VALIDACIÓN Y MONITOREO

Relapse prevention and sustained monitoring for digital mental health companions

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.

Relapse prevention and sustained monitoring for digital mental health companions

No digital mental health platform has published a validated step-down protocol for AI companions — making SentirIA a potential pioneer in this space. The clinical evidence, however, provides a robust foundation: relapse affects 30–50% of patients within 6 months of recovery, prodromal signals appear 1–3 months before full relapse, and EWMA-based monitoring of conversational features achieves 68% sensitivity and 74% specificity for detecting recurrence. This report synthesizes evidence across clinical psychiatry, NLP, digital health engineering, and Latin American health systems to provide SentirIA with actionable protocols, specific thresholds, and a technical architecture for relapse prevention on WhatsApp.

Depression is episodic but longitudinal — 85% of patients experience recurrence within a decade, and each successive episode increases risk by 16%. Latin America's 73.9% treatment gap makes digital monitoring not just valuable but essential. Chile's stepped-care program demonstrated that structured low-intensity interventions delivered by non-specialists achieve 70% recovery versus 30% with usual care. SentirIA's WhatsApp-native approach positions it to replicate and extend this model across the region.

---

1. How a digital companion should adapt when PHQ-9 scores drop

The evidence on stepping down

The stepped-care model is the dominant framework for adjusting intervention intensity, with an umbrella review of 38 primary studies showing improved depression response at 3–6 months (RR = 1.52, 95% CI 1.30–1.78). Yet a critical gap exists: a scoping review found that very few studies report criteria for stepping *down* — nearly all focus on step-up decisions. No major platform (Woebot, Wysa, Tess, Spring Health, or SilverCloud) has published a formal adaptive step-down protocol based on score improvement.

The antidepressant tapering literature provides a useful analogy. A Lancet Psychiatry meta-analysis found that slow tapering *with psychological support* prevented relapse as effectively as antidepressant continuation, while fast tapering or tapering without support performed poorly. This principle translates directly: reduce digital support gradually and simultaneously strengthen self-management skills.

The risk of reducing too quickly is concrete. Depression relapse rates reach 50–54% within two years post-remission when treatment is discontinued (Vittengl et al., 2007). Holländare et al. (2011) demonstrated that internet-based CBT for relapse prevention reduced relapse from 37.8% to 10.5% in partially remitted patients over 10 weeks. Conversely, over-monitoring carries its own

danger: Goldberg et al. (npj Digital Medicine, 2024) found a 6.7% deterioration rate across app conditions, with evidence that frequent mood monitoring may maintain depression through negative processing bias.

Clinical thresholds for transitioning

The PHQ-9's validated thresholds provide the scaffolding for SentirIA's step-down logic. Remission is defined as PHQ-9 < 5; treatment response requires ≥50% reduction from baseline AND score < 10. A 5-point change constitutes the Minimal Clinically Important Difference (Löwe et al., 2004), though Bauer-Staeb et al. (2021) showed this varies by baseline severity — patients starting at PHQ-9 ≥ 20 may need a 14-point change to perceive meaningful improvement.

The recommended graduated withdrawal framework for SentirIA:

Phase	PHQ-9 range	Check-in frequency	Content focus
Acute	≥15	Daily	Crisis support, CBT techniques, safety monitoring
Response	10–14	Every 2–3 days	Behavioral activation, cognitive restructuring
Partial remission	5–9	Twice weekly	Skill consolidation, relapse prevention planning
Remission	<5	Weekly → biweekly	Wellness, resilience, protective factors
Maintenance	<5 sustained ≥8 weeks	Monthly	Self-monitoring reminders, booster content

Step-down should require PHQ-9 < 10 sustained for ≥2 consecutive assessments (minimum 4 weeks apart). Re-escalation triggers when PHQ-9 increases by ≥5 points from the lowest recorded score or returns to ≥10.

Shifting from treatment to wellness mode

The content transition is as important as the frequency change. Hoorelbeke et al. (2019) used network analysis of 85 remitted patients and found that resilience — not merely reduced negative affect — formed the central hub connecting perceived cognitive control, emotion regulation, and residual symptomatology. A meta-analysis of 101 RCTs (npj Digital Medicine, 2024) confirmed that digital resilience interventions improve positive mental health (SMD = 0.27) and resilience factors (SMD = 0.31).

