Using Machine Learning for Real-Time BAC Estimation from a New-Generation Transdermal Biosensor in the Laboratory[1]

Catharine E. Fairbairn[a], Dahyeon Kang[a], and Nigel Bosch[b,c]

In press at *Drug and Alcohol Dependence*

[a] Department of Psychology, University of Illinois—Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820, USA.

[b] School of Information Sciences, University of Illinois—Urbana-Champaign, 501 East Daniel Street, Champaign, IL 61820, USA.

[c] Department of Educational Psychology, University of Illinois—Urbana-Champaign, 1310 South Sixth Street, Champaign IL 61820, USA.

Correspondence concerning this article should be addressed to Catharine Fairbairn, Ph.D., Department of Psychology, 603 East Daniel St., Champaign, IL 61820. Electronic mail: cfairbai@illinois.edu.

[1] Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi:

Abstract

Background: Transdermal biosensors offer a noninvasive, low-cost technology for the assessment of alcohol consumption with broad potential applications in addiction science. Older-generation transdermal devices feature bulky designs and sparse sampling intervals, limiting potential applications for transdermal technology. Recently a new-generation of transdermal device has become available, featuring smartphone connectivity, compact designs, and rapid sampling. Here we present initial laboratory research examining the validity of a new-generation transdermal sensor prototype. **Methods**: Participants were young drinkers administered alcohol (target BAC=.08%) or no-alcohol in the laboratory. Participants wore transdermal sensors while providing repeated breathalyzer (BrAC) readings. We assessed the association between BrAC (measured BrAC for a specific time point) and eBrAC (BrAC estimated based only on transdermal readings collected in the immediately preceding time interval). Extra-Trees machine learning algorithms, incorporating transdermal time series features as predictors, were used to create eBrAC. **Results**: Failure rates for the new-generation prototype sensor were high (16%-34%). Among participants with useable new-generation sensor data, models demonstrated strong capabilities for separating drinking from non-drinking episodes, and significant (moderate) ability to differentiate BrAC levels within intoxicated participants. Differences between eBrAC and BrAC were 60% higher for models based on data from old-generation vs new-generation devices. Model comparisons indicated that both time series analysis and machine learning contributed significantly to final model accuracy. **Conclusions**: Results provide favorable preliminary evidence for the accuracy of real-time BAC estimates from a new-generation sensor. Future research featuring variable alcohol doses and real-world contexts will be required to further validate these devices.

Keywords: Alcohol, biosensor, transdermal, blood alcohol concentration, machine learning, real-time

1. Introduction

The development of a wearable biosensor for assessing alcohol consumption has long been of interest to addiction scientists (Barnett, 2015; Leffingwell et al., 2013; Swift, 2003). A variety of factors might impact an individual's ability to accurately report on their alcohol consumption including variation in drink size/strength (i.e., variable "pours"; Barnett et al., 2009; Kerr et al., 2005) and neurocognitive effects of alcohol that might impact memory for consumption (Weissenborn and Duka, 2003; White, 2003). As awareness/monitoring can contribute substantially to the maintenance of health behaviors (Miller et al., 1994; Wharton et al., 2014), the development of a wearable alcohol sensor for use by broad populations of drinkers could have key implications for both prevention and intervention.

Researchers have explored a variety of wearable technologies for assessing alcohol use, including microneedle arrays and enzymatic sensors, with several of these technologies showing promise in early studies (Vinu Mohan et al., 2017; see Wang et al., 2019). Currently, however, transdermal sensors are the technology with the largest basis of empirical support for assessing alcohol use (Fairbairn and Kang, in press). Transdermal devices assess drinking by measuring the quantity of alcohol contained in sweat and insensible perspiration, a quantity known as Transdermal Alcohol Concentration (TAC). Research indicates that the relationship between TAC and blood alcohol concentration (BAC) is a complicated one, being influenced by a variety of both contextual- and individual-level factors (Fairbairn and Kang, in press; Luczak and Ramchandani, 2019) and involving some degree of lag time (Fairbairn and Kang, 2019; Marques and McKnight, 2009). Nonetheless, transdermal technology offers advantages in that it is relatively low-cost, non-invasive, and low-maintenance (Fairbairn and Kang, in press).

To date, the most widely researched transdermal device is the Secure Continuous Remote Alcohol Monitor (SCRAM™), an older-generation transdermal sensor (e.g., Sakai et al., 2006). SCRAM devices rely on a pump to actively generate airflow across the transdermal sensor, a feature that increases SCRAM's size and limits the automated sampling interval to a relatively sparse 30 minutes (see Wang et al., 2019). Weighing approximately 6 oz and the size of a large deck of cards (Figure 1), the ankle-worn SCRAM device is relatively bulky and can lead to embarrassment for some users (Barnett et al., 2011). Thus, although well suited to their primary application as abstinence monitors among criminal-justice involved populations, the usefulness of these ankle monitors for other applications is limited (e.g., as health behavior trackers among large populations of consumers). Further, the relationship between TAC and blood alcohol concentration (BAC) can vary depending on where on the body TAC is assessed (e.g., wrist vs. ankle; Swift, 1993), and the ankle positioning of SCRAM may lead to diminished temporal sensitivity to changes in BAC (Fairbairn and Kang, 2019). Although some studies have assessed alternative transdermal devices (e.g., Phillips et al., 1995; Roizen et al., 1990; Swift, 2000; Swift et al., 1992), the majority of prior transdermal validation studies have been conducted specifically using SCRAM. Thus our current knowledge of the complexity of the TAC–BAC conversion problem is based predominantly on data produced by this device (see Fairbairn and Kang, in press; Leffingwell et al., 2013).

Recently, a new-generation of transdermal alcohol sensor has become available to researchers (NIAAA, 2015; Wang et al., 2019). These devices feature smartphone connectivity and sleek designs intended to appeal to large voluntary populations of drinkers (Wang et al., 2019). One such device is BACtrack Skyn™, a small wrist-worn sensor similar in appearance to a Fitbit (see Figure 1). Skyn is similar to SCRAM in that it assesses drinking via transdermal

means. Unlike SCRAM, however, Skyn relies on passive rather than active airflow, a feature that decreases its size and permits more rapid TAC sampling, with Skyn prototypes permitting TAC measurement as rapidly as every 20 seconds (Wang et al., 2019). Preliminary research examining raw TAC values produced by Skyn indicated sensitivity to changes in alcohol consumption (Fairbairn and Kang, 2019). To date, however, limited human subjects research has been conducted with Skyn or any other new-generation sensor, so little is known of the validity of data produced by these devices.

