Original Paper

Platform Differences in Smartphone Sensing Features for Mental Health Across Four Annual College Cohorts: Secondary Analysis

Ariz Naqvi[1*], PhD; Farah Naqvi[2*], PhD

[1]College of Business Administration, American University of the Middle East, Egaila, Kuwait

[2]School of Psychology, University of Wolverhampton, Wolverhampton, England, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Ariz Naqvi, PhD
College of Business Administration, American University of the Middle East
Block 6, Building 1, Street 250
Egaila, 54200
Kuwait
Phone: 965 97952687
Email: ariz.naqvi@aum.edu.kw

Abstract

Background: Passive sensing for mobile mental health relies on observations that can differ across smartphone platforms. Distinguishing behavioral differences from differences in observation is necessary before using these features for symptom screening or monitoring.

Objective: We quantified platform-associated differences, feature availability, sensitivity to preprocessing, and the predictive contribution of connectivity features within the Generalization of Longitudinal Behavior Modeling (GLOBEM) dataset and its collection and processing pipeline.

Methods: We analyzed 4 annual US college cohorts from 2018 to 2021, comprising 705 annual records from 497 people reported by the dataset investigators. The feature sample contained 554 iOS and 151 Android records. Without verified cross-year identity links, inference and cross-validation were performed separately within each cohort. We compared 15 raw constructs and matched transformations using Cliff δ, participant resampling, and fixed logistic regression. Phone-derived location and screen features were separated from Fitbit activity and sleep measures. Analyses included coverage, complete cases, psychological adjustment, and retrospective depressive-symptom screening using the preceding 14 days.

Results: Total Bluetooth scan counts were lower on iOS in every cohort (Cliff δ −1.00 to −0.85). Connectivity counts differed more between platforms than core phone behavioral features did (difference in mean absolute Cliff δ 0.22-0.43; all 95% CIs above zero). Smaller observed-value differences did not imply equal coverage: 41 of 103 iOS records lacked all home-time values in 2020, compared with 0 of 34 Android records. Platform-classification area under the receiver operating characteristic curve (AUROC) was 0.94 to 0.997 for 13 raw constructs, 0.34 to 0.91 after normalization, and 0.65 to 0.80 after discretization. Bluetooth differences persisted after psychological adjustment. In 661 eligible screening windows, behavioral-model AUROC ranged from 0.41 to 0.56. All behavioral, connectivity-added, and platform-added models had worse Brier scores than the intercept-only benchmark. Adding connectivity changed AUROC by 0.004 to 0.034; every paired 95% CI included zero.

Conclusions: Within GLOBEM, selected connectivity features showed larger platform-associated differences than core phone behavioral features, while availability problems also affected location. Weak screening performance limits what these comparisons can establish about downstream harm. Platform-specific coverage and outcome calibration require assessment before interpreting sensing features as clinical markers.

(JMIR Mhealth Uhealth 2026;14:e103711) doi: 10.2196/103711

KEYWORDS

mobile health; passive sensing; smartphone; mental health; depressive symptoms; measurement; data quality; operating system; digital phenotyping; wearable electronic devices

XSL•FO
RenderX

Introduction

Measurement in Mobile Mental Health

A low smartphone-derived activity or communication value can reflect behavior, limited observation, or both. This ambiguity matters when mobile health (mHealth) apps use passive sensing to supplement depression screening or monitor changes between clinical assessments. Early research linked location and phone-use features to depressive symptoms [1]. Subsequent work used communication traces to model clinically assessed mood and anxiety symptoms [2], and screen and connectivity features to classify questionnaire-defined depression [3]. These studies motivate careful interpretation of the measurements supplied to a model, rather than treating a familiar feature name as evidence of comparable observation across devices.

Communication and proximity features are used as proxies for social engagement and withdrawal. Smartphone and Fitbit studies have related mobility, activity, and communication to loneliness [4]. Yet phone channels need not move together: Wang and colleagues [5] found fewer incoming contact calls but more frequent social-app use among Android users with depression-screen-positive scores. A call count is a recorded event count, and Bluetooth detections identify nearby devices rather than verified interpersonal contact. The distinction between a trace, a behavioral marker, and a clinical state is central to personal sensing models [6].

Monitoring an individual and ranking different individuals require different measurement evidence. Student studies have used repeated phone and wearable observations to model weekly symptoms and end-of-semester symptom changes [7,8]. A stable device-specific offset could affect between-person comparisons while preserving some within-person information. Intermittent observation or device replacement could affect both. Our annual platform comparisons address the former question; they do not evaluate personalized symptom trajectories.

Platform differences in observation are already documented. Halabi and colleagues [9] analyzed multimodal sensor data from approximately 3000 participants in real-world settings and found that data-quality features could identify Android versus iOS with accuracy as high as 0.98. The Reproducible Analysis Pipeline for Data Streams (RAPIDS), a toolkit for processing mobile sensing data, provides a common analysis framework, but its feature documentation describes platform-specific event handling [10]. Common output names therefore need not imply identical operational definitions. A systematic review of smartphone-only health sensing documented uneven operating-system coverage and incomplete platform reporting [11].

Prediction and measurement are distinct questions. Mobility-based depression prediction has performed differently across homogeneous student samples and heterogeneous populations [12]. The Generalization of Longitudinal Behavior Modeling (GLOBEM) benchmark also examined difficulties in generalizing longitudinal behavioral models across datasets [13]. Such findings do not identify platform as the cause. Conversely, high platform-classification performance establishes information about platform in a feature set, but does not establish that this information worsens a mental health model. Clinical relevance requires evaluation of the intended outcome and use case [14].

The GLOBEM benchmark restricted cross-dataset models to compatible feature classes and excluded Bluetooth, WiFi, and call features that were not common across its 2-institution datasets [13]. In its reimplementation of 2 earlier algorithms, most of the performance loss relative to the original studies was attributed to removing Bluetooth, WiFi, call, message, and battery features [13]. This practical loss of potentially informative communication and proximity measures motivates a feature-level comparability audit. That benchmark used 2 institutions across 2018 and 2019; the present analysis uses 4 annual cohorts from one institution and a different screening protocol.

Measurement Comparability and Predictive Use

Measurement comparability concerns observable events, feature definitions, and health inference. Goldsack and colleagues [15] distinguish technology verification, analytical validation, and clinical validation. Released features support an availability audit and outcome comparison, but simultaneous reference observations would be required to validate sensor accuracy or interchangeability.

Missingness may carry both technical and clinical information. Currey and Torous [16] found that coverage varied with app engagement and other factors in 1178 participants. In a small schizophrenia cohort, data-quality metrics were associated with symptom-related survey responses [17]. Comparing only observed means can therefore omit a relevant difference in the observation process.

Luijken and colleagues [18] showed analytically and through simulation that predictor-measurement differences between development and validation can affect performance. This establishes a possible pathway, not its presence in GLOBEM. We therefore evaluate platform recovery and symptom screening as separate tasks, with distinct outcomes and interpretations.

Between-person and within-person associations can differ. Yang and colleagues [19] separated app-use predictors into individual deviations and person averages and found different associations with loss of interest. Their Android study did not test cross-platform harmonization. It demonstrates why a classifier recovering platform from participant averages cannot establish whether deviations from an individual's baseline predict symptoms.

Objectives and Scope

We used 4 annual GLOBEM college cohorts to characterize platform-associated differences in released sensing features [20,21]. Our objectives were to locate these differences, distinguish phone-derived features from Fitbit measures, assess coverage and preprocessing, and evaluate whether adding selected connectivity features altered retrospective depressive-symptom screening. We use "platform-associated difference" for an observed contrast. We reserve "measurement artifact" for a documented incompatibility in collection or

XSL•FO
RenderX

feature meaning. We do not infer causal confounding, psychological independence, or measurement equivalence from platform classification alone.

We evaluated localization, contribution to platform prediction, sensitivity to transformation, and persistence after psychological adjustment. Coverage, matched-feature, and downstream analyses were exploratory extensions described in the "Methods" section. This secondary analysis was not preregistered and does not validate a clinical device.

Methods

Data Source and Analytical Units

The GLOBEM dataset, version 1.1, released as Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization, contains annual cohorts INS-W_1 through INS-W_4, collected during spring terms from 2018 through 2021 at a US research university. The dataset investigators report 155, 218, 137, and 195 annual participants, representing 705 person-years from 497 unique people. They collected phone data with an app built on the open-source AWARE mobile sensing framework and activity and sleep data with Fitbit, then released derived features processed with RAPIDS [20,21]. We obtained the supplied feature and survey files from the credentialed PhysioNet release [22]. No new participants were recruited for this secondary analysis.

The investigators recruited undergraduates by email and invited previous participants to return. Each study covered a 10-week spring quarter. Participants installed the app on their own phones, received a Fitbit, and completed prestudy, within-study, and poststudy surveys [20,21]. Smartphone ownership was not assigned. Individual demographic variables were unavailable in the supplied analytical files, preventing platform-specific demographic summaries and adjustment.

The release shifts dates by whole weeks to preserve privacy, within-person temporal order, and weekdays [20]. We used these dates for participant-specific survey alignment, not to compare external calendar events. Scheduled duration, released date span, and observed-day count are different quantities. An annual record represents one cohort contribution, not continuous observation for a calendar year.

