

What Conventional Severity Misses: Discordant Presentation and the Case for AI-Supported Severity Detection, with Evidence from an Ethiopian Rehabilitation Cohort

Yared Alemu

ya@aya-innovation.com

AYA-Innovation

Patrick Ohiomoba

AYA-Innovation

Octavian Balatel

AYA-Innovation

Yoeal Kahsay

AYA-Innovation

Fikirte Weldearegay

AYA-Innovation

Tsion Fekadu

AYA-Innovation

Lidya Hailu

AYA-Innovation

Delessa Neger

AYA-Innovation

Research Article

Keywords: masked psychiatric presentation, symptom under-detection, clinical defensiveness, response bias in self-report screening, complex trauma in women, measurement-based care, cultural idioms of distress, structural and gender-based violence, the detection gap in global mental health, voice-based artificial intelligence, sub-Saharan Africa, Ethiopia

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10696972/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

What Conventional Severity Misses: Discordant Presentation and the Case for AI-Supported Severity Detection, with Evidence from an Ethiopian Rehabilitation Cohort

Yared Alemu¹, Patrick Ohiomoba¹, Octavian Balatel¹, Yoeal Kahsay², Fikirte Weldearegay², Tsion Fekadu², Lidya Hailu², and Delessa Neger^{1,3}.

¹ AYA-Innovation, Addis Ababa, Ethiopia

² Unity University, Addis Ababa, Ethiopia

³ University of Gondar, Gondar, Ethiopia

Corresponding author: Yared Alemu, AYA-Innovation, Addis Ababa, Ethiopia. Email: ya@aya-innovation.com

Abstract

Background. Standardized symptom screeners are increasingly used to gauge psychiatric severity in low-resource African settings, including instruments validated for Ethiopian use, such as the PHQ-9, GAD-7, and ACE. Yet in populations defined by extreme structural adversity, low symptom endorsement may reflect defended or

constrained disclosure rather than the absence of disorder. We examined this discordance in a program population of women admitted to rehabilitation on the basis of severe poverty, histories of street-based sex work, and chronic unemployment — a profile in which clinically significant trauma is expected to be near-universal.

Methods. We analyzed first-completed and repeated behavioral-health screens from 3,617 female clients served through a digitally supported rehabilitation and screening program in Addis Ababa. We described baseline severity across PHQ-9, GAD-7, ACE, CAGE, and SDS; quantified discordant presentation (high-risk indicators with low or subthreshold mood/anxiety scores); examined item-level and survey-level safety signals against total-score thresholds; derived an exploratory five-profile model; and analyzed first-to-last symptom change among baseline-elevated clients using seven-day and 30-day windows.

Results. Conventional thresholds made the population appear low in severity: only 11.3% met moderate-or-higher PHQ-9/GAD-7 criteria, against 35.5% with high ACE exposure and 85.7% with any adverse childhood exposure. Among 1,250 clients with at least one high-risk signal, 78.6% scored below the moderate mood/anxiety threshold — a masked group 3.7 times larger than the group captured by conventional severity. Total-score thresholds concealed safety risk: 68.8% of clients with a positive PHQ-9 self-harm item, and 70.7% with a survey-level suicide flag, scored below the PHQ-9 moderate cut-off. Among baseline-elevated clients, repeated measures showed large improvements (median PHQ-9 change: -11.5; median GAD-7 change: -10.0). In practical terms, for every five clients carrying a documented high-risk signal, roughly four would be recorded as not warranting care, and about seven of every ten clients flagged for suicide risk would pass a total-score safety screen undetected.

Conclusions. In a population where complex trauma is structurally near-certain, conventional severity scoring systematically understates clinical and safety-relevant risk. We argue that the dominant low scores are best understood as defensive or constrained presentation — shaped by stigma, gendered shame, cultural and religious idioms of distress, poverty-normalized suffering, and a strength schema that discourages disclosure — rather than as evidence of resilience or low need. We critique the “resilient African woman” narrative as a documented hazard to women’s health when read off low self-report scores, and we argue that Africa’s mental health crisis comprises two distinct gaps: a severe workforce gap and a clinical defensive gap. A multimodal, voice-based AI infrastructure is uniquely positioned to address both — extending scarce expertise and adding an independent, stigma-resistant detection channel that recovers severity conventional instruments miss.

Keywords: masked psychiatric presentation; symptom under-detection; clinical defensiveness; response bias in self-report screening; complex trauma in women; measurement-based care; cultural idioms of distress; structural and gender-based violence; the detection gap in global mental health; voice-based artificial intelligence; sub-Saharan Africa; Ethiopia

1. Introduction

Measurement-based care has become a near-universal expectation in global mental health, and brief self-report instruments are its workhorses. The nine-item Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001), the seven-item Generalized Anxiety Disorder scale (GAD-7; Spitzer et al., 2006), and the Adverse Childhood

Experiences inventory (ACE; Felitti et al., 1998) are inexpensive, quick, and increasingly available in local languages. In Ethiopia, the PHQ-9 has been validated against a gold-standard psychiatric interview and found to be reliable and valid for diagnosing major depression among Ethiopian adults, and a primary care study similarly concluded that brief screening questionnaires, including the PHQ-9, were valid for detecting depression in that setting (Gelaye et al., 2013; Hanlon et al., 2015). The appeal of such instruments is obvious in a system with roughly 0.11 psychiatrists per 100,000 population — approximately one-ninth of the WHO-referenced benchmark of 0.90 — where specialist assessment remains structurally unavailable to the overwhelming majority of those who need it (Federal Ministry of Health, Ethiopia, 2020; World Health Organization, 2017).

