



OPEN Development of an alcohol biosensor non-wear algorithm: laboratory-based machine learning and field-based deployment

Nathan A. Didier^{1✉}, Rachel L. Gunn¹, Andrea C. King², Eric C. Polley³, Jennifer E. Merrill¹, Nancy P. Barnett¹ & Daniel J. Fridberg²

Wrist-worn alcohol biosensors can continuously track alcohol consumption, but their measurements are disrupted when the device is removed. Left unaddressed, non-wear data compromises observations of alcohol use and subsequent predictions of intoxication. To advance beyond commonly used temperature cutoffs and enable more precise detection of non-wear, we trained a random forest algorithm using laboratory ground truth data. Participants in Study One (N = 36) wore a wrist-worn alcohol biosensor (BACtrack Skyn) across 61 five-hour laboratory sessions, generating ground truth non-wear by removing and re-applying the device at specified times. Algorithm features included temperature, motion, and their time-series quadratic coefficients. According to device-based cross-validation, the algorithm performed with excellent sensitivity to detect non-wear (0.96) and specificity to confirm wear (0.99), out-performing all univariable temperature cutoffs from 25 to 30 °C. The algorithm was then used to evaluate biosensor adherence in Study Two, a four-week field study where participants (N = 114) wore the Skyn and self-reported non-wear intervals each day. The algorithm detected 1.6 h of daily non-wear per participant and had more agreement with self-report compared with the temperature cutoff method. This non-wear algorithm can assess biosensor adherence in field studies and may also facilitate precise data imputation, resulting in more objective models of alcohol-related outcomes.

Keywords Alcohol biosensors, Machine learning, Laboratory ground truth, Non-wear, Adherence

Alcohol biosensors are capable of continuously and passively recording transdermal alcohol concentration (TAC), a perspired biomarker useful for detecting alcohol consumption^{1–5} and severe intoxication^{6–8}. In the past decade, novel wrist-worn alcohol biosensors (e.g. BACtrack Skyn, first released in 2016) have reduced concerns about social stigma, physical discomfort, and low sampling rates associated with older alcohol biosensors, such as the SCRAM CAM ankle monitor commonly used in the judicial system^{1,2}. Given the predictive validity and established acceptability of wrist-worn biosensors, these devices have the potential to ameliorate high-risk drinking events by facilitating timely delivery of medical assistance⁹.

Before deploying wrist-worn biosensors for just-in-time interventions, signal processing systems must be able to reliably detect when the device is not worn. Wrist-worn alcohol biosensors are typically detachable and require charging, so individuals may avoid wearing or forget to wear the device. Whereas a powered-off device results in missing data, a powered-on device that is not firmly applied to the wrist (i.e., non-wear) results in readings that do not accurately reflect TAC. During non-wear, the biosensor may produce TAC readings that hover around zero, a false positive signal if exposed to environmental alcohol, or an unstable TAC signal if the device is loosely worn during a drinking event¹⁰. When left unaddressed, these muted or noisy signals lead to misclassifications of whether individuals consumed alcohol, diminishing the utility of using TAC features as model outcomes or predictors. However, if non-wear can be precisely detected by an algorithm, data quality can be improved through identification, removal and imputation^{3,11,12}. Further, a non-wear detection algorithm could be incorporated into real-time smartphone notifications to remind participants to equip or tighten the

¹Center for Alcohol and Addiction Studies, Brown University School of Public Health, 121 South Main Street, Providence, RI 02903, USA. ²Department of Psychiatry and Behavioral Neuroscience, The University of Chicago, 5841 S Maryland Ave, Chicago, IL 60637, USA. ³Department of Public Health Sciences, The University of Chicago, 5841 S Maryland Ave, Chicago, IL 60637, USA. ✉email: nathan_didier@brown.edu

device, leading to greater adherence and fewer missed drinking events in both natural environment studies and medical interventions.

Without an established non-wear algorithm, prior analyses of Skyn-derived TAC have probed for non-wear by visually assessing plots of temperature¹ or flagging biosensor readings below certain temperature cutoffs (26 °C⁴, 28 °C³, 29 °C²). Temperature cutoffs are straightforward but imprecise, with lower cutoffs overestimating and higher cutoffs underestimating biosensor adherence. A single temperature cutoff does not account for the natural variation in surface body temperature nor the sensor's temporal delays following device removal or re-application (i.e., temperature readings gradually fall following device removal and gradually rise following device re-application)^{13,14}. Therefore, to determine whether the device was worn during a particular biosensor reading, algorithms should evaluate temperature at that instance and systematic temperature trends before and after that instance. Further, motion readings recorded by the Skyn may also provide valuable information for algorithm decisions, yet motion and temperature features have not been used together to detect Skyn non-wear. Motion values are likely to be lower during non-wear than wear, especially when the device is resting on a stationary surface (versus, for example, in a backpack or pocket). Intervals of motionlessness might help an algorithm detect non-wear when temperature features are insufficient, such as in warm climates where ambient temperature can resemble body surface temperature¹⁰.

Guidance on non-wear algorithms can be found in work with other biosensors, such as accelerometers^{13–18}. In an analysis comparing several published non-wear algorithms, the top-performing algorithm included both temperature and motion time-series features, and it successfully identified 95% of wrist-worn accelerometer non-wear. Despite strong performance overall, for non-wear intervals less than an hour in duration, this algorithm only captured 66% of non-wear. The algorithm's limited ability to detect brief non-wear intervals may be due to the fact that ground truth labels were determined by error-prone plot annotations and self-report field diaries^{15,17,19,20}. To develop an algorithm sensitive to both short and long intervals of non-wear, it would be important to acquire high-resolution ground truth labels (i.e., minute-by-minute readings labeled “wear” and “non-wear” with near-perfect accuracy). One way to achieve this is by conducting a laboratory procedure that includes brief, scheduled non-wear intervals^{21–23}. To our knowledge, this approach has not been used previously to develop an algorithm for detecting TAC sensor non-wear.

The goals of the present work were to develop an algorithm that differentiates alcohol biosensor wear from non-wear using laboratory-based ground truth (Study One) and use the algorithm to retrospectively assess biosensor adherence in a separate field study (Study Two). For Study One, we hypothesized that, relative to temperature cutoffs, the newly developed algorithm would show better sensitivity to detect non-wear and better specificity to confirm wear. For Study Two, we hypothesized that the algorithm would agree more with self-report than with a temperature cutoff; we also used algorithm predictions to quantify biosensor adherence across participants and days, then compared adherence across two device models with different battery capacities. By developing a non-wear algorithm and using it to evaluate adherence in the field, this work addresses a critical issue in the validation of biosensor-based alcohol monitoring.

Study one: non-wear algorithm development using laboratory ground truth
Methods: laboratory study

Participants

Biosensor readings were collected during laboratory sessions from 36 participants (Table 1) enrolled in a larger investigation of subjective and objective responses to a fixed-dose oral alcohol challenge and their associations to future drinking outcomes^{24–26}.

