: 59, Q1: 31, Q3: 109 minutes).

Drinking magnitude: Overall, 146 out of the 206 self-reported drinking events were detected (i.e., the true positives). Our model predicted number of standard drinks consumed in each of these detected 146 drinking events. The MAE was 2.8, meaning on average the absolute difference between the self-reported and predicted number of standard drinks consumed in each of the drinking events was 2.8 drinks.

Model performance (exploratory external validation)

The mean age of participants in the prior study used for external validation was 21.6 years (51). During the five days of data collection, we sent out 25 daily surveys, one per participant per data collection day. All 25 daily surveys were completed. However, one participant was not wearing the device for one of the data collection days (no TAC data were collected) and consequently we removed that daily survey from validation analysis. In the remaining 24 daily surveys, participants reported drinking at least one drink on 15 days and no drinking on 9 days.

The overall sensitivity of our mode in detecting drinking events in the exploratory external validation dataset was 66.7%, and the overall specificity in not detecting any peak for a day when participants reported no drinking in that day was 100%. Sensitivity was 100% for three participants, 50% for one participants, and 0% for one participant. Specificity was 100% for all participants. Mean absolute time difference between the detected and self-reported drinking event start times for the 10 true positive values was 66 minutes (Median: 45 minutes, Q1: 22 minutes, Q3: 80 minutes). The MAE was 0.91.

DISCUSSION

We developed a model to identify drinking events, drinking event start time, and drinking magnitude using a large TAC dataset collected with Skyn alcohol monitors among a sample of undergraduate students. We developed the model using EMA data as our benchmark. The model’s outputs were moderately and strongly correlated with alcohol use data collected with daily surveys and breathalyzers, respectively. Model performance was comparable to daily surveys. Similar performance results were obtained in the exploratory external validation analyses.

Limitations

First, our sample size was small, we did not reach our calculated sample size for correlation analyses, and BrAC data were available only for 25 participants. However, the sample size was larger than other similar studies on wearable alcohol monitors (21, 23, 24, 53) and exceeded minimum sample size recommendations for validation studies (28). Second, data collected with our EMA app, daily surveys, or breathalyzers were prone to measurement error. This could have biased the model performance estimates in either direction. However, correlation and performance measures were similar when comparing the model to any of the reference standard tests. Third, participants could have changed their alcohol drinking patterns because they knew that their drinking behavior was being observed. However, alcohol use history reported on the baseline survey and drinking patterns captured using the reference standard tests in the data collection week were similar, suggesting participants did not change their alcohol drinking behavior in the week of data collection.

Fourth, more daily surveys were missing on the days when our model detected peaks compared to days that it did not detect any peaks. This could have biased the true sensitivity of our model towards worse values. Fifth, Skyn wrist-worn alcohol monitors need to be worn tightly to produce reliable data. Even though we asked participants to wear the Skyn monitors snugly, it is possible that they could not comply with this study procedure all the time, especially given the longer period of our study. This could have caused an underestimation of the true model performance. Sixth, we did not collect data on compliance and the compliance proxies that Skyn collects (temperature and motion) are yet to be validated. Lastly, external validation analysis was exploratory with a small sample size. The Skyn monitors used in the external validation study were earlier prototypes and different in firmware from the ones used in the current study. Larger external validation studies using the more recent version of the device are needed to better understand the external validity of our findings.

Interpretation

We conducted the first model development and validation study using TAC data produced by a new generation of wearable alcohol monitors in naturalistic drinking environments. We identified two other developed models that use TAC data produced by Skyn monitors to measure alcohol use, 1) changepoint detection model (CPDM) (23) and 2) TSFRESH and Extra-Trees model (24). When used with TAC dataset, the CPDM model finds the timestamp(s) when TAC value changes abruptly. The detected timestamp is a potential alcohol drinking event start time. This model has been used in laboratory setting (21). However, its performance in detecting drinking events has not been measured in field studies.