Yet the instrument that makes severity legible can also make it invisible. Self-report screening rests on a chain of assumptions: that the respondent can read or hear the item, understands the construct it names, recognizes the experience in herself, and is willing to disclose it to an assessor recording the answer. Each link can break, and they break most predictably in the populations carrying the heaviest burden — those facing stigma, gendered shame, low literacy, displacement, and the normalization of suffering under chronic poverty (Kohrt et al., 2016; Hanlon et al., 2015; Ventevogel et al., 2013). When the chain breaks, the instrument does not return a missing value. It returns a low score, which clinicians, programs, and funders read as low need. This conflation of a measurement artifact with a clinical fact is the central problem this paper addresses.

We examine a program population in which this failure mode should be at its most extreme. The clients are women admitted to a residential rehabilitation pathway on the basis of a deliberately severe set of criteria: extreme poverty, a history of street-based sex work, and chronic unemployability. This is not a general community sample in which a low symptom score might plausibly indicate wellness. It is a sample selected for cumulative, intersecting trauma — a population in which the prior probability of clinically significant psychiatric burden is very high before any instrument is administered (Alemu et al., 2026). If standardized severity scoring nonetheless renders this population as predominantly low in severity, the more defensible inference is not that the women are unburdened, but that the measurement is failing to register what is there.

Our argument proceeds in three movements. First, we situate the problem within the documented mental health crisis of sub-Saharan Africa — a crisis defined less by the absence of disorder than by the near-total failure to detect it — and within the cultural, religious, economic, and gendered forces that suppress symptom disclosure in this population. Second, we show empirically that conventional PHQ-9/GAD-7 severity categories capture only a small fraction of the high-risk clients in this sample, while trauma, substance, and safety indicators are far more prevalent than mood/anxiety thresholds alone would suggest, and we develop the interpretation that the dominant low scores reflect *defensive or constrained clinical presentation* rather than resilience or recovery. Third, we argue that this severity gap is precisely what multimodal, voice-based artificial intelligence is positioned to close — and that AI addresses two failures at once: the severe workforce gap and the clinical defensive gap documented in this analysis.

A note on language. Throughout, we distinguish two senses of “defensiveness.” The first is the client’s defended presentation — constrained or minimized disclosure shaped by safety, shame, and meaning. The second is the system’s defensive posture — the readiness to treat a reassuring number as a clinical fact and to stop looking. The data speak most directly to the first; the consequences fall hardest because of the second. We use “masked” or “discordant” presentation as the conservative empirical term and reserve “defensive” for the interpretive argument.

2. Background

2.1 Defensive and constrained presentation: a problem at the foundation of psychological measurement

The interpretive problem this paper confronts — that a low symptom score may reflect defended non-disclosure rather than wellness — is neither specific to Africa nor an artifact of any one instrument. It is among the oldest and most thoroughly documented problems in clinical psychology. From its psychoanalytic beginnings, the discipline has recognized that the mind protects itself from overwhelming affect through defense: repression, denial, suppression, dissociation, and reaction formation were catalogued as mechanisms by which distress is kept out of awareness and out of speech (Freud, 1936/1966; Vaillant, 1992). Whatever one's view of psychoanalytic theory, its empirical descendants are robust. A substantial body of research shows that many people who appear well on self-report scales are not, and that the appearance is produced by defensive processes rather than by health.

A particularly influential demonstration comes from the work of Shedler, Mayman, and Manis (1993) on the “illusion of mental health.” Using physiological criterion measures, they showed that standard mental health scales cannot distinguish genuine psychological health from its defensive facade: a substantial subgroup of people who score as healthy are in fact distressed individuals maintaining an illusion of health through defensive denial, and this illusory health carries physiological costs and elevated medical risk. Converging evidence on repressive coping shows the same split between channels — “repressors” report low anxiety on questionnaires while exhibiting high physiological and behavioral stress responses (Weinberger, Schwartz, & Davidson, 1979). The self-report channel, in other words, is systematically and predictably blind to defended distress, and it is blind in the reassuring direction.

This vulnerability is reinforced by two further, well-studied measurement phenomena. Social desirability and self-deceptive response styles lead respondents to under-report symptoms that are stigmatized or ego-threatening, biasing prevalence estimates downward (Paulhus, 1984; Van de Mortel, 2008). And alexithymia — a difficulty identifying and verbalizing one's own emotional states — leaves a meaningful subset of patients literally unable to render internal distress into the language a questionnaire requires, so that severe disturbance presents as low symptom endorsement or as somatic complaint (Sifneos, 1973). None of these mechanisms is exotic or culturally specific; they are baseline features of human psychology that any self-report instrument must contend with. The population we describe is not an exception to this global rule but an intensification of it: the same defensive processes operate, amplified by stigma, trauma, and structural adversity.

We therefore situate our findings within this global and historical frame rather than treating defensive presentation as a local curiosity. The question the data raise — whether a low score means wellness or defended distress — is the question clinical psychology has put to its instruments for a century. What is new here is the scale and the stakes: a full-population dataset in a setting where the consequences of getting the answer wrong are measured in untreated trauma and unrecognized suicidality.