Our feature sample comprised all annual identifiers in the released all-day feature tables that matched platform metadata. We validated one-to-one platform and survey joins and unique participant-date feature rows. The broader platform files contained 818 records, of which 113 had no corresponding feature record. Their absence is a coverage difference in the supplied files; we could not determine that it represented dropout. The "Results" section distinguishes metadata, feature, and screening samples.

The 705 annual records had 705 distinct literal identifier strings. Public overlap lists and their consuming code identified membership subsets but did not establish individual correspondence across years [23]. We therefore did not infer links from list position, identifier suffixes, or participant measurements. All inferential analyses and model splits were conducted within individual annual cohorts. We neither fitted pooled models assuming 705 independent people nor combined annual effects in an independence-based meta-analysis. Cross-year agreement is descriptive because some underlying people returned. A switcher count and a within-person platform-switching analysis could not be established from verified links.

Ethical Considerations

The original GLOBEM study received University of Washington Institutional Review Board approval (STUDY00003244), and participants provided informed consent before enrollment. Compensation was up to US $245 per quarter, depending on compliance [20,21]. The published consent forms described anonymized data sharing with university collaborators in 2018 and anonymous publication or disclosure to external researchers for further analysis and research in 2019-2021 [21]. We used the controlled-access, deidentified feature-level release and did not seek additional participant consent.

No separate ethics evaluation was undertaken for this secondary analysis. The rationale for not seeking additional review was that the analysis involved existing deidentified records, without participant interaction or access to direct identifiers. This rationale was informed by the definition of a human subject in the US Federal Policy for the Protection of Human Subjects (45 CFR 46.102(e)) [24]. The analysis was limited to existing, deidentified feature-level records released for research through controlled access, without participant recruitment, contact, or intervention. Access was obtained through PhysioNet registration and credentialing after completion of the required CITI Data or Specimens Only Research training, under the dataset's data-use agreement [20]. We did not have access to direct identifiers or attempt reidentification. Only aggregate findings are reported, and the supplementary materials exclude participant-level records and predictions. The source investigators restricted the release to derived features because sensor traces can reveal identities; deidentification does not eliminate all privacy risk [20].

Feature Definitions and Preprocessing

We retained the starting panel-selection rule: up to 6 all-day columns per sensor domain, selected alphabetically from keyword-matched candidates. This deterministic rule yielded 42 columns and did not use mental health outcomes or model performance. It contained 15 raw constructs, 13 normalized versions, and 14 discretized versions. We retained those raw constructs and added their 3 omitted counterparts for matched comparisons, yielding 45 columns. The raw constructs comprised 2 location features (home time and location entropy), 3 screen features (mean unlock duration overall, during exercise-classified locations, and during green-space locations), 2 Bluetooth counts, 2 incoming-call features, 2 connected-WiFi features, 2 Fitbit step summaries, and 2 Fitbit main-sleep summaries. Multimedia Appendix 1 defines each construct and records collection limitations, whereas Multimedia Appendix 2 supplies the complete column map and numerical outputs.

Alphabetical selection limits the panel's representativeness; other summaries could yield different domain contrasts. Holding

XSL•FO
RenderX

the 15 constructs fixed permits transformation comparisons without changing construct composition. We evaluated the 42-column panel as a supplementary diagnostic. The 45 matched columns represent 15 constructs in 3 forms, not independent measurements.

We distinguished 5 phone location/screen constructs from 4 Fitbit constructs. A core phone set contained home time, location entropy, and mean unlock duration without location context. The 2-event connectivity set contained total Bluetooth scans and incoming-call count. The 4-feature connectivity set additionally included scans of the least-frequently detected device and incoming calls from the most frequent contact. The latter contact feature was retained only as an artifact diagnostic: RAPIDS documentation states that iOS identifiers do not preserve repeated-contact identity [25]. It was excluded from downstream screening models.

Historical RAPIDS documentation also uses different screen-episode end points on Android and iOS [26]. Its Bluetooth provider documentation lists Android support, although the GLOBEM exports contain iOS values [27]. We could not verify the historical study-specific adaptation from derived features. Connected-WiFi documentation supports both platforms, but the supplied iOS WiFi values were entirely missing in the first 2 cohorts [28]. These observations preclude assuming that all iOS calls or Bluetooth values were absent, or attributing every observed contrast to a particular operating system restriction. Location estimates also depend on sampled records and feature construction [29].

Home time reflects sampled locations and inferred home, rather than self-reported withdrawal. Location entropy measures dispersion across observed locations in natural logarithm units. Mean unlock duration describes recorded episodes, not all attention to a phone. Bluetooth scans can repeatedly detect the same object; incoming-call counts depend on the events retained by the collection framework and do not measure all communication. Fitbit steps summarize recorded intraday intervals rather than total daily steps, and sleep measures describe wearable-estimated main sleep. The behavioral interpretation remains conditional on these definitions.

We averaged each feature over a participant's observed daily values, leaving an entirely unobserved feature missing. Each participant contributed equally to the between-person comparisons despite unequal observed-day counts. Zero remained a recorded value. Discretized labels (l/m/h or low/medium/high) were explicitly mapped to 0/1/2 before averaging; unrecognized labels triggered an error. The released discretized values are text labels, so generic numeric conversion would treat them as missing values rather than ordered categories. We used the supplied normalized and discretized features. The release defines normalization relative to each participant's median and 5th-to-95th percentile range, and discretization using participant-specific 33rd and 66th percentiles [20]. These are descriptions of released transformations, not transformations estimated from a prediction training fold.

These transformations are retrospective: their reference distributions may include observations outside a screening window, so they were excluded from downstream screening. Averaged ordinal codes describe the observed mix of low, middle, and high categories, not a recovered continuous behavioral scale. Ties and missing days can affect this summary.

Coverage and Feature Comparisons

For every feature and platform, we calculated the number of participants with any value, the proportion of missing released daily entries, observed days per participant, the proportion of observed values equal to zero, and within-person daily variance. The released tables contained 92, 97, 103, and 103 daily rows per participant across the 4 cohorts. These denominators describe the release, not verified sensor uptime or enrollment duration. A sensitivity analysis restricted dates to the interval between each participant's prestudy and poststudy questionnaires, inclusive. These questionnaire dates were not assumed to be exact enrollment boundaries.

Daily missingness divided missing entries by released daily rows; participant-level absence divided records lacking every daily value by the annual platform sample. Observed-day summaries included zero-day participants. Variance used at least 2 observations and the sample-variance denominator. Zero-valued observations remained observed, whereas missing values remained missing. Absent context-specific screen values could also reflect the absence of that context.

Within each cohort, we compared observed participant means with 2-sided Mann-Whitney tests and Cliff δ [30]. Cliff δ was defined as the probability that an iOS value exceeded an Android value minus the reverse probability; ties contributed zero. Negative values indicate lower iOS values. We used 2000 platform-stratified participant resamples for 95% percentile intervals. A bounded-difference interval replaced a degenerate bootstrap interval when all resampled estimates coincided. We applied Benjamini-Hochberg adjustment [31] separately within each cohort's raw, normalized, and discretized feature families; the latter 2 were exploratory. We did not classify small deltas as equivalence.

Cliff δ describes ordering on a −1 to 1 scale; medians and IQRs give original units. Entirely absent features were excluded from their observed-value comparisons.

Stratified resampling preserved platform counts and each person's joint feature vector. Complete empirical rank separation can yield a degenerate bootstrap distribution. The bounded-difference fallback avoids presenting that point mass as population certainty; its formula and derivation are in Multimedia Appendix 1. It does not correct selection or omitted-variable bias. Multiplicity adjustment was within cohort and feature family, without a claim of confirmatory testing or universal error control under arbitrary dependence.

To assess localization without treating correlated features as independent observations, we calculated the difference between the mean absolute Cliff δ in a fixed connectivity set and that in a fixed comparator set. Each participant resample preserved the full feature vector. The main phone-only contrast compared the 2-event connectivity set with the 3 core phone measures, excluding both Fitbit and the invalid contact-identity construct. Comparisons with all 5 phone behavioral measures, Fitbit, and

XSL•FO
RenderX

the 4 connectivity constructs were sensitivity analyses. Across-year averages of absolute deltas were descriptive and were kept distinct from absolute averages of signed deltas.

Platform Prediction and Missing-Data Sensitivity

We used fixed L2-penalized logistic regression (C=1; LIBLINEAR solver, as implemented in scikit-learn [32]) to predict iOS versus Android separately in each cohort. Models used 5 stratified folds, training-fold median imputation, and training-fold standardization. No hyperparameter selection was performed. An entirely missing training column was retained using the imputer's empty-feature handling. We calculated the area under the receiver operating characteristic curve (AUROC) from combined out-of-fold predictions, with iOS as the positive class. Values below 0.50 were reported without post hoc reversal.

Split seeds were identical across feature sets. Entirely missing training columns used the imputer's constant fallback, without held-out values.

Missingness-only models indicate entirely absent participant features, not intermittent daily absence. Missing-at-random assumptions cannot be established from observed rates alone [33].