	Laboratory (N = 36)	Field (N = 114)
Age (years)	49.9 (6.8)	23.2 (3.4)
Sex (female)	18 (50.0%)	65 (57.0%)
Ethnicity (Hispanic)	5 (13.9%)	28 (24.6%)
Race		
White	31 (86.1%)	82 (71.9%)
Black	2 (5.6%)	9 (7.9%)
Asian American	2 (5.6%)	4 (3.5%)
Native Hawaiian/Pacific Islander	0 (0%)	1 (0.9%)
Multiracial	1 (2.8%)	13 (11.4%)
Not Reported	0 (0%)	5 (4.4%)
Drinking Patterns ^a		
Drinks Per Week	16.4 (21.1)	18.5 (10.5)
Heavy Drinking ^b Days Per Month	6.4 (9.3)	8.8 (4.6)

Table 1. Sample characteristics. Mean (SD) or N (%), broken down by Study One (laboratory) and Study Two (field) samples. ^aDrinking patterns were based on a 1-month (28 day) timeline follow-back for the laboratory sample and a 2-month (60-day) timeline follow-back for the field sample. ^bHeavy drinking was defined as 4 + /5 + drinks for females/males on a given day; the total 28-day frequency was used for the laboratory sample while the 60-day frequency was halved for the field sample, to enable comparison.

At a screening visit, participants completed informed consent, verified their identity with a photo ID, provided biological samples (breath, urine, and blood), and filled out surveys on health, alcohol, and substance use^{27,28}. Participants met the study age criteria (40–65 years old) and were approved by the study physician based on their blood pressure, liver enzymes, medications, and substance/medical history. Full eligibility criteria can be found elsewhere²⁹.

Procedures

All study methods were fully approved by the University of Chicago Institutional Review Board. The study was conducted in accordance with governmental regulations and guidelines set by the Declaration of Helsinki. In brief, participants completed two individual double-blinded beverage administration sessions at the Clinical Addictions Research Laboratory at the University of Chicago. Upon arrival, each participant completed breathalyzer, drug, and pregnancy tests (for females) to ensure safety of participation. On separate study days and in random order, each participant consumed either a 0.8 g/kg alcohol beverage or a placebo beverage within a 15-min period²⁴. Before and several timepoints after beverage administration, participants completed a series of surveys and hand–eye coordination tasks with their non-dominant (biosensor-equipped) hand. Laboratory sessions took place in furnished, windowed rooms with an ambient temperature ranging from 19–24 °C (66–75°F). For completing all study components (screening, laboratory study, and an ambulatory phase beyond the scope of this report), participants were compensated \$400 (\$250 plus a \$150 completion bonus). Additional details on the laboratory protocol can be found elsewhere^{24,25,29–31}.

Alcohol biosensor The laboratory study used the BACtrack Skyn (BACtrack, San Francisco, CA, USA), a wrist-worn alcohol biosensor with a magnetic detachable band, a black external housing, a power button, and a small blue light that indicates whether the device is powered on (Fig. 1). On its underside, the device has a charging port and a protective sensor filter. Transdermal alcohol concentration (TAC) is measured by the device via an electrochemical fuel-cell that records micrograms of alcohol per liter of air (µg/L)^{32,33}. For assessing non-wear, the Skyn measures body surface temperature (degrees Celsius; °C) and motion (acceleration in g-forces; g). The device records TAC, temperature, and motion every 20 s. To upload the data to cloud storage, a smartphone must be connected to the device via Bluetooth and the BACtrack Skyn application must be opened¹⁰. Researchers can then download the raw data from BACtrack's online research portal. For more information on battery life and data storage, see the *Alcohol Biosensor* section in Study Two Methods.

Biosensor protocol Biosensor data were collected during 33 alcohol sessions and 28 placebo sessions. At the start of the baseline timepoint, participants equipped the wrist-worn biosensor to their non-dominant hand. Four unique Skyn devices, all Model T15 with firmware 4.13.1, were used during the laboratory study. These devices recorded data for a range of 44.4 to 73.1 h, totaling 245.0 h of Skyn data.

Ground truth non-wear There were two a priori determined non-wear intervals during each session at approximately 45 min and 180 min after the onset of beverage ingestion. At each of these intervals, the participant was instructed to remove the device and give it to the research assistant who placed it on a table for about 10–20 min. After this interval, the participant was asked to reapply the device to their wrist. At each instance of device removal and re-application, the research assistant simultaneously recorded the hour and minute to an online



Fig. 1. BACtrack Skyn Alcohol Biosensor. The Skyn wrist-worn alcohol biosensor (BACtrack, San Francisco, CA, USA) includes a power button and power indicator light on its front. The other side of the device features a charging port and protective sensor filter. The device is equipped using a magnetic band.

database³⁴. Throughout the study, there was a total of 120 non-wear intervals. See Fig. 2 (or Supplemental Fig. S1) for the device removal protocol and corresponding changes to temperature and motion.

Data were downloaded from the BACtrack research portal at one reading per minute (rather than one reading per 20 s) to simplify time-series interpretation and minimize data processing burden, consistent with prior work^{3,4}. Using ground truth timestamps recorded by research assistants, every biosensor reading was labeled “worn” or “not worn”. Per session, based on ground truth labels, the device was worn for an average ($\pm$ standard deviation) of 189.9 ± 20.0 min and not worn for 51.1 ± 13.3 min. Variation in wear and non-wear durations was due to varying rates of timepoint completion and minor procedural deviations (two sessions missed the first scheduled removal and nine sessions included removal at the end rather than the start of the 180-min timepoint). These deviations proved to be minimally consequential, as timestamps of non-wear intervals were still recorded and used to label ground truth for these sessions, resulting in more robust training data via increased variability of non-wear intervals.

Algorithm development

Algorithm features A random forest algorithm³⁵ was trained to differentiate between “worn” and “not worn” using *temperature* and *motion* recorded by the Skyn, as well as 20 additional features computed from temperature and motion: the difference between the current reading and the prior reading (*prior change*) and the difference between the current reading and the subsequent reading (*subsequent change*); the difference between the current reading and the mean of the 10 preceding readings (*prior mean change*) and the difference between the current reading and the mean of the 10 succeeding readings (*subsequent mean change*); the quadratic coefficients of the 10 preceding readings and the 10 succeeding readings (i.e., the *curvature* (a), *slope* (b), and *intercept* (c) of $y = ax^2 + bx + c$, where x is minutes relative to current timestamp and y is either temperature or motion). In total, the algorithm used 22 continuous features (Table 2).

Because time-series features depended on nearby values for their computation, time-series features were not computed when there were 5 or fewer available nearby readings. Limited nearby readings occurred before and after the laboratory session when the device was powered off. To enable algorithm predictions beside intervals of missing data, we configured three additional random forests: (1) a pre-gap algorithm built without *subsequent* features, designed to make predictions on data before an interval of missing data; (2) a post-gap algorithm built without *prior* features, designed to make predictions on data after an interval of missing data; (3) a between-gap algorithm built only with current temperature and motion, designed to make predictions on readings surrounded by missing data. The final non-wear algorithm defaults to the 22-feature random forest, but if certain features cannot be computed due to missing data, the algorithm will utilize predictions from the appropriate reduced-feature algorithm. To assess performance of the final combined algorithm, all available predictions from the 22-feature algorithm were used (which accounted for 99.1% of laboratory data) and predictions from the reduced-feature algorithms were used for the remaining 0.9%.

