Digital Phenotyping Is Not a Shared Language: A Construct Audit of Public Mental-Health Datasets

Zefeng Sun
School of Educational Psychology and Counselling
Faculty of Education, Monash University
zsun0074@student.monash.edu

Abstract

Public digital-phenotyping datasets share names for behaviors and mental-health outcomes while measuring them through different devices, processing rules, and instruments. We audit StudentLife, the College Experience Study (CES), CrossCheck, and Depresjon across feature definitions, temporal aggregation, population and study design, and outcome labels. Eight CES–CrossCheck daily correspondences yield raw-unit Hellinger distances from 0.104 for time in vehicle to 0.517 for unlock duration. Sensitivity analyses change the apparent similarity of several features: mobility distance rises from 0.148 to 0.307 after soft-outlier removal. Labels span PHQ-9, PHQ-2/PHQ-4, single-item and composite EMA, and clinician-rated MADRS. The effect audit separates between-person summaries from within-person variation. Between-person phone-feature correlations are small and uncertain; within-person stillness associations are 0.082 in CES and 0.053 in CrossCheck, with different outcome definitions. Depresjon shows lower mean activity in clinical cases than controls ($d = -0.547$). These findings make feature definition, preprocessing, label content, and statistical estimand explicit parts of cross-dataset evaluation. We provide field-level crosswalks, aggregate results, and reconstruction scripts for multimodal mental-health research.

1 Introduction

Digital phenotyping uses personal-device traces to characterize everyday behavior (Jain et al., 2015; Onnela and Rauch, 2016; Insel, 2017). Phones and wearables provide repeated observations of movement, routines, communication, and device use alongside self-report and clinical assessments (Torous et al., 2015, 2016; Montag et al., 2020). Reusing these observations across studies requires a precise account of the measurements being transferred.

A feature called *activity* may denote a phone-inferred state, its accumulated duration, a sample proportion, or wrist accelerometer counts. A label called *depression* may denote a two-item screen, a nine-item symptom inventory, momentary mood, or clinician-rated severity. Their source, unit, window, item content, and sampling schedule determine what a pooled or transferred model learns.

This problem also arises in multimodal mental-health NLP. A language representation can be paired with behavioral features and symptom labels collected under different protocols. Shared text encoders preserve neither the meaning of a sensing variable nor the target instrument. An explicit compatibility record makes it possible to distinguish language-model transport from changes in the behavioral inputs and outcome definition. This extends the provenance and intended-use questions emphasized by data statements and datasheets (Bender and Friedman, 2018; Gebru et al., 2021).

We conduct a purposive audit of four public resources: StudentLife, CES, CrossCheck, and Depresjon (Wang et al., 2014; Nepal et al., 2024; Wang et al., 2016; Garcia-Ceja et al., 2018). They provide contrasts across student and clinical populations, phone and wrist sensing, raw and processed releases, and several outcome instruments. We ask:

1. **RQ1:** Which released behavioral measurements support documented cross-dataset correspondences?
2. **RQ2:** How do their empirical distributions and preprocessing sensitivities compare?
3. **RQ3:** How do outcome content, administration, and availability affect label alignment?
4. **RQ4:** How do behavior–outcome associations vary by dataset, label, and between-person or within-person estimand?

© 2026 Association for Computational Linguistics (ACL). All Rights Reserved. This work is licensed under a Creative Commons Attribution 4.0 International License.

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 26–34
November 9, 2026 ©2026 Association for Computational Linguistics

The contributions are a field-level feature and label crosswalk, a participant-weighted distribution diagnostic, an association audit with explicit time alignment, and a reporting record for cross-dataset computational studies.

2 Related Work

Passive sensing studies have linked depressive symptoms with mobility, routines, and social behavior (Saeb et al., 2015; Canzian and Musolesi, 2015; Wang et al., 2018; Huckins et al., 2020). Devices, feature extraction, samples, and validation practices vary substantially across this literature (Cornet and Holden, 2018). EMA adds frequent self-reports whose interpretation depends on question wording, recall period, and prompt schedule (Shiffman et al., 2008).

Construct validity connects observed measurements to the constructs used in scientific claims (Cronbach and Meehl, 1955; Messick, 1995); measurement invariance examines the stability of that relationship across settings (Vandenberg and Lance, 2000). Depression instruments overlap only partially in symptom content (Fried, 2017; Fried and Nesse, 2015). For computational studies, these differences enter both the input representation and the supervised target.