Validation used one record per annual participant, matching the intended between-person comparison [34]. Bootstrap duplicates stayed together in a fold. We excluded leave-one-year-out validation because unidentified returning participants could occur in both training and evaluation years. No reported estimate represents validation in entirely new people at a new site or time.

The main matched comparison used 13 raw constructs excluding WiFi, their 13 normalized counterparts, and their 13 discretized counterparts. A distinct removal sequence contained 13 constructs, 12 after removing the contact feature, 10 after also removing both Bluetooth features, and 9 after removing the remaining incoming-call count. The final set comprised 5 phone behavioral and 4 Fitbit features. We also evaluated the 3 core phone measures, Fitbit alone, and indicators that participant-level features were entirely missing. WiFi-inclusive and 42-column models were diagnostic analyses rather than the primary comparison.

We repeated raw-panel classification in complete cases and compared all 3 transformations in participants complete on all 39 matched columns. Complete-case results refer to a selected population, so their difference from full-sample results cannot isolate the effect of imputation. For the main raw, normalized, discretized, and 9-feature behavioral models, uncertainty used 500 participant bootstrap samples with preprocessing and cross-validation refitted. Missingness-only and raw complete-case classifiers also used refitted resamples. Duplicate copies of a resampled participant stayed together in a fold. Other diagnostic models, including common-complete-case transformation comparisons, used conditional resampling of fixed out-of-fold predictions, explicitly labeled in the supplementary outputs. No interval treated 5-fold scores as 5 independent observations. Intervals are approximate, conditional on the observed stratum sizes, and can be unstable with small platform groups or near-perfect rankings.

Requiring all 39 columns holds participants constant across transformations. Complete-record logistic coefficients can be unbiased under specified missingness structures [35], but this does not guarantee representative AUROC. Platform-specific retention therefore remains part of the interpretation.

Overlapping training sets make fold errors dependent [36]. The outputs identify the resampling method and usable replicates; conditional intervals omit model-refitting variation.

Psychological Comparisons and Adjustment

We compared prestudy loneliness, perceived stress, resilience, emotion-regulation reappraisal and suppression, and depressive symptoms across platforms using the released common-version scores. Cohort-specific anxiety scores were compared separately: the State-Trait Anxiety Inventory–State (STAI-S) score in the first cohort and the released State-Trait Anxiety Inventory (STAI) score in the remaining cohorts. Benjamini-Hochberg adjustment covered the 7 score comparisons within each cohort. We did not interpret nonsignificant tests as psychological equivalence.

The common adjustment set comprised the 10-item University of California, Los Angeles Loneliness Scale (UCLA-10), the 10-item Perceived Stress Scale (PSS-10), the Brief Resilience Scale (BRS), Emotion Regulation Questionnaire (ERQ) reappraisal and suppression scores, and the released 9-item Center for Epidemiologic Studies Depression (CESD-9) score. Higher scores indicate more of the named construct: greater loneliness, perceived stress, resilience, reappraisal, suppression, or depressive symptoms, respectively. Anxiety used the cohort-specific released instrument rather than assuming that the state and other anxiety versions were identical. We analyzed the supplied scores, without reconstructing questionnaire items or claiming new psychometric validation in this sample.

For total Bluetooth scans and incoming-call counts, we regressed the standardized log (1 + participant mean) on an iOS indicator, first without and then with the 6 common standardized psychological scores. Both models used the same complete-case sample within each cohort. We used ordinary least squares with heteroskedasticity-consistent SEs using the HC3 covariance estimator and t-based intervals [37]. An additional specification included the cohort-specific anxiety score. The contact feature was analyzed separately as an artifact diagnostic. We checked model rank, conditioning, and leverage. This analysis tests sensitivity to measured psychological adjustment; it does not control for income, age, device price, or other unavailable ownership-related characteristics.

Android was the reference category. Coefficients describe SD differences in log-transformed participant means, not raw-count differences. Psychological coefficients were not treated as mediators.

We quantified sensitivity of the adjusted platform coefficients to omitted variables using the point-null robustness value of Cinelli and Hazlett [38]. We refitted the same regressions with classical ordinary least squares covariance for this calculation;

XSL•FO
RenderX

the main coefficient intervals retain HC3 covariance. The robustness value summarizes the equal partial R-squared an omitted factor would need with platform and the outcome, conditional on the specified controls, to explain away the coefficient. It does not estimate actual demographic confounding or test the measurement mechanism. Multimedia Appendix 1 gives the formula and assumptions.

Exploratory Depressive-Symptom Screening

We used the released end-of-term Beck Depression Inventory-II (BDI-II) label, defined by a score greater than 13, as a symptom-screening outcome rather than a clinical diagnosis. We verified agreement between the released label and the score threshold. Predictors were means of raw daily features from the 14 calendar days strictly before the questionnaire date. A feature required at least 7 observed days in that window. Participants were eligible if at least 1 of the 9 candidate predictors met this criterion. Models therefore shared an eligibility rule, with remaining missing features imputed within training folds.

A common end-of-term measure avoided mixing heterogeneous weekly targets, including the first cohort's derived labels [20]. Neither the 14-day window nor the 7-day rule was tuned against performance. Requiring one qualifying predictor retained partial multimodal records, making missingness part of the evaluated use case.

The behavioral model included the 3 core phone measures and 4 Fitbit measures. We compared this with a model adding total Bluetooth scans and incoming-call count, then a model additionally including platform. Intercept-only, platform-only, and core-phone models were supplementary comparators. Contact identity, across-dataset device-selection features, context-specific screen features, and WiFi were excluded. We used the same fixed classifier and participant-level folds, stratifying jointly by outcome and platform in 2019 to 2021 and by outcome alone in 2018, when the smallest joint subgroup had only 2 records.

We reported overall and platform-specific AUROC, average precision, Brier score, log loss, symptom prevalence, and mean predicted probability. Calibration-in-the-large was estimated only in groups with at least 20 participants and at least 5 positive and 5 negative outcomes. Paired participant resampling with model refitting estimated changes after adding connectivity and platform. Resampling preserved outcome-by-platform stratum counts, limiting the uncertainty represented for prevalence and calibration. We used Fisher exact tests for platform versus screening-label associations and, separately, for metadata records lacking all feature records. The latter 4 exploratory tests are reported without multiplicity adjustment. These exploratory models evaluate retrospective screening; presurvey feature windows do not establish prospective validity because upstream released feature construction can use broader participant context.

AUROC measures ranking, whereas average precision depends on outcome prevalence and summarizes precision across recall [39]. Brier score measures squared probability error; log loss penalizes confident errors. Lower values of both are preferable. Metrics used held-out predictions, without selecting a clinical referral threshold.

We estimated calibration-in-the-large by fitting a logistic intercept with the logit of held-out probability as an offset, fixing its coefficient at 1 [40]. We also reported observed/expected ratios, calculated as observed positive outcomes divided by the sum of predicted probabilities. Ideal values are 0 and 1, respectively. These are descriptive estimates without CIs. The minimum-size rule applied to the intercept fit; ratios remained reportable in smaller groups but were treated as unstable. Jointly estimated calibration slopes are retained in the electronic outputs as a separate diagnostic.

Reproducibility

The analysis used Python (version 3.12; Python Software Foundation), *NumPy* 2.5.3, *pandas* 3.0.6, *SciPy* 1.18.1, *scikit-learn* 1.9.1, and *statsmodels* 0.15.0. Multimedia Appendix 2 contains the executable pipeline, pinned dependencies, concise execution instructions, table-export scripts, and aggregate results. We checked sample counts, ordinal decoding, joins, Cliff δ, HC3 estimates, end point windows, and out-of-fold predictions through separate reconstructions. The "Methods" section describes the exploratory design and statistical assumptions; the study was not preregistered. Participant-level source data and intermediate predictions are not redistributed. Multimedia Appendix 3 maps the reporting items to their manuscript locations.

Results

Sample and Feature Availability

The feature sample contained 554 iOS and 151 Android annual records; Android represented 21.4% of 705 records. Its annual share ranged from 6.5% to 27.1% (Table 1). There were 669 records with end-of-term labels and 661 eligible screening windows.