For SentirIA's maintenance mode, the conversational tone should shift from empathic/supportive to empowering/coaching. Content should move from symptom management to values-based living, self-compassion exercises, and social connection reinforcement — protective factors identified by network analysis as forming a distinct community from risk factors. Behavioral activation prompts should evolve from prescriptive ("Try scheduling one pleasant activity today") to reflective ("What activities have been bringing you energy lately?").

How existing platforms handle improvement

Woebot delivers structured CBT through 10-minute daily check-ins but is rule-based and does not dynamically adjust pacing or content when users improve. Only 22% of users complete ≥4 weeks. Wysa can flag users crossing risk thresholds for human review (NHS deployment) but lacks a published step-down protocol. Tess/X2AI adjusts "ping" frequency (daily vs. biweekly) and routes users between modules based on emotional content, but has no systematic step-down tied to PHQ-9 improvement. Spring Health uses measurement-based care with panel management tools flagging overdue assessments and high-risk members, allowing stepping from therapy to coaching to digital experiences. SilverCloud (now Amwell) explicitly positions itself as a "recovery toolkit post-therapy" — essentially functioning as a step-down from higher-intensity care, with 49% of users with clinical depression showing clinical recovery (PHQ-9 moving from >9 to <9).

---

2. Early warning signs detectable through conversation

Linguistic markers that predict relapse

The most reliable text-based signals emerge from three domains. Absolutist language — words like "always," "nothing," "completely," "never" — is approximately 50% more prevalent in depression forums and 80% more prevalent in suicidal ideation forums compared to controls (Al-Mosaiwi & Johnstone, 2018; Cohen's d > 3.14). Critically, recovered individuals still show elevated absolutist word use, making this both a trait vulnerability marker and a potential relapse predictor.

First-person singular pronouns (I, me, my) reflect the self-focused attention characteristic of depression, confirmed by Edwards & Holtzman's (2017) meta-analysis as a more consistent marker than negative emotion words. Positive emotion word reduction correlates with PHQ-9 scores and predicts longitudinal deterioration (Burkhardt et al., 2021). A 2024 PNAS study (N=467) demonstrated that language sentiment evaluated by GPT-4 predicted changes in depressive symptoms at 3-week follow-up even after controlling for current mood — while LIWC-based bag-of-words sentiment did *not* predict future changes, suggesting that contextual semantic understanding substantially outperforms dictionary-based approaches.

An important caveat: a 2025 Nature study (N=218) found that linguistic features were not consistently associated with depressive symptoms across all text tasks, and platform matters enormously. Waterloo et al. (2014) documented significant cross-platform variation, with SMS-like messages more useful than social media posts for depression detection. WhatsApp conversations may therefore yield more reliable markers than publicly studied platforms.

The prodromal phase offers a detection window

Clinical evidence confirms a substantial prodromal period before full relapse. The DEEP-IN study

found 93.2% of patients reported a prodromal phase with mean duration of 7.9 months. A systematic review (Benasi et al., 2021) identified the most frequent prodromal symptoms: anxiety/tension, irritability (45%), sleep disturbance (45%), loss of interest, and reduced energy (43.8%). These symptoms are readily detectable through conversational patterns — increased references to sleep difficulties, reduced mentions of social activities, heightened rumination themes, and decreased future-oriented language.

The temporal dynamics matter for SentirIA's alert system. Using EWMA control charts on EMA data, Smit & Snippe (2019, Psychotherapy & Psychosomatics) detected increases in mean restlessness more than 2 months before onset of depressive symptoms. Confirmatory replication (2023, Psychological Medicine) showed that repetitive negative thinking was the most sensitive early sign (detected in 82% of cases before recurrence vs. 42% in stable remission), while high-arousal negative affect (stress, irritation, restlessness) was the most specific (45% before recurrence vs. 11% in remission).

Vocal biomarkers and behavioral signals

Voice analysis captures signals invisible in text. Depression reduces F0 variability (pitch becomes monotonic), slows speech rate, and increases pause duration. Kintsugi Voice achieved 71.3% sensitivity for detecting moderate-to-severe depression from ≥25 seconds of free-form speech (Annals of Family Medicine, January 2025), and their technology is notably language-agnostic — analyzing *how* someone speaks rather than *what* they say, making it directly applicable to Spanish-speaking populations. Kintsugi is releasing its models as open source in 2025.