The current study represents what is, to our knowledge, the first examination of the validity of real-time estimates of alcohol consumption from a new-generation transdermal sensor. This represents the largest study to examine the validity of transdermal data using objective BAC assessments (see Fairbairn and Kang, 2019 for review) thus, for the first time, permitting the application of more "data-hungry" analytic approaches to transdermal sensor output (i.e., machine learning; Geurts et al., 2006). Since a variety of factors are theorized to influence the relationship between TAC and BAC, we opted to conduct this first validation effort under controlled laboratory dosing conditions (target peak BAC .08%, or no-alcohol; see Sirlanci et al., 2018, 2019). We used breathalyzer readings to validate TAC measures, a noninvasive measure with a strong and well-characterized relationship with BAC (Bendtsen et al., 1999; Jones and Andersson, 1996). A primary aim of this study was to examine the extent to which BAC estimated via Skyn might be used to distinguish instances of alcohol consumption from sober instances. In light of the potential utility of a comfortable, compact wristband for distinguishing non-abstinent moments (e.g., just-in-time adaptive interventions; Nahum-Shani et al., 2017), this aim was viewed as important not only as a first step in sensor validation but also as an important end in its own right. Additional aims of this study included examining the extent

to which models were able to identify differential BAC levels among participants consuming alcohol, as well as a comparison of BAC estimates produced by Skyn vs. SCRAM.

2. Method

2.1. Participants

The study recruited young healthy social drinkers (ages 21-30). The study population was chosen in line with guidelines for the administration of alcohol in humans (National Advisory Council on Alcohol Abuse and Alcoholism, 1989), and for the purposes of the parent study examining etiological factors in alcohol use disorder (e.g., see Fairbairn et al., 2015, 2018). A total of 110 individuals underwent experimental procedures. The final sample consisted of the 73 individuals for whom we were able to obtain Skyn readings (see section 3.2). Of these individuals, 49 were randomly assigned to the alcohol condition and 24 to the no-alcohol condition. Participants were 55% female. Sixty-four percent of participants were White, 23% Asian, 6% African-American, and 7% multiracial.

2.2. Procedure

Upon arriving at the laboratory, all participants signed informed consent. Participants were breathalyzed (Intoximeters Alco-Sensor IV) to verify a 0.00 breath alcohol concentration (BrAC). Next, Skyn devices were positioned on the inside of participants' wrists. SCRAM devices were positioned on participants' ankles. After a baseline period (1-2 hours), beverages were administered in 3 equal parts over 36 minutes. Participants assigned to receive alcohol received a dose intended to bring them up to the legal driving limit (target peak BAC=.08%), with the exact dose adjusted for participants' approximate body water (see Curtin and Fairchild, 2003 for formulas). Participants in the no-alcohol condition were administered a non-alcoholic beverage.

Following beverage administration, participants in the alcohol condition provided breathalyzer readings at approximately 30-minute intervals until they left the lab. Participants in the no-alcohol condition were breathalyzed upon arriving in the lab and then again immediately post-drink. No-alcohol participants were allowed to leave after study tasks were completed (5-6 hour sessions). Alcohol participants were required to remain until BrACs dropped below .025% and also SCRAM output registered at least one descending value (6-9 hour sessions).[2]

2.3. Data Analysis Plan

Analyses were conducted to predict BrAC values (serving as "ground truth") from data derived from Skyn. Our approach leveraged the relatively high-frequency TAC readings produced by Skyn (~1 minute interval, although in units that do not correspond to BAC). In particular, we estimated BrAC for a precise time point using TAC time series features (e.g., mean, trends, periodicity) extracted from Skyn during the immediately preceding 30-minute time interval. Time series features were then entered into machine learning algorithms to produce BrAC estimates in "real time." Note that all of the models presented here were constructed such that they could be run rapidly (within 1-2 seconds) using the computing power of the average smartphone. Details of data pre-processing are provided in supplemental materials.[3] Figure 2 provides a visual depiction of the complete data analysis plan.

2.3.1. Time Series Feature Extraction

We extracted features from 30 minutes of TAC data leading up to each BrAC reading, thus forming a set of 1,092 instances (input/output pairs) for a machine learning model. The 30-

[2] Given the relatively substantial dose of alcohol administered in the current study, and the time required for alcohol metabolism, it was not feasible to keep participants in the lab to 0.00% BrAC. However, using the current procedures, we were able to capture the majority of the descending BAC limb for all participants.

[3] Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi:

minute window was selected as equivalent to the interval separating BrAC readings for participants assigned to the alcohol condition in the current study. To create instances for the no-alcohol condition, we inserted synthetic (artificial) 0.00% BrAC readings (see supplemental materials).[4] In cases where less than 30 minutes had elapsed since the prior instance, 30-minute intervals were allowed to overlap. The model was constrained such that no predictions were made until at least 30 minutes of TAC data had accrued. We used TSFRESH (Christ et al., 2018), a Python software package, to extract time series features from TAC data (see Table 1). Importantly, to produce a model that might be applied for real-time BrAC estimation (not just retrospective prediction), we only included TAC time series *preceding* (not following) BrAC readings.

2.3.2. Machine Learning Methods

The machine learning model type employed was Extra-Trees (Geurts et al., 2006), which is a tree-based ensemble regression algorithm similar to random forests (Breiman, 2001). Extra-Trees is particularly useful for cases when there may be non-linear relationships between features and output variables, many features, and too few instances to leverage deep neural network methods for big data. No other variables were entered into machine learning models, so that models estimated BAC based on TAC data (TSFRESH features) alone.

We employed 4-fold, participant-independent cross-validation to ensure that predictions were not over-fit to specific data points or participants. To do so, we randomly divided participants into four groups, trained a model using data from three of those groups (the training set), tested it on the fourth group (the testing set), and repeated the process three more times so

[4] Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi:

that each participant was in the testing set once. We selected Extra-Trees hyperparameters based on training data only, via nested 4-fold cross-validation.

2.3.3. Model Evaluation

We evaluated model results with mean absolute error (*MAE*; i.e., *L1* distance) and, as an additional metric, root mean squared error (*RMSE*; i.e., *L2* distance). *MAE* represents the average absolute difference between actual BrAC values and estimates of BrAC from transdermal data (eBrAC). We calculated *MAE* and *RMSE* per-participant and report the mean of these participant-level measures, calculating 95% confidence intervals for the means via bootstrapping with 10,000 iterations, thus accounting for the non-normal distributions of *MAE* and *RMSE* (Efron, 1987). In addition to *MAE* and *RMSE*, we also present the correlation (Pearson's *r*) between BrAC and eBrAC across all observations, provided as a standardized effect size corresponding to those presented in prior transdermal publications (Davidson et al., 1997; Sakai et al., 2006). Correlations are supplemented with mixed models, which assess the association between eBrAC, entered as the predictor, and BrAC, entered as the outcome, while accounting for participant-level clustering via random effects estimation (Raudenbush and Bryk, 2002).

3. Results

3.1. BrAC Descriptives

An average of 10 BrAC readings were collected from alcohol participants after beverage administration. Average maximum BrAC was .084% (*SD*=.011), and average (post-baseline) minimum was .026% (*SD*=.014). Of post-baseline alcohol condition BrAC values, 14.0% were <.03%, 23.4% were between .03%-.05%, 30.3% were between .05%-.07%, 25.5% were between .07-.09, and 6.9% were ≥.09%.