Developed by Fairbairn et al. (24), the TSFRESH and Extra-Trees model was built in Python programming language on TAC data collected by earlier prototypes of Skyn devices. The MAE between predicted and true BrAC values was small (MAE: 0.010%) (24) but increased in an external validation field pilot study (MAE: 0.041%) (22). Our model does not predict BrAC; instead, it predicts number of standard drinks. Nonetheless, assuming each standard drink could roughly increase BrAC level by 0.020%, our model accuracy in predicting drinking magnitude was similar and slightly better than the TFRESH and Extra-Tress model. Compared to our model, the TSFRESH and Extra-Trees model seemed to have a higher sensitivity, though the definitions for true/false negative/positive values were different from that in our study (24). Similar to the strong Pearson correlation between predicted and true BAC values in Fairbairn et al. study (r=0.9) (24) and an external validation pilot field study (r=0.8) (22), the correlations between detected AUC and BrAC values in our study were strong (r=0.88). At the time of our study, we could not evaluate the external validity of the

TSFRESH and Extra-Trees model in our TAC dataset because this model is not currently publicly available.

Our model is unable to detect alcohol use at the time participants start drinking because a peak needs to be formed before our model can detect it. A peak forms hours after drinking start time. Detecting alcohol use at drinking start time is important particularly when developing just-in-time adaptive interventions (54) or EMIs (55). With the use of advanced machine learning approaches (56) it might be possible to improve our model and account for this limitation by predicting formation of a peak before it is actually formed, and then use our peak detection model to detect the predicted peak (drinking event) at the drinking start time. Our model, in its current form, could be used by other researchers when they aim to passively collect objective and real-time (i.e., EMA) alcohol use data among their study participants.

Even though we found similar results in a small external validation analysis among five IUB undergraduate students, the accuracy of our model in other settings and populations remains unknown. It is possible to include more features (covariates such as BMI, sex, or mealtime) to improve model’s performance potentially further. Researchers can use our model with their data or even recalibrate our model to fit their data collection needs. For example, if specificity of detecting drinking events is more important than sensitivity, researchers can increase the prominence hyperparameter value in the peak detection algorithm to increase the model’s specificity.

Conclusion

BACTrack® Skyn wearable alcohol monitors provide high frequency TAC data. We developed and validated a model to translate the raw TAC data into measures that approximate commonly known alcohol use measures. Our model can be used for ecological momentary assessment/intervention of alcohol use, at this time, it cannot be used for EMIs that aim to deliver an intervention at the time of alcohol drinking. Additionally, it is possible to recalibrate the model to adjust the model performance. More external validation studies are needed to better understand the validity of our model and to replicate our findings in other populations. The developed model is included in supplementary materials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

I (SK) am deeply grateful to my PhD mentor, Dr. Molly Rosenberg, for her invaluable guidance and support throughout my PhD program. The current study was part of SK’s dissertation in his PhD program (57). SK was a pre-doctoral trainee of an NIH/NIAAA funded program while working on this study [NIAAA grant # R25DA051249, 2021]. NIAAA had no role in the design, analysis, interpretation, or publication of this study. The content is solely the responsibility of the authors. Prevention Insights at the Indiana University Bloomington School of Public Health (Big Idea Challenge, awarded to SK) provided part of the research expenses. The BACtrack® Skyn devices were provided by Research Equipment Funding through the IU Office of the Vice Provost for Research to MR and CL.

Role of Funding Source:

JA (co-author) is the Deputy Director for Research at Prevention Insights, which provided funding for parts of this study through a competitive grant opportunity for students at Indiana University. Prevention Insights is a translational research center within the Indiana University School of Public Health and issued the grant opportunity (the “Big Idea Challenge”) in 2019 as a means of supporting and collaborating with innovative student scholars.