2.2 The psychodynamic lens: trauma and the distance between what is said and what is meant

A second framework clarifies why the problem is most acute in trauma. The psychodynamic tradition rests on a premise directly relevant to screening: that things are frequently not what they appear, and that the account a person can give of herself is shaped, and often constrained, by what she cannot bear to know or to say (Shedler, 2010). Clinically, this is not mysticism; it is the routine observation that presenting complaints, surface affect, and self-report frequently understate or misrepresent the disturbance beneath them, and that the work of assessment is to read past the surface. Contemporary psychodynamic psychotherapy has a substantial evidence base, and its core assessment stance — that the manifest and the latent diverge — is precisely what a brief self-report screen cannot accommodate (Shedler, 2010).

Trauma further intensifies this divergence. The defining feature of complex and post-traumatic presentations is the active management of unbearable experience: avoidance, numbing, emotional constriction, and dissociation are diagnostic features, not incidental ones (Bryant, 2019; Herman, 1992). A traumatized woman may have organized her entire functioning around not feeling and not disclosing the index events, and the dissociative subtype of post-traumatic stress disorder is characterized precisely by detachment and emotional shutdown that can read, on a mood checklist, as the absence of symptoms (Lanius et al., 2010). Van der Kolk's (2014) synthesis makes the mechanism vivid: traumatic experience is frequently stored and expressed somatically and behaviorally rather than narrated, so that the body and the clinical relationship carry information the questionnaire never receives. In a population defined by sexual violence, exploitation, and chronic adversity, this is not a fringe consideration; it is the expected presentation.

The convergence of the measurement and trauma literatures yields a clear expectation for our data. In a population this heavily traumatized, low self-report scores should be common and should not be read at face value, because both the defensive architecture of the mind in general and the specific phenomenology of trauma predict exactly this pattern of under-disclosure. The empirical sections that follow confirm the prediction at scale, and the remainder of this Background section examines the cultural, religious, economic, and gendered forces that, in this setting, drive the same suppression still harder.

2.3 The mental health crisis in Africa is a detection crisis

The scale of unmet mental health need in sub-Saharan Africa is now well documented, and its defining feature is not invisibility of disorder but invisibility of the person to the care system. An estimated 150 million people across Africa live with a mental health condition, against a care infrastructure that reaches only a small minority (World Health Organization Regional Office for Africa, 2025). In sub-Saharan Africa, fewer than one in ten people living with mental illness access professional mental health services (Komu et al., 2025; Sankoh et al., 2018). The treatment gap is, at root, a detection gap: most people who need care are never identified at the point of first contact. Studies of depression recognition in Ethiopian primary care have found detection rates approaching zero without structured screening (Fekadu et al., 2017; Hanlon et al., 2008), and across the continent the workforce that might close that gap does not exist — Africa has roughly 1.4 mental health professionals per 100,000 population, against a global average near 9 and a WHO benchmark far higher (Wondimagegn et al., 2023; World Health Organization, 2017).

Ethiopia sits at the severe end of this distribution. The country has approximately 0.11 psychiatrists, 0.58 psychiatric nurses, and 0.045 clinical psychologists per 100,000 population (Federal Ministry of Health, 2020; World Health Organization, 2017). The burden these few providers face is not static: it has been amplified by years of armed conflict, mass displacement, and economic shock. Among internally displaced persons in conflict-affected northwest Ethiopia, prevalence estimates for depression and anxiety have been reported at 79.6% and 74.6%, respectively (Bekeko et al., 2025). Similarly, studies among war-affected populations in Ethiopia report substantial levels of post-traumatic stress disorder (PTSD) symptoms, with gendered vulnerability shaping post-traumatic stress in these settings (Deyessa et al., 2025; Simie et al., 2025; Tinsae et al., 2024). Trauma-related depression, anxiety, and PTSD in these settings are not independent diagnoses but a tightly comorbid cluster that amplifies risk and impairs functioning, treatment adherence, and recovery (Charlson et al., 2019; Patel et al., 2018).

Two features of this crisis are decisive for the present paper. First, the detection gap is the binding constraint: a system that cannot identify need at first contact cannot triage, treat, or follow up, however well-designed its downstream services (Kakuma et al., 2011; Saraceno et al., 2007). Second, the instruments deployed to close the detection gap — literate, self-report symptom screeners — are themselves vulnerable to the same forces that drive the crisis, because stigma, trauma normalization, poverty, and culturally specific idioms of distress all

suppress the very disclosures the screeners depend on (Kohrt et al., 2014; Lund et al., 2018; Thornicroft et al., 2017). The crisis and the inadequacy of its standard measurement tools are therefore entangled. This is the context in which our population must be understood, and the context in which the case for an alternative detection infrastructure becomes urgent rather than speculative.

2.4 Cultural idioms of distress and the grammar of disclosure

Distress in Ethiopia is frequently expressed and understood through explanatory frameworks that a depression or anxiety scale was never designed to capture. Seizure-like episodes, possession states, and somatic complaints may be interpreted through Zar (spirit affliction), Buda (the evil eye), Azurit, and related idioms, and addressed through holy water, prayer, religious ritual, or traditional healing rather than psychiatric care (Grisaru et al., 1997; Kleinman, 1980). Across sub-Saharan Africa, religious and traditional pathways frequently precede or substitute for biomedical care for emotional and functional disorders (Kalu, 2008). These frameworks are not errors; they carry meaning, reduce isolation and self-blame, and locate suffering within a shared moral world.