Training Each random forest was trained using device-based cross-validation^{36,37}, also called leave-source-out cross-validation. As there were four unique devices used throughout the laboratory study, the data were split four times such that, at each split, three devices’ data were used for training and one device’s data were reserved for testing. Device-based cross-validation was chosen to evaluate how well the algorithm generalized to discrete devices³⁷ and to limit wasting data as in single-split approaches³⁸. At the conclusion of the fourth split, each device had a turn being left out and every biosensor reading had a corresponding prediction of “worn” or “not

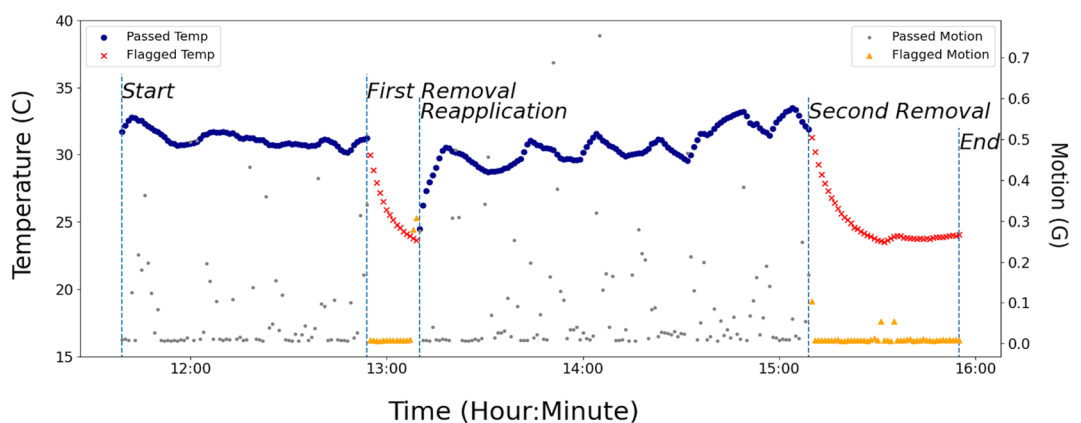


Fig. 2. Example of Laboratory Ground Truth Non-Wear. This plot illustrates one participant’s temperature and motion data during the laboratory protocol. The data show a characteristic change following device removal and re-application for a single participant. Removal produces a downward slope in temperature and a ceasing of motion signals while re-application produces a sharp rise in temperature and scattering of motion signals. The red temperature values and orange motion values correspond to non-wear intervals. These intervals were determined in the laboratory, where research assistants recorded the timestamps upon device removal and re-application; these timestamps are indicated by the dashed vertical lines.

Rank	Feature	MDI
1	Subsequent temperature intercept	0.249
2	Temperature	0.219
3	Prior mean temperature change	0.111
4	Prior temperature intercept	0.088
5	Prior temperature curvature	0.049
6	Subsequent temperature change	0.048
7	Prior temperature change	0.035
8	Motion	0.030
9	Prior temperature slope	0.025
10	Subsequent motion intercept	0.023
11	Subsequent temperature slope	0.022
12	Subsequent temperature curvature	0.018
13	Subsequent mean temperature change	0.017
14	Prior motion intercept	0.013
15	Subsequent mean motion change	0.011
16	Prior motion curvature	0.009
17	Subsequent motion curvature	0.008
18	Prior mean motion change	0.008
19	Subsequent motion slope	0.005
20	Prior motion slope	0.004
21	Prior motion change	0.003
22	Subsequent motion change	0.002

Table 2. Feature importance rankings of the random forest algorithm. Rankings of feature importance according to mean decrease in impurity (MDI), a measure of relative contribution toward random forest predictions. Change is calculated as the difference between current reading and the reading before (prior) or after (subsequent). Mean change is calculated as the difference between current reading and the average across 10 readings occurring either before (prior) or after (subsequent) the current reading. Curvature (a), slope (b), and intercept (c) are based on quadratic equations ($ax^2 + bx + c$, where x is time) fitted across 10 readings occurring either before (prior) or after (subsequent) the current reading. Raw temperature, as well as information about nearby temperature trends, were greatly influential towards algorithm predictions. Motion was influential but to a lesser extent.

worn”. A probability threshold of 0.5 was used to classify wear and non-wear readings. For details on hyperparameter tuning, see Supplementary Table S1.

Performance metrics Given the precedent set by prior non-wear algorithm research^{17,18}, non-wear was considered a positive classification. Therefore, correct predictions of non-wear were considered true positives, correct predictions of wear were considered true negatives, incorrect non-wear predictions were considered false positives, and incorrect wear predictions were considered false negatives. From these counts, sensitivity (proportion of detected non-wear), specificity (proportion of confirmed wear), and accuracy were calculated, as well as their 95% confidence intervals (CI) based on the Wilson score method. The area under the receiver operator curve (AUC-ROC) was also calculated. These metrics were calculated separately for each device and collectively for the entire dataset.

Algorithm performance was compared to the predictive performance of five different temperature cutoffs (26, 27, 28, 29, and 30 °C), three of which (26, 28, and 29 °C) have been used in prior research to detect Skyn non-wear^{2–4}. For each cutoff, biosensor readings above or equal to the cutoff were labeled “worn” and readings below the cutoffs were labeled “not worn”. True positives, true negatives, false positives, false negatives, sensitivity, specificity, and accuracy were calculated for each temperature cutoff.

Feature importance Mean decrease in impurity (MDI) and split selection percentage were calculated for each of the features in the complete 22-feature random forest. MDI provides an indication of relative importance toward predictive accuracy, with larger values indicating greater ability to differentiate wear versus non-wear^{35,39}. Split selection percentage represents the proportion of all decision splits that incorporated a particular feature. More information on MDI and split selection is provided in Supplementary Table S2.

Code Data processing and machine learning software was built using the Scikit-learn³⁹ and Pandas⁴⁰ modules in Python. The trained algorithm and temperature cutoffs have been added to the Skyn Data Manager³ as optional strategies for flagging non-wear. To replicate algorithm development, see repository here: <https://zenodo.org/records/16914785>. For the complete Skyn Data Manager software, access may be granted upon request to the corresponding author.

Results: laboratory study

Sample

The laboratory sample consisted of 36 individuals (18, 50.0% female) with an average age of 50.6 years. Participants were mostly White ($n=31$, 86.1%) and consumed an average of 16.4 drinks per week (Table 1).

Temperature and motion descriptives

Throughout the laboratory study, temperature ranged from 20.6 °C to 36.5 °C and motion ranged from 0.004 g to 0.98 g. Based on laboratory ground truth, the average ($\pm$ standard deviation) temperature and motion were, respectively, 31.2 ± 1.7 °C and 0.11 ± 0.14 g while the device was worn, compared to 25.7 ± 2.5 °C and 0.02 ± 0.04 g while it was not worn.

Temperature cutoff performance

For temperature cutoff predictions in the laboratory, the greatest sensitivity ($\pm 95\%$ CI) to detect non-wear occurred at the 30 °C cutoff (0.934 ± 0.009 CI), but this cutoff had the least specificity to confirm when the device was worn (0.774 ± 0.008 CI). Conversely, the greatest specificity to confirm when the device was worn occurred at the 25 °C cutoff (0.994 ± 0.001), but this cutoff had the worst sensitivity to detect non-wear (0.463 ± 0.018). The temperature cutoff of 28 °C had the highest accuracy of all temperature cutoffs (0.940 ± 0.004), showing strong specificity (0.974 ± 0.003) while also retaining moderate sensitivity to detect non-wear (0.815 ± 0.014).

Algorithm performance

As hypothesized, the algorithm performed with greater specificity and greater sensitivity than each temperature cutoff (Fig. 3). Across the four devices used for laboratory testing, sensitivities ranged from 0.917–0.978, specificities ranged from 0.989–0.999, and AUC-ROCs ranged from 0.998–0.999 (Table 3). Overall, the algorithm demonstrated high sensitivity to detect non-wear (0.960 ± 0.007) and high specificity to detect when the device was worn (0.995 ± 0.001), with an AUC-ROC of 0.999.