REFERENCES

1. Danaei G, Ding EL, Mozaffarian D, Taylor B, Rehm J, Murray CJ, et al. The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. PLoS Med. 2009;6(4):e1000058. [PubMed: 19399161]
2. Merrill JE, Carey KB. Drinking Over the Lifespan: Focus on College Ages. Alcohol Res. 2016;38(1):103–14. [PubMed: 27159817]
3. Carter AC, Brandon KO, Goldman MS. The college and noncollege experience: A review of the factors that influence drinking behavior in young adulthood. Journal of studies on alcohol and drugs. 2010;71(5):742–50. [PubMed: 20731981]
4. Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the United States: Results from the 2018 National Survey on Drug Use and Health (HHS Publication No. PEP19–5068, NSDUH Series H-54). Rockville, MD: Center for Behavioral Health Statistics and Quality. Substance Abuse and Mental Health Services Administration. 2019.
5. National Institute on Alcohol Abuse and Alcoholism. Fall Semester—A Time for Parents To Discuss the Risks of College Drinking 2019 [cited 20020 01.21.2020]. Available from: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/time-for-parents-discuss-risks-college-drinking.
6. Hingson R, Zha W, Smyth D. Magnitude and trends in heavy episodic drinking, alcohol-impaired driving, and alcohol-related mortality and overdose hospitalizations among emerging adults of college ages 18–24 in the United States, 1998–2014. Journal of studies on alcohol and drugs. 2017;78(4):540–8. [PubMed: 28728636]
7. Hingson R, Heeren T, Winter M, Wechsler H. Magnitude of alcohol-related mortality and morbidity among US college students ages 18–24: Changes from 1998 to 2001. Annual review of public health. 2005;26.
8. Centers for Disease Control and Prevention. Alcohol & Substance Misuse 2018 [updated February 1, 2018. Available from: https://www.cdc.gov/workplacehealthpromotion/health-strategies/substance-misuse/index.html.
9. Leffingwell TR, Cooney NJ, Murphy JG, Luczak S, Rosen G, Dougherty DM, et al. Continuous objective monitoring of alcohol use: twenty‐first century measurement using transdermal sensors. Alcoholism: Clinical and Experimental Research. 2013;37(1):16–22. [PubMed: 22823467]
10. Weissenborn R, Duka T. Acute alcohol effects on cognitive function in social drinkers: their relationship to drinking habits. Psychopharmacology. 2003;165(3):306–12. [PubMed: 12439627]
11. Luczak SE, Rosen IG. Estimating Br AC from Transdermal Alcohol Concentration Data Using the Br AC Estimator Software Program. Alcoholism: clinical and experimental research. 2014;38(8):2243–52. [PubMed: 25156615]
12. Swift R Direct measurement of alcohol and its metabolites. Addiction. 2003;98 Suppl 2:73–80. [PubMed: 14984244]
13. Wang Y, Fridberg DJ, Leeman RF, Cook RL, Porges EC. Wrist-worn alcohol biosensors: Strengths, limitations, and future directions. Alcohol. 2019;81:83–92. [PubMed: 30179709]
14. Campbell AS, Kim J, Wang J. Wearable Electrochemical Alcohol Biosensors. Curr Opin Electrochem. 2018;10:126–35. [PubMed: 30859141]
15. Bond JC, Greenfield TK, Patterson D, Kerr WC. Adjustments for drink size and ethanol content: new results from a self-report diary and transdermal sensor validation study. Alcohol Clin Exp Res. 2014;38(12):3060–7. [PubMed: 25581661]
16. Piasecki TM. Assessment of alcohol use in the natural environment. Alcoholism: clinical and experimental research. 2019;43(4):564–77. [PubMed: 30748019]

17. Karns-Wright TE, Roache JD, Hill-Kapturczak N, Liang Y, Mullen J, Dougherty DM. Time Delays in Transdermal Alcohol Concentrations Relative to Breath Alcohol Concentrations. Alcohol Alcohol. 2017;52(1):35–41. [PubMed: 27522029]

18. Dougherty DM, Hill-Kapturczak N, Liang Y, Karns TE, Lake SL, Cates SE, et al. The Potential Clinical Utility of Transdermal Alcohol Monitoring Data to Estimate the Number of Alcoholic Drinks Consumed. Addict Disord Their Treat. 2015;14(3):124–30. [PubMed: 26500459]