But they also reroute experience away from the vocabulary presupposed by PHQ-9 and GAD-7 items. Anthropological research in global mental health has shown that distress is commonly expressed through culturally specific idioms of suffering that do not map onto standardized psychiatric questionnaires, and that somatic presentation — bodily complaint rather than reported mood — is a dominant mode in many low- and middle-income settings (Hinton & Lewis-Fernández, 2010; Kohrt et al., 2014). A woman who organizes her suffering around spiritual affliction or bodily symptom can endorse a few mood and anxiety items and remain severely unwell. The instrument records absence; the distress has simply taken a form that the instrument cannot read.

This is compounded by the social grammar of disclosure. Self-report screening presumes that naming one's symptoms to an assessor is safe and appropriate. Where mental illness is stigmatized, and disclosure can carry consequences for marriage, family standing, and livelihood, under-reporting becomes a learned protective strategy rather than conscious deception (Thornicroft et al., 2017; Tomlinson et al., 2009). The resulting bias is systematic and direction-specific: the instrument is most likely to be wrong, and wrong in the reassuring direction, precisely in the people who are most unwell. Standard social-desirability and self-report research confirms that sensitive, stigmatized experiences are differentially under-reported on direct questionnaires, biasing prevalence estimates downward (Bauhoff, 2011; Van de Mortel, 2008).

2.5 Religious framing and the meaning of suffering

Religious meaning-making is a profound source of strength for many Ethiopian women, and we do not pathologize it. But certain religious framings of suffering interact with measurement in ways that suppress symptom endorsement. When hardship is understood as a trial to be endured with faith, when complaint is experienced as ingratitude or spiritual weakness, and when relief is sought through prayer and holy water before — or instead of — clinical care, the act of endorsing depressive or anxious symptoms to a screener can feel both spiritually inappropriate and practically pointless (Tadesse, 2025; Kleinman, 1980).

Endurance is reframed as a virtue and disclosure as its failure. This is not a marginal consideration in a country where religious institutions are often the first and most trusted point of contact for people in distress, and where religious institutions, and holy-water sites function as *de facto* mental health settings (Baheretibeb et al., 2021). A scale that rewards admission of symptoms will read a faith-shaped stance of endurance as wellness. It is closer to a silence with reasons — and the silence is reinforced, not relieved, by the meaning system that surrounds it. A culturally responsive measurement approach must therefore treat low endorsement in this context as a question to be investigated rather than an answer to be accepted.

2.6 Economic adversity and the normalization of distress

For clients admitted under the criteria of extreme poverty and chronic unemployability, distress is not an episode against a calm baseline; it is the baseline. Chronic adversity recalibrates what counts as normal. Symptoms that would alarm a clinician — pervasive low mood, sleep disturbance, hopelessness about the future — may be experienced by the client as unremarkable features of a hard life rather than as a condition warranting endorsement (Felitti et al., 1998; Patel et al., 2018).

This matters mechanically for self-report. PHQ-9 and GAD-7 items ask the respondent to judge how much a problem has bothered her, relative to an implied norm of feeling well. Where feeling unwell is the unbroken norm, the implicit comparison collapses, and endorsement falls. Poverty does not only cause distress; it also normalizes it, and normalization is invisible to a self-report cut-off. The economic dimension is therefore not merely a risk factor sitting beside the measurement problem — it is one of the engines of the measurement problem. The same chronic adversity that produces the psychiatric burden also dampens its expression on the instruments meant to detect it, which is why the burden in this population is so reliably underestimated by symptom totals.

2.7 Gender, shame, and the secrecy of sexual violence

Gendered vulnerability runs through every layer of this population, and it is the layer most consequential for measurement. Sexual and reproductive health and rights remain unmet for millions of women and girls worldwide, a gap that widens sharply under poverty, displacement, and conflict (Starrs et al., 2018). Histories of street-based sex work and the conditions that produce it — displacement, destitution, exploitation, and frequently sexual and gender-based violence (SGBV) — are precisely the experiences least likely to be disclosed on a brief screen. The Ethiopian evidence-based on SGBV in conflict-affected settings is stark: in a community-based study of 5,171 women of reproductive age in war-affected Tigray, 43.3% had experienced at least one form of gender-based violence, and among sexual-violence survivors, 82.2% were raped, of whom 68.4% were subjected to gang rape (Fisseha et al., 2023). Independent studies in conflict-affected Amhara and North Shewa report similarly elevated burdens across emotional, physical, and sexual domains (Asefa et al., 2024; Dellie et al., 2024), and research documenting the mass-trauma and mental health sequelae of conflict-related sexual violence in Tigray corroborates the severity of this pattern (Abreha et al., 2025). Critically, services almost never reach survivors: 89.7% of survivors in the Tigray study received no post-violence medical or psychological support of any kind (Fisseha et al., 2023), and trauma severity is seldom even identified, because routine pathways screen for physical injury rather than psychological harm.

Shame, fear of judgment, fear of consequences, and the protective habit of secrecy that survival under exploitation demands all push disclosure downward (Kizilhan et al., 2020). Women in conflict-affected Ethiopia carry more than twice the odds of PTSD relative to men, driven substantially by sexual-violence exposure (Simie et al., 2025), yet that exposure is the experience most heavily defended against in face-to-face self-report. The empirical signature in our data is consistent with this: among clients with a recorded sexual-abuse flag, 72.9% nonetheless scored below the moderate mood/anxiety threshold (Section 4). The gap between a documented history of sexual violence and a low anxiety score is not plausibly explained by the absence of distress; it is far better explained by the predictable silence that gendered shame imposes — a silence the instrument cannot distinguish from health.