3.2. Device Failures

Skyn devices used in this research were early, hand-assembled prototypes. In total, this research produced 37 missing Skyn files—18 due solely to device malfunction, and 19 that involved a combination of device and user issues. Of the 18 failures attributable solely to the devices themselves, 9 files were completely blank for unknown reasons, 3 consisted of an entirely flat line with no oscillation, and 6 were blank or severely truncated due to battery failure. An additional 19 Skyn files were lost as our team learned to work with these delicate prototypes. All participants for whom we had useable Skyn data were included in our final sample of participants.

Of the final sample of 73 participants for whom we had Skyn files, 66 of these individuals also had useable SCRAM files. Six SCRAM files were missing due to procedural issues associated with SCRAM assignment, and one due to device malfunction.

3.3. Model Evaluation

Across all participants and both alcohol and no-alcohol conditions, the average difference between actual BrAC and eBrAC (i.e., *MAE*) was .010 [0.008, 0.012]. Model accuracy tended to be higher in the no-alcohol vs. the alcohol condition (see Table 2) —effects that are likely partially attributable to the bounded nature of BrAC and resulting floor effects at lower BrACs. When subdivided according to different BrAC levels, the distance between BrAC and eBrAC was smaller at lower BrAC values, and increased as BrAC increased—for BrACs .00%-.03%, *MAE*=.009, 95% CI [0.007, 0.011]; for BrACs .03%-.06%, *MAE*=.011, 95% CI [0.010, 0.012]; for BrACs over .06%, *MAE*=.015, 95% CI [0.012, 0.017].

The model demonstrated strong capabilities for distinguishing episodes of drinking from non-drinking. Among participants assigned to the no-alcohol condition, rates of false positives were low, with only 1.8% of eBrAC values falling above .02%. Thus, in more than 98% of cases,

sober individuals were correctly identified as having consumed less than the equivalent of one alcoholic beverage (see Watson et al., 1981). The model also demonstrated strong ability to correctly detect episodes of alcohol consumption, with 98.5% of post-baseline alcohol condition eBrAC values falling above .02%.

MAE did not differ significantly as a function of participant gender, age, race, or drinking patterns. Minor discrepancies emerged across different Skyn prototype devices (see Table 3 for full results). Graphs for "best", "worst", and "average" prediction cases appear in Figure 3.

3.4. *Model Comparison*

Next, we evaluated the incremental utility of the specific model employed (referred to here as the "full model") beyond more parsimonious models. As noted previously, the full model involved two key elements: the extraction of time series features from TAC data and the implementation of machine learning algorithms that included these as predictors. But to what extent do these elements drive the accuracy of the models—i.e., does the model need to be so complex, or would a simpler model suffice? To address this question, we constructed two additional models: 1) A linear regression model including a single TAC value (TAC-reading taken immediately preceding BrAC reading) as a predictor—providing a basic point of comparison involving neither machine learning nor time series analysis; 2) The same Extra-Trees machine learning model used above, but this time including only the immediately preceding TAC value as a predictor—allowing us to assess the incremental utility of our time series feature-extraction beyond machine learning alone. All models were trained and tested using the same 4-fold participant-level cross-validation procedures. Complete model results are presented in Table 2. The basic linear regression approach produced a *MAE* (i.e., error) that was more than double that of the full model. The model employing machine learning methods—but no time

series feature extraction—produced a *MAE* that was 27% lower (better) than the basic linear regression model but still 60% higher than our full model. Confidence intervals for *MAE* were overlapping for none of these models. In sum, the model integrating both time series feature extraction and machine learning methodology outperformed other methods by a substantial margin. Shapley feature importance values (Lundberg and Lee, 2017) for the "full" model are presented in supplemental materials.[5]

3.5. Predicting Differential BrAC Levels

Based on analyses presented to this point, it is unclear the extent to which the model's accuracy is explained solely by its ability to differentiate intoxicated from non-intoxicated participants. To explore this further, we ran analyses examining only those readings that were collected post-baseline within the alcohol condition. Despite the limited BrAC range in this subsample (Average min=.026%; Average max=.084%; see section 3.1), we nonetheless found a significant correlation between eBrAC and BrAC that was moderate in magnitude, r=.495, 95% CI [0.424, 0.559]. Next, to evaluate the model's ability to accurately predict *MAE* among intoxicated participants, we randomly shuffled post-baseline eBrAC values within participants in the alcohol condition. Specifically, for observations taken on alcohol sessions post-baseline, we compared the *MAE* for our final model in which eBrAC values were correctly matched with BrAC values with a model in which these eBrAC values were randomly shuffled within participants, permitting us to examine the extent to which the *MAE* for our final model outperformed what might be expected based on random chance. Again, despite the restricted BrAC range in this study subsample and thus the relatively low ceiling for *MAE* values, the *MAE* for the randomly shuffled model, *MAE*=0.021, 95% CI [0.020, 0.023], was substantially higher

[5] Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi:

than the non-randomly shuffled model, *MAE*=0.015, 95% CI [0.013, 0.016]. Confidence intervals for these values did not overlap. Taken together, these results suggest that, despite a restricted BrAC range post-baseline among alcohol participants in the current study, the accuracy of our final model was not driven solely by the model's ability to differentiate drinking episodes from non-drinking episodes.

3.6. Session Phase Learning Analyses

For the purposes of analysis, each 30-minute interval was treated as a discrete analytic unit (instance), independent of any information that might contextualize it within the broader arc of the session for a given participant. Nonetheless, although the shape of BrAC curves did vary somewhat across alcohol participants, this variation was not large (see descriptive statistics). Thus, given that BrAC curves among alcohol participants in our study were relatively consistent in their shape, one possibility is that our machine learning model was simply learning to recognize specific patterns of TAC values characteristic of session epochs produced by our specific dosing paradigm and outputting predictions on the basis of these epochs. Some amount of such epoch learning is likely unavoidable and, in fact, a more sophisticated version of such learning might have direct practical applications for predicting BAC values in the real-world, given a volume of data sufficient for characterizing most types of drinking patterns. Nonetheless, several data points led us to believe that such epoch learning was unlikely to be a primary factor driving effects in the current study: 1) We constructed a machine learning model including time series TAC features as input and % Drinking Episode Elapsed as output—in other words, we constructed the same final model used above for predicting BrAC, but used the model to predict session epoch (% Drinking Episode Elapsed) instead. Results indicated that session epoch is difficult to predict directly from TAC input (*MAE*=24% of the session). 2) Visual inspection of

graphs indicated that eBrAC values produced by our models appear to respond to distinctive characteristics of raw TAC data vs. simply predicting the characteristic shape of lab-based drinking episodes (e.g., see Figure 3, alcohol condition "Maximum *MAE*"). Taken together, these results suggest that session epoch alone is an unlikely explanation for the success of models in predicting BrAC.