Data statements, datasheets, model cards, and measurement-focused fairness research make such choices explicit (Bender and Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019; Jacobs and Wallach, 2021). Work on data cascades and abstraction describes how upstream decisions affect downstream systems (Sambasivan et al., 2021; Selbst et al., 2019; Paullada et al., 2021). GLOBEM evaluates cross-dataset depression detection and domain generalization across harmonized longitudinal sensing datasets (Xu et al., 2022). Our audit describes the measurement relationships that a harmonization and evaluation protocol must specify.

3 Datasets and Methods

3.1 Data releases and case selection

The selection criteria were public access, passive behavioral sensing, a mental-health outcome, and variation in measurement design. StudentLife offers Android streams from a 10-week university study in 2013, with PHQ-9 pre/post assessments. We use its Zenodo RDS conversion. CES follows college students over four years including the pandemic and releases Android/iOS sensing features and repeated PHQ-4 items. CrossCheck's public smartphone arm contains daily/hourly Eureka features and EMA from a year-long schizophrenia-spectrum study. Depresjon contains minute-level Actiwatch recordings from 23 patients with unipolar or bipolar depression and 32 controls, plus MADRS totals for the patient group. These four cases represent selected measurement contrasts; the source manifest records the releases retrieved on 5 September 2026.

3.2 Feature crosswalk and analytic days

The author audited the released dictionaries and fields along four axes: sensing and feature definition, temporal aggregation, population and study design, and outcome measurement. A *documented daily correspondence* has a common described behavior, unit, and whole-day window. An *approximate bridge* relates different operationalizations of a broader construct. Table 2 gives the eight CES–CrossCheck correspondences. Correspondence describes the available release documentation; source extraction algorithms can still differ.

We preserve missing values and exclude invalid IDs, dates, and negative or nonfinite feature values. CrossCheck contains 69 duplicate study-ID/day keys. For each feature we select the row with highest relevant quality, then the other quality field, then lexicographically smallest Eureka ID. Event fields use activity quality followed by location quality. Twenty-seven duplicate keys have conflicting EMA values; the same feature-specific rule selects the label row. Positive extremes remain in the base distribution. Values outside $Q_1 - 3\mathrm{IQR}$ to $Q_3 + 3\mathrm{IQR}$, calculated within each dataset/feature among eligible days, are flagged for a separate sensitivity analysis.

StudentLife activity states are 0=still, 1=walking, 2=running, and 3=unknown. We derive daily state proportions using UTC dates because the converted stream lacks a verified local-time convention. State 3 remains an unknown-state feature. Depresjon daily means, medians, 90th percentiles, and zero-count fractions require at least 720 valid minutes. Observed-day averages define the participant-level summaries for these two releases.

3.3 Distribution diagnostics

For each CES–CrossCheck feature, participants contribute at least three eligible days. Each participant's day histogram sums to one; averaging these histograms gives participants equal weight. We

Table 1: Released measurement layers and outcome availability. Study enrollment, source identifiers, and analytic samples are distinct denominators. CES demographics contain 216 identifiers; sensing and general EMA contain 220. StudentLife sensing contains 49 identifiers, compared with 48 enrolled students.

Dataset	Population and source layer	Sensing	Outcome observations
StudentLife	48 enrolled students; raw/inferred RDS streams	Android phone	PHQ-9: 46 pre, 38 post; 38 paired
CES	College students; released daily/hourly features	Android/iOS phone	35,348 complete PHQ-4 rows, 218 people
CrossCheck	61 valid study IDs; public smartphone arm	Android phone, Eureka features	6,176 complete ten-item EMA rows after ID/date exclusions
Depresjon	23 clinical cases, 32 controls; minute records	Wrist Actiwatch AW4	Two MADRS totals for each of 23 cases

Table 2: CES–CrossCheck field crosswalk. Both releases use the listed `ep_0` fields for the full day. Activity/location quality is measured in hours. CES call fields cover an Android subsample. Complete source and eligibility records accompany the paper.