Digital behavioral signals compound the predictive power. A systematic review (JMIR, 2023) found that the rate of behavioral anomalies was 71% higher in the 2 weeks before relapse compared to other periods. An LSTM model combining smartphone telemetry achieved AUC 0.88, sensitivity 0.83, specificity 0.81 for adolescent depression relapse prediction over a 14-day horizon, with severe reductions in geospatial mobility variance being the most significant predictor.

For WhatsApp specifically, SentirIA should track: response latency increases (withdrawal), message length shortening, reduced emoji diversity, shifts in time-of-day engagement (late-night activity spikes correlate with circadian disruption), decreased conversation initiation, and increased missed check-ins.

Critical slowing down before mood transitions

Van de Leemput et al. (2014, PNAS) provided the theoretical foundation by demonstrating that before transitions into or out of depression, temporal autocorrelation increases, variance increases, and correlations between different emotions increase — the same critical slowing down phenomenon observed before tipping points in ecosystems. The confirmatory TRANS-ID study (Smit, Helmich et al., 2025; N=37, ~524 assessments per person) found early warning signals preceded recurrence in 32.9% of participants — moderate sensitivity that is nonetheless clinically meaningful for identifying subgroups at imminent risk.

The response shift problem

SentirIA must account for the response shift phenomenon: after treatment, patients may paradoxically report higher symptoms on standard scales because they have become better at recognizing their own symptoms (recalibration response shift). Fokkema et al. (2013) showed that apparent worsening in a psychotherapy group disappeared once response shift was statistically adjusted. Practically, a slight PHQ-9 uptick after sustained improvement may reflect increased self-awareness, not actual relapse. SentirIA should incorporate longitudinal anchoring questions ("Compared to your best week this month…") and weight behavioral/linguistic signals alongside self-report scores.

---

3. Stepped-care frameworks and their application to SentirIA

International guidelines converge on similar thresholds

NICE (UK, 2022) uses a 4-step model: active monitoring for subthreshold depression (PHQ-9 5–9), low-intensity interventions for mild-moderate (5–14), high-intensity for moderate-severe (≥15), and specialist care for severe/complex cases (≥20). The 2022 update established PHQ-9 = 16 as the dividing line between "less severe" and "more severe" depression. Active monitoring review occurs at 2 weeks; low-intensity intervention review at 4–6 weeks.

The Dutch model adds a 5th step and includes a liaison-consultation function where a psychiatrist advises the GP every 6 weeks. The Dutch Depression Breakthrough Collaborative (2006–2008) showed 28% recovery within 3 months and another 27% between 3–6 months. Australia mandated stepped care nationally in 2015 through Primary Health Networks, with StepCare implementing systematic notification of deterioration. WHO's mhGAP-IG v3.0 (November 2023) uses an Assess-Decide-Manage flowchart emphasizing task-shifting to non-specialist health workers — directly relevant to SentirIA's positioning.

Escalation protocols require precision

The critical decision — when to escalate from digital to human care — requires specific thresholds:

- PHQ-9 Item 9 (suicidality): Any score ≥1 triggers additional screening; score ≥2 requires clinician contact within 24–48 hours; score = 3 demands same-day crisis intervention
- Total PHQ-9 ≥ 20: Immediate professional referral
- PHQ-9 increase ≥ 10 points between assessments: Urgent clinical review
- Non-response (< 5-point decrease after 8 weeks of guided intervention): Routine escalation
- PHQ-9 ≥ 15 persisting after 6–8 weeks of digital intervention: Professional referral

The Reliable Change Index for PHQ-9 — approximately 5–6 points based on test-retest reliability

of r = 0.82–0.84 — determines whether a change exceeds measurement error and constitutes clinically significant worsening or improvement.

Chile's program proves stepped care works in Latin America

Chile's National Depression Detection and Treatment Program stands as the gold standard. Araya et al. (Lancet, 2003) randomized 240 low-income women with major depression to stepped care versus usual care: 70% recovery in the stepped-care group versus 30% in usual care at 6 months, producing 50 additional depression-free days at an extra cost of ~$1.04 USD per depression-free day. Since 2006, Chile has provided universal access to depression treatment, with 80% of depressed patients now treated by primary care teams.