3.7. SCRAM Analyses

Finally, we constructed machine learning models using data from the SCRAM ankle bracelet to predict BrAC. Ninety percent of participants in this study had useable SCRAM files (N=66; see section 3.2). Note that the SCRAM device, which relies on a pump to generate airflow across the sensor, cannot be programmed to produce automated readings at a rate faster than 30 minutes due to issues of battery life (Rosales, 2020). Thus, in light of this fixed sampling interval, only the machine learning and not the time series portion of our model was applied. Specifically, we applied Extra-Trees machine learning to readings taken from SCRAM devices, incorporating as inputs the closest SCRAM reading preceding a BrAC reading. Results indicated strong accuracy for SCRAM in the no-alcohol condition (Table 2), unsurprising given the precision-level of raw SCRAM output. However, across all conditions, the error for the SCRAM model was over 60% higher vs. the error for Skyn (see Table 2).

4. Discussion

The present laboratory study employed high-frequency TAC readings from a new-generation transdermal alcohol sensor, leveraging a combination of time series analysis and machine learning to produce real-time BAC estimates. We have previously published research examining a subset of individuals from this same dataset (N=30; Fairbairn and Kang, 2019). While this prior research examined correlations between BAC and *raw* TAC values, the current

study uses machine learning to transform these values into estimates of BAC, thus for the first time permitting the examination of the accuracy of transdermal BAC estimates from a new-generation sensor. Among participants with useable data, the model demonstrated strong capabilities for distinguishing episodes of drinking from non-drinking, and also indicated significant (moderate) ability to differentiate BrAC levels specifically among intoxicated participants. Correlations between BrAC and eBrAC were high, and the average absolute difference between BrAC and eBrAC was .010. Estimates were particularly precise at lower BrAC levels, with the precision of these estimates tending to decrease as BrAC increased. Results indicated that both time series feature-extraction and machine learning contributed to model accuracy, with the "full" model incorporating both of these elements outperforming more parsimonious models. Finally, analyses indicated that models built with data from the new-generation Skyn sensor outperformed similar models built with data from the older-generation SCRAM.

Strengths and limitations of this research should be noted. Scientists have long been interested in algorithms that might permit the translation of TAC output from transdermal sensors into estimates of alcohol consumption (Dougherty et al., 2012; Luczak and Rosen, 2014; Sirlanci et al., 2019; Swift and Swette, 1992). The current research is the first to test the validity of BrAC estimates from a new-generation transdermal sensor. This study is also the first, to our knowledge, to test the validity of transdermal BrAC estimates produced in "real-time"—created for a precise time point based only on TAC readings collected in the immediately preceding time interval. Regarding limitations, the current study employed fixed alcohol dosing procedures, resulting in relatively uniform BAC curves in the alcohol condition, and was further conducted in a controlled laboratory context. Research examining variable alcohol doses and real-world

drinking contexts is required to validate TAC-BAC conversion algorithms. Further, although procedures employed in this study did capture the majority of BAC and TAC curves, participants nonetheless left the laboratory before their TACs reached zero. Finally, although the current study does provide evidence that the new-generation Skyn device produces more accurate BAC estimates than the older-generation SCRAM, data produced by this study are incapable of providing a firm answer surrounding the reason for this differential accuracy. Of note, however, the accuracy of the SCRAM model is similar to that of the Skyn model omitting the time series component, indicating Skyn's higher sampling rate as a potential factor driving final model accuracy.

Regarding the Skyn devices themselves, although these data indicate promise, device development will be required before they are suitable for most applications. This research employed early hand-assembled Skyn prototypes, and failure rates ranged from 16% (device issues alone) to 34% (device/user issues). Data used in analyses were ideal in that they excluded those with device error. Further, the specific Skyn prototype we examined produced output only in terms of raw electrical current detected at the transdermal sensor, thus necessitating data standardization prior to analysis (see also Fairbairn and Kang, 2019; issue addressed with new Skyn prototypes). Finally, Skyn devices have limited water resistance, a feature that is suboptimal for real-world test conditions. More robust prototypes would be important in facilitating large-scale field testing. Thus, at the current time, SCRAM remains the most reliable transdermal alcohol sensor.

The task of predicting BAC from transdermal sensor data across contexts certainly represents a challenge, and the success of such an undertaking is currently unclear. Importantly, however, several additional tools are available to researchers that might aid in this endeavor.

First, in the current project, we used a combination of time series feature extraction and tree-based machine learning analyses. The approaches employed here are well-suited to the current dataset, which involved relatively few instances. With larger datasets, additional modeling frameworks become available. For example, individual data points from all preceding 30 minutes might be directly entered into models, bypassing time series feature extraction entirely and allowing for additional model complexity/flexibility. Beyond tree-based approaches, deep neural network methods have shown great promise for learning even more complex associations between low-level variables, given very large datasets (LeCun et al., 2015). Furthermore, new-generation transdermal devices incorporate sensors beyond those assessing TAC (e.g., skin temperature, accelerometer). Within the context of larger datasets, data from these additional sensors might be used to refine the precision of estimates across diverse environments.

In sum, transdermal sensors offer a passive, noninvasive method for assessing alcohol use likely to be attractive to broad populations of drinkers. Results of the current study provide evidence for the validity of data produced by a new-generation of transdermal sensor, and further indicate that real-time transdermal estimation of BAC is possible under specific conditions. Future research employing varying doses and contexts is needed to further clarify the place of transdermal sensors in our arsenal of techniques for assessing, preventing, and treating problem drinking.

References

Barnett, N.P., 2015. Alcohol sensors and their potential for improving clinical care. Addiction 110, 1–3.

Barnett, N.P., Tidey, J., Murphy, J.G., Swift, R., Colby, S.M., 2011. Contingency management for alcohol use reduction: A pilot study using a transdermal alcohol sensor. Drug Alcohol Depend. 118, 391–399.

Barnett, N.P., Wei, J., Czachowski, C., 2009. Measured alcohol content in college party mixed drinks. Psychol. Addict. Behav. 23, 152–156.

Bendtsen, P., Hultberg, J., Carlsson, M., Jones, A.W., 1999. Monitoring ethanol exposure in a clinical setting by analysis of blood, breath, saliva, and urine. Alcohol. Clin. Exp. Res. 23, 1446–1451.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Christ, M., Braun, N., Neuffer, J., Kempa-Liehr, A.W., 2018. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). Neurocomputing 307, 72–77.

Curtin, J.J., Fairchild, B.A., 2003. Alcohol and cognitive control: Implications for regulation of behavior during response conflict. J. Abnorm. Psychol. 112, 424–436.

Davidson, D., Camara, P., Swift, R., 1997. Behavioral effects and pharmacokinetics of low-dose intravenous alcohol in humans. Alcohol. Clin. Exp. Res. 21, 1294–1299.

Dougherty, D.M., Charles, N.E., Acheson, A., John, S., Furr, R.M., Hill-Kapturczak, N., 2012. Comparing the detection of transdermal and breath alcohol concentrations during periods of alcohol consumption ranging from moderate drinking to binge drinking. Exp. Clin. Psychopharmacol. 20, 373–381.

Efron, B., 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. 82, 171–185.