Behavior	Field in both daily releases	Unit	Day eligibility
Stillness	`act_still_ep_0`	seconds	activity quality $\geq$ 12 h
Time in vehicle	`act_in_vehicle_ep_0`	seconds	activity quality $\geq$ 12 h
Time bicycling	`act_on_bike_ep_0`	seconds	activity quality $\geq$ 12 h
Mobility distance	`loc_dist_ep_0`	meters	location quality $\geq$ 12 h
Unlock count	`unlock_num_ep_0`	count	observed finite value
Unlock duration	`unlock_duration_ep_0`	seconds	observed finite value
Incoming calls	`call_in_num_ep_0`	count	observed finite value
Outgoing calls	`call_out_num_ep_0`	count	observed finite value

pool the two datasets' eligible values, construct 50 equal-width bins between their pooled 1st and 99th percentiles, and add two tail bins. We calculate

$$H(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}.$$

The base comparison preserves the recorded units. A fixed-bin KS diagnostic uses the maximum difference of the same binned cumulative distributions. We resample participants independently within each dataset 1,000 times, holding bin edges fixed, and report percentile intervals.

Sensitivity analyses change participant/day weighting, minimum days (1/3/7), quality thresholds (6/12/18 hours), soft-outlier inclusion, and distribution scaling. Dataset-specific robust scaling subtracts the median and divides by IQR/1.349, with MAD and unit-scale fallbacks. It compares standardized shape. Approximate StudentLife–phone and phone–Actiwatch activity bridges use this shape comparison; their outputs retain the original units and definitions. Supplementary bin-count checks use 25, 50, and 100 bins.

3.4 Labels and association estimands

The label crosswalk records item content, rater, score range, recall period, schedule, and observed coverage. PHQ-9 items are scored 0–3 and summed over nine items (Kroenke et al., 2001). In CES we reconstruct PHQ-2 from items `phq4-3` and `phq4-4`, GAD-2 from `phq4-1` and `phq4-2`, and PHQ-4 from all four (Kroenke et al., 2009). CrossCheck's primary association outcome is its negative EMA composite: depressed mood, harm, seeing things, stress, and voices. The single depressed-mood item and net EMA score are sensitivity outcomes. Depresjon supplies MADRS totals (Montgomery and Asberg, 1979) and the case/control grouping.

CES and CrossCheck features are joined to labels on participant and released day, followed by a three-paired-day minimum. Between-person Spearman correlations compare each participant's mean feature and mean label over paired days. Within-person Pearson correlations pool paired daily values after subtracting each participant's feature and label means; each paired day has equal weight. The participant bootstrap resamples whole centered series. Person-centered Spearman results accompany these estimates as a sensitivity analysis. CES PHQ items retain their two-week recall period even when joined by day.

StudentLife pairs the full observed-period still-state fraction with pre and post PHQ-9 separately.

Table 3: Recorded-unit CES–CrossCheck distances. Participants are equally weighted and contribute at least three eligible days. The CI is a 95% participant-bootstrap interval.

Feature	n CES/CC	H	95% CI
Time in vehicle	218/60	0.104	[0.071, 0.163]
Incoming calls	34/61	0.130	[0.092, 0.188]
Mobility distance	216/60	0.148	[0.138, 0.171]
Time bicycling	218/60	0.166	[0.139, 0.206]
Outgoing calls	34/61	0.232	[0.188, 0.286]
Stillness	218/60	0.361	[0.299, 0.433]
Unlock count	218/61	0.446	[0.394, 0.504]
Unlock duration	218/61	0.517	[0.479, 0.564]

Depresjon pairs observed-period mean activity with MADRS1 in the 23 cases and reports a pooled-SD Cohen's d for case minus control activity. Its bootstrap resamples cases and controls separately. These participant-level analyses use the release's assessment identifiers rather than daily assessment dates. All intervals use 1,000 resamples with recorded seeds; they are unadjusted for multiple comparisons. Associations are unadjusted for demographic and clinical covariates.

4 Results

4.1 RQ1: Correspondence depends on the released measurement

CES and CrossCheck support the eight daily correspondences in Table 2. Their analytic coverage varies by feature: calls represent 34 CES Android participants, whereas activity summaries represent 218 CES participants. Phone-inferred still duration, StudentLife's sampled still-state fraction, and Actiwatch zero-count fraction support approximate inactivity bridges. They measure related behaviors at different sensor and aggregation layers. Sleep requires a separate derivation and validation protocol: the available releases provide different combinations of self-reports, inferred routines, and actigraphy. The crosswalk records those available sources and the resulting comparison tier.