Colombia's DIADA project used digital screening tools in primary care, increasing detection rates from <0.1% to >10% at ~$2 USD per treated person. A scoping review (Lancet Regional Health Americas, 2024) identified 18 integrated mental health programs across 6 Latin American countries. Key challenges include: concentrated specialist availability in urban centers (leaving ≥45% of population unserved), cultural stigma, and treatment-resistant depression prevalence of 29% across 4 countries (ranging from 21% in Mexico to 40% in Brazil).

The PHQ-9 has been validated in Spanish with sensitivity 90.6% and specificity 84.5% for moderate depression (Urtasun et al.). WhatsApp's near-universal penetration across Latin America and Colombia's DIADA results demonstrating 100x increases in detection via digital tools together validate SentirIA's platform strategy.

A decision algorithm for SentirIA

```
SCREENING → PHQ-2 via WhatsApp (2 items, 30 seconds)
  IF PHQ-2 ≥ 3 → Full PHQ-9

PHQ-9 0–4 → Psychoeducation + wellness monitoring; re-screen 4 weeks
PHQ-9 5–9 → Active monitoring (weekly check-ins); re-assess 2–4 weeks
  IF no improvement after 4 weeks → Step up to guided intervention
PHQ-9 10–14 → Guided digital CBT intervention; weekly monitoring
  IF <5-point decrease after 8 weeks → Escalate to professional care
  IF ≥50% reduction sustained → Begin step-down protocol
PHQ-9 15–19 → Recommend professional evaluation within 1–2 weeks
  Provide supported digital intervention while awaiting care
PHQ-9 ≥ 20 → IMMEDIATE escalation; daily check-in until connected to clinician

AT ANY TIME: Item 9 ≥ 1 → Suicidality assessment protocol
  Item 9 ≥ 2 → Human contact within 24 hours
```

4. Keeping users engaged after they feel better

Attrition is steep but partially by design

Real-world retention in mental health apps is brutally low: median 30-day retention is 3.3% for Android mental health apps (Frontiers in Psychiatry, 2022). The JAMA Psychiatry 2025 meta-analysis of 79 RCTs found uptake was high (92%) but user attrition reached 18.6% during intervention and 28.4% by follow-up. Chien et al. (2020, JAMA Network Open) identified 5 engagement patterns in 54,604 digital CBT patients: only 10.6% were high persistent engagers, while 36.5% were low engagers.

However, some attrition reflects success. The concept of "e-attainment" (npj Digital Medicine, 2025) describes discontinuation because personal goals have been met. Distinguishing this from risky dropout requires tracking symptom trajectories alongside engagement patterns: declining engagement with stable/improving PHQ-9 scores suggests healthy graduation; declining engagement with plateauing or worsening scores signals risk.

Dropout prediction models using random forest achieved 89% precision, and machine learning using baseline plus usage data achieved AUROC 0.72 with sensitivity 76% — exceeding the 65–70% threshold where clinicians are willing to act (Moshe et al., JMIR 2022). Key risk factors for dangerous dropout include higher baseline severity, male gender, lower education, and poor early engagement.

What actually brings users back

Human support has the strongest evidence for maintaining engagement, with meta-analyses showing medium positive effect sizes for human-supported versus unsupported digital interventions. Hybrid models with AI plus minimal guidance achieve 60–70% completion rates versus 30–40% for fully automated apps. Personalized messages significantly outperform generic ones across 8 studies.

Gamification, surprisingly, does not help depressed populations. A meta-analysis of 38 studies (N=8,110) in JMIR Mental Health found gamification was not a significant moderator of depression outcomes ($\beta = -0.03$, $p = .38$) or adherence ($\beta = -1.93$, $p = .40$). The likely explanation: depression's core feature of anhedonia renders reward-based mechanics less effective. SentirIA should instead rely on simple progress visualization, habit-anchoring to daily routines, and personalized micro-acknowledgments.

The most directly relevant evidence comes from Malins et al. (2020, British Journal of Clinical Psychology): personalized smart-messaging after CBT completion, where patients wrote advice for themselves across three scenarios (doing well, early warning signs, full relapse), yielded only 11% relapse over 25 weeks. This self-authored approach is highly adaptable to WhatsApp.