Fairbairn, C.E., Bresin, K.W., Kang, D., Rosen, I.G., Ariss, T., Luczak, S.E., Barnett, N.P., Eckland, N.S., 2018. A multimodal investigation of contextual effects on alcohol's emotional rewards. J. Abnorm. Psychol. 127, 359–373. https://doi.org/10.1037/abn0000346

Fairbairn, C.E., Kang, D., 2019. Temporal dynamics of transdermal alcohol concentration measured via new-generation wrist-worn biosensor. Alcohol. Clin. Exp. Res. 43, 2060–2069. https://doi.org/10.1111/acer.14172

Fairbairn, C.E., Kang, D., in press. Transdermal alcohol monitors: Research, applications, and future directions, in: Frings, D., Albery, I. (Eds.), The Handbook of Alcohol Use and Abuse. Elsevier.

Fairbairn, C.E., Sayette, M.A., Wright, A.G.C., Levine, J.M., Cohn, J.F., Creswell, K.G., 2015. Extraversion and the rewarding effects of alcohol in a social context. J. Abnorm. Psychol. 124, 660–673. https://doi.org/10.1037/abn0000024

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.

Jones, A.W., Andersson, L., 1996. Variability of the blood/breath alcohol ratio in drinking drivers. J. Forensic Sci. 41, 916–921.

Kerr, W.C., Greenfield, T.K., Tujague, J., Brown, S.E., 2005. A drink is a drink? Variation in the amount of alcohol contained in beer, wine and spirits drinks in a US methodological sample. Alcohol. Clin. Exp. Res. 29, 2015–2021.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436–444.

Leffingwell, T.R., Cooney, N.J., Murphy, J.G., Luczak, S., Rosen, G., Dougherty, D.M., Barnett, N.P., 2013. Continuous objective monitoring of alcohol use: Twenty-first century

measurement using transdermal sensors. Alcohol. Clin. Exp. Res. 37, 16–22. https://doi.org/10.1111/j.1530-0277.2012.01869.x

Luczak, S.E., Ramchandani, V.A., 2019. Special issue on alcohol biosensors: Development, use, and state of the field. Alcohol 81, 161–165.

Luczak, S.E., Rosen, I.G., 2014. Estimating BrAC from transdermal alcohol concentration data using the BrAC estimator software program. Alcohol. Clin. Exp. Res. 38, 2243–2252. https://doi.org/10.1111/acer.12478

Lundberg, S.M., Lee, S.L., 2017. A unified approach to interpreting model predictions, in: Guyon, L., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. pp. 4765–4774.

Marques, P.R., McKnight, A.S., 2009. Field and laboratory alcohol detection with 2 types of transdermal devices. Alcohol. Clin. Exp. Res. 33, 703–711.

Miller, W.R., Zweben, A., DiClimente, C., Rychtarik, R., 1994. Motivational enhancement therapy manual: A clinical research guide for therapists treating individuals with alcohol abuse and dependence. NIAAA, Bethesda MD.

Nahum-Shani, I., Smith, S.N., Spring, B.J., Collins, L.M., Witkiewitz, K., Tewari, A., Murphy, S.A., 2017. Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. Ann. Behav. Med. 52, 446–462.

National Advisory Council on Alcohol Abuse and Alcoholism, 1989. Recommended Council Guidelines on Ethyl Alcohol Administration on Human Experimentation.

NIAAA, 2015. Wearable alcohol biosensor challenge [WWW Document]. URL https://www.nih.gov/news-events/news-releases/niaaa-selects-winners-its-wearable-alcohol-biosensor-challenge

Phillips, M., Greenberg, J., Andrzejewski, J., 1995. Evaluation of the Alcopatch, a transdermal dosimeter for monitoring alcohol consumption. Alcohol. Clin. Exp. Res. 19, 1547–1549.

Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical linear models: Applications and data analysis methods. Sage Publications, Thousand Oaks, CA.

Roizen, M.F., Lichtor, L., Lane, B., 1990. A “Band-Aid” to detect alcohol levels in the blood. Anesthesiology 73, A510.

Rosales, C., 2020. SCRAM Sampling Interval.

Sakai, J.T., Mikulich-Gilbertson, S.K., Long, R.J., Crowley, T.J., 2006. Validity of transdermal alcohol monitoring: Fixed and self-regulated dosing. Alcohol. Clin. Exp. Res. 30, 26–33.

Sirlanci, M., Luczak, S.E., Fairbairn, C.E., Kang, D., Pan, R., Yu, X., Rosen, I.G., 2019. Estimating the distribution of random parameters in a diffusion equation forward model for a transdermal alcohol biosensor. Automatica 106, 101–109.

Sirlanci, M., Rosen, I.G., Luczak, S.E., Fairbairn, C.E., Bresin, K., Kang, D., 2018. Deconvolving the input to random abstract parabolic systems: A population model-based approach to estimating blood/breath alcohol concentration from transdermal alcohol biosensor data. Inverse Probl. 34, 125006.

Swift, R.M., 2003. Direct measurement of alcohol and its metabolites. Addiction 98, 73–80.

Swift, R.M., 2000. Transdermal alcohol measurement for estimation of blood alcohol concentration. Alcohol. Clin. Exp. Res. 24, 422–423.

Swift, R.M., 1993. Transdermal measurement of alcohol consumption. Addiction 88, 1037–1039.

Swift, R.M., Martin, C.S., Swette, L., Laconti, A., Kackley, N., 1992. Studies on a wearable, electronic, transdermal alcohol sensor. Alcohol. Clin. Exp. Res. 16, 721–725.

Swift, R.M., Swette, L., 1992. Assessment of ethanol consumption with a wearable, electronic ethanol sensor/recorder, in: Litten, R.Z., Allen, J.P. (Eds.), Measuring Alcohol Consumption: Psychosocial and Biochemical Methods. Humana Press, Totowa, NJ, pp. 189–202.

Vinu Mohan, A.M., Windmiller, J.R., Mishra, R.K., Wang, J., 2017. Continuous minimally-invasive alcohol monitoring using microneedle sensor arrays. Biosens. Bioelectron. 91, 574–579. https://doi.org/10.1016/j.bios.2017.01.016

Wang, Y., Fridberg, D.J., Leeman, R.F., Cook, R.L., Porges, E.C., 2019. Wrist-worn alcohol biosensors: Strengths, limitations, and future directions. Alcohol Biomed. J. 81, 83–92.

Watson, P.E., Watson, I.D., Batt, R.D., 1981. Prediction of blood alcohol concentrations in human subjects. Updating the Widmark Equation. J. Stud. Alcohol 42, 547–556.

Weissenborn, R., Duka, T.A., 2003. Acute alcohol effects on cognitive function in social drinkers: their relationship to drinking habits. Psychopharmacology (Berl.) 165, 306–312.

Wharton, C.M., Johnston, C.S., Cunningham, B.K., Sterner, D., 2014. Dietary self-monitoring, but not dietary quality, improves with use of smartphone app technology in an 8-week weight loss trial. J. Nutr. Educ. Behav. 46, 440–444.

White, A.M., 2003. What happened? Alcohol, memory blackouts, and the brain. Alcohol Res. Health 27, 186–196.

Figure Legend

Figure 1. AMS SCRAM ankle bracelet (left) and BACtrack Skyn wrist monitor (right) displayed side-by side. The approximate weights of the devices are 6oz (SCRAM) and 1oz (Skyn prototype).