4.2 RQ2: Similarity depends on feature and preprocessing

Table 3 reports base distances with effective participant counts and uncertainty. Time in vehicle has the lowest base distance ($H = 0.104$); unlock duration has the highest (0.517). The call rows describe a platform-specific CES subset. Participant-day counts and missingness are provided with the aggregate results.

Figure 1 shows that this ordering depends on preprocessing. Removing soft outliers changes mobility distance from 0.148 to 0.307 and bicycling from 0.166 to 0.528. The unit audit documents extreme released distance values in both datasets. After robust scaling, time-in-vehicle distance is 0.557, compared with 0.104 in recorded units. Raw-unit support and standardized shape therefore yield different diagnostics. Unlock duration remains high across these settings (0.507–0.592).

At 25/50/100 bins, bicycling distances are $0.122/0.166/0.209$, while unlock-duration distances are $0.514/0.517/0.519$. Other checks locate particular sources of sensitivity. Time-in-vehicle H changes from 0.084 to 0.139 as the quality threshold rises from 6 to 18 hours; mobility is approximately 0.148, 0.148, and 0.145 at 6/12/18 hours. Incoming-call distance falls from 0.130 overall to 0.087 on positive days, while unlock-count distance changes from 0.446 to 0.399. These comparisons show how coverage and zero-day mass enter the observed distance. Approximate stillness bridges yield $H = 0.188$ for StudentLife–CES, 0.383 for StudentLife–CrossCheck, and 0.500 for CrossCheck–Depresjon after robust scaling.

4.3 RQ3: Outcomes differ in content, time, and coverage

Table 4 summarizes the outcome relationships. PHQ-9 and CES PHQ-4 share the anhedonia and depressed-mood item concepts. CES's two remaining items assess anxiety. CrossCheck's negative composite additionally includes psychosis-related experiences, while its depressed-mood item supports a narrower semantic link. The public CrossCheck documentation leaves its full instrument version and recall anchor unspecified. MADRS differs in rater and item structure.

Availability is a further measurement property. CES has 35,348 complete PHQ-4 rows among 217,155 general EMA rows (16.3%); PHQ items are absent on the other rows. The stored PHQ-4 total disagrees with the item sum on 182 complete rows, so the audit uses reconstructed scores. PHQ-2 is zero on 52.3% of complete PHQ-4 rows. CrossCheck has 6,176 complete ten-item EMA rows among 22,079 rows after ID/date exclusions (28.0%). These row proportions describe availability in the released tables; prompt scheduling and nonresponse both contribute. Depresjon provides MADRS for 23 cases and leaves it blank for 32