19. Hill-Kapturczak N, Lake SL, Roache JD, Cates SE, Liang Y, Dougherty DM. Do variable rates of alcohol drinking alter the ability to use transdermal alcohol monitors to estimate peak breath alcohol and total number of drinks? Alcoholism: clinical and experimental research. 2014;38(10):2517–22. [PubMed: 25335857]

20. Dougherty DM, Charles NE, Acheson A, John S, Furr RM, Hill-Kapturczak N. Comparing the detection of transdermal and breath alcohol concentrations during periods of alcohol consumption ranging from moderate drinking to binge drinking. Exp Clin Psychopharmacol. 2012;20(5):373–81. [PubMed: 22708608]

21. Fairbairn CE, Kang D. Temporal Dynamics of Transdermal Alcohol Concentration Measured via New-Generation Wrist-Worn Biosensor. Alcohol Clin Exp Res. 2019;43(10):2060–9. [PubMed: 31469451]

22. Ariss T, Fairbairn CE, Bosch N. Examining new-generation transdermal alcohol biosensor performance across laboratory and field contexts. Alcohol Clin Exp Res. 2023;47(1):50–9.

23. Killick R, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association. 2012;107(500):1590–8.

24. Fairbairn CE, Kang D, Bosch N. Using machine learning for real-time BAC estimation from a new-generation transdermal biosensor in the laboratory. Drug and Alcohol Dependence. 2020;216:108205. [PubMed: 32853998]

25. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clinical chemistry. 2015;61(12):1446–52. [PubMed: 26510957]

26. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11):e012799.

27. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement. Circulation. 2015;131(2):211–9. [PubMed: 25561516]

28. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1–73. [PubMed: 25560730]

29. Cohen R, Havlin S, Ben-Avraham D. Efficient immunization strategies for computer networks and populations. Phys Rev Lett. 2003;91(24):247901. [PubMed: 14683159]

30. BACtrack. WEARABLE ALCOHOL BIOSENSOR BACtrack Skyn Track your alcohol level directly from your wrist, in near real-time 2022 [1.25.2020]. Available from: https://skyn.bactrack.com/.

31. National Institute of Health. NIAAA selects winners of its Wearable Alcohol Biosensor Challenge 2016 [Available from: https://www.nih.gov/news-events/news-releases/niaaa-selects-winners-its-wearable-alcohol-biosensor-challenge.

32. National Institutes of Health. What’s a “standard” drink? [Available from: https://www.rethinkingdrinking.niaaa.nih.gov/How-much-is-too-much/what-counts-as-a-drink/whats-A-Standard-drink.aspx.

33. Del Boca FK, Darkes J. The validity of self-reports of alcohol consumption: state of the science and challenges for research. Addiction. 2003;98:1–12.

34. Ekholm O Influence of the recall period on self-reported alcohol intake. Eur J Clin Nutr. 2004;58(1):60–3. [PubMed: 14679368]

35. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O’Neal L, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform. 2019;95:103208. [PubMed: 31078660]

36. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. Journal of biomedical informatics. 2009;42(2):377–81. [PubMed: 18929686]

37. Kianersi S, Parker M, Christina L, Agley J, Rosenberg M. Introduction and validation of an ecological momentary assessment methodology to measure alcohol use among college students. in submission. 2022.

38. Higgins-Biddle JC, Babor TF. A review of the Alcohol Use Disorders Identification Test (AUDIT), AUDIT-C, and USAUDIT for screening in the United States: Past issues and future directions. The American journal of drug and alcohol abuse. 2018;44(6):578–86. [PubMed: 29723083]

39. Control CfD, Prevention. Planning and implementing screening and brief intervention for risky alcohol use: A step-by-step guide for primary care practices. Atlanta: Centers for Disease Control and Prevention. 2014.

40. Babor T, Higgins-Biddle J, Robaina K. The alcohol use disorders identification test, adapted for use in the United States: a guide for primary care practitioners. Geneva: World Health Organization. 2014.