XSL•FO
RenderX

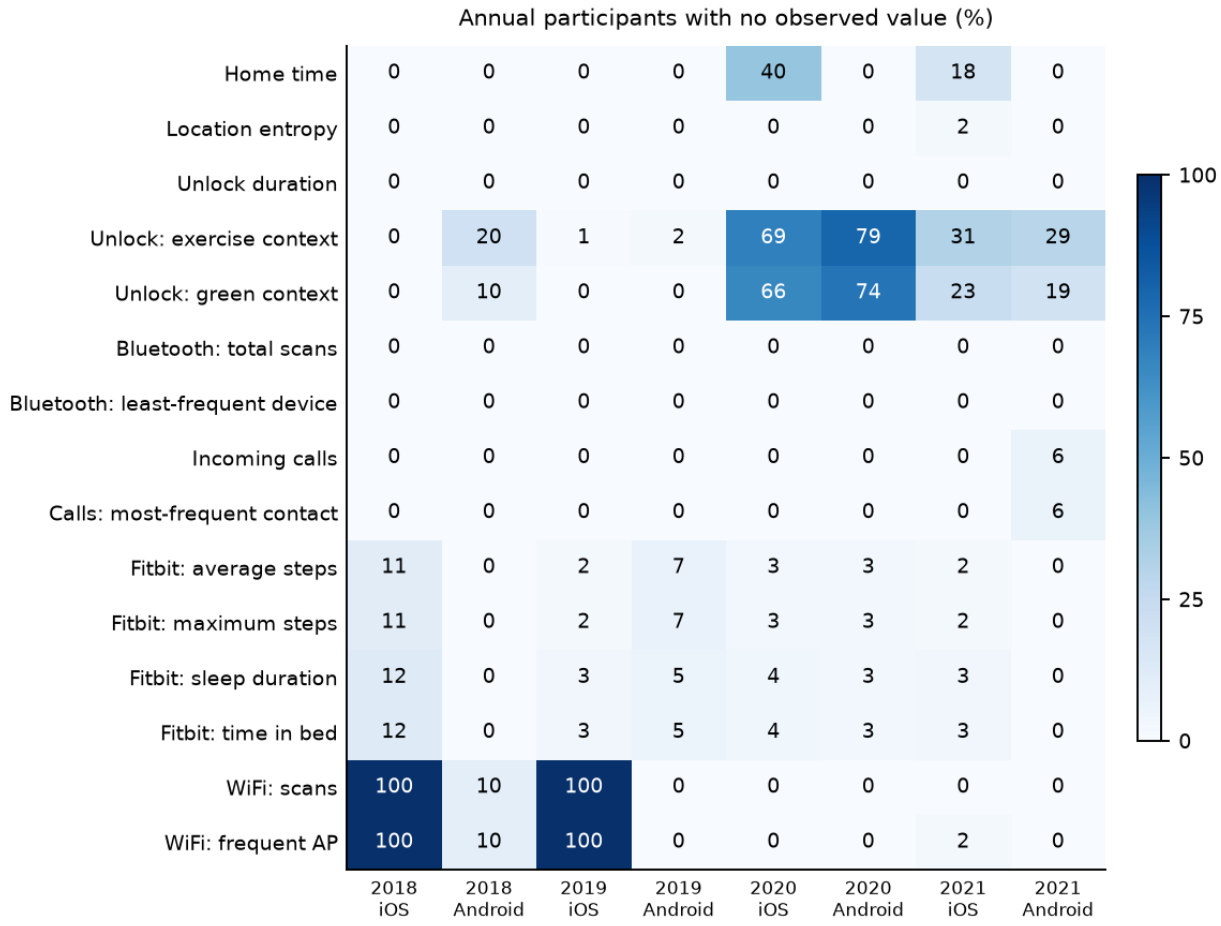
Table 1. Sample flow by platform in the secondary analysis of 4 Generalization of Longitudinal Behavior Modeling US college cohorts, 2018 to 2021[a].

Year	Platform	Metadata, n	Features, n	End-term label, n	Eligible screening, n
2018	iOS	160	145	132	131
2018	Android	41	10	10	10
2019	iOS	172	159	153	152
2019	Android	76	59	50	49
2020	iOS	114	103	102	98
2020	Android	44	34	33	33
2021	iOS	153	147	141	140
2021	Android	58	48	48	48

[a]Feature sample: 705 annual records, including 554 iOS and 151 Android (21.4%). The source reports 497 underlying people. Of 818 metadata records, 113 had no feature record; the reason is unknown. End-term labels were present for 669 feature records; 661 met the 14-day screening eligibility rule. Annual totals are not counts of distinct people across years. End-term label: Beck Depression Inventory-II score greater than 13.

Missingness differed by feature, platform, and cohort, rather than following a uniform phone-versus-wearable pattern (Figure 1; Multimedia Appendices 1 and 2). In Figure 1, absence means no value across all released daily rows, with denominators set by the annual iOS/Android samples of 145/10, 159/59, 103/34, and 147/48 for 2018 to 2021. Percentages are rounded to whole numbers, so 0% does not imply complete daily coverage. Fitbit features are wearable-derived; all others are phone-derived. This measure of missingness does not capture sensor uptime or verified enrollment coverage.

Figure 1. Participants with no observed daily value for each raw feature in a secondary analysis of Generalization of Longitudinal Behavior Modeling US college cohorts, 2018 to 2021. AP: access point.



Home time was entirely missing for 41 of 103 iOS records in 2020 and 26 of 147 in 2021, compared with 0 of 34 and 0 of 48 Android records, respectively. Both WiFi constructs were entirely absent from iOS observations in 2018 and 2019,

XSL•FO
RenderX

whereas at least one value was available for 9 of 10 and 59 of 59 Android records. Context-specific screen features also had limited coverage. Observed-day counts and within-person variance are reported separately from participant means; none of these summaries establishes equal measurement reliability.

Metadata-to-feature loss was greater for Android in every cohort. Missing feature records numbered 15/160 iOS versus 31/41 Android in 2018 (9.4% vs 75.6%; $P<.001$), 13/172 versus 17/76 in 2019 (7.6% vs 22.4%; $P=.002$), 11/114 versus 10/44 in 2020 (9.6% vs 22.7%; $P=.04$), and 6/153 versus 10/58 in 2021 (3.9% vs 17.2%; $P=.002$). These unadjusted Fisher tests describe availability in the supplied files; the cause of absence is unknown (Table S12 in Multimedia Appendix 1).

Total Bluetooth scans were observed for every annual participant on both platforms, with no zeros among observed daily values. Participant-level median imputation was therefore never triggered for this feature. Daily missingness was 31.0% versus 32.0% (iOS vs Android) in 2018, 22.1% versus 28.9% in 2019, 54.2% versus 31.5% in 2020, and 33.5% versus 28.1% in 2021 (Table S11 in Multimedia Appendix 1). Intermittent missingness could still affect participant means. Missingness-only classifier AUROCs ranged from 0.50 to 0.69. The 2020 observed-value home-time contrast uses 62 of 103 iOS records; smaller differences among retained observations do not resolve the missing-data problem.

Magnitude and Localization of Platform Differences

Total Bluetooth scan counts were lower on iOS in all 4 cohorts, with signed Cliff δ from −1.00 to −0.85. Home-time Cliff δ were also negative in every cohort, ranging from −0.32 to −0.21. The broader pattern contained smaller, variable, and sometimes reversed differences in other features. We did not combine these annual estimates into a pooled inferential effect.

The phone-only contrast comparing total Bluetooth and incoming-call counts with home time, entropy, and unlock duration was positive in each cohort. Differences in mean absolute Cliff δ ranged from 0.22 to 0.43, with all 4 participant-resampling intervals excluding zero (Table 2). This result does not depend on including Fitbit in the comparator or on the contact-identity artifact. It describes the selected feature panels, not all possible connectivity and behavioral constructs.

Table 2. Bluetooth counts and a phone-only feature contrast in 4 Generalization of Longitudinal Behavior Modeling US college cohorts, 2018 to 2021[a].

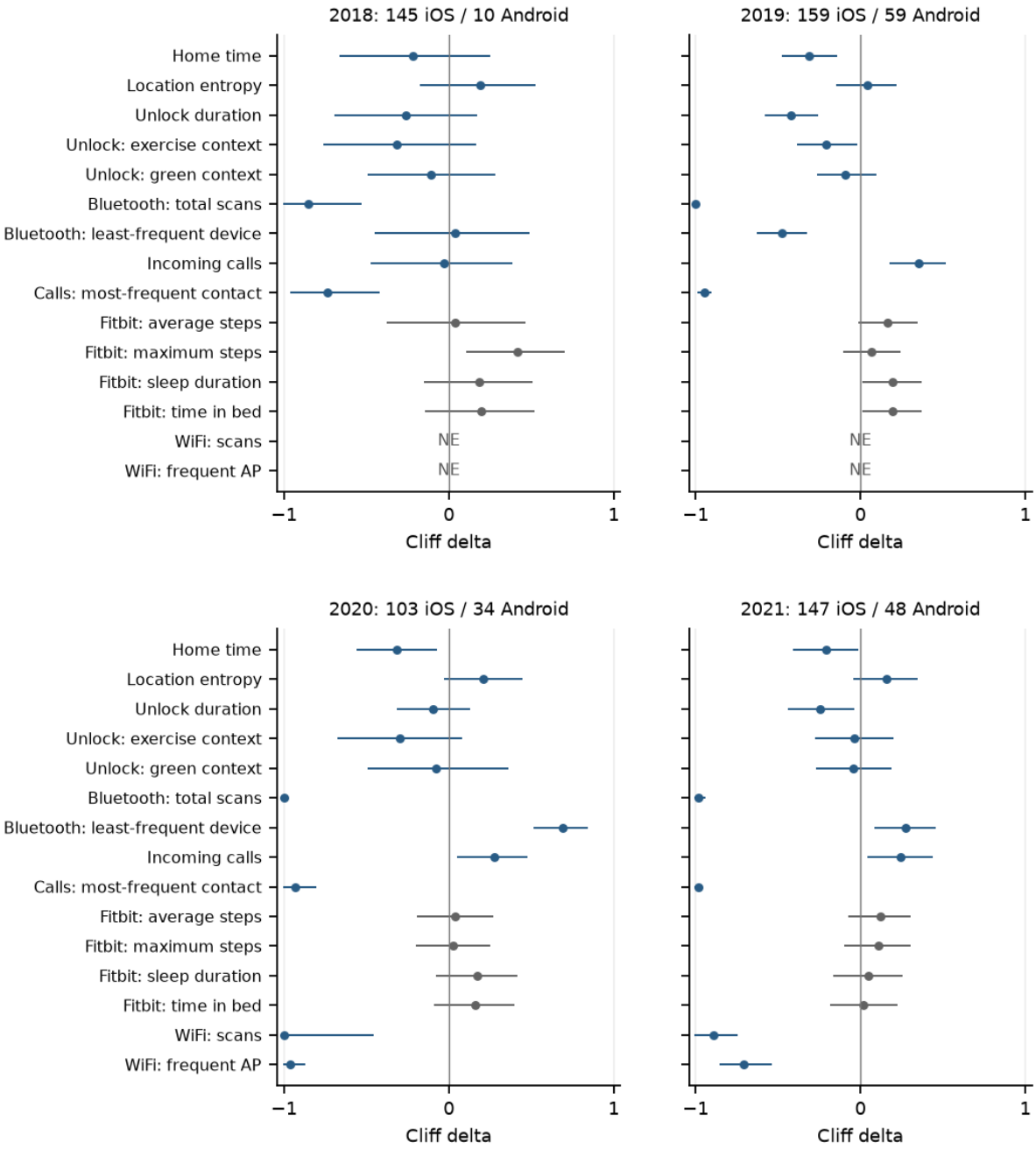
Year	iOS, n; median (IQR)	Android, n; median (IQR)	Bluetooth, Cliff δ (95% CI)	Phone-only contrast, mean absolute Cliff δ (95% CI)
2018	145; 39.3 (28.4 to 57.5)	10; 720.9 (380.8 to 1145.8)	−0.850 (−1.000 to −0.532)	0.216 (0.021 to 0.475)
2019	159; 126.8 (91.5 to 177.3)	59; 4413.3 (2735.6 to 5439.4)	−0.999 (−1.000 to −0.997)	0.416 (0.291 to 0.515)
2020	103; 5.2 (3.0 to 21.4)	34; 588.4 (445.7 to 1434.5)	−0.997 (−1.000 to −0.990)	0.427 (0.242 to 0.582)
2021	147; 12.5 (6.1 to 26.6)	48; 877.5 (324.6 to 1567.5)	−0.980 (−1.000 to −0.943)	0.409 (0.253 to 0.562)