Figure 2. A visual representation of the data analysis plan employed in the current project. Data analysis involved the extraction of multiple time series features (e.g., mean, trends, periodicity) from the 30 minutes of raw TAC data that preceded each breathalyzer (BrAC) reading. These time series features were then entered as predictors into Extra-Trees machine learning algorithms to create estimates of BrAC from transdermal data (eBrAC). The top panel provides a broad visual depiction of the entire analysis process, the bottom left panel provides examples of a subset of time series features extracted (see Table 1 for additional features), and the bottom right panel provides a flow chart of machine learning modeling procedures.

Figure 3. Graphs for participants with the "best" (minimum *MAE*), "worst" (maximum *MAE*), and average (Median *MAE*) prediction accuracy from both alcohol and no-alcohol (control) conditions in the current study. Precise average *MAE*s for alcohol condition graphs shown above are as follows: best case *MAE*=0.006; worst case *MAE*=0.028; median case *MAE*= 0.013. Precise average *MAE*s for the no-alcohol (control) condition graphs are as follows: best case *MAE*=0.000; worst case *MAE*=0.011; median case *MAE*=0.001. Baseline standardization procedures were applied to all Skyn data, as described in the Data Analysis Plan. For the purposes of graphs displayed here, data from Skyn was transformed (divided by 20,000) such that it could be visualized on approximately the same scale as eBrAC and BrAC.

Table 1. A partial list of time series features extracted by the software package TSFRESH, consisting of the most important features in the main model. For a complete list of all features extracted by this package, together with more detailed feature explanations, see below link:
https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html

Feature Name	**Feature Description**
agg_linear_trend(x, f_agg, chunk_len, attr)	Aggregate timeseries *x* into chunks of length *chunk_len* using function *f_agg*, fit a line to the result, and return *attr* (the slope, p-value, or other attribute of the fitted linear model)
change_quantiles(x, ql, qh, isabs, f_agg)	Aggregate consecutive (absolute) differences in x using f_agg for values that fall between quantiles ql and qh
fft_coefficient(x, coeff, attr)	Attributes of a specific coefficient from the Fourier transform of x (frequencies represented in x)
linear_trend(x, attr)	Fit a line to x and return an attribute of the fitted linear model
maximum(x)	Maximum value in x
mean_abs_change(x)	Mean of absolute differences between consecutive values in x
percentage_of_reoccurring_datapoints_to_all_datapoints(x)	Proportion of unique data points in x that occur more than once
percentage_of_reoccurring_values_to_all_values(x)	Proportion of values in x that occur more than once
quantile(x, q)	Value of x at quantile q
ratio_value_number_to_time_series_length(x)	Proportion of data points in x that are unique
standard_deviation(x)	Standard deviation of x
sum_of_reoccurring_data_points(x)	Sum of non-unique data points in x
sum_of_reoccurring_values(x)	Sum of non-unique values in x (non-unique data points with duplicates removed)

Table 2. Comparison of models for translating transdermal data into estimates of BrAC for Skyn (Models 1-3) and SCRAM (Model 4)

	Model SM. (*SCRAM Model*) Machine Learning applied to SCRAM readings	**Model 1.** (*Comparison Model*) Linear Regression without Time Series Features	**Model 2.** (*Comparison Model*) Machine Learning without Time Series Features	**Model 3.** (*Full Model*) Machine Learning with Time Series Features
MAE [95% CI]	.018 [.016, .020]	.022 [.020, .024]	.016 [.014, .018]	.010 [.008, .012]
RMSE [95% CI]	.018 [.016, .020]	.025 [.022, .028]	.018 [.016, .021]	.013 [.011, .015]
r [95% CI]	.021 [.019, .024]	.637 [.601, .671]	.776 [.751, .799]	.907 [.896, .917]
All Conditions				
% within .01 of BrAC	62.1%	12.7%	61.5%	70.8%
% within .02 of BrAC	76.2%	73.6%	74.9%	86.4%
% within .03 of BrAC	87.2%	83.6%	87.3%	94.5%
Alcohol Condition				
% within .01 of BrAC	32.1%	24.4%	30.7%	44.1%
% within .02 of BrAC	57.3%	48.9%	57.0%	73.3%
% within .03 of BrAC	77.1%	65.6%	80.0%	89.6%
No-Alcohol Condition				
% within .01 of BrAC	99.8%	2.1%	89.7%	95.1%
% within .02 of BrAC	99.8%	96.1%	91.2%	98.2%
% within .03 of BrAC	99.8%	100.0%	93.9%	98.9%

MAE, *RMSE*, and *r* values are presented for data aggregated across all conditions (i.e., alcohol and no-alcohol). 95% confidence intervals are presented within brackets for *MAE*, *RMSE*, and *r* values above.

Model SM employed Extra-Trees machine learning to SCRAM readings, incorporating the closest SCRAM reading preceding a BrAC reading as a predictor (due to the sparse sampling interval of SCRAM, calculating TAC time series features was not an option). Model 1 employed linear regression including a single Skyn TAC value (TAC-reading taken immediately preceding BrAC reading) as a predictor. Model 2 employed Extra-Trees machine learning incorporating only the immediately preceding Skyn TAC value as a predictor. Model 3—the "full" (final) Skyn model—incorporated Extra-Trees machine learning with Skyn TAC time series features as

predictors. All models were trained and tested using 4-fold participant-level cross-validation. N=67 SCRAM model, N=73 Skyn models.