41. SciPy developers. Numpy and Scipy Documentation 2021 [Available from: https://docs.scipy.org/doc/.

42. Hutter F, Lücke J, Schmidt-Thieme L. Beyond manual tuning of hyperparameters. KI-Künstliche Intelligenz. 2015;29(4):329–37.

43. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal of machine learning research. 2012;13(2).

44. scikit-learn. 3.1. Cross-validation: evaluating estimator performance [Available from: https://scikit-learn.org/stable/modules/cross_validation.html.

45. scikit-learn. 3.3. Metrics and scoring: quantifying the quality of predictions [Available from: https://scikit-learn.org/stable/modules/model_evaluation.html.

46. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.

47. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods. 2020;17(3):261–72. [PubMed: 32015543]

48. scikit-learn. Developing scikit-learn estimators [Available from: https://scikit-learn.org/stable/developers/develop.html.

49. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825–30.

50. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesthesia & Analgesia. 2018;126(5):1763–8. [PubMed: 29481436]

51. Rosenberg M, Ludema C, Kianersi S, Luetke M, Jozkowski K, Guerra-Reyes L, et al. Wearable alcohol monitors for alcohol use data collection among college students: feasibility and acceptability in a pilot study. medRxiv. 2021:2021.02.17.21251959.

52. Kianersi S, Ludema C, Macy JT, Colato EG, Chen C, Luetke M, et al. A Cross-Sectional Analysis of Demographic and Behavioral Risk Factors of Severe Acute Respiratory Syndrome Coronavirus 2 Seropositivity Among a Sample of US College Students. Journal of Adolescent Health. 2021.

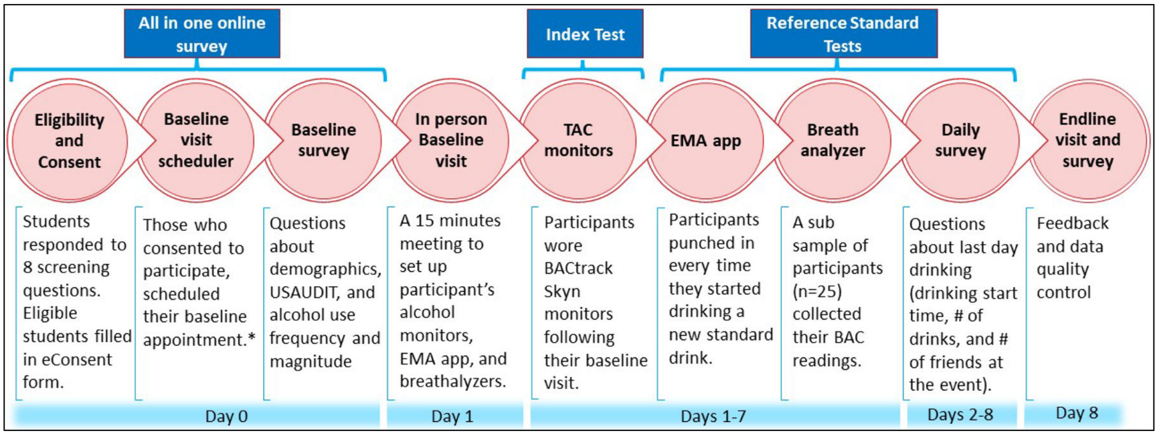
53. Wang Y, Fridberg DJ, Shortell DD, Leeman RF, Barnett NP, Cook RL, et al. Wrist-worn alcohol biosensors: Applications and usability in behavioral research. Alcohol. 2021;92:25–34. [PubMed: 33609635]

54. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. Ann Behav Med. 2018;52(6):446–62. [PubMed: 27663578]

55. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. British journal of health psychology. 2010;15(1):1–39. [PubMed: 19646331]

56. Fairbairn CE, Bosch N. A new generation of transdermal alcohol biosensing technology: practical applications, machine -learning analytics and questions for future research. Addiction. 2021;116(10):2912–20. [PubMed: 33908674]

57. Kianersi S Accuracy of Skyn Wearable Alcohol Monitors in Measuring Alcohol Consumption in Naturalistic Drinking Environments: A Network Sampling Approach: Indiana University; 2022.