[a]Medians and IQRs are across participant means of daily total Bluetooth scans. Cliff δ is Pr(iOS>Android) minus Pr(iOS<Android); negative values indicate lower iOS values. The phone-only contrast is mean absolute Cliff δ for total Bluetooth scans and incoming-call count minus that for home time, location entropy, and mean unlock duration. Fitbit and contact identity are excluded. Intervals use 2000 within-cohort participant resamples; overlapping people preclude independent cross-year pooling.

Figure 2 summarizes observed-value contrasts across the selected phone (blue) and Fitbit (gray) constructs. Each point is a Cliff δ, positive when iOS values are higher, and each bar is a within-cohort 95% participant-resampling interval, with a conservative bounded fallback where the bootstrap interval was degenerate. Annual platform counts appear above each panel; observed counts vary by feature and are supplied in Multimedia Appendix 2. WiFi contrasts are not estimable where iOS values are wholly missing. Because the annual cohorts share people, the panels are not independent replications.

XSL•FO
RenderX

Figure 2. Platform differences in observed participant means of raw sensing features in Generalization of Longitudinal Behavior Modeling US college cohorts, 2018 to 2021. AP: access point; NE: not estimable.



Platform Prediction and Transformations

The 13 raw constructs classified platform with out-of-fold AUROC of 0.94, 0.99, 0.997, and 0.98 across the 4 cohorts (Table 3). The 2020 estimate was below 1. Tables distinguish near-perfect rankings from exact observed rank separation. Removing all 4 connectivity constructs yielded 9-feature behavioral AUROCs of 0.78, 0.68, 0.41, and 0.48. These results do not support describing all behavioral-only models as chance classifiers or the retained features as platform-neutral.

XSL•FO
RenderX

Table 3. Out-of-fold platform area under the receiver operating characteristic curve and 95% intervals for matched transformations and behavioral features in 4 Generalization of Longitudinal Behavior Modeling US college cohorts, 2018 to 2021[a].

Year	Raw, 13	Normalized, 13	Discretized, 13	Behavioral, 9
2018	0.94 (0.78-1.00)	0.34 (0.08-0.76)	0.65 (0.24-0.84)	0.78 (0.36-0.90)
2019	0.99 (0.98-1.00)	0.91 (0.80-0.95)	0.72 (0.57-0.80)	0.68 (0.53-0.76)
2020	0.997 (0.907-1.000)	0.87 (0.70-0.94)	0.73 (0.62-0.85)	0.41 (0.26-0.64)
2021	0.98 (0.92-1.00)	0.91 (0.79-0.97)	0.80 (0.70-0.88)	0.48 (0.34-0.66)

[a]Cohort sizes are 155, 218, 137, and 195. Positive class: iOS. The 13-construct panels exclude WiFi; raw, normalized, and discretized columns represent the same constructs. The 9-feature panel contains 5 phone location/screen and 4 Fitbit features. All use fixed L2 logistic regression, 5 folds, and training-fold imputation/scaling. Intervals use 500 participant resamples with preprocessing and cross-validation refitted. They are approximate, conditional on observed platform counts. Estimates at or above 0.995 use 3 decimals; an asterisk marks an exact AUROC of 1 in the evaluated predictions. A boundary interval is conditional on the observed sample and does not certify future performance.

Matched normalized AUROCs were 0.34, 0.91, 0.87, and 0.91; discretized AUROCs were 0.65, 0.72, 0.73, and 0.80. Discretization yielded lower platform discrimination in every cohort. The first normalized estimate had a wide interval and only 10 Android records.

Missingness-only AUROC ranged from 0.50 to 0.69. Complete cases for the raw 13-construct panel numbered 135, 206, 23, and 115; the 2020 subset contained only 18 iOS and 5 Android records. Multimedia Appendix 1 reports complete-case models, the distinct removal steps, and sensitivity analyses using the 42-column panel. Multimedia Appendix 2 includes all conditional and refitted intervals and the questionnaire-bounded sensitivity results. The complete-case restriction markedly changed the available sample, especially in 2020.

Raw complete-case AUROCs were 0.89 (95% CI 0.70-1.00), 1.000* (0.994-1.000), 0.97 (0.82-1.00), and 0.99 (0.89-1.00), with refitted intervals. The asterisk marks an exact AUROC of 1 in the evaluated 2019 predictions. Requiring all 39 columns retained the same people, but used conditional intervals (Table S5 in Multimedia Appendix 1). In 2020, normalized AUROC was 0.48 among 23 complete cases versus 0.87 among all 137 records.

Restricting aggregation to dates between each participant's prestudy and poststudy questionnaires retained 143, 205, 131, and 189 annual records. Raw-panel AUROCs remained 0.93, 1.00, 0.99, and 0.98, with the 2019 unrounded estimate below 1. The corresponding 9-feature behavioral AUROCs were 0.74, 0.69, 0.42, and 0.53. Table S9 in Multimedia Appendix 1 supplies conditional intervals for this sensitivity analysis. Core-phone and Fitbit-only classifier results are reported separately in Table S4 in Multimedia Appendix 1, preserving the distinction between phone-derived behavior and wearable measurements.

Paired comparisons qualify the transformation results. In 2021, normalized minus raw AUROC was −0.066 (95% CI −0.191 to 0.014), so a reduction was not established in that cohort. Table S13 in Multimedia Appendix 1 reports every paired transformation contrast with its sign defined consistently. Below-chance estimates, including normalized AUROC 0.34 in 2018, are evaluated without reversing predictions or relabeling classes; small samples and unstable cross-validation can yield such values without establishing a reproducible inverse relationship.

Psychological Measures and Adjusted Feature Differences

None of the 7 psychological-score comparisons met the within-cohort 5% Benjamini-Hochberg threshold. The estimated differences and intervals remain relevant because the absence of a detected difference is not equivalence. Multimedia Appendix 1 reports distributions and effect sizes.

After adjustment for the 6 common psychological scores, the iOS coefficient for standardized log Bluetooth counts was −2.95, −2.12, −1.99, and −2.00 across the 4 cohorts, with all 4 intervals below zero (Table 4). The corresponding samples were 153, 216, 133, and 194. Incoming-call results differed from the contact-specific diagnostic and are reported separately. These estimates establish persistence under the specified adjustment, not that platform choice or sensing behavior is independent of personal characteristics.

XSL•FO
RenderX

Table 4. Cohort-specific platform coefficients before and after psychological adjustment in Generalization of Longitudinal Behavior Modeling US college students, 2018 to 2021[a].