The structural point is that in this population, the very characteristics that define eligibility — poverty, sexual exploitation, social exclusion — are simultaneously the strongest drivers of trauma and the strongest suppressors of its disclosure. Adversity and under-detection are not independent; they are produced by the same conditions, which is why severity here is so reliably masked. This is the conceptual core of the paper: in a

population built from intersecting structural harms (Collins, 2000; Crenshaw, 1991), low symptom scores are the expected output of the measurement system, not evidence about the women's mental health.

2.8 The “resilient African woman” narrative: a critical appraisal

Into the interpretive space created by low scores steps a familiar and powerful story: that African women possess an extraordinary, almost constitutional resilience that allows them to absorb poverty, exploitation, and oppression without sustaining the psychological injury those conditions would inflict elsewhere. The narrative is attractive. It is affirming, it resonates with genuine strength and survivorship, and it relieves systems of the obligation to look harder. But the empirical and theoretical literature on the closely related “Strong Black Woman” or “Superwoman” schema shows that this image, far from being protective, is itself a documented risk to women's health — and that when low self-report scores are read through it, the narrative actively conceals pathology.

The schema has been characterized across two decades of research as an obligation to project strength, suppress emotion, resist vulnerability, and prioritize others' needs over one's own (Beauboeuf-Lafontant, 2007; Woods-Giscombé, 2010). Its endorsement is consistently associated with adverse outcomes. Emotional suppression and self-silencing — core components of the schema — mediate elevated depression and anxiety (Abrams et al., 2019; Liao et al., 2020), and schema endorsement predicts greater psychological distress and, in some studies, suicidal thoughts and behaviors (Castelin & White, 2022; Donovan & West, 2015). Crucially for measurement, the schema is inversely associated with help-seeking and with the willingness even to acknowledge psychological problems: the stronger the endorsement, the less open women are to recognizing distress in themselves or disclosing it to others (Nelson et al., 2016; Nicolaidis et al., 2010; Watson & Hunter, 2015). The schema also carries physical health costs, including premature deterioration and maladaptive coping (Woods-Giscombé et al., 2019). In short, the cultural ideal of the unbreakably strong woman is a recognized pathway to under-recognition and under-treatment, not a marker of wellness.

Three problems, therefore, undermine any reading of low scores as resilience in this population. First, it confuses measurement failure with the absence of pathology: a number suppressed by stigma, idiom, normalization, gendered silence, and a strength-schema that explicitly discourages disclosure cannot be used as proof of clinical wellness without committing a basic inferential error, in which a limitation of the tool is mistaken for a property of the person. Second, the narrative is contradicted by the convergent risk data. Resilience, properly understood, is positive adaptation in the presence of adversity, not the absence of the markers of injury; in our population, the markers of injury — high adverse childhood exposure, documented sexual abuse and suicide-risk flags, substance-use risk, and item-level self-harm signals — coexist abundantly with the low mood/anxiety totals. A construct of resilience that must ignore suicidality and trauma exposure to sustain itself is not resilience; it is selective attention to the reassuring half of the record.

Third, and most consequentially, the narrative does real and measurable harm. When low scores are read as resilience, the clinical conclusion is that the woman does not need care, and scarce resources are directed elsewhere. The resilience story thus functions, in practice, as a rationalization for non-treatment: it converts a failure of detection into a flattering attribute of the patient and, in doing so, licenses the diagnostic overshadowing and clinical neglect documented in the data in this paper. The narrative's toxicity is not rhetorical. It operates through the same self-silencing and help-seeking suppression that the schema literature has tied to depression, untreated distress, and premature physical decline in the very women it claims to celebrate (Watson & Hunter, 2015; Woods-Giscombé et al., 2019). The romance of resilience and the reality of under-treatment are not in tension; the first underwrites the second.

The reframing we propose. These women are not resilient in the sense of being unharmed. Many are surviving — a different and harder thing — while carrying severe, under-detected psychiatric burden that conventional screening fails to register and that the resilience narrative then explains away. Naming this accurately is a precondition for treating it.

3. Methods

3.1 Rationale and relation to prior research

Discordance between standardized symptom scores and other indicators of clinical need is not a novel observation in global mental health, but the way it has been studied leaves a specific gap that this dataset is unusually well placed to fill. Cross-cultural psychiatry has long documented that brief symptom screeners under-detect distress in settings shaped by stigma, somatic idioms, and self-silencing (Hinton & Lewis-Fernández, 2010; Kohrt et al., 2014; Thornicroft et al., 2017), and validation studies of the PHQ-9 and related tools in Ethiopia have reported imperfect sensitivity against diagnostic interviews, especially where holy-water and traditional-healing populations are included (Gelaye et al., 2013; Hanlon et al., 2015).

What is largely absent from the African literature is a large, full-population, multi-instrument analysis that quantifies the discordance directly — that is, one that measures, within the same clients, how often documented high-risk indicators co-occur with low symptom endorsement, and that examines safety risk at the item level rather than through total scores. Most African mental health studies report single-instrument prevalence (depression or PTSD rates) rather than cross-instrument concordance, and few link standardized screens to safety flags and social-vulnerability records in one analytic frame. To our knowledge, an analysis of this structure — five standardized instruments, item-level safety logic, repeated measures, and social-vulnerability subsets across more than three thousand clients — has not previously been reported in an Ethiopian or comparable sub-Saharan population. This is the methodological contribution that accompanies the substantive one.