Students responded to 8 screening questions. Eligible students filled in eConsent form.

Those who consented to participate, scheduled their baseline appointment.*

Questions about demographics, USAUDIT, and alcohol use frequency and magnitude

A 15 minutes meeting to set up participant's alcohol monitors, EMA app, and breathalyzers.

Participants wore BACtrack Skyn monitors following their baseline visit.

Participants punched in every time they started drinking a new standard drink.

A sub sample of participants (n=25) collected their BAC readings.

Questions about last day drinking (drinking start time, # of drinks, and # of friends at the event).

Feedback and data quality control

Figure 1.

Study procedures

EMA: Ecological Momentary Assessment

*Following the baseline visit scheduler, participants from the random cluster sample provided names and contact information of 2–3 friends.

Index test: the measurement tool under evaluation (26).

Reference standard test: "the best available method for establishing the presence or absence of the target condition" (25).

TAC monitoring, EMA app data collection, and breathalyzer readings were conducted in parallel.

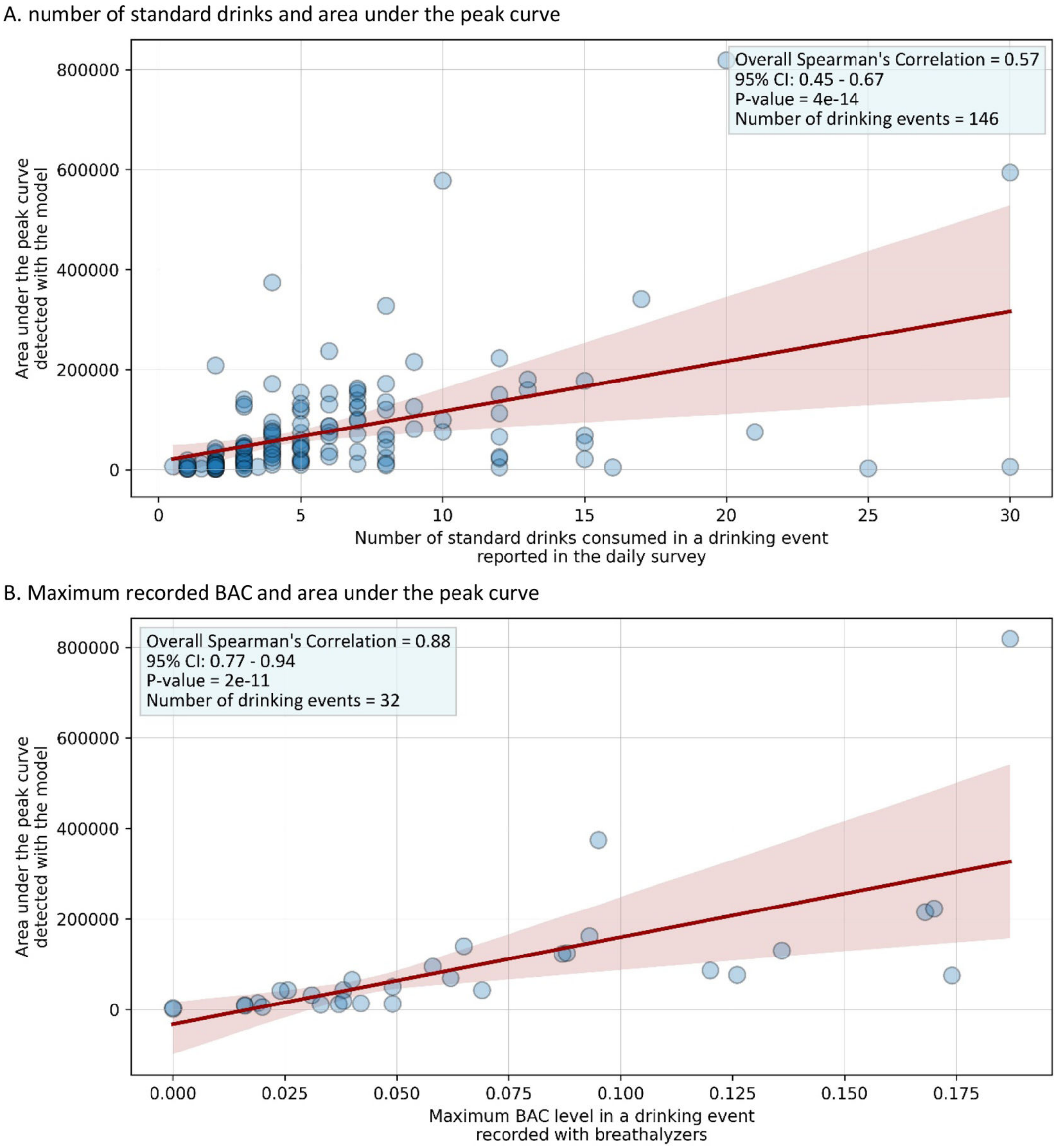


Figure 2. Correlation analysis: model compared to daily surveys and breathalyzers

-Red line shows the linear regression between the x and y axes and its 95% CI. Darker circles indicate overlapping data points.