Out of the 120 device removals throughout the laboratory study, 119 had at least one concurrent non-wear prediction (i.e., only one removal went entirely undetected). Out of 245.0 h of laboratory data, there were 101 discrete intervals of incorrect predictions, with the average duration being 2.4 min and the longest interval being 24 min. See Supplementary Materials for additional plots (Fig. S2) and results (Tables S3 and S4).

Algorithm feature importance

Out of the 22 algorithm features (Table 2), the four most important were subsequent temperature intercept (MDI=0.249), temperature (MDI=0.219), prior mean temperature change (MDI=0.111), and preceding temperature intercept (MDI=0.088). While motion features were less important overall than temperature features, motion (MDI=0.030) was ranked eighth most important. Across all features, split selection percentages ranged from 2.75% to 6.99%. For more results on feature importance, see Supplementary Table S2.

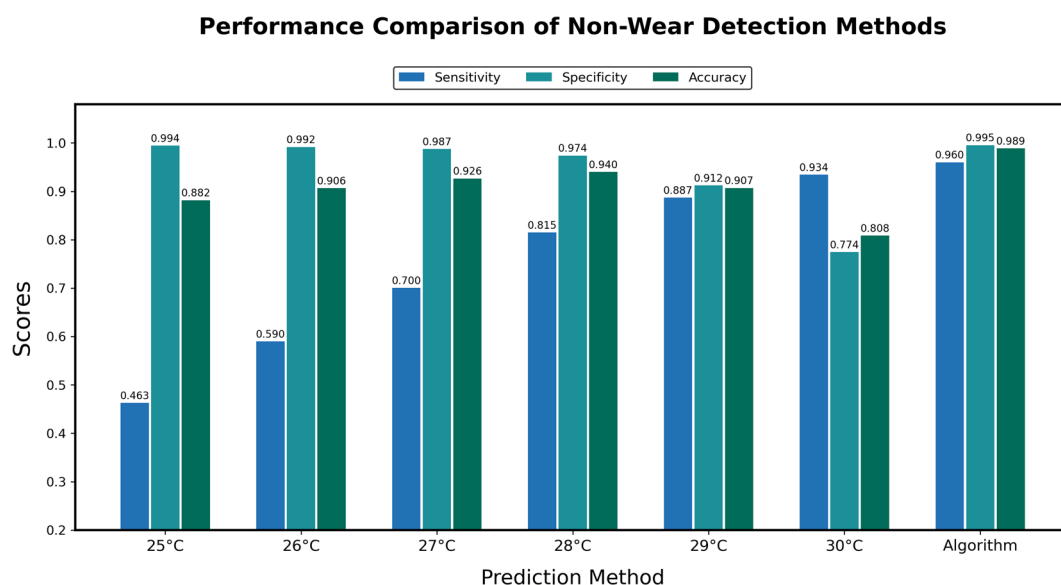


Fig. 3. Laboratory Comparison of Non-Wear Detection Methods. Bar plots of sensitivity, specificity, and accuracy across the various temperature cutoffs (25–30 Celsius) and the trained algorithm. The algorithm outperformed the temperature cutoffs on each performance metric.

Metric	Device 1	Device 2	Device 3	Device 4	Total
True Positives (correct non-wear prediction)	941	668	567	817	2,993
True Negatives (correct wear prediction)	3,421	2,439	2,075	3,592	11,527
False Positives (incorrect non-wear prediction)	5	27	8	14	54
False Negatives (incorrect wear prediction)	21	16	14	74	125
Specificity	0.999	0.989	0.996	0.996	0.995
Sensitivity	0.978	0.977	0.976	0.917	0.960
Accuracy	99.4%	98.6%	99.2%	98.0%	98.8%
AUC-ROC	0.999	0.999	0.999	0.998	0.999

Table 3. Non-wear algorithm performance with laboratory ground truth data. Positive indicates a prediction of “not worn” and negative indicates a prediction of “worn”. Predictions were conducted for each biosensor reading; data had a sampling rate of one reading per minute. True indicates that the prediction was confirmed by laboratory-based ground truth, while false indicates that the prediction was incorrect according to laboratory-based ground truth. AUC-ROC: Area under the received operator curve. Algorithm performance was strong overall and reliable across devices, although with slightly weakened performance for Device 4.

Study two: field-based deployment of non-wear algorithm
Methods: field study

Participants

The algorithm was next used to assess non-wear in field data collected in an independent young adult sample (N = 114, Table 1). All study methods were approved by the Brown University Institutional Review Board. The study was conducted in accordance with governmental regulations and guidelines set by the Declaration of Helsinki. Participants wore the Skyn for 28 days in the natural environment as part of an ecological momentary assessment (EMA) protocol designed to evaluate the effects of alcohol-cannabis co-use⁴¹. To confirm eligibility, participants completed demographic, health, and substance use measures²⁷. Participants were non-treatment seeking young adults who reported alcohol consumption at least twice per week. All participants completed informed consent. Full inclusion criteria can be found elsewhere⁴¹.

Study orientation

At the initial study orientation, participants installed two applications onto their smartphones: the TigerAware⁴² application for EMA data collection and the BACtrack Skyn application for uploading biosensor data. Next, they were taught how to fill out EMA surveys, apply the Skyn to their wrist, upload their Skyn data, and charge the device. They were instructed to wear the Skyn for the entirety of the 28-day study, except for when charging the device, bathing, or swimming. They were also encouraged to contact the study team in the event of technical difficulties.

Alcohol biosensor

Participants began wearing the BACtrack Skyn at the end of the orientation session and returned it after the 28-day study. The first 33 participants used T10 Skyn models (7 unique devices; firmware 2.0.8) while the last 81 participants used T15 Skyn models (64 unique devices; firmware 4.12.1–4.13.1; the same model used in the laboratory study). Regarding battery specifications, T10 models had a three-day battery life while T15 models had a twenty-day battery life. Both T10 and T15 devices could store up to three days of recent data at a time. Failing to upload data within these timeframes resulted in missing data, as the oldest data would be erased to provide storage space for recent data. Missing data also occurred when participants manually powered off the device or allowed it to run out of battery power.

Morning surveys

As a part of the EMA protocol, morning surveys asked the participant whether they removed the biosensor at any time the prior day. For those who endorsed removing the device, the smartphone survey allowed for up to three non-wear time intervals using dropdowns for the hour (00–23) and minute (00–59) of device removal and re-application. For each non-wear interval, the participant indicated a reason (“bathing”, “charging”, or “other”). To incentivize uploading data and charging the Skyn, participants were paid a \$5 bonus for each day that had 19 or more hours of biosensor data. Depending on survey and biosensor completion, participants earned up to \$450. The full study protocol is detailed elsewhere⁴¹.

Data analysis

Days without any biosensor readings (n = 19, 0.6%) and days without a corresponding morning report (n = 680, 18.6%) were removed, resulting in 2658 days across 114 participants for analysis. All biosensor readings from the field were labeled “worn” or “not worn” using each of the following methods: the trained algorithm, a temperature cutoff at 28 °C, and self-report. The temperature cutoff of 28 °C was chosen because it showed better accuracy in the laboratory relative to other temperature cutoffs.

Comparison of non-wear detection methods To test the hypothesis of the field study, the duration of non-wear per day was calculated for each non-wear detection method (algorithm, temperature cutoff, self-report). Then

the duration of discrepancy, defined as the hours per day when non-wear predictions disagreed, was calculated between the algorithm and self-report, between the cutoff and self-report, and between the algorithm and cutoff. Since discrepancies were normally distributed across days, a paired t-test was used to test whether the algorithm (versus temperature cutoff) had more agreement with self-report.