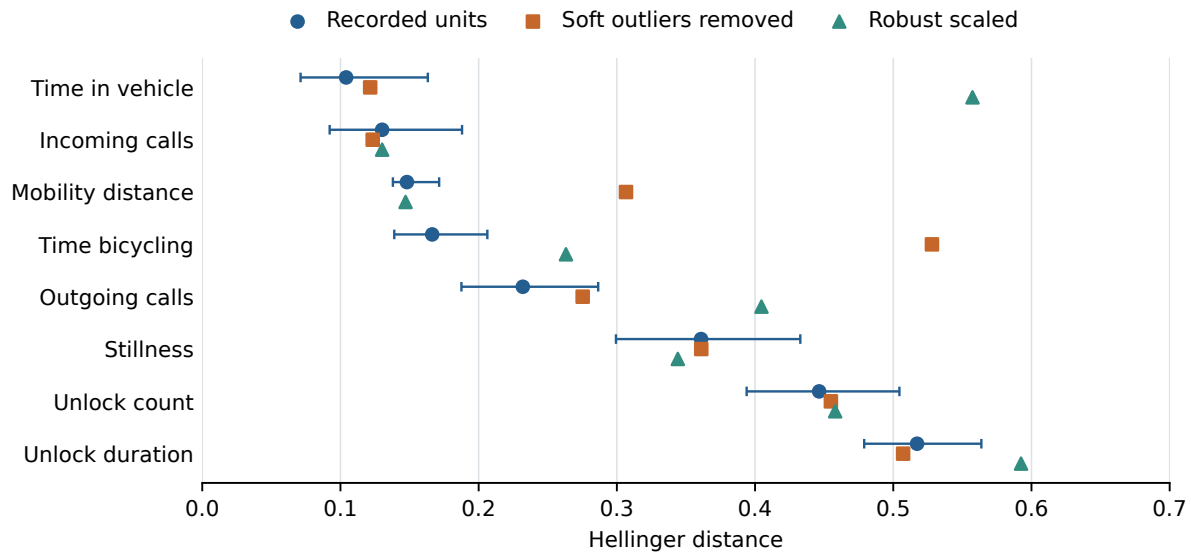


Figure 1: CES–CrossCheck distribution diagnostics under three preprocessing settings. Base estimates preserve recorded units; horizontal intervals show their participant-bootstrap 95% intervals. Soft-outlier removal and dataset-specific robust scaling each recompute the common bins for that setting. All settings give participants equal weight and require at least three base-eligible days.

Table 4: Outcome content and time structure. CrossCheck's composite and single-item outcomes are kept distinct. Depresjon's control-group MADRS values are structurally absent. The item crosswalk and score-quality tables provide the full record.

Dataset	Instrument / outcome	Scale	Time and crosswalk
StudentLife	Self-reported PHQ-9	0–27	Two-week recall; pre/post. Two item concepts shared with PHQ-2.
CES	PHQ-2, GAD-2, PHQ-4 reconstructed from items	0–6 / 0–6 / 0–12	Repeated records with two-week recall; depression and anxiety subsets are separately available.
CrossCheck	Five-item negative EMA; depressed-mood item	0–15 / 0–3	Repeated EMA; composite combines negative mood and psychosis-related experiences. Single item has mood-domain overlap.
Depresjon	Clinician-rated MADRS1/2; case/control	0–60 / binary	Two totals for cases; assessment dates absent from the released score table.

controls. Consequently, its severity association and group contrast use different analytic samples and targets.

4.4 RQ4: Associations differ by estimand and outcome

Table 5 reports the revised day-aligned analysis. Between-person CES and CrossCheck phone-feature correlations range from approximately -0.088 to 0.075 for the primary labels, with intervals spanning zero. Their signs supply little evidence of cross-dataset disagreement at these magnitudes. Within-person stillness associations are positive in both datasets: $r = 0.082$ for CES PHQ-2 and 0.053 for CrossCheck negative EMA. Unlock duration has a positive within-person association in CES (0.080) and a negative one in CrossCheck (-0.041). These estimates refer to distinct outcomes and populations.

The single CrossCheck depressed-mood item yields within-person stillness $r = 0.053$ and unlock-duration $r = -0.035$ under the same alignment, preserving the primary composite's directions for these features. The complete sensitivity results include net EMA, PHQ-4, GAD-2, and person-centered rank associations. Within CES, stillness also has a positive within-person association with GAD-2 ($r = 0.068$, 95% CI $[0.045, 0.091]$). The two subscales thus share a small positive stillness association under the same alignment. For unlock duration, the person-centered rank sensitivity is 0.069 in CES and

Table 5: Behavior–outcome associations with 95% participant-bootstrap intervals. Between-person estimates are Spearman ρ; within-person estimates are Pearson r after person centering. The final row reports Cohen's d for 23 cases versus 32 controls. Dashes indicate participant-level assessment without a daily estimand.