Year	Outcome	n (iOS/Android)	Unadjusted β (95% CI)	Adjusted β (95% CI)	Adjusted *P*
2018	Bluetooth scans	153 (143/10)	−2.85 (−3.79 to −1.92)	−2.95 (−3.87 to −2.03)	<.001
2018	Incoming calls	153 (143/10)	−0.28 (−1.28 to 0.72)	−0.32 (−1.30 to 0.67)	.52
2019	Bluetooth scans	216 (159/57)	−2.12 (−2.24 to −2.01)	−2.12 (−2.24 to −2.01)	<.001
2019	Incoming calls	216 (159/57)	0.64 (0.38 to 0.89)	0.56 (0.30 to 0.82)	<.001
2020	Bluetooth scans	133 (99/34)	−2.00 (−2.17 to −1.83)	−1.99 (−2.15 to −1.83)	<.001
2020	Incoming calls	133 (99/34)	0.47 (0.13 to 0.81)	0.38 (0.05 to 0.70)	.02
2021	Bluetooth scans	194 (146/48)	−2.01 (−2.20 to −1.82)	−2.00 (−2.19 to −1.81)	<.001
2021	Incoming calls	191 (146/45)	0.38 (0.06 to 0.71)	0.34 (0.00 to 0.67)	.048

[a]Outcome units are SDs of log (1 + participant-mean count), standardized in the analysis sample. Android is the reference category. Models include an intercept; adjusted models add standardized prestudy 10-item University of California, Los Angeles Loneliness Scale (UCLA-10), 10-item Perceived Stress Scale (PSS-10), Brief Resilience Scale, Emotion Regulation Questionnaire reappraisal, Emotion Regulation Questionnaire suppression, and 9-item Center for Epidemiologic Studies Depression (CESD-9). Unadjusted and adjusted models use identical complete cases. Intervals and *P* values use HC3 covariance and t inference within cohort. No pooled model or cross-year person clustering is claimed. Anxiety sensitivity and contact-identity diagnostic models are in Multimedia Appendix 2.

The incoming-call coefficient was −0.32 (95% CI −1.30 to 0.67) in 2018 and positive in the 3 later cohorts after adjustment (Table 4). The 2021 incoming-call model included 191 records because 3 Android records lacked the outcome feature; its Bluetooth counterpart included 194. All primary adjusted design matrices had full rank, with 8 parameters, and residual degrees of freedom ranged from 125 to 208. Electronic outputs report anxiety-adjusted models and the most-frequent-contact diagnostic separately.

Point-null robustness values for the adjusted Bluetooth coefficient were 0.636, 0.892, 0.809, and 0.793, compared with 0.075, 0.234, 0.164, and 0.135 for incoming calls (Table S14 in Multimedia Appendix 1). These values concern equal-strength omitted associations under linear regression, not the probability of an unmeasured explanation.

Across the primary adjusted Bluetooth and incoming-call models, design-matrix condition numbers ranged from 4.57 to 9.56 and maximum observation leverage from 0.129 to 0.181. The complete diagnostic records accompany the coefficient estimates in Multimedia Appendix 2.

Exploratory Screening Performance

The 661 eligible annual windows comprised 141, 201, 131, and 188 records. Within these samples, the platform-versus-symptom-label Fisher tests had *P* values of .33, .87, .42, and .31. The first cohort's Android subgroup contained only 2 positive labels among 10 records. Subgroup estimates were therefore particularly uncertain, and its calibration-in-the-large was not estimated.

Positive symptom screens numbered 51 of 141 (36.2%), 86 of 201 (42.8%), 52 of 131 (39.7%), and 75 of 188 (39.9%). For the model including connectivity, the corresponding mean held-out predicted probabilities were 36.8%, 43.4%, 40.6%, and 40.0%. These summaries describe annual screening records; they do not count distinct diagnosed people across cohorts.

The 7-feature behavioral screening models had overall AUROCs of approximately 0.41 to 0.56. Adding total Bluetooth and incoming-call counts changed AUROC by 0.029, 0.023, 0.004, and 0.034; all paired 95% intervals included zero (Table 5). Adding platform to those models also produced intervals spanning zero. Supplementary results report Brier scores, log loss, average precision, platform-specific performance, and calibration. The models did not show consistently useful discrimination or establish a detrimental effect of connectivity features on screening.

XSL•FO
RenderX

Table 5. Exploratory end-of-term depressive-symptom screening in eligible Generalization of Longitudinal Behavior Modeling US college students, 2018 to 2021[a].

Year	n	Behavioral model AUROC[b]	Behavioral + connectivity model AUROC	Behavioral + connectivity + platform model AUROC	Connectivity AUROC change (95% CI)
2018	141	0.56	0.59	0.58	0.029 (−0.045 to 0.088)
2019	201	0.41	0.43	0.41	0.023 (−0.059 to 0.087)
2020	131	0.54	0.54	0.54	0.004 (−0.077 to 0.072)
2021	188	0.52	0.55	0.54	0.034 (−0.036 to 0.101)

[a]Outcome: BDI-II >13. Raw feature means cover the preceding 14 days, excluding the survey date, with at least 7 observations per feature. The behavioral model uses 3 core phone and 4 Fitbit features; connectivity adds total Bluetooth scans and incoming-call count; the final model also adds platform. Cohort models use the same eligible records and folds. Change is connectivity model minus behavioral model. Intervals refit preprocessing and models in 500 paired participant resamples. These are retrospective screening comparisons, not diagnoses, equivalence tests, or causal platform effects. Subgroup and calibration results are supplied in Multimedia Appendices 1 and 2.

[b]AUROC: area under the receiver operating characteristic curve.

All 3 substantive screening models had worse Brier scores than the intercept-only benchmark in each cohort. For the model including connectivity, scores were 0.233, 0.257, 0.260, and 0.253, compared with 0.231, 0.245, 0.240, and 0.240 for the benchmark. These descriptive comparisons do not establish population-level inferiority. Paired Brier-score intervals for connectivity and platform additions included zero throughout (Table S8 in Multimedia Appendix 1). Joint calibration slopes for the connectivity model were 0.39, −1.03, −0.07, and 0.04, indicating unstable or weak probability ordering.

Adding connectivity shifted Android mean predicted risk without an established AUROC gain. Android observed/expected ratios moved from 0.504 to 0.953 in 2018 (10 records), from 0.893 to 0.965 in 2019, and from 0.779 to 0.996 in 2021, but from 0.875 to 0.800 in 2020. The direction was inconsistent across cohorts, and these fixed-prediction estimates have no uncertainty intervals. They do not establish better subgroup calibration in a new sample. Table S7 in Multimedia Appendix 1 reports calibration-in-the-large and observed/expected ratios for both models and all platform groups.

Discussion

Principal Findings

Within GLOBEM, selected connectivity features carried more platform-associated variation than core phone behavioral measures, including after Fitbit features were considered separately. For mHealth apps that interpret communication or proximity as indicators of social engagement, recorded values must therefore be considered alongside their collection methods. A low Bluetooth count cannot, by itself, distinguish limited nearby-device exposure from limited observation. This finding supports feature-specific assessment before clinical interpretation, rather than treating all passive sensing as unreliable.

Smaller location and screen contrasts did not establish comparable monitoring across platforms. Home-time differences were directionally consistent, behavioral features retained platform information in some cohorts, and location-feature coverage differed between iOS and Android. In mobile monitoring between clinical assessments, an unavailable home-time feature must be distinguished from an observed low value. Similar averages among available records do not establish that an app observes behavior equally consistently on both systems.

The depressive-symptom screening experiment did not demonstrate deterioration after adding connectivity features, but weak baseline discrimination and wide paired intervals prevented conclusions about clinical usefulness or absence of harm. For mobile and ubiquitous mental health monitoring, these results support evaluating platform-specific coverage alongside performance for the intended clinical outcome. Before sensing-derived risk estimates inform follow-up or referral, their calibration and usefulness require assessment in the intended population and use case [14,15].

What the Platform Contrast Identifies

Bluetooth counts combine nearby-device exposure, scan opportunity, and feature construction. Ownership-related behavior could affect exposure; hardware and software could affect observation. These components were not observed separately. Near separation of released values therefore does not identify how much of the contrast belongs to operating-system policy, the app, the provider, or behavior.

Contact identity presents a documented construct mismatch. Without stable identifiers, repeated contacts cannot support the same social-network statistic across platforms. Exclusion from shared substantive models follows from the feature definition, not its coefficient size. Total incoming events do not require persistent contact identity. Their direction also differed: iOS values were higher in 2019 to 2021, inconsistent with a blanket claim that iOS call observations were absent. Vendor and AWARE-iOS documentation distinguish restrictions on scanning or identity from the ability to observe call events (Multimedia Appendix 1). The exact historical implementation remains unverified; third-party calling services cannot be assigned responsibility from these exports.

Implications for mHealth Measurement

Individualization can remove a stable additive offset through baseline subtraction, or a stable multiplicative difference through rescaling under suitable assumptions. It cannot restore

XSL•FO
RenderX

unrecorded events or lost contact identity. If observation changes with charging, phone state, or behavior, a fixed baseline may remain inadequate. The daily exports do not distinguish these mechanisms. With far fewer recorded Bluetooth events on iOS, centering also leaves different observation counts. It cannot be assumed to equalize measurement noise or preserve equally precise deviations from each person's baseline.

Discretization can reduce scale differences while discarding distances relevant to symptoms. Lower platform AUROC is therefore insufficient grounds for choosing a mental health representation. Validation should compare retained outcome information and measurement stability using a baseline available before prediction. Our retrospective transformation comparison leaves that prospective question open, especially after a device replacement or software update.