Comparison of device models To explore whether battery differences may have affected the frequency of missing data or non-wear, the T10 and T15 models were compared across three non-adherence outcomes: daily frequency of missing data (i.e., no data recorded either due to device being powered off or data not uploaded), daily frequency of algorithm-detected non-wear (i.e., data were available but the algorithm predicted that the device was not worn), and total non-adherence (i.e., frequency of either missing data or detected non-wear). Since these non-adherence outcomes were right-skewed (smaller values were more frequent than larger values), Mann–Whitney U tests were used to compare non-adherence across the T10 and T15 models.

Non-wear interval durations Last, to understand how long non-wear intervals lasted, the average duration of non-wear intervals was calculated, with interval defined as one or more consecutive non-wear readings separated by at least one wear reading.

Results: field study

Sample

The field sample consisted of 114 individuals (65, 57.0% female) with an average age of 23.2 years (Table 1). They consumed an average of 18.5 drinks per week and more than half were non-Hispanic White ($n = 67$, 58.8%).

Biosensor adherence

Of the 2658 days used in analyses, the biosensor was powered on for an average of 22.25 h per day and worn for an average of 20.63 h per day according to the non-wear algorithm (Fig. 4). On more than half of the study days (57.4%), the biosensor was powered on for all 24 h. Nearly half ($n = 55$, 48.2%) of participants had a daily average of 21 or more hours with adherent biosensor data (i.e., the device was worn and powered on). Participants endorsed bathing/showering as the most common reason for non-wear (63.6%), followed by charging (34.0%) and other (2.4%).

Temperature and motion descriptives

Throughout the field study, temperature ranged from 5.8 °C to 36 °C and motion ranged from 0.0 g to 2.53 g. Based on algorithm predictions, temperature and motion readings, respectively, averaged 32.8 °C and 0.10 g while the device was worn and 23.7 °C and 0.05 g while the device was not worn.

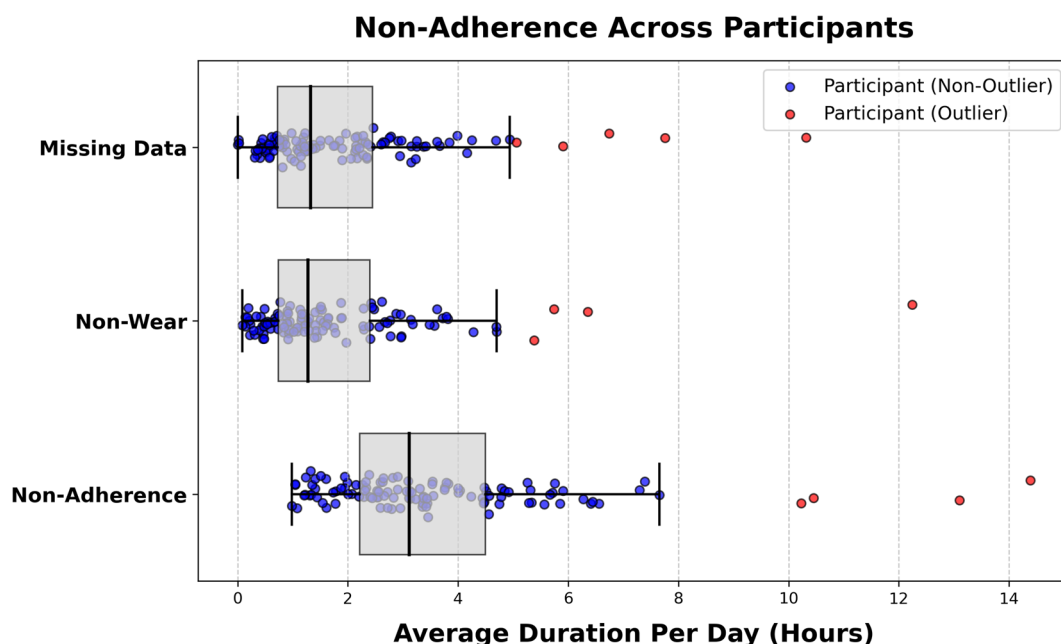


Fig. 4. Daily Adherence Across Participants in the Field. For each box-and-whisker plot, a dot represents a participant. The middle line is the median, the box captures the inter-quartile range (IQR; i.e., the middle 50%), and the outlier whiskers extend 1.5 times the IQR (or until the farthest datapoint in that direction). Outlier participants are represented by red dots. Non-adherence is the sum of missing data and non-wear durations.

Non-wear detection comparison

The various non-wear detection methods (algorithm, temperature cutoff, and self-report) resulted in sample averages of 1.62, 1.87, and 1.14 h of daily non-wear, respectively. In support of the Study Two hypothesis, self-reported non-wear agreed with the algorithm more than the temperature cutoff (13 more minutes per day, on average; $p < 0.0001$). Notably, the algorithm and temperature cutoff disagreed with self-report for an average of 2.02 and 2.24 h per day, respectively, and these were significantly greater than their disagreement with each other (0.53 h per day, on average; $ps < 0.0001$).

Comparison of Skyn models

Missing data was more common for T10 models than T15 models (2.6 > 1.4 h per day, $p < 0.0001$) (Fig. 5). Meanwhile, there was less non-wear for T10 models than T15 models (1.0 < 1.9 h per day, $p < 0.0001$). When considering missing data and non-wear together, T10 models were associated with more total non-adherence than T15 models (3.6 > 3.3 h per day, $p < 0.0001$).

Non-wear intervals

The algorithm detected 20,749 discrete intervals of non-wear (7.8 non-wear intervals per day, on average) during the field study, but more than half of these lasted three minutes or less (Fig. 6). The average duration for a non-wear interval was 17.0 min and the longest observed non-wear interval was 3,827.0 min (63.8 h).

Discussion

In this work, we used laboratory training data to develop a reliable non-wear algorithm (Study One), then used this algorithm to assess biosensor adherence in a separate field study (Study Two). The algorithm showed excellent specificity (0.99) and sensitivity (0.96) to identify scheduled intervals of wear and non-wear in the laboratory, out-performing temperature cutoffs from 25 to 30 °C. For the field study, the algorithm made predictions on every reading across 71 Skyn biosensors, regularly detected brief non-wear intervals, and showed more agreement with self-reported non-wear compared to a 28 °C temperature cutoff.

With a sensitivity of 0.96, the present algorithm is comparable to top-performing accelerometer non-wear algorithms¹⁷. The algorithm's excellent sensitivity, even for brief intervals of non-wear—a notable weakness of accelerometer non-wear algorithms—can be attributed to novel features that capture time-series trends (e.g., quadratic coefficients) and laboratory-controlled ground truth (i.e., precisely recorded non-wear intervals of 10–20 min). According to MDI, the most important algorithm feature was the quadratic intercept of subsequent temperature readings. This feature provides foresight into where temperature readings are headed, enabling the algorithm to identify the early moments of non-wear (when temperature readings have begun their decline but are still in a biologically plausible range) and the early moments following re-application (when temperature readings have begun their incline but are still below a biologically plausible range). By employing device-based cross-validation, we observed that the algorithm was consistently excellent for three devices with slightly weakened performance for one device (sensitivities of 0.977–0.980 versus 0.917). Reduced performance in the latter device was likely due to inadequate training data; this device recorded the most laboratory data compared to the other three devices, meaning that this iteration of cross-validation had the least amount of training data relative to other devices. Further, a third of the false negatives associated with this device occurred during a single 24-min non-wear interval where temperature only dropped by about two degrees. This may have been caused by environmental influences (e.g., exposure to direct sunlight or body heat) or a transient period of poor sensor performance. Fortunately, for the final algorithm trained on all data, these previously misclassified data likely helped it to recognize non-wear intervals with subtle changes in temperature⁴³.