Dataset	Feature	Outcome	n	Between person (95% CI)	Within person (95% CI)
StudentLife	Still fraction	PHQ-9 pre	46	$+0.115$ [-0.180, +0.429]	—
StudentLife	Still fraction	PHQ-9 post	38	$+0.274$ [-0.064, +0.555]	—
CES	Still duration	PHQ-2	217	-0.007 [-0.135, +0.133]	$+0.082$ [+0.057, +0.106]
CES	Distance	PHQ-2	215	-0.019 [-0.152, +0.113]	-0.006 [-0.015, +0.001]
CES	Unlock count	PHQ-2	217	-0.088 [-0.234, +0.052]	-0.022 [-0.056, +0.021]
CES	Unlock duration	PHQ-2	217	$+0.026$ [-0.110, +0.163]	$+0.080$ [+0.042, +0.126]
CrossCheck	Still duration	Negative EMA	59	$+0.075$ [-0.199, +0.361]	$+0.053$ [+0.024, +0.080]
CrossCheck	Distance	Negative EMA	59	$+0.014$ [-0.262, +0.294]	$+0.048$ [-0.049, +0.106]
CrossCheck	Unlock count	Negative EMA	60	-0.068 [-0.331, +0.187]	-0.037 [-0.075, +0.001]
CrossCheck	Unlock duration	Negative EMA	60	$+0.007$ [-0.249, +0.253]	-0.041 [-0.071, -0.008]
Depresjon	Mean activity	MADRS1	23	$+0.092$ [-0.367, +0.511]	—
Clinical group contrast: raw activity, case minus control, pooled-SD Cohen's d					
Depresjon	Mean activity	Case/control	55	$d = -0.547$ [-1.096, -0.037]	—

-0.024 in CrossCheck; the latter interval spans zero ($[-0.058, 0.014]$).

StudentLife stillness correlations are 0.115 for pre and 0.274 for post PHQ-9; their broad intervals reflect the smaller samples and whole-period behavioral aggregation. In Depresjon, mean activity is lower in cases than controls ($d = -0.547$), while its case-only correlation with MADRS1 is 0.092. Group separation and severity association describe different variation in this release. Taken together, the results show why between-person summaries, within-person co-variation, and clinical group contrasts should each retain their own outcome and time definition.

5 Implications for Multimodal Mental-Health Research

A shared feature table is a useful representation whose measurement choices should travel with it. For CES–CrossCheck, the documented daily correspondences support distribution checks in common recorded units. Approximate phone–wrist or sample-fraction–duration bridges support construct-level comparisons after an explicit transformation. A multimodal model that combines language with these inputs should identify which kind of bridge enters its experiment.

The label crosswalk has a parallel role. A language model supervised with PHQ-2 targets a short depression screen; one supervised with CrossCheck negative EMA targets a composite that includes psychosis-related experiences. A case/control classifier uses another target. Keeping these definitions explicit makes a model's evaluation interpretable across cohorts and links performance changes to a recorded change in inputs, labels, or both.

The association audit also makes the estimand part of that record. Weak between-person correlations can coexist with small within-person associations. A model predicting a person's long-term average and a model tracking changes around that person's average should use evaluation designs matched to those questions. This distinction complements cross-dataset prediction studies such as GLOBEM (Xu et al., 2022).

We recommend a five-part compatibility record: (1) source fields, transformations, units, windows, and eligibility rules; (2) cohort, platform, collection context, and feature-specific analytic denominators; (3) outcome items, rater, recall period, cadence, and reconstruction rules; (4) raw and sensitivity distribution diagnostics; and (5) target estimand, participant split, and intended transfer claim. This record gives data statements and model cards concrete measurement content for mental-health applications (Mitchell et al., 2019; Sambasivan et al., 2021; Selbst et al., 2019; Paullada et al., 2021; Bender et al., 2021; Blodgett et al., 2020; Raji et al., 2020).

6 Limitations and Ethics

The four datasets provide a purposive case set. Population, platform, collection period, and processing vary together, so their separate contributions to a distribution or association difference remain unidentified. The feature audit uses release documentation and one author's judgments. Raw-unit distances depend on histogram support; the sensitivity outputs expose that dependence. Future validation can compare reconstructed features against

reference measurements and evaluate the audit's correspondence decisions with independent raters.

The association analysis uses available observations and unadjusted estimates. Missingness and scheduled item availability can select the paired days. Within-person centering pools observations across study periods and preserves each scale's recall window; calendar trends, treatment, and temporal lags require additional models. StudentLife and Depresjon have participant-level assessment identifiers without verified daily alignment. These design properties define the scope of the reported estimates.

This study reanalyzes public, de-identified releases and involves no new recruitment or contact. Published materials contain aggregate statistics, dictionaries, and code; participant-level inputs remain in a separate private reconstruction directory. The source manifest records access conditions and required citations. Behavioral and symptom records are sensitive, so reconstruction scripts retain the released pseudonymous keys locally and publish only the aggregates used in the paper.