The analytic strategy is deliberately multi-instrument because no single scale can adjudicate the question at issue. A depression score alone cannot reveal masking; only the comparison of that score against independent trauma, substance, and safety indicators can. We therefore treat the five instruments not as interchangeable proxies for severity but as distinct measurement channels — mood (PHQ-9), anxiety (GAD-7), trauma exposure (ACE), alcohol risk (CAGE), and substance severity (SDS) — whose disagreement is itself the signal of interest. This framing follows the logic of multimodal and cross-verification approaches in measurement-based care, in which concordance increases confidence and discordance triggers further evaluation rather than averaging (Alemu et al., 2024).

3.2 Setting and population

Data were drawn from a digitally supported rehabilitation and behavioral health screening program serving vulnerable women in Addis Ababa, Ethiopia, delivered in collaboration with a faith-based social service network and a women’s rehabilitation center. Admission to the rehabilitation pathway was based on a profile of severe structural adversity: extreme poverty, a history of street-based sex work, and chronic unemployability. Screening was conducted by trained program staff using tablet-based digital forms, with instruments adapted for Ethiopian use, including the PHQ-9 (Kroenke et al., 2001), validated for Ethiopian use (Gelaye et al., 2013); the GAD-7 (Spitzer et al., 2006); the ACE inventory (Felitti et al., 1998); the CAGE alcohol screen (Ewing, 1984); and the Severity of Dependence Scale (SDS; Gossop et al., 1995), supplemented by admission and integrated psychosocial forms developed by the program for routine clinical documentation (English-language versions are

provided as Additional files 1 and 2). All five core screening instruments are previously published standardized measures; no new questionnaires were developed for this study.

3.3 Analysis sample and screening coverage

The starting client table contained 3,627 records. After removing obvious test or duplicate entries, 3,617 clients were retained for aggregate analysis. The analysis cascade from raw records to the complete battery subset is shown in Table 1. The program serves women exclusively, and all clients in the analysis are female. A biological-sex field was completed for 2,433 records and contained 12 male-coded entries, which we treat as data-entry errors that, given the women-only admission criteria, do not represent male clients; no sex comparison is therefore attempted.

Item	N	Interpretation
Raw client records	3,627	Starting client table before removing obvious test/dummy records.
Analysis clients	3,617	All non-test clients retained for aggregate analysis.
Any core screen	2,669	At least one completed PHQ-9, GAD-7, ACE, CAGE, or SDS.
All five core screens	1,592	Complete baseline battery used for five-scale profile analysis.
Mood + ACE eligible	2,586	Clients with PHQ-9, GAD-7, and ACE baseline scores.
Mood + substance eligible	1,595	Clients with PHQ-9, GAD-7, CAGE, and SDS baseline scores.
Admission form coverage	836	Used for poverty/social-vulnerability subset analysis.
Integrated form coverage	158	Used for richer clinical-context subset analysis only.

Table 1. Analyze the sample cascade from raw records to the complete battery subset.

Coverage differs markedly across instruments. The standardized mood and trauma screens have usable coverage for a full-population analysis, while substance screens and social-vulnerability forms have substantially lower coverage (Table 2; Figure 1). All analyses, therefore, use available-case denominators, reported transparently for each measure, with the complete-five-screen subset reserved for sensitivity analysis and the admission/integrated forms reserved for exploratory subset analysis rather than full-population inference.

Item	Clients	Analysis clients	Coverage
PHQ-9	2,647	3,617	73.2%
GAD-7	2,634	3,617	72.8%
ACE	2,623	3,617	72.5%
CAGE	1,659	3,617	45.9%
SDS	1,642	3,617	45.4%
All five core screens	1,592	3,617	44.0%
Admission social form	836	3,617	23.1%
Integrated psychosocial form	158	3,617	4.4%

Table 2. Coverage of core screening and social-vulnerability forms among 3,617 analysis clients.

Figure 1 presents the same coverage information graphically, making the divide explicit: the standardized mood and trauma screens cluster at high coverage, whereas the substance screens and social-vulnerability forms fall well below them. This divide is not incidental to the analysis. It determines which denominators support full-population inference and which support only the exploratory subset claims reported later, and it is the reason coverage is reported alongside every result rather than assumed.