After training the final algorithm, we used it to identify non-wear in all data recorded by 71 devices in the field. Seven of these were older T10 devices, not the newer T15 devices used to train the algorithm. The T15 models were associated with more overall adherence (i.e., less frequency of either missing data or non-wear), demonstrating the benefits of prolonged battery life and upgrades to sensor hardware. When considering missing data and non-wear separately, the T10 devices had more frequent missing data, but the newer T15 devices had more frequent non-wear. The higher frequency of missing data in the T10 devices was probably due to its limited battery life relative to T15 devices (3 days versus 20 days, respectively), causing it to lose power and require charging more frequently. Meanwhile, the higher frequency of non-wear in the T15 devices may be because the algorithm was trained using T15 data, making it more sensitive to detect non-wear in T15 versus T10 devices. Related, while the present algorithm was trained with and designed for the T15 Skyn, it may be able to detect non-wear with other biosensors that also measure temperature and motion. However, it is important to consider whether the sampling rate or sensor technology have substantial differences, as the algorithm may not generalize due to systematic differences in temperature and motion readings. Prior to deploying the algorithm for another biosensor, researchers should test algorithm performance on a ground truth dataset from that biosensor. If performance is inadequate, then it is recommended to train a new algorithm, and this could be done by replicating the machine learning approach described in this paper.

The present non-wear algorithm may enhance field work with alcohol biosensors in several ways. First, the algorithm is integrated within the Skyn Data Manager software which streamlines feature engineering, non-wear detection, and imputation³, limiting the computational burden on research staff and providing a standardized approach that can be used across laboratories. Second, the algorithm could eventually be built into a smartphone application that delivers automatic, real-time notifications when participants avoid wearing or forget to wear the Skyn, leading to greater adherence and data quality. For immediate use, research staff can evaluate adherence by employing the algorithm on a regular basis via the Skyn Data Manager, then contact participants, as needed. Third, the algorithm will enable precise removal or imputation of non-wear data, leading to higher data quality for detecting drinking events and computing TAC features. Brief intervals of non-wear (e.g., less than five

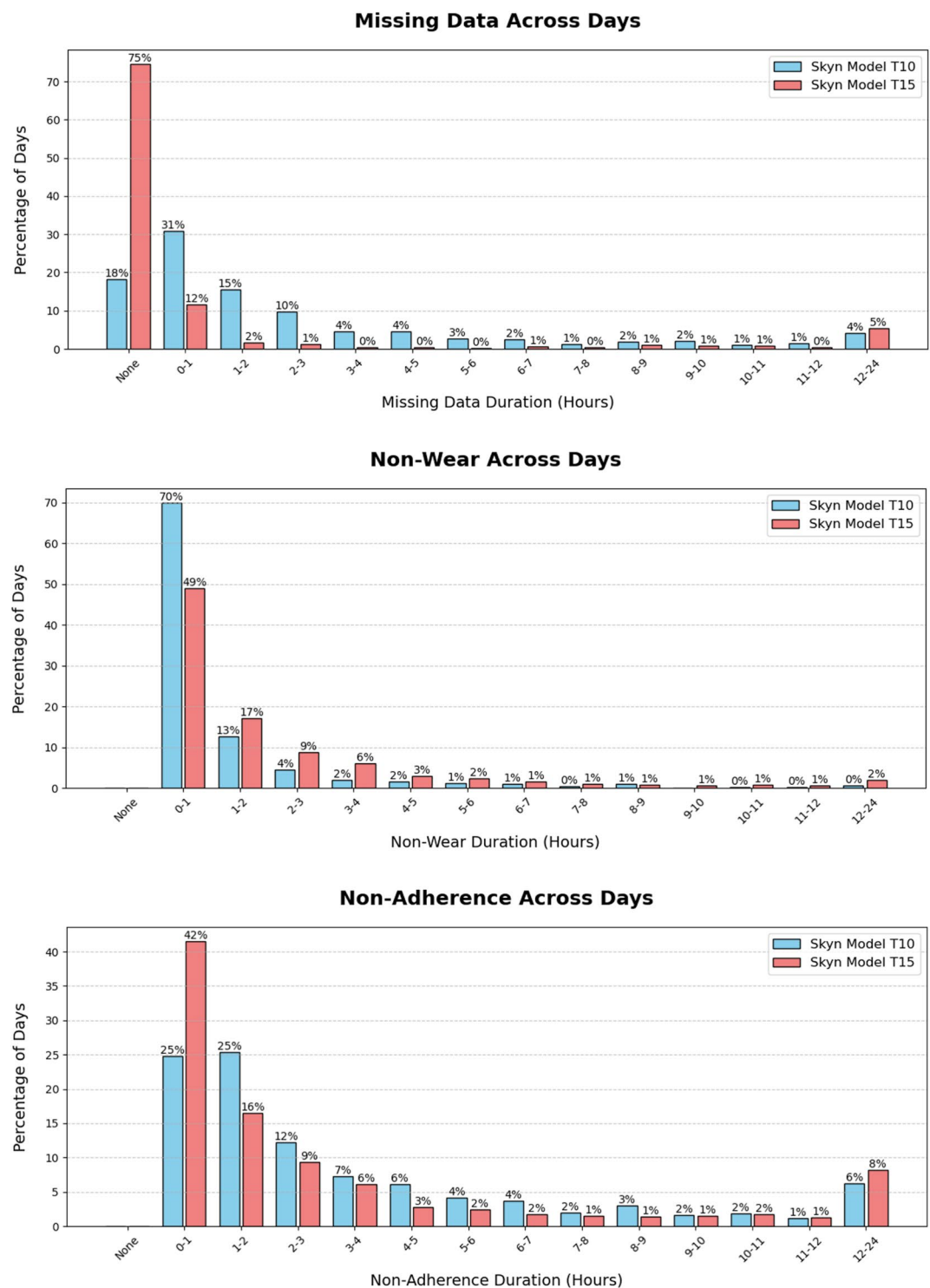


Fig. 5. Day-Level Adherence Metrics in the Field. These three histograms illustrate the distribution of biosensor non-adherence at the day-level. The top histogram captures the daily duration of missing data (i.e., device powered off or data not uploaded). The middle histogram captures the daily duration of when the device was powered on but not worn (i.e., non-wear). The bottom histogram captures the daily duration of either type of non-adherence. The y-axis represents the percentage of days relative to the total number of days recorded by either the T10 or T15 Skyn models. Likely a consequence of battery life differences, missing data was more common for T10 models, while non-wear was more common for T15 models.