Smaller behavioral contrasts can prioritize features for validation without establishing interchangeability. Paired-device or reference observations would separate agreement from merely similar averages, consistent with measurement-comparison methodology [41]. Observed daily variance mixes behavioral fluctuation with measurement variation. It cannot substitute for repeated reference measurements of the same underlying behavior.

Textbox 1 summarizes practical checks that follow from these findings. Bluetooth scan counts should be excluded from a shared clinical interpretation when comparable observation cannot be justified, or retained as an explicitly evaluated sensitivity input pending measurement validation. This conditional recommendation differs from automatic removal of every connectivity feature. Software and collection choices restrict which scientific questions released traces can answer [42].

These are measurement and reporting checks, not evidence that any resulting screening model is clinically validated.

Textbox 1. Platform audit steps for passive-sensing mobile health studies.

Audit step
• Report platform counts at enrollment, feature availability, and outcome-model eligibility; account for losses between these samples.
• Separate observed zeros, intermittently missing days, entirely absent participant features, and structural noncollection. Report their denominators by platform.
• Verify event definitions and collection capabilities. Exclude or redefine constructs that cannot represent the same quantity on both platforms.
• Use platform classification as a measurement diagnostic, with matched constructs, participant-disjoint validation, and explicit missing-data checks.
• Evaluate the intended health outcome against an intercept benchmark. Compare calibration and probability error by platform alongside discrimination and uncertainty.
• Record operating system, app, hardware, and feature-provider versions, and repeat validation after consequential changes.

Relation to Prior Evidence

Platform variation is a candidate explanation for generalization difficulties, not a demonstrated cause of them [12,13]. Platform-specific modeling also has precedent: Lu and colleagues [43] fitted separate iOS and Android depression models to handle device-platform differences. Earlier connectivity-based symptom studies establish relevance, rather than validating every implementation [2,3]. Location-feature associations have also varied with weekday versus weekend aggregation [44]. Replication needs comparable sources, operational definitions, and observation windows, not just matching column labels.

Currey and Torous [45] found weak passive-only prediction and variation with analytical subset and coverage restriction. In an mHealth study involving depression and diabetes, complete-case deletion retained 950 of 2694 participant-days [46]. Neither study tests GLOBEM's mechanism. They illustrate why apparent feature quality, evaluated sample, and prediction performance must be considered together.

Cross-study coverage findings also caution against ranking platforms universally. Kiang and colleagues [47] found lower GPS noncollection on iOS across 6 digital-phenotyping studies, unlike the iOS home-time absence observed here. Raw GPS noncollection and absence of a derived home-time feature are different endpoints, collected with different systems. This contrast supports conditional interpretation, not a contradiction that either dataset alone can resolve. Shared Fitbit hardware removes one phone-sensor difference by design but does not validate every wearable output in GLOBEM.

Interpreting the Downstream Experiment

The fixed screening model made feature additions comparable but left nonlinear and personalized models untested. Weak baseline discrimination means the behavioral panel did not clearly recover the target symptom signal. Small incremental changes could reflect weak relevance, redundancy, sampling variation, or model misspecification. The null cannot distinguish these explanations. The GLOBEM benchmark also reported weak generalization, but its 2-institution datasets, feature sets, and validation tasks differ from this end-of-term experiment [13]. It cannot establish that weak screening performance here is an inevitable property of these cohorts.

A ranking improvement can coexist with poor calibration or higher probability error [40]. Clinical usefulness also depends on the consequences of decisions at specified thresholds [48]. Our models justify neither a referral cutoff nor a claim of improved patient care. Such evaluation requires a defined clinical action and an appropriate reference outcome.

XSL•FO
RenderX

Small platform groups and missing demographics constrain subgroup interpretation. Similar point estimates would not establish equal errors or benefit without adequate evaluation of specified clinical decisions.

Limitations

The platform was not randomized, and income, age, device price, and related demographic attributes were unavailable. For example, if a characteristic associated with iOS ownership also increased time away from home, it could contribute to the negative home-time contrast; the opposite behavioral association could attenuate that contrast. Direction cannot be assigned without those measurements. The Bluetooth robustness values quantify the strength of omitted associations required under the fitted linear specification, but cannot label socioeconomic explanations implausible or identify a device mechanism. They neither cover arbitrary nonlinear misspecification nor remedy selection into the available-feature sample. Demographics are available from the custodians by special request; none was supplied.

Verified cross-year person links were unavailable. Conducting analyses separately by cohort avoids fitting models that assume repeated annual records are independent, but prevents estimation of person-level random intercepts across years and within-person platform changes. The same individuals may contribute to several annual findings, so consistency across years is not independent replication. We cannot infer that there were no platform switchers. A custodian-provided cross-walk is needed for those analyses.

This is a selected college population from one US institution using one collection and processing system during 2018 to 2021. Platform ratios, hardware, operating system versions, app configurations, and user behavior may differ elsewhere. Android metadata records were more often absent from the feature files, particularly in 2018. If this reflects collection failure, the retained Android sample would selectively represent devices with recorded data. The cause is unknown, so contrasts characterize retained records rather than all enrolled phone owners. The first cohort included only 10 Android records. Complete-case selection was especially severe in 2020 and could change both effect magnitude and classifier performance. These factors limit precision and transportability, including to Android-majority settings.

The release provides derived features rather than the raw observation process. Observed days and within-person variance are coverage diagnostics, not estimates of sensor accuracy or test-retest reliability. The released date range does not identify exact enrollment, and upstream feature construction can depend on participant-level context beyond a prediction window. The historical Bluetooth implementation and context-specific screen construction could not be fully verified. Our results apply to these exports and cannot attribute every difference to the operating system alone.

The alphabetical panel rule did not prioritize commonly modeled unlock counts, unique Bluetooth devices, or outgoing calls. Holding constructs fixed makes transformation comparisons interpretable, but cannot establish that these constructs represent their domains. The screening experiment used one fixed model family, a limited feature panel, one thresholded symptom measure, and one retrospectively defined window. It did not assess diagnosis, onset forecasting, intervention effects, or deployment under a new platform distribution. Small outcome-by-platform groups constrained calibration and subgroup comparisons. Bootstrap intervals are approximate and conditional on observed stratum counts; they do not capture all uncertainty in prevalence, model specification, or a new setting. Nonsignificant paired differences therefore cannot establish clinical equivalence or fairness.

Conclusions

Within GLOBEM's 4 annual cohorts, selected connectivity features showed larger platform-associated differences than core phone behavioral features, while coverage differences also affected location. Matched transformations changed platform predictability. Because the screening models lacked demonstrated predictive usefulness, their incremental comparisons cannot establish or exclude downstream harm from platform-associated measurement differences. A shared feature name is insufficient evidence of comparable measurement. The unresolved question is which differences alter clinically useful monitoring when observation quality, owner characteristics, and device changes can be distinguished.

Acknowledgments

The authors used the Generalization of Longitudinal Behavior Modeling (GLOBEM) datasets, made available by the University of Washington and collaborating institutions through PhysioNet, and thank the original investigators for releasing this resource. The authors thank colleagues at the American University of the Middle East and the University of Wolverhampton for helpful discussions.

Data Availability

The datasets analyzed during this study are available in the PhysioNet repository (GLOBEM Dataset, version 1.1) [20]. They are not publicly available because passive-sensing data carry a risk of reidentification; access requires PhysioNet credentialing, completion of the required training, and a signed data-use agreement. Demographic variables are available only from the data custodians on special request. The analysis code and aggregate results generated during this study are included in Multimedia Appendix 2; participant-level records and predictions are not redistributed.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

XSL•FO
RenderX

Generative AI Use

ChatGPT 5.5 assisted with editing, proofreading, and citation check. The authors are responsible for the analysis and final text and for the accuracy and integrity of the work.

Authors' Contributions

Conceptualization: AN
Data curation: AN
Formal analysis: AN
Methodology: AN, FN
Software: AN
Validation: AN
Visualization: AN
Writing – original draft: AN
Writing – review & editing: AN, FN

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary methods, feature provenance, psychological comparisons, and sensitivity analyses for platform-associated sensing differences in 4 US college cohorts, 2018 to 2021.
[DOCX File , 87 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Executable analysis code, aggregate result tables, pinned software dependencies, and instructions for reproducing the analyses, tables, and figures.
[ZIP File (Zip Archive), 132 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Reporting checklist mapping for observational analyses and exploratory prediction models.
[DOCX File , 24 KB-Multimedia Appendix 3]

References

1. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res. 2015;17(7):e175. [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
2. Place S, Blanch-Hartigan D, Rubin C, Gorrostieta C, Mead C, Kane J, et al. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. J Med Internet Res. 2017;19(3):e75. [FREE Full text] [doi: 10.2196/jmir.6678] [Medline: 28302595]
3. Opoku Asare K, Terhorst Y, Vega J, Peltonen E, Lagerspetz E, Ferreira D. Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. JMIR Mhealth Uhealth. 2021;9(7):e26540. [FREE Full text] [doi: 10.2196/26540] [Medline: 34255713]
4. Doryab A, Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data. JMIR Mhealth Uhealth. 2019;7(7):e13209. [FREE Full text] [doi: 10.2196/13209] [Medline: 31342903]
5. Wang Y, Ren X, Liu X, Zhu T. Examining the correlation between depression and social behavior on smartphones through usage metadata: empirical study. JMIR Mhealth Uhealth. 2021;9(1):e19046. [FREE Full text] [doi: 10.2196/19046] [Medline: 33404512]
6. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol. 2017;13:23-47. [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
7. Wang R, Wang W, Dasilva A, Huckins JF, Kelley WM, Heatherton TF, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2018;2(1):43. [doi: 10.1145/3191775] [Medline: 39449996]