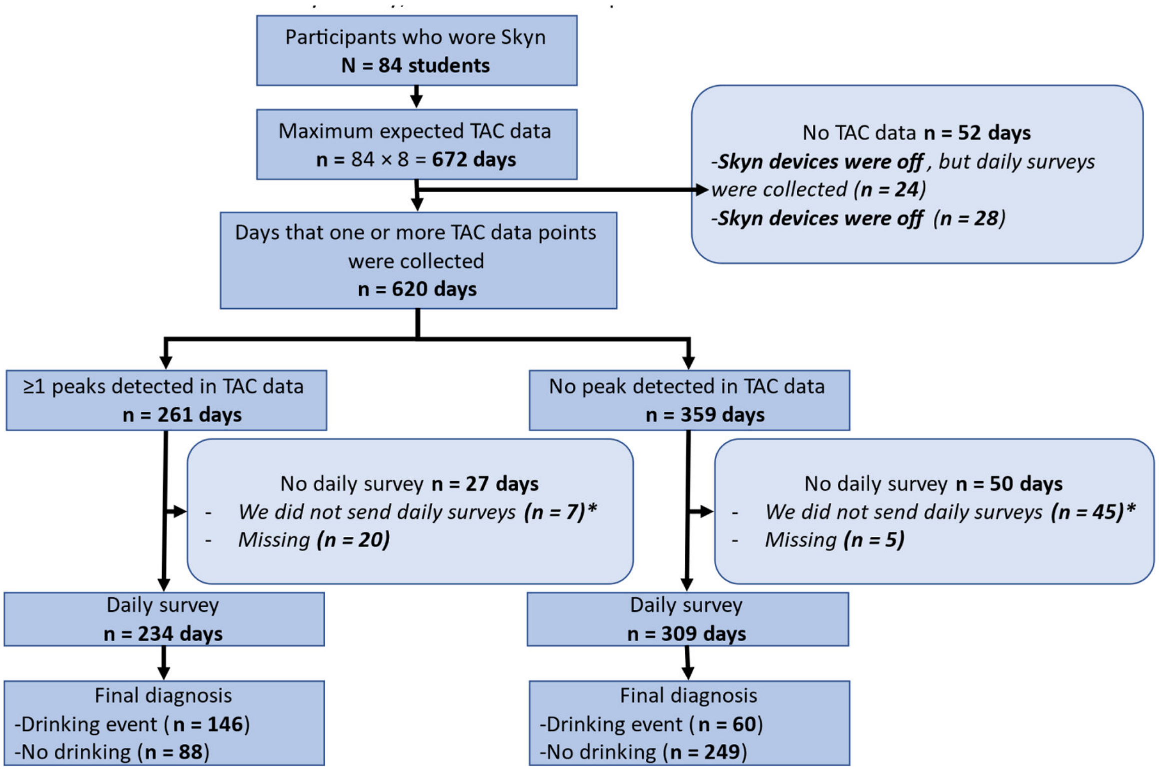

Participants who wore Skyn
N = 84 students
Maximum expected TAC data
n = 84 × 8 = 672 days
No TAC data n = 52 days
-Skyn devices were off, but daily surveys were collected (n = 24)
-Skyn devices were off (n = 28)
Days that one or more TAC data points were collected
n = 620 days
≥1 peaks detected in TAC data
n = 261 days
No peak detected in TAC data
n = 359 days
No daily survey n = 27 days
- We did not send daily surveys (n = 7)*
- Missing (n = 20)
No daily survey n = 50 days
- We did not send daily surveys (n = 45)*
- Missing (n = 5)
Daily survey
n = 234 days
Daily survey
n = 309 days
Final diagnosis
-Drinking event (n = 146)
-No drinking (n = 88)
Final diagnosis
-Drinking event (n = 60)
-No drinking (n = 249)

Figure 3. STARD flow diagram (model validation)

Reference Standard test: Daily survey, index test: developed model

* By study design, participants could have TAC data on the endline visit days. However, we did not send daily surveys for these days. Therefore, the maximum number of days that TAC data could have been collected (672 days) was larger than the maximum number of days daily surveys could have been collected (588 days).

- For daily survey, the unit of analysis was day.

Table 1.

Participants' baseline characteristics collected in an online baseline survey, spring 2021

Covariate	**Completed baseline visit** **N = 84**	**IUB Undergraduate population, Spring 2021** **N=31,364**[a]	**IU Serosurvey Study** **N = 1,267**[b]
Categorical variables: n (%)			
Age			
18 years	15 (19.2)	11,818 (37.7)	254 (22.2)
19 years	22 (28.2)		258 (22.6)