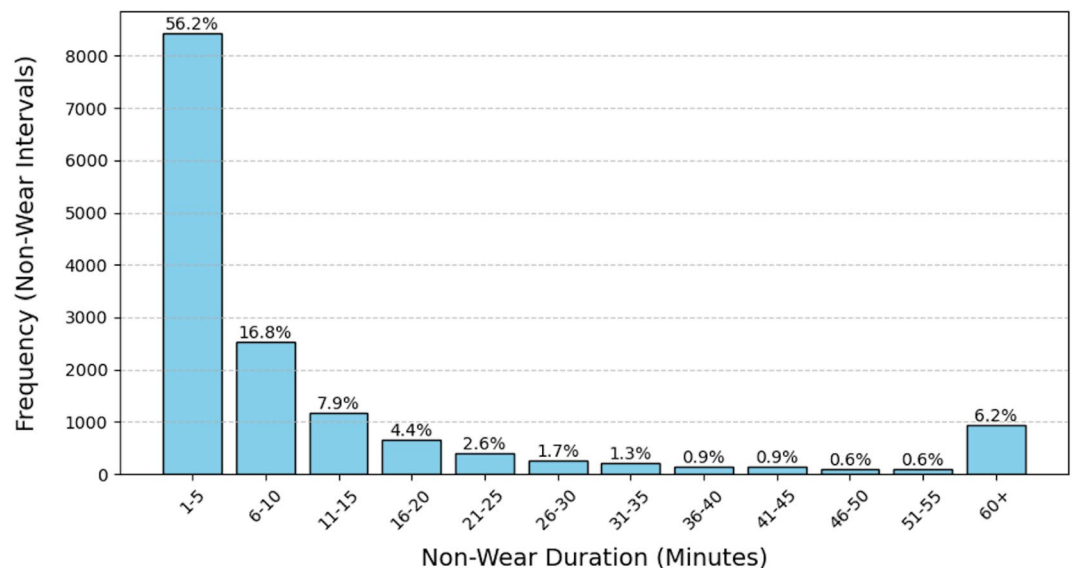


Fig. 6. Distribution of Non-Wear Interval Durations in the Field. Distribution of non-wear interval durations; an interval consists of one or more consecutive biosensor readings labeled as non-wear (separated by at least one wear reading), according to algorithm predictions. Most non-wear intervals had very short durations.

minutes) were regularly detected in the field study; to inform whether such brief non-wear intervals should be imputed or left as-is, future work should explore how TAC is affected by non-wear intervals of varying duration, as imputation may only be impactful for longer intervals of non-wear. In the meantime, for non-wear intervals surrounded by high-quality data (i.e., minimal missing data or non-wear; TAC not disrupted by artifacts), an effective and cautious approach would be to use the surrounding data to train an imputation algorithm¹², then use this algorithm to impute TAC at each non-wear reading, as well as several readings before and after the non-wear interval.

Strengths of this study include the use of ground truth data from a tightly controlled laboratory protocol, robust algorithm features and training, field application to quantify adherence metrics across days and individuals, and an evaluation of algorithm performance across devices and model versions. However, the results should be interpreted considering some important limitations. While the algorithm was evaluated using laboratory-based ground truth, the field study lacked a ground truth. Self-reported non-wear could hypothetically be used for this purpose, but these data showed significant discrepancy with both the algorithm and a temperature cutoff, likely due to memory and data entry errors. These discrepancies highlight the difficulty of acquiring ground truth via self-report²⁰. Ground truth non-wear in future field studies may be acquired through EMA event marking tools (e.g., timestamped photo submissions at device removal/re-application), corroboration by another individual, and comparison with signals from a second biosensor. Without ground truth data to confirm algorithm predictions in the present field study, it remains unclear how well the algorithm would have generalized from the laboratory (older adults wearing the device in climate-controlled conditions) to the field (younger adults wearing the device in natural environments where extreme climates and activities may influence the performance of the algorithm). Last, field adherence outcomes were based on missing data and non-wear, but collaborative non-adherence (i.e., giving the device to someone else to wear) is a form of non-adherence that is not yet addressed and may require additional biofeedback to enable detection (e.g., blood volume pulse readings which have idiosyncratic patterns⁴⁴). In the present work, given that the biosensors were not used for forensic purposes or as part of a contingency management program that financially incentivized sobriety, it is unlikely that participants were motivated to disguise their drinking through collaborative non-adherence.

In conclusion, we developed an algorithm that accurately detects non-wear of a wrist-worn alcohol biosensor and then used the algorithm to quantify adherence in a field study. Future work should explore how to use the algorithm for data cleaning and for delivering notifications when participants forget to wear the device. Ultimately, this algorithm will improve TAC data quality and lead to more reliable detection of intoxication, increasing the viability of using alcohol biosensors for just-in-time interventions.

Data availability

Data for testing the algorithm can be found within the codebase repository: <https://zenodo.org/records/16914785>. For access to training data, please send a request to the corresponding author.

Code availability

Code for replicating algorithm development can be found here: <https://zenodo.org/records/16914785>. For full signal processing software that includes the trained non-wear algorithm, access to the Skyn Data Manager may be granted upon request to the corresponding author.