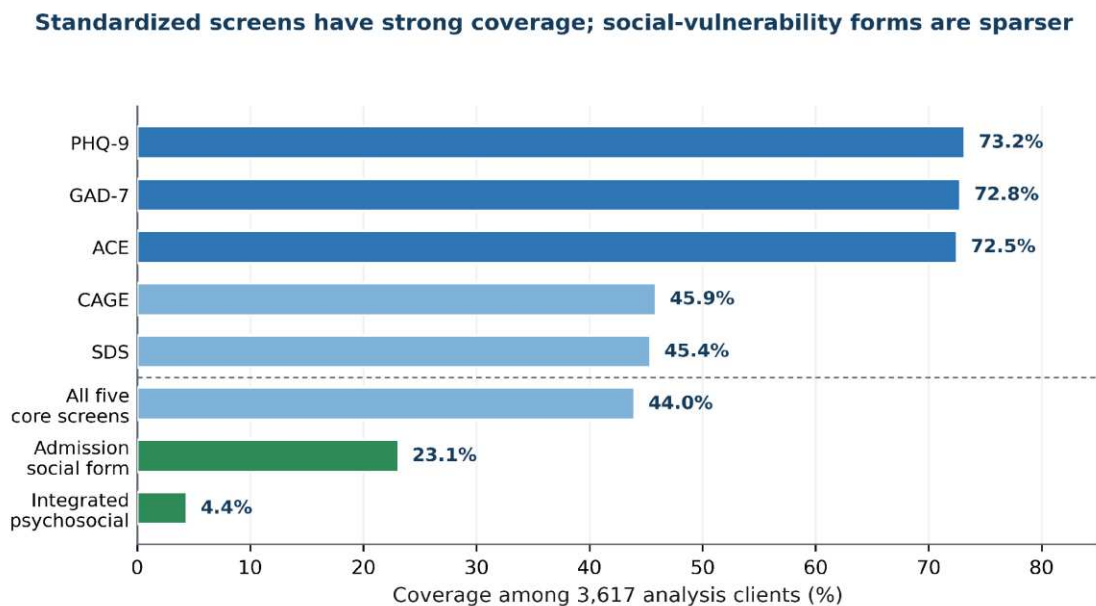


Figure 1. Coverage of core screens and social-vulnerability forms. Standardized mood and trauma screens are well covered; substance and social forms are sparser.

3.4 Measures and operational definitions

- PHQ-9 and GAD-7 were used to index depression and anxiety severity. “Moderate or higher” denotes a total score of ≥ 10 ; “subthreshold” denotes 5–9; “low” denotes < 5 . Unless stated otherwise, masked presentation requires low on both scales.
- ACE (Adverse Childhood Experiences Scale) provided an adverse-childhood/trauma-exposure context; “high ACE” denotes ≥ 4 .
- CAGE screened alcohol-use risk (with a positive screen defined at the standard cut-off); SDS screened substance-use severity/dependence in this dataset and is not interpreted as a functional-disability index.
- Safety indicators comprised the PHQ-9 self-harm item (item 9) and a survey-level suicide-risk flag, analyzed separately from total scores.
- A “high-risk signal” denotes any of the following: high ACE, substance-risk elevation, suicide-risk flag, or sexual-abuse flag.

3.5 Analytic approach

We (a) described baseline severity across all five screens; (b) compared the general severity profile against a masked-presentation profile, including the severity distribution of clients with versus without a high-risk signal; (c) quantified discordant presentation under several operational definitions; (d) compared item-level and survey-level safety signals against total-score thresholds; (e) derived an exploratory five-profile description of cross-scale patterning and examined cross-scale correlations; (f) analyzed first-to-last symptom change among baseline-elevated clients using 7-day and 30-day follow-up windows; and (g) examined poverty and social-vulnerability indicators in the admission and integrated subsets. We follow a strict wording discipline: discordant

scores are treated as evidence of masked or constrained presentation, not as proof of intentional defensiveness, and repeated-measure change is described as change consistent with improvement, rather than causal in the absence of intervention-exposure data.

3.6 Ethics and AI-use statement

The underlying clinical care, case identification, and service-delivery activities were conducted as routine practice under the memorandum of understanding between AYA-Innovation and the Lenegewa Women’s Rehabilitation and Skills Development Center. All records were de-identified before analysis; the dataset contains no client-level identifiers, voice files, or row-level free text, and this manuscript reports only aggregate findings. Generative AI assisted with language editing, structuring of author-provided analyses and tables, and figure preparation; all interpretations, analyses, and conclusions were reviewed and approved by the authors, and no AI tool created, altered, or fabricated data, findings, or references.

The de-identified records were accessed for research purposes on 25 May 2026. During routine service delivery, authorized clinical and program personnel had access to identifiable information as necessary for the provision of care. Before the research dataset was created, all direct identifiers were removed. The authors conducting the retrospective analysis received only the de-identified analytic dataset and did not have access to information that could identify individual participants during the research analysis or afterward.

4. Results

We present the results in a sequence designed to make a single comparison unavoidable: what the population looks like through conventional severity categories, and what it looks like once the full multi-instrument record is brought to bear. The descriptive baseline (Section 4.1) establishes the apparent picture of modest burden. The central comparison (Section 4.2) sets that picture against the masked-presentation view and quantifies the gap between them. The domain-specific (Section 4.3), multiple-definition (Section 4.4), and safety-level (Section 4.5) analyses establish that the gap is robust across operational definitions and most pronounced precisely where it is most clinically significant, namely around suicidality. The profile model and correlations (Section 4.6) show the cross-scale structure underlying the gap, the repeated measures (Section 4.7) demonstrate that the elevated subgroup is treatment-responsive once detected, and the social-vulnerability subsets (Section 4.8) link the pattern to the structural adversity in the population. Throughout, available-case denominators are reported explicitly, and complete-case analyses are used for sensitivity checks.

4.1 The general severity profile makes the population look well

When interpreted through conventional severity categories, this population appears to carry a modest psychiatric burden. Moderate-or-higher depression and anxiety were uncommon, while trauma exposure was the dominant feature of the sample, and a meaningful substance-risk subgroup was present (Table 3). The contrast between low overt mood and anxiety severity and very high trauma exposure — 8.7% meeting PHQ-9 moderate-or-higher criteria compared with 35.3% classified as high-ACE and 85.7% with any ACE exposure — is the first indication that the conventional severity lens is missing clinically relevant risk captured elsewhere in the dataset.