Therapeutic alliance with AI predicts continued use

The Therabot RCT (Heinz et al., NEJM AI, 2025, N=210) — the first clinical trial of a generative AI

therapy chatbot — found participants rated therapeutic alliance comparable to human therapists, with large effect sizes for depression (d = 0.845–0.903). Users engaged for over 6 hours across 4 weeks and frequently initiated conversations spontaneously, including at unusual hours, suggesting genuine relationship formation.

A meta-analysis of 117 studies (279,791 participants) confirmed that stronger therapeutic relationship with a digital mental health intervention correlated with higher engagement (d = 0.59, p = .019). Social-oriented chatbots outperform task-oriented ones. For SentirIA, this means prioritizing empathic, personalized exchanges over directive task delivery, and maintaining relational continuity by referencing past conversations.

WhatsApp-specific engagement advantages

WhatsApp messages achieve >90% open rates versus 2–5% for app push notifications — a transformative advantage. Voice notes are culturally preferred in Latin America and serve as low-barrier mood check-ins while simultaneously capturing vocal biomarkers. However, over-messaging risks the user blocking SentirIA's number entirely — an irreversible loss. Brazilian studies found WhatsApp health engagement declined significantly after 4 weeks, suggesting that content variety and personalization are essential for sustained participation.

Effective notification principles: 64% of users delete an app receiving ≥5 notifications weekly; gain-framed messages ("Attending your appointment can help you maintain your mental health") significantly outperform loss-framed messages for mental health engagement. During maintenance, SentirIA should limit messages to 2–3 per week maximum, varied in format (text, emoji-based mood scales, voice note prompts, images), and timed to culturally appropriate hours.

Framing matters enormously. Rather than "You need ongoing monitoring because depression can return," maintenance engagement should be positioned as emotional fitness: "Your wellness check-in," "Building your resilience toolkit," or "Your emotional fitness routine" — analogous to not stopping gym attendance because you are fit.

---

5. How long to monitor after recovery

Clinical guidelines establish minimum durations

The convergence across guidelines is clear. APA recommends 4–9 months of continuation pharmacotherapy at full dose after remission for first-episode MDD, with indefinite maintenance after a 3rd episode (or 2nd with aggravating factors). NICE specifies ≥2 years for patients with 2+ episodes and significant functional impairment, with reviews every 6 months minimum. CANMAT (2023 update) extended their recommendation to up to 12 months after remission, with longer durations for higher-risk patients. WHO consensus suggests 4–12 months for first

episodes, ≥2 years for recurrent depression, and indefinite treatment for 3+ episodes.

Relapse rates peak in the first year, then plateau slowly

The NIMH Collaborative Depression Study (Mueller et al., 2000; N=318, 10-year follow-up) provides the most granular longitudinal data: approximately 30% cumulative recurrence at 6 months, 40% at 12 months, and two-thirds experiencing recurrence within a decade. Each successive episode increases recurrence risk by 16% (OR = 1.16, 95% CI 1.03–1.31). Hardeveld et al. (2010) found 60% recurrence at 5 years and 85% at 15 years in specialized care settings.

Median time to new episode is approximately 40 months with continued treatment versus 13 months without (Baldessarini et al., 2016). This 3:1 ratio quantifies the protective value of sustained monitoring.

Residual symptoms are the strongest modifiable predictor

Paykel et al.'s landmark 1995 study found that 76% of patients with residual symptoms (HDRS ≥ 8) relapsed within 10 months versus 25% without — a 3× difference. Nierenberg et al. established that 82.4% of acute-phase responders retain at least one residual symptom; the most common are sleep problems (44%), fatigue (38%), and anhedonia. On the PHQ-9, a score of 5–9 represents the "residual symptom" zone warranting heightened vigilance — any score ≥5 in a remitted patient signals elevated relapse risk.