Received: 23 May 2025; Accepted: 18 August 2025

Published online: 25 August 2025

References

- Ash, G. I. et al. Sensitivity, specificity, and tolerability of the BACtrack Skyn compared to other alcohol monitoring approaches among young adults in a field-based setting. *Alcohol. Clin. Exp. Res.* **46**, 783–796 (2022).
- Courtney, J. B., Russell, M. A. & Conroy, D. E. Acceptability and validity of using the BACtrack skyn wrist-worn transdermal alcohol concentration sensor to capture alcohol use across 28 days under naturalistic conditions—A pilot study. *Alcohol* **108**, 30–43 (2023).
- Didier, N. A., King, A. C., Polley, E. C. & Fridberg, D. J. Signal processing and machine learning with transdermal alcohol concentration to predict natural environment alcohol consumption. *Exp. Clin. Psychopharmacol.* **32**, 245–254 (2024).
- Fairbairn, C. E., Han, J., Caumiant, E. P., Benjamin, A. S. & Bosch, N. A. Wearable Alcohol biosensor: Exploring the accuracy of transdermal drinking detection. *Drug Alcohol Depend.* 112519 (2024) <https://doi.org/10.1016/j.drugalcdep.2024.112519>.
- Gunn, R. L., Steingrimsson, J. A., Merrill, J. E., Souza, T. & Barnett, N. Characterising patterns of alcohol use among heavy drinkers: A cluster analysis utilising alcohol biosensor data. *Drug Alcohol Rev.* **40**, 1155–1164 (2021).
- Richards, V. L. et al. Transdermal alcohol concentration features predict alcohol-induced blackouts in college students. *Alcohol Clin. Exp. Res.* **48**, 880–888 (2024).
- Richards, V. L., Mallett, K. A., Turrissi, R. J., Glenn, S. D. & Russell, M. A. Profiles of transdermal alcohol concentration and their prediction of negative and positive alcohol-related consequences in young adults' natural settings. *Psychol. Addict. Behav.* **39**, 163–172 (2025).
- Russell, M. A. et al. Profiles of alcohol intoxication and their associated risks in young adults' natural settings: A multilevel latent profile analysis applied to daily transdermal alcohol concentration data. *Psychol. Addict. Behav.* **39**, 173–185 (2025).
- Wang, Y., Porges, E. C., DeFelice, J. & Fridberg, D. J. Integrating alcohol biosensors with ecological momentary intervention (EMI) for alcohol use: A synthesis of the latest literature and directions for future research. *Curr. Addict. Rep.* **11**, 191–198 (2024).
- Gunn, R. L. et al. Use of the BACtrack Skyn alcohol biosensor: Practical applications for data collection and analysis. *Addiction* **118**, 1586–1595 (2023).
- Ae Lee, J. & Gill, J. Missing value imputation for physical activity data measured by accelerometer. *Stat. Methods Med. Res.* **27**, 490–506 (2018).
- Jafrasteh, B., Hernández-Lobato, D., Lubián-López, S. P. & Benavente-Fernández, I. Gaussian Processes for Missing Value Imputation. Preprint at <https://doi.org/10.48550/arXiv.2204.04648> (2022).
- Vert, A. et al. Detecting accelerometer non-wear periods using change in acceleration combined with rate-of-change in temperature. *BMC Med. Res. Methodol.* **22**, 147 (2022).
- Zhou, S.-M. et al. Classification of accelerometer wear and non-wear events in seconds for monitoring free-living physical activity. *BMJ Open* **5**, e007447 (2015).
- Ahmadi, M., Nathan, N., Sutherland, R., Wolfenden, L. & Trost, S. Non-wear or sleep? Evaluation of five non-wear detection algorithms for raw accelerometer data. *J. Sports Sci.* **38**, 399–404 (2020).
- Pagnamenta, S., Grönvik, K. B., Aminian, K., Vereijken, B. & Paraschiv-Ionescu, A. Putting temperature into the equation: Development and validation of algorithms to distinguish non-wearing from inactivity and sleep in wearable sensors. *Sensors* **22**, 1117 (2022).
- Skovgaard, E. L. et al. Generalizability and performance of methods to detect non-wear with free-living accelerometer recordings. *Sci. Rep.* **13**, 2496 (2023).
- Syed, S., Morseth, B., Hopstock, L. A. & Horsch, A. A novel algorithm to detect non-wear time from raw accelerometer data using deep convolutional neural networks. *Sci. Rep.* **11**, 8832 (2021).
- Skovgaard, E. L., Pedersen, J., Møller, N. C., Grøntved, A. & Brønd, J. C. Manual annotation of time in bed using free-living recordings of accelerometry data. *Sensors* **21**, 8442 (2021).
- Ainsworth, B. E. et al. ecommendations to improve the accuracy of estimates of physical activity derived from self report. *J. Phys. Activity Health* <https://doi.org/10.1123/jpah.9.s1.s76> (2012).
- Fairbairn, C. E. & Bosch, N. A new generation of transdermal alcohol biosensing technology: Practical applications, machine-learning analytics and questions for future research. *Addiction* **116**, 2912–2920 (2021).
- Merrill, M. A. et al. Homekit2020: A benchmark for time series classification on a large mobile sensing dataset with laboratory tested ground truth of influenza infections. In *Proceedings of the Conference on Health, Inference, and Learning* 207–228 (PMLR, 2023).
- Luo, S. et al. Deep learning-enabled imaging flow cytometry for high-speed cryptosporidium and giardia detection. *Cytometry A* **99**, 1123–1133 (2021).
- King, A. C., de Wit, H., McNamara, P. J. & Cao, D. Rewarding, stimulant, and sedative alcohol responses and relationship to future binge drinking. *Arch. Gen. Psychiatry* **68**, 389–399 (2011).
- King, A. C., McNamara, P. J., Hasin, D. S. & Cao, D. Alcohol challenge responses predict future alcohol use disorder symptoms: A 6-year prospective study. *Biol. Psychiatry* **75**, 798–806 (2014).
- King, A. C. et al. Subjective responses to alcohol in the development and maintenance of alcohol use disorder. *Am. J. Psychiatry* **178**, 560–571 (2021).
- Sobell, L. C. & Sobell, M. B. Timeline follow-back. In *Measuring Alcohol Consumption: Psychosocial and Biochemical Methods* (eds Litten, R. Z. & Allen, J. P.) 41–72 (Humana Press, Totowa, NJ, 1992). https://doi.org/10.1007/978-1-4612-0357-5_3.
- Rueger, S. Y., Trela, C. J., Palmeri, M. & King, A. C. Self-administered web-based timeline followback procedure for drinking and smoking behaviors in young adults. *J. Stud. Alcohol Drugs* **73**, 829–833 (2012).
- Didier, N., Cao, D. & King, A. C. The eyes have it: Alcohol-induced eye movement impairment and perceived impairment in older adults with and without alcohol use disorder. *Alcohol Clin. Exp. Res.* **49**, 437–447 (2025).
- Didier, N., Vena, A., Feather, A. R., Grant, J. E. & King, A. C. Holding your liquor: Comparison of alcohol-induced psychomotor impairment in drinkers with and without alcohol use disorder. *Alcohol Clin. Exp. Res.* **47**, 1156–1166 (2023).
- Roche, D. J. O., Palmeri, M. D. & King, A. C. Acute alcohol response phenotype in heavy social drinkers is robust and reproducible. *Alcohol Clin. Exp. Res.* **38**, 844–852 (2014).
- Campbell, A. S., Kim, J. & Wang, J. Wearable electrochemical alcohol biosensors. *Curr. Opin. Electrochem.* **10**, 126–135 (2018).
- Wang, Y. et al. Wrist-worn alcohol biosensors: Applications and usability in behavioral research. *Alcohol* **92**, 25–34 (2021).
- Harris, P. A. et al. The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
- Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
- Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Methodol.* **36**, 111–133 (1974).
- Leinonen, T. et al. Empirical investigation of multi-source cross-validation in clinical ECG classification. *Comput. Biol. Med.* **183**, 109271 (2024).
- Steyerberg, E. W. Validation in prediction research: the waste by data splitting. *J. Clin. Epidemiol.* **103**, 131–133 (2018).
- Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

40. McKinney, W. Data structures for statistical computing in python. in 56–61 (Austin, Texas, 2010). <https://doi.org/10.25080/Majora-92bf1922-00a>.
41. Gunn, R. L. et al. Examining the impact of simultaneous alcohol and cannabis use on alcohol consumption and consequences: Protocol for an observational ambulatory assessment study in young adults. *JMIR Res. Protoc.* **13**, e58685 (2024).
42. Morrison, W., Guerdan, L., Kanugo, J., Trull, T. & Shang, Y. TigerAware: An innovative mobile survey and sensor data collection and analytics system. in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)* 115–122 (2018). <https://doi.org/10.1109/DSC.2018.00025>.
43. Daoudi, N., Allix, K., Bissyandé, T. F. & Klein, J. Guided retraining to enhance the detection of difficult android malware. in *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis* 1131–1143 (Association for Computing Machinery, New York, NY, USA, 2023). <https://doi.org/10.1145/3597926.3598123>.
44. Singh, R., Lewis, B., Chapman, B., Carreiro, S. & Venkatasubramanian, K. A machine learning-based approach for collaborative non-adherence detection during opioid abuse surveillance using a wearable biosensor. *Biomed. Eng. Syst. Technol. Int. Jt. Conf. BIOSTEC Revis. Sel. Pap. BIOSTEC Conf.* **5**, 310–318 (2019).

Acknowledgements

Research was supported by funding from the National Institutes of Health. Grant identification numbers are R01-AA013746 (ACK) & R21-AA029746 (DJF) for the laboratory study, K08-AA027551 (RLG) for the field study.

Author contributions

All authors have substantially contributed to the creation and write-up of this work. All authors have approved this manuscript for submission and take ownership of the following contributions. Laboratory Study Concept and Design: NAD, DJF, ACK, ECP; Laboratory Data Acquisition: ACK, DJF; Field Study Concept and Design: NAD, RLG; Field Study Data Acquisition: RLG; Analysis: NAD, ECP; Software: NAD; Interpretation: NAD, DJF, ACK, ECP, RLG, JEM, NPB Drafting: NAD, RLG, DJF, ACK, ECP, JEM, NPB.

Funding

Research was supported by funding from the National Institutes of Health. Grant identification numbers are R01-AA013746 (ACK) & R21-AA029746 (DJF) for the laboratory study, K08-AA027551 (RLG) for the field study.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-16640-y>.

Correspondence and requests for materials should be addressed to N.A.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025