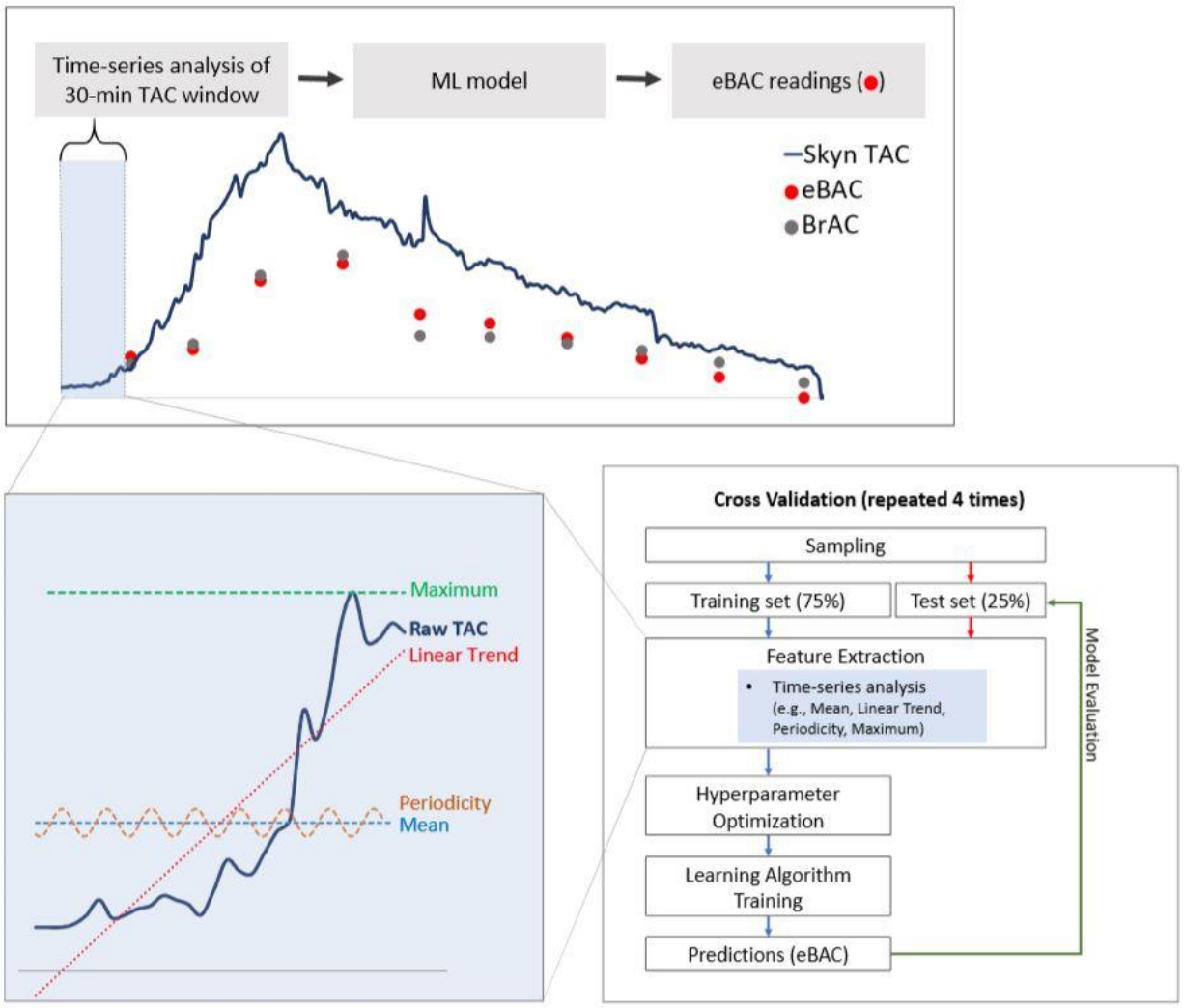
MAE=Average absolute distance between measured BrAC and eBrAC, calculated per-participant and then averaged across participants; *RMSE*=Root mean squared error, also calculated per-participant and averaged across participants. "*r*" refers to the Pearson correlation between eBrAC and BrAC. Mixed models accounted for participant-level clustering of BrAC values. "% within XX of BrAC"=percentage of eBrAC values that fall within XX of measured BrAC (or, put differently, % *MAE* values <XX)

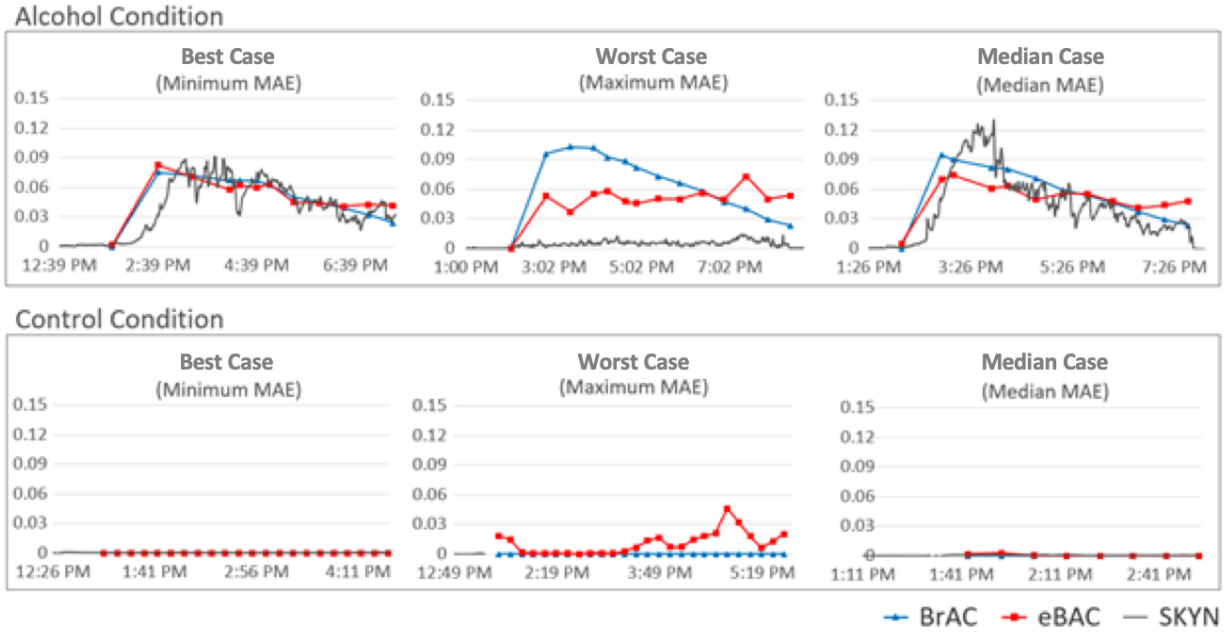
Table 3. *MAE* as a function of participant and device characteristics.

	b	***SE***	***t***	***p***
Gender	-0.001	0.001	-0.72	0.473
Age	0.000	0.000	0.91	0.366
Days Drink/30	0.000	0.000	-1.49	0.142
Race				
African American	0.002	0.003	0.73	0.467
Asian	0.000	0.001	0.25	0.807
Multiracial	-0.001	0.001	-1.14	0.260
Skyn Device ID				
0BB4	-0.003	0.001	-3.09	0.003
0DB5	-0.003	0.001	-2.10	0.040
18	-0.001	0.002	-0.86	0.394
7AB3	-0.004	0.002	-2.44	0.017
9	0.000	0.002	0.13	0.900

The above represent coefficients derived from multilevel models predicting *MAE* (average absolute distance between measured BrAC and eBrAC) while accounting for clustering of observations within participants. All variables were entered into separate models. All models control for beverage condition assignment. Gender was coded such that Female=1 and Male=0. "Days Drink/30"=number of days reported drinking at baseline out of past 30; Race was coded as a set of dummy codes, with "White" as the reference group; Skyn Device ID was coded as a set of dummy codes, with device B6B3 as the reference. Of the Skyn devices used here, 0BB4, 7AB3, B6B3, and 0DB5 represent older (2016) Skyn prototypes, whereas devices 18 and 9 represent newer (2018) prototypes.