This has direct implications for monitoring duration. The proposed risk-stratified framework:

Risk level	Profile	Monitoring duration
Low	First episode, full remission (PHQ-9 <5), no risk factors	6–9 months
Moderate	First episode with residual symptoms (PHQ-9 5–9) OR 2nd episode with full remission	12–18 months
High	2+ episodes with residual symptoms, comorbid anxiety, childhood maltreatment	≥24 months
Very high	3+ episodes, chronic history, multiple risk factors	Indefinite

Digital monitoring appears cost-effective, especially in Latin America

Internet-delivered CBT showed a "dominant" ICER (lower costs with similar effectiveness) versus standard care in a UK NHS study of 27,540 patients (Catarino et al., Nature Mental Health, 2023). WhatsApp-based monitoring has near-zero marginal technology cost per user, with the primary cost driver being any human support component. In Latin American settings where face-to-face mental health resources are scarce, automated digital monitoring is likely highly cost-effective.

For relapse prevention specifically, psychological interventions reduce 12-month relapse risk with HR = 0.60 (95% CI 0.48–0.74) versus control (Breedvelt et al., Nature Mental Health, 2024), with the strongest effect for patients with 3+ prior episodes. App-based interventions for moderate-severe depression achieve a medium effect size (SMD = 0.50, 95% CI 0.40–0.61) in a meta-

analysis of 13 RCTs.

---

6. Technical architecture for SentirIA on WhatsApp

Adaptive scheduling within WhatsApp's constraints

The WhatsApp Business API imposes specific constraints: new business portfolios start at 250 unique users per 24-hour rolling period, scaling to 1K → 10K → 100K → Unlimited based on quality and volume. Business-initiated messages outside the 24-hour customer service window require pre-approved template messages. SentirIA should pre-approve multiple templates across risk levels and assessment types, categorized as "utility" messages (lower cost than marketing).

The recommended architecture uses an event-driven pipeline (AWS EventBridge + SQS or RabbitMQ) that processes scheduling decisions based on risk-level changes. Cron-based schedulers trigger routine check-ins; event-driven escalation activates when anomaly detection flags risk increases. The three-tier scheduling protocol:

- Low risk (PHQ-9 < 5): PHQ-2 biweekly + 1 conversational prompt weekly (~2 min/week user burden)
- Medium risk (PHQ-9 5–14): PHQ-2 weekly, PHQ-9 biweekly, 2–3 conversational prompts weekly, weekly voice note request (~5 min/week)
- High risk (PHQ-9 ≥ 15): Daily single-item check-in + weekly PHQ-9 + daily voice note encouraged + immediate clinician notification (~3–5 min/day)

Anomaly detection algorithms validated for depression monitoring

EWMA (Exponentially Weighted Moving Average) control charts are the strongest recommendation, having been directly validated for depression relapse detection (Psychological Medicine). Applied to EMA restlessness data, EWMA achieved sensitivity 68.2% and specificity 73.7% for recurrence — and detected changes >2 months before onset. The formula is straightforward: EWMA_t = λ × xt + $(1 - \lambda) \times EWMA$_{t−1}, with λ = 0.05–0.20 (smaller values weight history more, better for detecting gradual shifts).

CUSUM (Cumulative Sum) detects small persistent shifts in mean — capable of identifying a 1σ shift in ~10 observations. Self-starting CUSUM variants calibrate without pre-specified baselines, crucial for new users. Bayesian Online Change-Point Detection is ideal for streaming WhatsApp data, processing one observation at a time. SentirIA should run dual EWMA + CUSUM monitoring during high-risk periods.

For baseline establishment, 4–6 observations suffice for initial calibration; 10+ observations enable reliable trend detection; and 30+ observations support idiographic prediction models. Group-based trajectory modeling has identified 6 PHQ-9 subgroups: high/nonresponse, high/response, moderate-severe/nonresponse, moderate-severe/response, moderate/remission,

and moderate/nonresponse.

Building personalized relapse signatures

The N-of-1 framework (Scientific Reports, 2023) using LSTM Encoder-Decoder Anomaly Detection achieved balanced accuracy ≥71% with false alarm rate ≤2.3 alarms/patient/year and median relapse detection 2–3 weeks in advance, validated across 277 MDD patients with ≥1 year follow-up. The WARN-D study (Leiden University) demonstrated that rising autocorrelation (r=0.51), variance (r=0.53), and network connectivity (r=0.42) significantly increased a month before symptom transitions.

For SentirIA's implementation: during weeks 1–4, collect daily conversational affect ratings, response latency, message length, emoji patterns, and voice note characteristics. After 30+ observations, train individualized models. Apply EWMA to key features with 14–21 day moving windows. Flag alerts when ≥2 features simultaneously cross control limits.

