DETECCIÓN Y BIOMARCADORES

Digital phenotyping for depression: from smartphone sensors to WhatsApp-based detection

SentirIA Research Papers · 2026

Infraestructura de detección temprana en salud mental

Este documento es parte de la base científica de SentirIA,
plataforma de detección temprana y monitoreo continuo de deterioro en salud mental.
No constituye diagnóstico clínico. La evaluación es responsabilidad del profesional.

Digital phenotyping for depression: from smartphone sensors to WhatsApp-based detection

Passive smartphone data can detect depressive symptoms with moderate accuracy, but the field remains pre-clinical. The strongest validated signals — GPS mobility reduction (meta-analytic r = −0.25), sleep disruption, and keystroke dynamics — correlate with PHQ-9 severity at small-to-medium effect sizes. Population-level prediction models perform poorly (median $R^2 \approx 0$), while personalized models show promise. For a WhatsApp-only approach like SentirIA, the most informative digital phenotyping signals — language, voice, temporal patterns, and social engagement — remain accessible, though GPS and accelerometer data are lost. Latin America's 92%+ WhatsApp penetration, voice-note culture, and 75% mental health treatment gap make it an ideal deployment context, but no digital biomarker for depression has yet achieved full clinical validation.

The science of behavior captured through phones

Digital phenotyping — defined by Torous, Kiang, Lorme, and Onnela in *JMIR Mental Health* (2016) as "the moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices" — transforms smartphones into continuous behavioral observatories. The concept emerged from Onnela and Rauch's 2016 framework paper in *Neuropsychopharmacology*, which distinguished between active data (requiring user input, such as surveys) and passive data (collected automatically from sensors without user involvement).

The passive signals used for depression detection span multiple sensor modalities. GPS data yields location variance, entropy, circadian movement patterns, and homestay duration — all proxies for the mobility restriction characteristic of depression. Accelerometer readings capture physical activity levels and psychomotor retardation. Keystroke dynamics reveal cognitive slowing through inter-key delay, typing variability, and error rates. Call and SMS logs track social withdrawal through contact diversity and communication frequency. Screen-on/off patterns infer sleep-wake cycles, while app usage metadata reflects behavioral engagement patterns. Multiple open-source platforms enable this research: Beiwe (Onnela Lab, Harvard), RADAR-base (RADAR-CNS consortium), mindLAMP (Beth Israel/Harvard), Purple Robot (Northwestern), BiAffect (University of Illinois Chicago), and StudentLife (Dartmouth).

The evidence base, synthesized across multiple systematic reviews published between 2022 and 2026, paints a picture of promising but immature science. A 2024 systematic review by Leaning et al. in *Neuroscience & Biobehavioral Reviews* analyzed 24 studies (9,801 participants) and found models achieving "moderate prediction performance" with accuracy ranging from 0.52 to

0.98 — a range so wide it underscores the field's inconsistency. A critical finding from Pratap et al. (2019, *Depression and Anxiety*, n = 271) demonstrated that sample-wide models yielded median $R^2 \approx 0$, while personalized idiographic models achieved AUC > 0.80 for 11.8% of participants. This individual variability is the field's central challenge: depression manifests differently in each person's digital behavior.

GPS mobility leads the evidence, but effect sizes remain small

The first meta-analysis of GPS features and depression — Terhorst et al. (2024, *JMIR*, k = 19 studies, N = 2,930) — established the quantitative benchmark for the field. Total distance traveled showed the strongest and most homogeneous correlation with depressive symptoms (pooled r = −0.25, 95% CI −0.29 to −0.21, I^2 = 0%). Other GPS features showed smaller effects: normalized entropy (r = −0.17), location variance (r = −0.17), entropy (r = −0.13), number of location clusters (r = −0.11), and homestay (r = +0.10). Circadian movement, transition time, and speed while moving were not significant in the pooled analysis.

Individual studies report stronger correlations, though with smaller samples that raise generalizability concerns. Saeb et al.'s foundational 2015 study in *JMIR* (n = 28) found circadian movement correlated with PHQ-9 at r = −0.63 (p = .005), with normalized entropy and location variance both at r = −0.58. Their 2016 replication in *PeerJ* (n = 48) confirmed these patterns (r = −0.43 to −0.46) and showed weekend mobility was more strongly associated with depression than weekday mobility. GPS features predicted depressive symptom severity up to 10 weeks before formal assessment. Meyerhoff et al.'s 2024 large-scale LifeSense study (n = 1,013, *npj Mental Health Research*) found homestay was an early within-person predictor of PHQ-8 severity (distal β = 0.219, p = .012), while circadian movement was only a concurrent correlate.