20 years	20 (25.6)	NA	258 (22.6)
21–22 years	21 (26.9)	NA	374 (32.7)
Missing	6		
Female	59 (70.2)	15,673 (50.0)	800 (63.4)
White	61 (72.6)	21,640 (69.0)	975 (77.3)
Year in school			
1st	27 (32.1)	4,824 (15.4)	286 (22.7)
2nd	27 (32.1)	7,434 (23.7)	284 (22.5)
3rd	20 (23.8)	7,455 (23.8)	306 (24.3)
4th-5th	10 (11.9)	11,377 (36.3)	384 (30.5)
Off-campus	59 (70.2)	NA	850 (67.4)
Greek members (missing = 2)	25 (30.5)	NA	303 (24.1)
Income <$25,000 (missing = 1)	73 (88.0)	NA	NA
Body mass index			
Underweight (BMI <18.5)	3 (3.6)	NA	NA
Normal weight (18.5≤BMI≤24.9)	50 (59.5)	NA	NA
Overweight (25≤BMI≤29.9)	25 (29.8)	NA	NA
Obesity (BMI≥30)	6 (7.1)	NA	NA
Continuous variables: mean (SD)			
BrAC after a typical night of drinking >0.08	51 (60.7)	NA	NA
High-risk drinking based on USAUDIT	72 (85.7)	NA	394 (47.6)[c,d]
Heavy drinking	53 (63.1)	NA	NA
No. of days in a week drinking alcohol	2.5 (1.1)	NA	2.3 (1.2)[c]
No. of drinks consumed in a drinking night	5.1 (3.1)	NA	4.1 (2.3)[c]
Total number of drinks consumed in a week	13.7 (12.9)	NA	NA
Total USAUDIT score	13.1 (5.6)	NA	8.1 (4.3)[c]

[a] Data in this column were retrieved from the following official IU website: https://uirr.iu.edu/facts-figures/enrollment/index.html

[b] Data obtained from references (52). This study used a random sample of IUB undergraduate population and was conducted in Fall 2020.

[c] Among students who reported drinking at least once a week (i.e., one of the inclusion criteria of the current dissertation study).

[d] Measured with AUDIT (AUDIT≥8 vs. AUDIT<8)

IUB: Indiana University Bloomington, NA: Not available