Multimodal fusion of voice and text

The optimal approach for WhatsApp voice notes: receive the opus/ogg file → convert to 16-kHz WAV → extract features using openSMILE (eGeMAPS 88-dimensional feature set) or Wav2vec2.0 → run depression detection → fuse with text features. Late/decision-level fusion is recommended over early fusion because it handles missing modalities gracefully — critical when voice notes are optional. LSTM-CNN hybrid + BERT decision-level fusion achieves accuracy 94.3% and F-score 94.5%.

Kintsugi's technology, being released as open source in 2025, analyzes acoustics only (not speech content), providing inherent privacy protection and language-agnostic operation applicable to Spanish-speaking populations. Voice-psychopathology correlations reach $r \geq 0.80$ with just 2-minute speech recordings at 2-day intervals (Stassen et al.).

Privacy architecture for Latin American data protection

SentirIA must comply with Brazil's LGPD (AES-256 encryption, explicit consent for health data, penalties up to 2% of revenue), Colombia's Law 1581 (all databases registered in National Register, SIC enforcement), and Mexico's LFPDPPP (explicit written consent for health data). The recommended privacy-by-design architecture:

- On-device processing: Extract voice features locally; transmit only 88-dimensional feature vectors, never raw audio
- Data minimization: Store derived features and scores only; delete raw conversation text after 30 days; retain PHQ-9 scores and trend metrics indefinitely
- Pseudonymization: Separate identity data from health data using cryptographic tokens
- Geofenced deployment: Region-specific storage (Brazilian data stays in Brazil)
- Federated learning: FL-PDP framework with differential privacy via Laplacian noise achieves ~97.9% accuracy versus 97.4% centralized, with minimal accuracy loss

The minimum viable monitoring protocol

The least-intrusive approach that still catches relapses combines the PHQ-2 as a screening gate (sensitivity 83%, specificity 92% at cutoff ≥3; Kroenke et al., 2003) with passive behavioral monitoring. For stable patients: PHQ-2 biweekly, one open-ended conversational prompt weekly, passive tracking of response latency and message patterns, and a monthly optional voice note — total burden approximately 2 minutes per week.

Single-item daily tracking ("On a scale of 0–3, how much have you been feeling down today?") has been validated for detecting early warning signals in depression transitions (Wichers et al., replicated). This can serve as the ultra-brief daily assessment for medium-risk users, while passive WhatsApp signals (message frequency changes, response time patterns, conversation initiation patterns) provide supplementary data with zero user burden.

Conclusion

SentirIA has an opportunity to pioneer what no existing digital mental health platform has published: a validated, adaptive step-down protocol for AI companions with personalized relapse detection. The clinical evidence strongly supports several core design decisions. First, the 5-point PHQ-9 change threshold and the <5 remission criterion should govern all step-up/step-down transitions, with mandatory sustained improvement across ≥2 assessments before stepping down. Second, EWMA-based monitoring of conversational features — particularly repetitive negative thinking, absolutist language, and response latency — can detect deterioration 1–2 months before clinical relapse, giving SentirIA an actionable early warning window.

Three findings deserve special emphasis for implementation priority. The smart-messaging approach (Malins et al.) where patients write self-advice for three scenarios achieves only 11% relapse and is immediately implementable on WhatsApp. Gamification should be avoided for depressed populations due to anhedonia-mediated reward insensitivity. And the Kintsugi open-source release creates a direct path to language-agnostic vocal biomarker integration without licensing barriers.

The Latin American context provides both challenge and validation. Chile's stepped-care program achieving 70% recovery at ~$1/depression-free day proves the model works in resource-constrained settings. Colombia's DIADA project demonstrating 100x increases in detection through digital screening validates the platform approach. WhatsApp's >90% message open rates and cultural ubiquity make it the optimal delivery mechanism. SentirIA's core technical stack — EWMA + CUSUM anomaly detection, late-fusion multimodal analysis, PHQ-2 screening gating, and privacy-by-design architecture — is supported by converging evidence and feasible within WhatsApp's API constraints. The minimum viable monitoring burden of ~2 minutes per week for stable patients makes sustained engagement realistic rather than aspirational.