7 Data and Code Availability

The accompanying supplement contains feature and item crosswalks, quality and missingness summaries, all association estimates and bootstrap seeds, distribution sensitivities, a source-file manifest with SHA-256 hashes, and scripts to reconstruct private inputs from the cited releases. A documented command sequence regenerates the analysis tables and figure. StudentLife is accessed through its Zenodo RDS release; CES through its Kaggle release; CrossCheck through Cornell's public data page; and Depresjon through Simula. Direct access links and terms are recorded in the supplement.

8 Conclusion

Across four public digital-phenotyping datasets, comparability depends on the released feature, preprocessing, outcome definition, and statistical estimand. CES–CrossCheck offers daily correspondences with markedly different distribution distances and sensitivities. The label audit distinguishes symptom screens, repeated mood and negative-experience reports, clinical severity, and case/control membership. The effect audit shows small between-person phone-feature estimates alongside several small within-person associations. Recording these properties makes cross-dataset and multimodal mental-health evaluation more precise and reproducible.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Su Lin Blodgett, Solon Barocas, Hal Daume, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1293–1304.

Victor P. Cornet and Richard J. Holden. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of Biomedical Informatics*, 77:120–132.

Lee J. Cronbach and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Eiko I. Fried. 2017. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, 208:191–197.

Eiko I. Fried and Randolph M. Nesse. 2015. Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, 13:72.

Enrique Garcia-Ceja, Michael Riegler, Petter Jakobsen, Jim Torresen, Tine Nordgreen, and Ketil J. Oedegaard. 2018. Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 472–477.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Jeremy F. Huckins, Alex W. daSilva, Weichen Wang, Emma Hedlund, Courtney Rogers, Subigya Nepal, Jialu Wu, Masahiro Obuchi, Eilis I. Murphy, Meghan L. Meyer, David D. Wagner, Paul E. Holtzheimer, and Andrew T. Campbell. 2020. Mental health and behavior of college students during the early phases of the COVID-19 pandemic: Longitudinal smartphone and ecological momentary assessment study. *Journal of Medical Internet Research*, 22(6):e20185.

Thomas R. Insel. 2017. Digital phenotyping: Technology for a new science of behavior. *JAMA*, 318(13):1215–1216.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385.

S. H. Jain, B. W. Powers, J. B. Hawkins, and J. S. Brownstein. 2015. The digital phenotype. *Nature Biotechnology*, 33(5):462–463.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Kurt Kroenke, Robert L. Spitzer, Janet B. W. Williams, and Bernd Lowe. 2009. An ultra-brief screening scale for anxiety and depression: The PHQ-4. *Psychosomatics*, 50(6):613–621.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Christian Montag, Cornelia Sindermann, and Harald Baumeister. 2020. Digital phenotyping in psychological and medical sciences: A reflection about necessary prerequisites to reduce harm and increase benefits. *Current Opinion in Psychology*, 36:19–24.

Stuart A. Montgomery and Marie Asberg. 1979. A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, 134(4):382–389.

Subigya Nepal, Wenjun Liu, Arvind Pillai, Weichen Wang, Vlado Vojdanovski, Jeremy F. Huckins, Courtney Rogers, Meghan L. Meyer, and Andrew T. Campbell. 2024. Capturing the college experience: A four-year mobile sensing study of mental health, resilience, and behavior of college students during the pandemic. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):38:1–38:37.

Jukka-Pekka Onnela and Scott L. Rauch. 2016. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, 41:1691–1696.

Amandalyn Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44.

Sohrab Saeb, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, 17(7):e175.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68.

Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. 2008. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32.

John Torous, Mathew V. Kiang, Justine Lorme, and Jukka-Pekka Onnela. 2016. New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*, 3(2):e16.

John Torous, Patrick Staples, and Jukka-Pekka Onnela. 2015. Realizing the potential of mobile mental health: New methods for new data in psychiatry. *Current Psychiatry Reports*, 17(8):61.

Robert J. Vandenberg and Charles E. Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1):4–70.

Rui Wang, Min S. H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A. Scherer, Vincent W. S. Tseng, and Dror Ben-Zeev. 2016. CrossCheck: Toward passive sensing and

detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 886–897.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14.

Rui Wang, Weichen Wang, Alex W. DaSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):43:1–43:26.

Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Kumar Nepal, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2022. GLOBEM: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):190:1–190:32.