Beyond GPS, keystroke dynamics show strong signal. The BiAffect study (Vesel et al., 2020, *JAMIA*, n = 250 users, 14 million+ keypresses) found that more severe depression correlated with more variable typing speed (p < .001), shorter keyboard sessions (p < .001), and lower typing accuracy (p < .05). Zulueta et al. (2018, *JMIR*) predicted clinician-rated depression severity (HDRS-17) from keystroke metadata with conditional R^2 = 0.63 (p = .01) in bipolar patients.

Classification accuracy in the most optimistic studies reaches 86.5–87% for binary depression detection (Saeb et al., 2015; Choudhary et al., 2022, *JMIR Formative Research*, n = 558). Choudhary's three-class model (none/mild/severe) achieved 78% accuracy. However, Müller et al. (2021, *Scientific Reports*) issued a critical warning: GPS-based models that achieved AUC = 0.82 in a homogeneous student sample (n = 57) dropped to AUC = 0.57 in a large heterogeneous U.S. sample (n = 5,262). The meta-analysis by Terhorst et al. found no study adhered to STROBE reporting guidelines, 79% were underpowered, and evidence of publication bias was detected.

Three programs that shaped the field — and what they learned

Onnela Lab and the Beiwe platform

JP Onnela's lab at the Harvard T.H. Chan School of Public Health, enabled by a 2013 NIH Director's New Innovator Award, built Beiwe — an open-source, HIPAA-compliant smartphone platform for digital phenotyping (released under BSD-3 license in 2017). Named after a Nordic goddess of sunlight and mental health, Beiwe collects GPS, accelerometer, call/text logs, and screen activity at costs ranging from $3.50 to $80 per subject-month. The lab also developed Forest, a Python library for analyzing digital phenotyping data.

Key contributions include Barnett et al.'s 2018 pilot in *Neuropsychopharmacology* (n = 17 schizophrenia patients) showing statistically significant behavioral anomalies in the days preceding relapse, and Barnett and Onnela's 2020 *Biostatistics* paper developing methods for inferring mobility measures from GPS traces with missing data. Pellegrini et al. (2021, *Brain and Behavior*) demonstrated estimation of longitudinal depressive symptoms from passively collected smartphone data in a transdiagnostic cohort. The lab remains active, with Beiwe used globally and Forest under continuous development.

RADAR-CNS: Europe's €26 million consortium

The Remote Assessment of Disease and Relapse – Central Nervous System (RADAR-CNS) project represents the largest coordinated digital phenotyping effort for depression. Funded by the Innovative Medicines Initiative 2 (€26 million, 2016–2022), the consortium comprised 22 organizations led by King's College London (Matthew Hotopf, Richard Dobson) and Janssen Pharmaceutica. Their RADAR-base platform, published in *JMIR mHealth and uHealth* (Ranjan et al., 2019, cited 168 times), built on Apache Kafka for real-time data streaming and was open-sourced under the Apache 2.0 license.

The flagship RADAR-MDD study enrolled 623 individuals with recurrent MDD across three European sites (London, Amsterdam, Barcelona) for up to 24 months, collecting Fitbit data, smartphone sensor data, and biweekly PHQ-8 assessments. Sun et al. (2023, *JMIR*) reported that at least 8 days of data were needed per 14-day window for reliable feature calculation, and only 110 of 623 participants had >50% data across all types. Zhang et al.'s 2024 summary of five RADAR-MDD investigations confirmed that elevated depression associated with diminished sleep quality, reduced sociability (Bluetooth-approximated), decreased physical activity, slower walking cadence, and circadian rhythm disturbances. Laiou et al. (2022, *JMIR mHealth uHealth*) found high PHQ-8 scores associated with longer homestay on weekdays but not weekends. The RADAR-base platform now supports studies in Alzheimer's disease, ADHD, and autism.

Mindstrong Health: $160 million and a cautionary tale

Mindstrong Health, founded in 2014 by Paul Dagum and later joined by former NIMH Director Tom Insel as president, raised $160 million in venture capital to develop cognitive biomarkers from smartphone touchscreen interactions — typing patterns, scrolling, swiping, and tapping. Their core innovation used content-free HCI metadata as continuous passive biomarkers. Dagum's foundational paper (2018, *npj Digital Medicine*, n = 27) reported digital biomarkers predicting neuropsychological test scores for working memory, executive function, and language with correlations significant at $p < 10^{-4}$. A 2023 *JMIR Formative Research* study (n = 142 patients) found significant longitudinal associations between smartphone interaction patterns and depressive symptom severity.

The company shut down in early 2023 after pivoting from digital biomarker R&D to virtual therapy provision. SonderMind acquired the technology assets. The failure taught the field several lessons: the foundational science rested on a pilot of only 27 subjects over 7 days; investors pressured premature commercialization before adequate validation; and as John Torous observed, "Americans value mental health extremely highly until they have to pay for it." A 2022 analysis found 44% of digital health companies had a clinical robustness score of zero, and digital health funding collapsed from $29.3 billion in 2021 to roughly half in 2022.

CrossCheck, StudentLife, and LAMP anchor the broader evidence

Beyond the three major programs, several studies form the field's empirical backbone. The CrossCheck study (Wang et al., 2016, *UbiComp*, Dartmouth) tracked 61 individuals with schizophrenia using passive smartphone sensing and found significant behavioral changes 30 days before relapse, with personalized rather than universal predictive patterns. The publicly released dataset has been used in 10+ subsequent ML studies.

The StudentLife study (Wang et al., 2014, *UbiComp*) monitored 48 Dartmouth students for 10 weeks, finding significant correlations between automatically sensed conversation frequency, sleep, activity, mobility, and PHQ-9 change scores. A Lasso regression model predicted GPA with $r = 0.81$. The 2018 extension tracked 200 high school students through four years at Dartmouth — the longest longitudinal mobile sensing study conducted.

The mindLAMP platform (Vaidyam, Halamka, and Torous, 2022, *JMIR mHealth and uHealth*) emerged from Beth Israel Deaconess Medical Center as a fully open-source alternative, now deployed across 54 sites worldwide through the LAMP Consortium. Its modular architecture (Learn, Assess, Manage, Prevent) supports passive sensing, active assessment, and just-in-time adaptive interventions, with multilingual support and FHIR interoperability.

Wahle et al.'s MOSS study (2016, *JMIR mHealth and uHealth*, ETH Zürich) was among the first to

combine passive sensing with context-sensitive CBT delivery, though classification accuracy was modest (Random Forest: 60.1%). The study demonstrated significant PHQ-9 reduction (p = .01) in clinically depressed participants who adhered for ≥8 weeks.

Translating digital phenotyping to WhatsApp-only signals

For SentirIA's WhatsApp-based approach, the critical question is which signals survive the loss of full sensor access. The answer is encouraging: the highest-specificity signals for depression detection — language content, voice characteristics, temporal patterns, and social engagement metrics — are all extractable from WhatsApp interactions. What is lost is GPS/mobility data, accelerometer-based activity measurement, app usage diversity, and ambient light sensing.

Text-derived signals carry strong evidence. First-person singular pronoun use ("I/me") is a robust depression marker across multiple NLP studies. BiAffect's keystroke research confirms that response latency, message length trends, and typing variability all correlate with depression severity. Within a WhatsApp chatbot interaction, SentirIA can measure response latency to chatbot prompts (psychomotor retardation proxy), message length changes over time (social withdrawal/anhedonia indicator), emoji usage shifts (affective state tracking), and linguistic complexity changes.

Voice notes offer perhaps the richest single signal. A landmark 2026 study by Otani et al. (*PLOS Mental Health*) directly evaluated depression detection through voice analysis of WhatsApp audio messages from 160 Brazilian Portuguese speakers, demonstrating feasibility using 68 acoustic features across multiple ML models. The broader vocal biomarker literature is robust: a study of 14,898 adults achieved sensitivity 71.3% and specificity 73.5% for detecting PHQ-9 ≥ 10 from ≥25 seconds of free-form speech. Key acoustic features include reduced vocal pitch (F0), flattened pitch modulation, increased pauses, reduced speech rate, and increased jitter and shimmer. Latin Americans' extensive use of voice notes provides ecologically valid speech samples without special recording tasks.

Timestamp-derived circadian patterns are validated as depression proxies. A 2023 *JMIR* study of the "Rhythm" app found that smartphone-use timestamp variability (Intradaily Variability) correlated significantly with PHQ-9 scores in 33 insomnia patients. Late-night messaging (2–5 AM), shifts in peak messaging hours, and irregular day-to-day messaging onset/offset times can serve as sleep-wake cycle proxies.

The single-platform limitation is real but manageable. WhatsApp-only loses approximately 40% of traditional digital phenotyping signals (mobility, physical activity) but retains the highest-specificity signals. Insel himself noted in *World Psychiatry* (2017) that "activity and geolocation data are non-specific and noisy" for individual-level prediction. A single-platform approach also gains consistency and reduces noise from cross-platform data heterogeneity. Existing precedents validate the platform choice: Crisis Text Line acquired Aquí Estoy in 2025 to expand WhatsApp-

based mental health support across 20 Spanish-speaking countries, and PsyBot (2025) demonstrated significant loneliness reduction in a WhatsApp-based RCT.

Ethics, cultural adaptation, and regulatory pathways for LATAM

Ethical frameworks for digital phenotyping are emerging but incomplete. The most authoritative guidance comes from Martinez-Martin et al.'s 2021 Delphi consensus (*JMIR mHealth and uHealth*), which identified five priority domains: privacy, transparency, dynamic consent, accountability, and algorithmic fairness. Existing regulations (HIPAA, GDPR) were deemed inadequate for the granularity of digital phenotyping data. For SentirIA, data minimization is essential — processing raw data on-device, transmitting only aggregate features, and implementing federated learning for model training. FedTherapist, a federated learning system for mobile mental health monitoring, achieved 0.15 AUROC improvement over non-language features while keeping raw data on edge devices.

The PHQ-9 is well-validated in Spanish across multiple LATAM populations. A 2023 *JAMA Network Open* systematic review by Martinez et al. confirmed both PHQ-2 and PHQ-9 as reliable screening instruments for Spanish-speaking adults. Country-specific validations exist for Peru (n = 30,449), Chile (n = 1,327), Mexico (n = 55,555 women), Costa Rica (n = 1,162), and Colombia. However, optimal cut-off points may vary by country, and cultural factors significantly influence depression expression. Latinos are more likely to present somatic symptoms ("nervios," fatigue, headaches), the cultural syndrome "ataques de nervios" overlaps with but differs from DSM categories, familismo creates both protective support and stigma-driven concealment, and machismo leads men to underreport. A 2016 systematic review found stigma remains a principal barrier to mental health service access in Latin America, with approximately 50% of Colombians citing personal stigma.

Several digital mental health interventions have been tested in LATAM. SilverCloud Health culturally adapted its depression/anxiety program for Colombian and Mexican university students, achieving 86.5–94.7% user satisfaction (2024, *JMIR Formative Research*). The Tess chatbot was tested in a pilot RCT with 181 Argentinian students — the first chatbot mental health study in Latin America. ChatBot-Juntos in Peru deployed COVID-era mental health screening reaching 21 regions. Less than 3% of national health budgets in the region are allocated to mental health, creating urgent need for scalable digital solutions.

Regulatory pathways vary by country. A depression monitoring app would likely classify as Software as a Medical Device (SaMD) in most jurisdictions. Brazil's ANVISA follows IMDRF guidelines with four risk classes and may require local clinical trials; its LGPD provides GDPR-equivalent data protection. Mexico's COFEPRIS introduced an abbreviated regulatory pathway in September 2025, accepting prior authorizations from FDA and other recognized bodies with

target 30-business-day review. Colombia's INVIMA has a robust but developing digital health framework. No LATAM-specific framework exists for digital phenotyping or passive mental health monitoring. The V3 framework (Goldsack et al., 2020, *npj Digital Medicine*) — Verification, Analytical Validation, and Clinical Validation — with the 2025 V3+ extension adding Usability Validation, provides the methodological roadmap for systematic biomarker validation. Critically, no digital biomarker for depression has achieved full V3 clinical validation and regulatory qualification as a clinical endpoint.

Conclusion: building SentirIA on evidence, not hype

The digital phenotyping field has accumulated meaningful evidence across approximately 60 studies and nearly 10,000 participants that passive smartphone signals correlate with depression severity. But the correlations are small-to-medium, models fail to generalize across heterogeneous populations, and Mindstrong's $160 million collapse demonstrates the cost of commercializing before the science is sound. For SentirIA, four strategic insights emerge from this evidence base.

First, prioritize personalized over population models. The most consistent finding across studies is that individual-level models vastly outperform group-level prediction. SentirIA should establish within-person baselines and detect deviations from each user's own patterns, not attempt universal thresholds.

Second, WhatsApp voice notes may be the single most valuable signal. The 2026 Otani et al. study directly validates WhatsApp audio for depression detection in Brazilian Portuguese, and the vocal biomarker literature is among the most robust in digital phenotyping. Latin America's voice-note culture provides this data naturally.

Third, follow the V3/V3+ validation framework rigorously. Start with the most validated passive biomarkers (temporal patterns, voice features, linguistic markers, social engagement), conduct analytical validation against PHQ-9 in LATAM populations with country-specific cut-offs, and achieve clinical validation before making diagnostic claims. The Mindstrong lesson is clear: promising pilot data is not proof.

Fourth, design for cultural context from day one. Somatic symptom expression, familismo dynamics, stigma barriers, and gender-specific underreporting patterns in LATAM populations are not edge cases — they are the primary use context. The PHQ-9's validation across LATAM provides a reliable anchor, but SentirIA's conversational interface must be culturally adapted to elicit authentic engagement rather than socially desirable responses.