SCRAM
CONTINUOUS ALCOHOL
MONITORING


Time-series analysis of 30-min TAC window
ML model
eBAC readings (●)
Skyn TAC
eBAC
BrAC
Maximum
Raw TAC
Linear Trend
Periodicity
Mean
Cross Validation (repeated 4 times)
Sampling
Training set (75%)
Test set (25%)
Feature Extraction
Time-series analysis (e.g., Mean, Linear Trend, Periodicity, Maximum)
Hyperparameter Optimization
Learning Algorithm Training
Predictions (eBAC)
Model Evaluation


Alcohol Condition
Best Case
(Minimum MAE)
Worst Case
(Maximum MAE)
Median Case
(Median MAE)
0.15
0.12
0.09
0.06
0.03
0
12:39 PM
2:39 PM
4:39 PM
6:39 PM
1:00 PM
3:02 PM
5:02 PM
7:02 PM
1:26 PM
3:26 PM
5:26 PM
7:26 PM
Control Condition
Best Case
(Minimum MAE)
Worst Case
(Maximum MAE)
Median Case
(Median MAE)
12:26 PM
1:41 PM
2:56 PM
4:11 PM
12:49 PM
2:19 PM
3:49 PM
5:19 PM
1:11 PM
1:41 PM
2:11 PM
2:41 PM
BrAC
eBAC
SKYN

Fairbairn, Kang, & Bosch

Supplemental Material

This material supplements, but does not replace, the peer-reviewed paper in *Drug and Alcohol Dependence*

Data Processing: The output provided by the Skyn prototype employed in this research represents the measurement of raw current detected at the sensor and involves no meaningful zero metric. Thus, to approximate a more standardized metric, we subtracted the average of the first five minutes of TAC readings of each session from the entire session as a simple baseline, and one which could be easily implemented in a practical application. [Note that the most recent Skyn prototype also provides measurements in terms of units of alcohol per volume of air, thus providing a standardized metric with a meaningful zero value.] Regarding breathalyzer readings, we obtained regular BrAC measurements from participants in the alcohol condition, but more sparse readings within the no-alcohol condition (see methods section). Participants in both no-alcohol and alcohol conditions were monitored continuously throughout their experimental sessions, and they were further not allowed to keep any possessions with them during their study participation. It was thus possible to infer 0.00% BrAC at times when no alcoholic beverage had been administered by experimenters. Thus, to create instances for the no-alcohol condition, we inserted synthetic (artificial) 0.00% BrAC readings every 10 minutes, so that predictions for no-alcohol participants could be made as well. For the experimental condition, we also added a single synthetic 0.00% baseline reading 1 minute before drinking began in each session. In total, these procedures created 571 instances for the no-alcohol condition and 521 for the alcohol condition.

Feature Importance Analysis: We calculated Shapley feature importance values for each instance in the testing set (Lundberg & Lee, 2017). Shapley values describe what a model has learned about the relationships between features and the response variable in terms of the effect each feature had on the prediction for each instance. For example, the value of a feature such as standard deviation of Skyn may have a positive effect on the predicted value in some cases, a negative effect in some, and may have no effect at all in others. Shapley values can be calculated for every instance by tracing the path taken in every decision tree learned by Extra-Trees and recording the influence of the feature on the final prediction. We calculated feature importance by finding the mean absolute Shapley value across all instances for each feature, thus quantifying its total (positive and negative) influence.

We examined mean absolute Shapley values across all instances to discover what features the model found to be the most important in determining the final model predictions. Many features appeared at least once in the model during cross-validation (338), so we examined only the 25 most important (see Figure S1 and Table 1 for expanded descriptions). The three most important were related to the uniqueness of values in the timeseries, which is likely a good indicator of drinking vs. not drinking behavior (if most Skyn readings are identical, it indicates a flat line). Conversely, most of the important features captured change over time in various ways. For example, the four “fast Fourier transform” (FFT) features represent frequency characteristics of the Skyn signal (e.g., repeating patterns), the eight “change quantile” features measure variation restricted to specific quantiles of the data, and the two “linear trend” features capture linear change. Thus, the model appeared to distinguish drinking episodes from non-drinking activity by measuring flat-line Skyn readings, then estimated BrAC during drinking episodes from slope, variation, and frequency-related features.

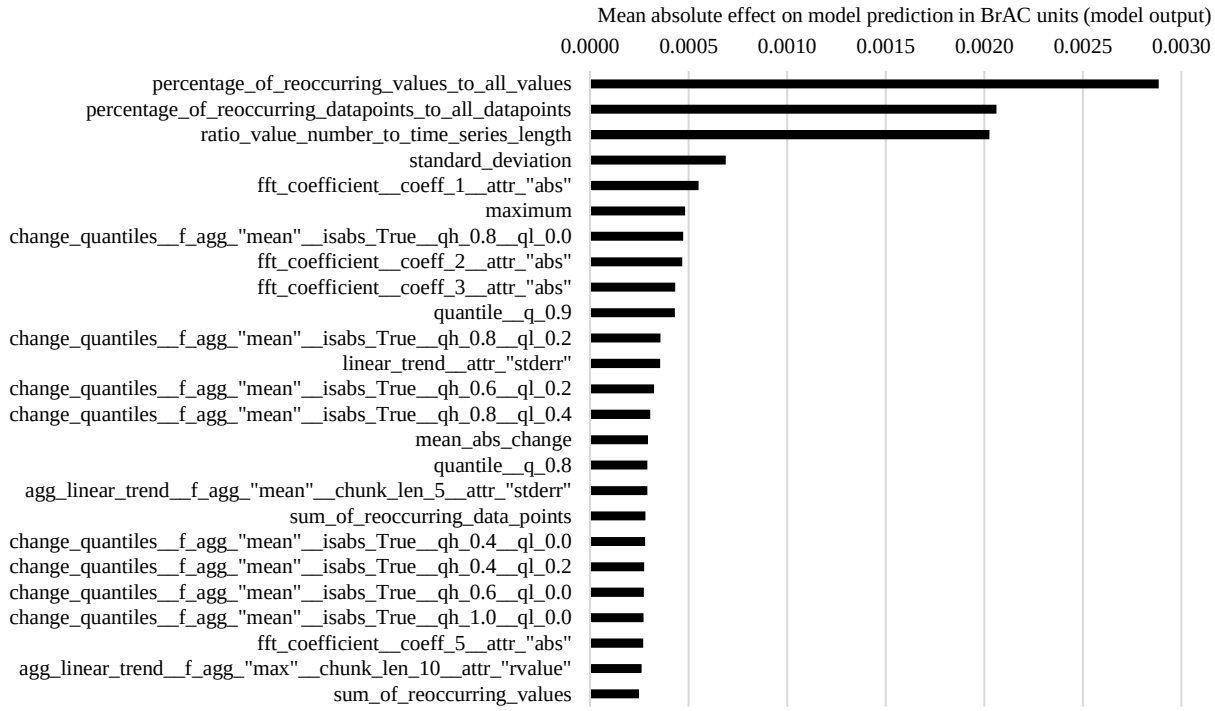


Figure S1. Shapley feature importance (mean absolute effect on predicted value) for the top 25 most important features from the main model (TSFRESH features with Extra-Trees machine learning regression). Feature names are from TSFRESH to enable exact matching in TSFRESH documentation; more intuitive descriptions of features are provided in Table 1. In cases where variables may be highly collinear, Extra-Trees will select one variable (essentially at random) when creating each branch in each tree in the model. Thus, total feature importance may be distributed across correlated features.