XSL•FO
RenderX

8. Chikersal P, Doryab A, Tumminia M, Villalba DK, Dutcher JM, Liu X, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. ACM Trans Comput-Hum Interact. 2021;28(1):1-41. [doi: 10.1145/3422821]
9. Halabi R, Selvarajan R, Lin Z, Herd C, Li X, Kabrit J, et al. Comparative assessment of multimodal sensor data quality collected using android and iOS smartphones in real-world settings. Sensors (Basel). 2024;24(19):6246. [FREE Full text] [doi: 10.3390/s24196246] [Medline: 39409286]
10. Vega J, Li M, Aguillera K, Goel N, Joshi E, Khandekar K, et al. Reproducible analysis pipeline for data streams: open-source software to process data collected with mobile devices. Front Digit Health. 2021;3:769823. [FREE Full text] [doi: 10.3389/fdgth.2021.769823] [Medline: 34870271]
11. Trifan A, Oliveira M, Oliveira JL. Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. JMIR Mhealth Uhealth. 2019;7(8):e12649. [FREE Full text] [doi: 10.2196/12649] [Medline: 31444874]
12. Müller SR, Chen XL, Peters H, Chaintreau A, Matz SC. Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. Sci Rep. 2021;11(1):14007. [FREE Full text] [doi: 10.1038/s41598-021-93087-x] [Medline: 34234186]
13. Xu X, Liu X, Zhang H, Wang W, Nepal S, Sefidgar Y, et al. GLOBEM: cross-dataset generalization of longitudinal human behavior modeling. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2022;6(4). [doi: 10.1145/3569485]
14. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med. 2019;2:88. [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
15. Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). NPJ Digit Med. 2020;3:55. [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
16. Currey D, Torous J. Increasing the value of digital phenotyping through reducing missingness: a retrospective review and analysis of prior studies. BMJ Ment Health. 2023;26(1):e300718. [FREE Full text] [doi: 10.1136/bmjment-2023-300718] [Medline: 37197799]
17. Torous J, Staples P, Barnett I, Sandoval LR, Keshavan M, Onnela JP. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. NPJ Digit Med. 2018;1:15. [FREE Full text] [doi: 10.1038/s41746-018-0022-8] [Medline: 31304300]
18. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. Stat Med. 2019;38(18):3444-3459. [FREE Full text] [doi: 10.1002/sim.8183] [Medline: 31148207]
19. Yang X, Knights J, Bangieva V, Kambhampati V. Association between the severity of depressive symptoms and human-smartphone interactions: longitudinal study. JMIR Form Res. 2023;7:e42935. [FREE Full text] [doi: 10.2196/42935] [Medline: 36811951]
20. Xu X, Zhang H, Sefidgar Y, Ren Y, Liu X, Seo W, et al. GLOBEM dataset: multi-year datasets for longitudinal human behavior modeling generalization (version 1.1). PhysioNet. 2023. URL: https://physionet.org/content/globem/1.1/
21. Xu X, Zhang H, Sefidgar Y, Ren Y, Liu X, Seo W, et al. GLOBEM dataset: multi-year datasets for longitudinal human behavior modeling generalization. 2022. Presented at: Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track; November 28 to December 9, 2022:24655-24692; New Orleans, LA. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c7e8a0821dfcb58a9a83cbd37cc8131-Abstract-Datasets_and_Benchmarks.html [doi: 10.52202/068431-1790]
22. Pollard T, Moody BE, Lehman LH, Gow BJ, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. Nat Health. 2026;1(8):792-795. [doi: 10.1038/s44360-026-00096-z] [Medline: 42841059]
23. UW-EXP. GLOBEM. GitHub. URL: https://github.com/UW-EXP/GLOBEM/tree/4f140fc5290dc97204298cca28b956165aa0a29f [accessed 2026-09-22]
24. Definitions for purposes of this policy. Electronic Code of Federal Regulations (eCFR). URL: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.102 [accessed 2026-10-25]
25. Phone calls. RAPIDS. URL: https://www.rapids.science/1.6/features/phone-calls/ [accessed 2026-09-22]
26. Phone screen. RAPIDS. URL: https://www.rapids.science/1.6/features/phone-screen/ [accessed 2026-09-22]
27. Phone bluetooth. RAPIDS. URL: https://www.rapids.science/1.6/features/phone-bluetooth/ [accessed 2026-09-22]
28. Phone WiFi connected. RAPIDS. URL: https://www.rapids.science/1.6/features/phone-wifi-connected/ [accessed 2026-09-22]
29. Phone locations. RAPIDS. URL: https://www.rapids.science/1.6/features/phone-locations/ [accessed 2026-09-22]
30. Cliff N. Dominance statistics: ordinal analyses to answer ordinal questions. Psychol Bull. 1993;114(3):494-509. [doi: 10.1037/0033-2909.114.3.494]
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Methodol. 1995;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
32. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: a library for large linear classification. J Mach Learn Res. 2008;9:1871-1874. [FREE Full text]
33. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-592. [doi: 10.2307/2335739]

XSL•FO
RenderX

34. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. Gigascience. 2017;6(5):1-9. [FREE Full text] [doi: 10.1093/gigascience/gix019] [Medline: 28327985]
35. Bartlett JW, Harel O, Carpenter JR. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. Am J Epidemiol. 2015;182(8):730-736. [FREE Full text] [doi: 10.1093/aje/kwv114] [Medline: 26429998]
36. Bengio Y, Grandvalet Y. No unbiased estimator of the variance of K-fold cross-validation. J Mach Learn Res. 2004;5:1089-1105. [FREE Full text]
37. MacKinnon JG, White H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J Econom. 1985;29(3):305-325. [doi: 10.1016/0304-4076(85)90158-7]
38. Cinelli C, Hazlett C. Making sense of sensitivity: extending omitted variable bias. J R Stat Soc Series B Stat Methodol. 2020;82(1):39-67. [doi: 10.1111/rssb.12348]
39. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
40. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. [FREE Full text] [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
41. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;327(8476):307-310. [doi: 10.1016/S0140-6736(86)90837-8]
42. Onnela JP. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology. 2021;46(1):45-54. [FREE Full text] [doi: 10.1038/s41386-020-0771-3] [Medline: 32679583]
43. Lu J, Shang C, Yue C, Morillo R, Ware S, Kamath J, et al. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2018;2(1):1-21. [doi: 10.1145/3191753]
44. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ. 2016;4:e2537. [FREE Full text] [doi: 10.7717/peerj.2537] [Medline: 28344895]
45. Currey D, Torous J. Digital phenotyping correlations in larger mental health samples: analysis and replication. BJPsych Open. 2022;8(4):e106. [FREE Full text] [doi: 10.1192/bjo.2022.507] [Medline: 35657687]
46. Sarda A, Munuswamy S, Sarda S, Subramanian V. Using passive smartphone sensing for improved risk stratification of patients with depression and diabetes: cross-sectional observational study. JMIR Mhealth Uhealth. 2019;7(1):e11041. [FREE Full text] [doi: 10.2196/11041] [Medline: 30694197]
47. Kiang MV, Chen JT, Krieger N, Buckee CO, Alexander MJ, Baker JT, et al. Sociodemographic characteristics of missing data in digital phenotyping. Sci Rep. 2021;11(1):15408. [FREE Full text] [doi: 10.1038/s41598-021-94516-7] [Medline: 34326370]
48. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. [FREE Full text] [doi: 10.1177/0272989X06295361] [Medline: 17099194]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BDI-II: Beck Depression Inventory-II
BRS: Brief Resilience Scale
CESD-9: 9-item Center for Epidemiologic Studies Depression score
ERQ: Emotion Regulation Questionnaire
GLOBEM: Generalization of Longitudinal Behavior Modeling
mHealth: mobile health
PSS-10: 10-item Perceived Stress Scale
RAPIDS: Reproducible Analysis Pipeline for Data Streams
STAI: State-Trait Anxiety Inventory
STAI-S: State-Trait Anxiety Inventory–State
UCLA-10: 10-item UCLA Loneliness Scale

XSL•FO
RenderX

Edited by A Stone; submitted 05.Jun.2026; peer-reviewed by A Tai, D Hidalgo-Mazzei; comments to author 11.Sep.2026; revised version received 25.Oct.2026; accepted 26.Oct.2026; published 16.Nov.2026

Please cite as:
Naqvi A, Naqvi F
Platform Differences in Smartphone Sensing Features for Mental Health Across Four Annual College Cohorts: Secondary Analysis
JMIR Mhealth Uhealth 2026;14:e103711
URL: https://mhealth.jmir.org/2026/1/e103711
doi: 10.2196/103711
PMID:

©Ariz Naqvi, Farah Naqvi. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 16.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX