VALIDACIÓN Y MONITOREO

PHQ-9 validation, digital administration, and designing SentirIA's clinical study

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.

PHQ-9 validation, digital administration, and designing SentirIA's clinical study

The PHQ-9 is among the most extensively validated depression screening instruments worldwide, with robust evidence across 70+ languages and 200+ validation studies — but no published study has yet validated PHQ-9 administration via WhatsApp, the dominant messaging platform in Latin America. This gap represents both SentirIA's core opportunity and its central validation challenge. Existing evidence from chatbot-administered PHQ-9 studies shows strong concordance with standard formats (ICC up to 0.91), and Spanish-language validations across Latin America confirm the instrument's reliability — though optimal cutoff scores vary by country (≥7 in Colombia, ≥8 in Argentina, ≥10 in Costa Rica). A validation study with 100 patients is feasible as a preliminary pilot but insufficient for definitive validation, which requires 250–300 participants. Below is a thorough synthesis of the evidence across all five requested domains.

The PHQ-9's foundational validation and global reach

The PHQ-9 was developed by Kurt Kroenke, Robert Spitzer, and Janet Williams and published in the *Journal of General Internal Medicine* in 2001. The original study enrolled 6,000 patients across 8 primary care and 7 obstetrics-gynecology clinics, with 580 patients undergoing independent structured mental health professional interviews for criterion validation. At a cutoff of ≥10, the instrument achieved 88% sensitivity and 88% specificity for major depressive disorder, with an AUC of 0.95, Cronbach's α of 0.86–0.89, and test-retest reliability of r = 0.84.

The instrument's nine items map directly to DSM-IV (now DSM-5) criteria for major depression, scored 0–3 on a frequency scale from "not at all" to "nearly every day," yielding total scores of 0–27. The severity categories — minimal (0–4), mild (5–9), moderate (10–14), moderately severe (15–19), and severe (20–27) — correlate strongly with functional impairment, disability days, and healthcare utilization, establishing robust construct validity.

Subsequent meta-analyses have consolidated this evidence at scale. The landmark DEPRESSD individual participant data meta-analysis (Levis, Benedetti, and Thombs, *BMJ*, 2019) pooled 58 studies with 17,357 participants (2,312 major depression cases) and confirmed that a cutoff of ≥10 against semistructured interviews yields sensitivity of 0.88 (95% CI 0.83–0.92) and specificity of 0.85 (95% CI 0.82–0.88). An updated IPDMA by Negeri et al. (*BMJ*, 2021) expanded to 100 studies and 44,503 participants, confirming these estimates. A reliability generalization meta-analysis (2025, *Discover Mental Health*) across 60 studies and 232,147 participants found a pooled Cronbach's α of 0.86. Bianchi et al. (2022) demonstrated essential unidimensionality and measurement invariance across gender and nationality in over 58,000 participants from 29 samples, 7 countries, and 5 languages.

The PHQ-9 has been validated across specialty populations including oncology (Al-Shamsi et al.: AUC 0.91, optimal cutoff ≥9 in Arabic-speaking cancer patients), neurology (Prisnie et al., 2016: cutoff >13 yielded 81.8% sensitivity, 97.1% specificity in stroke patients), cardiology, and perinatal care (pooled meta-analytic sensitivity 0.84, specificity 0.81 at ≥10; Kroenke et al., 2022, *General Hospital Psychiatry*). In adolescents, Richardson et al. (2010) found PHQ-9 ≥11 optimal (sensitivity 89.5%, specificity 77.5% vs. DISC-IV), and in elderly populations, Lamers et al. (2008) recommended a lower cutoff of ≥6 for chronically ill elderly (sensitivity 95.6%, specificity 81.0%). Measurement invariance across U.S. racial/ethnic groups has been confirmed in large studies (Keum et al., 2018; Patel et al., 2019; Harry et al., 2023), with all PHQ-9 items showing statistically significant but clinically negligible differential item functioning across groups.

Latin American validations reveal lower optimal cutoffs

For SentirIA's target population, the evidence is substantial and growing. Martinez, Teklu, Tahir, and Garcia published the first comprehensive systematic review and meta-analysis of the Spanish-language PHQ-9 in *JAMA Network Open* (2023), confirming that the PHQ-2 and PHQ-9 are reliable and valid for MDD screening among Spanish-speaking populations. Country-specific validation studies paint a nuanced picture:

- Colombia (Cassiani-Miranda et al., 2021, *Revista Colombiana de Psiquiatría*): n=243 primary care adults; Cronbach's α=0.80; AUC=0.92; optimal cutoff ≥7 yielded sensitivity 90.4%, specificity 81.7%, NPV 96.9% (gold standard: MINI)
- Argentina (Daray et al., 2019, *BMC Psychiatry*): n=169 ambulatory adults; α=0.87; AUC=0.87; optimal cutoff ≥8 (gold standard: MINI)
- Peru (Villarreal-Zegarra et al., 2019, *PLOS ONE*): n=30,449 from the national ENDES survey; measurement invariance confirmed across sex, age, education, SES, and region — the largest Latin American population-based validation
- Chile (Saldivia et al., 2020; Caneo et al., 2022): unidimensional structure confirmed; lower cutoff ≥8 recommended for elderly; validated in Spanish-speaking immigrants (n=897, AUC=0.93)
- Mexico (Arrieta and Aguerrebere, 2017; Lara-Muñoz et al., 2019): validated in rural Chiapas primary care and in 55,555 women from the Mexican Teachers' Cohort
- Costa Rica (Sánchez-Villena et al., 2023, *Scientific Reports*): n=1,162; α=0.928; cutoff ≥10 yielded sensitivity 78.6%

The critical finding for SentirIA: optimal cutoffs in Latin American studies tend to be lower (≥7 to ≥8) than the standard ≥10, likely reflecting cultural differences in symptom expression, the use of fully structured reference standards (MINI, CIDI) that systematically yield lower sensitivity than semistructured interviews, and population-specific prevalence patterns. SentirIA's validation study should explicitly test multiple cutoffs and determine the population-specific optimum via

ROC analysis.

Diagnostic accuracy varies dramatically by cutoff and reference standard

The choice of cutoff score involves a fundamental trade-off between sensitivity and specificity that must be calibrated to clinical purpose. The Levis et al. (2019) IPDMA provides the most authoritative estimates across cutoffs, stratified by reference standard type:

Cutoff	Sensitivity (semistructured)	Specificity (semistructured)	Sensitivity (MINI)	Specificity (MINI)
≥5	0.98	0.55	—	—
≥8	0.95	0.75	—	—
≥10	0.88	0.85	0.77	0.87
≥12	0.79	0.91	—	—
≥15	0.56	0.96	0.42	0.97

The sensitivity gap between reference standard types is striking: semistructured interviews (like the SCID) yield 5–22% higher sensitivity than fully structured interviews, and 2–15% higher than the MINI across all cutoffs. This has direct implications for SentirIA's validation: using the MINI as gold standard will produce lower apparent sensitivity than using the SCID, and this must be acknowledged in reporting.

Positive predictive value (PPV) depends heavily on prevalence. At the standard cutoff of ≥10 (sensitivity 0.88, specificity 0.85), in a primary care population with 12% depression prevalence, approximately half of all positive screens are false positives (PPV ≈44%). However, negative predictive value exceeds 97% across all realistic prevalence scenarios, meaning a negative screen reliably excludes major depression. For SentirIA's intended use as a screening and early detection tool, high sensitivity at a lower cutoff (≥7 or ≥8) may be preferable, accepting more false positives in exchange for missing fewer cases.

The Manea et al. (2012, *CMAJ*) meta-analysis of 18 studies (n=7,180) found no significant differences in diagnostic properties for cutoffs between 8 and 11, suggesting flexibility in cutoff selection. The updated Negeri et al. (2021) IPDMA confirmed that cutoff ≥10 maximizes combined sensitivity and specificity across all demographic subgroups, with specificity 3–5% higher in men and in adults over 60, and no significant sensitivity differences by age or sex.

Digital and chatbot-based PHQ-9 administration shows strong equivalence

The evidence supporting automated PHQ-9 administration is robust and directly relevant to SentirIA. Three categories of evidence are most pertinent: chatbot-based, smartphone app-based, and IVR-based.

Chatbot-based PHQ-9 represents the closest analog to SentirIA. The Perla conversational agent (Arrabales, 2020) — built on Google Dialogflow and administered in Spanish — achieved Pearson's r = 0.91 against traditional electronic PHQ-9, with sensitivity of 96%, specificity of 90%, and Cohen's κ of 0.77 (substantial agreement). Users preferred the chatbot format 2.5 times more than traditional questionnaires. This is the only validated Spanish-language chatbot for PHQ-9 and represents the strongest precedent for SentirIA. The HopeBot system (Guo et al., 2025/2026), powered by GPT-4o with retrieval-augmented generation, demonstrated ICC = 0.91 against self-administered PHQ-9 in 132 participants, with 45% identical scores and no systematic bias; 71% of participants reported greater trust in the chatbot than in self-administration. The Tess chatbot (Dosovitsky, Kim, and Bunge, 2021, *Frontiers in Digital Health*) achieved a remarkable 99.82% completion rate (3,895/3,902 participants) with Cronbach's α = 0.896, substantially higher than smartphone app completion rates (73.9%).

Smartphone app concordance with paper PHQ-9 is well-established. Zhen et al. (2020, *Neuropsychiatric Disease and Treatment*) found ICC = 0.951 between paper and smartphone versions in 110 Chinese depressed outpatients. Torous et al. (2015, *JMIR Mental Health*) reported r = 0.84, though app scores were 3.02 points higher on average and captured more suicidal ideation disclosures. Bush et al. (2013) demonstrated comparable psychometric properties across paper, computer, and smartphone formats.

IVR (Interactive Voice Response) systems show moderate concordance. Turvey et al. (2012) found ICC = 0.65 in 51 veterans, with IVR less sensitive to higher severity levels. The TLC-PHQ-9 system showed stronger results: weighted κ = 0.76 across 5 administrations over 3 months in 80 subjects. Notably, Piette et al. (2016) tested IVR depression monitoring in Bolivia using PHQ-8 in Spanish, Aymara, and Quechua, achieving 54% call completion — one of the few digital mental health studies in a Latin American LMIC.

Vocal biomarker research shows promise but remains pre-clinical. Voice acoustic features achieve approximately 90% classification accuracy in controlled settings (Frontiers in Psychiatry, 2022; 71 MDD patients vs. 62 controls), with MFCCs predictive of PHQ-9 scores. However, a 2024 systematic review concluded that no vocal biomarker system yet achieves performance comparable to the PHQ-9 (sensitivity ≈0.85 benchmark). For SentirIA, vocal biomarkers should be treated as a supplementary signal rather than a primary scoring mechanism. A 2026 meta-analysis of NLP-based depression detection from text (123 studies, 40,983 samples, *npj Digital Medicine*) found pooled accuracy of 0.80 and AUC of 0.79.

The critical gap in the literature: no published study validates PHQ-9 administration specifically

via WhatsApp, despite it being the dominant messaging platform across Latin America with over 400 million users in the region. This positions SentirIA's validation study as potentially first-in-class.

Concordance metrics consistently support cross-modal equivalence

The broader concordance literature provides strong support for the premise that PHQ-9 scores are stable across administration modalities. Two large meta-analyses of patient-reported outcomes (not PHQ-9 specific) establish the baseline: Gwaltney, Shields, and Shiffman (2008, *Value in Health*; 65 studies, 278 scales) found average paper-computer differences of just 0.2% of scale range, with pooled correlation of 0.90 across 207 coefficients. Muehlhausen et al. (2015, *Health and Quality of Life Outcomes*; 72 studies) confirmed a pooled ICC of 0.90 (95% CI 0.88–0.92) for paper-electronic PRO equivalence.

PHQ-9-specific concordance data across modalities:

Comparison	Study	N	Key Metric	Equivalence
Self vs. telephone	Kroenke (2001)	580	r = 0.84; mean diff = 0.05	□ Confirmed
Self vs. telephone	Pinto-Meza (2005)	346	ICC excellent; all item κ ≥ 0.58	□ Confirmed
Paper vs. computer	Erbe (2016)	130	r = 0.92; α equivalent	□ Confirmed
Paper vs. smartphone	Zhen (2020)	110	ICC = 0.951	□ Confirmed
Self vs. LLM chatbot	HopeBot (2025)	132	ICC = 0.91; no bias	□ Confirmed
Chatbot vs. electronic	Perla (2020)	—	r = 0.91; κ = 0.77	□ Confirmed
Paper vs. IVR	Turvey (2012)	51	ICC = 0.65	⚠ Moderate
AI model vs. PHQ-9	Weisenburger (2024)	393	r = 0.73	□ Acceptable

ISPOR ePRO Good Research Practices Task Force guidelines (Coons et al., 2009, *Value in Health*; updated 2023) classify modifications during paper-to-electronic migration by magnitude. Minor modifications (changing "circle" to "select," one question per screen) require only cognitive debriefing and usability testing — no formal equivalence study. Moderate modifications (IVR, conversational format) require equivalence testing. SentirIA's conversational AI approach, which uses NLP to interpret free-text responses and maps them to PHQ-9 Likert scores, represents a substantial modification requiring full psychometric validation — the approach outlined in the study design below.

A rigorous validation protocol for SentirIA

Study design and phases

A cross-sectional diagnostic accuracy study with a repeated-measures reliability component is the recommended design, following STARD 2015 reporting guidelines and COSMIN measurement properties standards. The study should proceed in two phases:

Phase 1 (Month 1, n=20): Pilot feasibility testing covering usability (System Usability Scale), technical stability of WhatsApp delivery, conversation flow assessment, time-to-completion, and cultural appropriateness review. These participants are excluded from the primary analysis.

Phase 2 (Months 2–3, n=100): Primary validation. Each participant completes three assessments within 72 hours: (a) SentirIA conversational PHQ-9 via WhatsApp, (b) standard paper or clinician-administered PHQ-9 within 48 hours, and (c) gold standard diagnostic interview within 72 hours, administered by a blinded clinician. A subset of at least 50 participants completes SentirIA a second time at 7–14 days for test-retest reliability.

Gold standard selection

The MINI 7.0.2 (Spanish version) is recommended over the SCID-5 as the gold standard, based on three considerations: the MINI has been the predominant reference standard in Latin American PHQ-9 validation studies (Colombia, Chile, Peru), it requires only 15–30 minutes versus 45–90 for the SCID, and it can be administered by trained non-clinicians, enhancing scalability. However, the MINI systematically over-classifies depression by approximately 46% relative to the SCID (Wu et al., 2022, *Psychotherapy & Psychosomatics*, 3-IPDMA synthesis, 212 studies, n=69,405). This will inflate SentirIA's apparent sensitivity and deflate specificity — a limitation that must be disclosed. All MINI administrators must be blinded to SentirIA and paper PHQ-9 results.

Sample size: 100 patients is a pilot, not a definitive validation

Using Buderer's (1996) formula for diagnostic accuracy studies, with expected sensitivity of 0.85, prevalence of 20%, and desired 95% CI width of ±5%: the required number of diseased subjects is 49, necessitating a total sample of 245 patients. With n=100 and 20% prevalence (~20 cases), the 95% CI for sensitivity spans approximately 62%–97% — too wide for definitive conclusions. The study should be framed as a "preliminary validation" and an enrichment strategy is recommended: oversampling depressed patients from mental health clinics to achieve ~40% prevalence, yielding approximately 40 cases and narrowing the sensitivity CI to ±12%. For the test-retest ICC subset (n=50, 2 measurements, expected ICC=0.80), power exceeds 99% to detect ICC ≥0.80 versus a null of 0.50.

Statistical analysis plan

- Primary endpoints: Sensitivity, specificity, and AUC of SentirIA scores versus MINI diagnosis, with ROC analysis across all cutoffs (0–27) and optimal cutoff determined via Youden's J statistic

- Secondary endpoints: ICC(2,1) for test-retest reliability (target ≥0.75); ICC(3,1) and Bland-Altman plots for SentirIA versus paper PHQ-9 concordance (target: mean bias <1.5 points, 95% limits of agreement within ±5 points); weighted Cohen's κ for severity category agreement (target ≥0.60); Pearson correlation for convergent validity (target r ≥0.80); Cronbach's α for internal consistency (target ≥0.70)
- Subgroup analyses by sex, age (18–30, 31–50, 51–65), depression severity, education, and digital literacy
- Missing data: Complete-case primary analysis; multiple imputation sensitivity analysis if missing data exceeds 10%
- Pre-registration on ClinicalTrials.gov and/or local registry (ReBEC in Brazil, RNEC in Mexico) is mandatory

Ethical safeguards and the Item 9 protocol

The suicidal ideation safety protocol is the study's most ethically critical element. PHQ-9 Item 9 has 87.6% sensitivity but only 66.1% specificity and 28.6% PPV for suicidal risk against the Columbia Suicide Severity Rating Scale (Na et al., 2018, *Journal of Affective Disorders*). SentirIA must implement a real-time tiered response system:

- Score = 1: Empathic automated message plus clinical coordinator notification within 15 minutes; phone follow-up with C-SSRS screener within 2 hours
- Score = 2: Crisis hotline numbers delivered immediately plus coordinator notification; phone follow-up within 1 hour
- Score = 3: Urgent safety message, crisis hotline, immediate coordinator call; emergency contact protocol if unreachable within 30 minutes

Every SentirIA session must display visible crisis resources (Línea de la Vida 800-911-2000 in Mexico; CVV 188 in Brazil; Línea 106 in Colombia). The system requires clinical backup during all operating hours or must restrict availability to hours when backup is active.

Data privacy compliance spans multiple frameworks: Brazil's LGPD classifies health data as sensitive personal data requiring explicit highlighted consent, with anonymization required for research purposes and breach notification within 3 business days. Mexico's federal data protection law requires express written consent for sensitive data, with COFEPRIS oversight. WhatsApp conversations must be extracted and stored in a separate encrypted, compliant database, with phone numbers pseudonymized and conversation logs deleted after extraction.

Regulatory pathways across Latin America

SentirIA would likely be classified as a Class II Software as Medical Device (SaMD) in most jurisdictions. Key regulatory considerations include COFEPRIS's new Abbreviated Regulatory Pathway (activated September 2025), which accepts prior authorizations from FDA, EMA, and other reference agencies with a 30-business-day review target — a significant opportunity for SentirIA. ANVISA (Brazil) follows IMDRF-aligned frameworks under RDC N° 751/2023 and may

require local clinical trial data. The EU AI Act (2024) would classify mental health AI as a high-risk system requiring conformity assessment, bias mitigation, and human oversight. WHO guidelines (2019, 2023) emphasize that digital health tools must be embedded within broader care pathways with appropriate clinical oversight — SentirIA should never be positioned as a standalone diagnostic tool.

Conclusion: what SentirIA must prove

SentirIA sits at a well-evidenced intersection: the PHQ-9 is validated across Latin American populations, chatbot-administered PHQ-9 achieves ICC values of 0.77–0.91 against standard formats, and Spanish-language conversational agents (notably Perla) demonstrate strong concordance with sensitivity of 96% and specificity of 90%. The WhatsApp delivery channel and conversational-AI mapping approach are novel and unvalidated — this is simultaneously SentirIA's differentiator and its evidentiary burden.

Three priorities emerge from this evidence synthesis. First, SentirIA must establish that its indirect conversational mapping to PHQ-9 domains produces scores concordant with direct PHQ-9 administration (target ICC ≥0.80). Second, the validation study should test population-specific cutoffs rather than assuming the standard ≥10, given that Latin American studies consistently find optimal cutoffs of ≥7 to ≥8. Third, the 100-patient study should be explicitly framed as a preliminary validation, with a pre-planned Phase 3 study of 250–300 patients powered for definitive sensitivity/specificity estimates. The vocal biomarker component should be evaluated as a supplementary feature with its own concordance analysis against PHQ-9 total scores, not as a replacement for structured assessment. If designed rigorously — pre-registered, STARD-compliant, with blinded MINI administration and a robust Item 9 safety protocol — this study can establish SentirIA as the first validated WhatsApp-based depression screening tool for Latin American populations.