Metric	Clients	Denominator	Percent
PHQ-9 moderate or higher	229	2,647	8.7%
GAD-7 moderate or higher	168	2,634	6.4%
Either PHQ-9 or GAD-7 moderate or higher	296	2,618	11.3%

Metric	Clients	Denominator	Percent
PHQ-9 and GAD-7 both low	1,552	2,618	59.3%
PHQ-9/GAD-7 subthreshold	770	2,618	29.4%
ACE high risk (≥ 4)	931	2,623	35.5%
Any ACE exposure	2,247	2,623	85.7%
CAGE high probability	209	1,659	12.6%
SDS moderate/dependency	78	1,642	4.8%
Any substance elevation	354	1,613	21.9%
Any high-risk signal (with PHQ/GAD data)	1,250	2,618	47.7%

Table 3. Baseline severity profile across screening instruments. Mood and anxiety severity appear low, while trauma exposure and high-risk signals are common.

4.2 Severity profile versus masked-presentation profile: the central comparison

The decisive analysis compares the general severity view with the masked-presentation view within the same population. The general view, using moderate-or-higher PHQ-9/GAD-7, identifies 11.3% of the baseline population as meeting moderate-or-higher severity criteria. The masked-presentation view asks a different question: among clients carrying a high-risk signal, how many remain below the moderate threshold? Among the 1,250 such clients, 78.6% scored in the low or subthreshold range, and only 21.4% were captured by conventional severity (Figure 2). The masked group is therefore 3.7 times larger than the group the conventional lens identifies. For roughly four of five high-risk clients, a clinician relying on PHQ-9/GAD-7 severity categories would observe non-elevated scores despite the presence of independently documented risk.

Conventional symptom thresholds capture only one in five high-risk clients

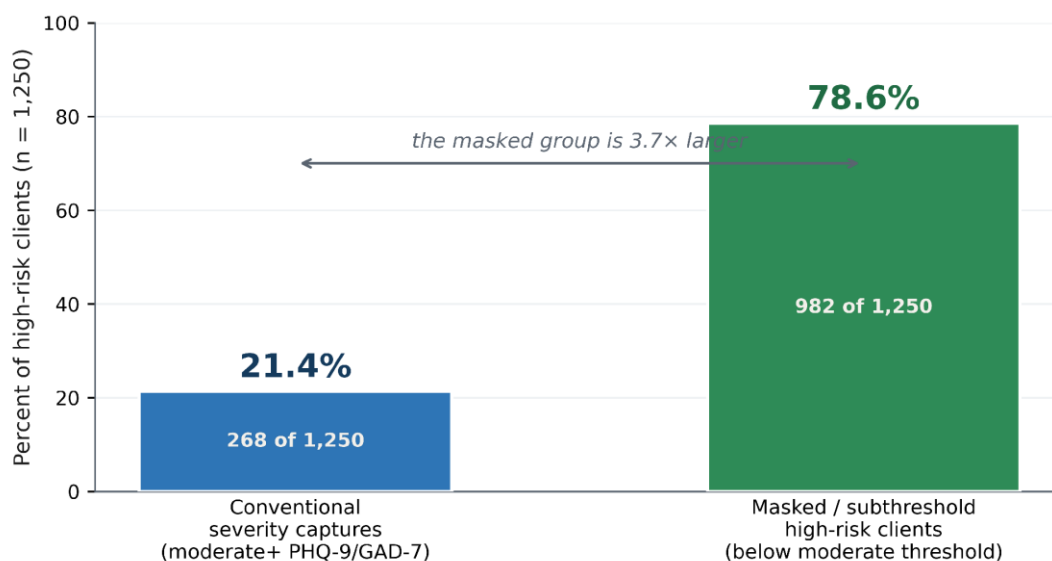


Figure 2. Conventional severity thresholds identify 21.4% of high-risk clients; 78.6% fall below the moderate threshold despite carrying a high-risk signal.

Table 4 makes the comparison explicit by stratifying the full PHQ/GAD baseline population into clients with and without an observed high-risk signal. The two strata diverge sharply (Figure 3). Among clients with no observed

high-risk signal, 82.1% scored low, and only 2.0% reached the moderate threshold — a distribution consistent with that of a lower-risk reference group. Among clients with a high-risk signal, the low fraction falls to 34.3%, and the subthreshold fraction rises to 44.2%, yet the moderate-or-higher fraction reaches only 21.4%. In other words, carrying a documented high-risk signal shifts clients out of the low band and into the subthreshold band to a much greater extent than it shifts them into the moderate-or-higher category, which is the threshold on which the conventional severity framework depends.

Group	Mood/anxiety status	Clients	% within group
Any high-risk signal	Low	429	34.3%
Any high-risk signal	Subthreshold	553	44.2%
Any high-risk signal	Moderate/higher	268	21.4%
No observed high-risk signal	Low	1,123	82.1%
No observed high-risk signal	Subthreshold	217	15.9%
No observed high-risk signal	Moderate/higher	28	2.0%
All PHQ/GAD baseline clients	Low	1,552	59.3%
All PHQ/GAD baseline clients	Subthreshold	770	29.4%
All PHQ/GAD baseline clients	Moderate/higher	296	11.3%

Table 4. Mood/anxiety severity categories among clients with and without an observed high-risk signal, and in the full baseline population.

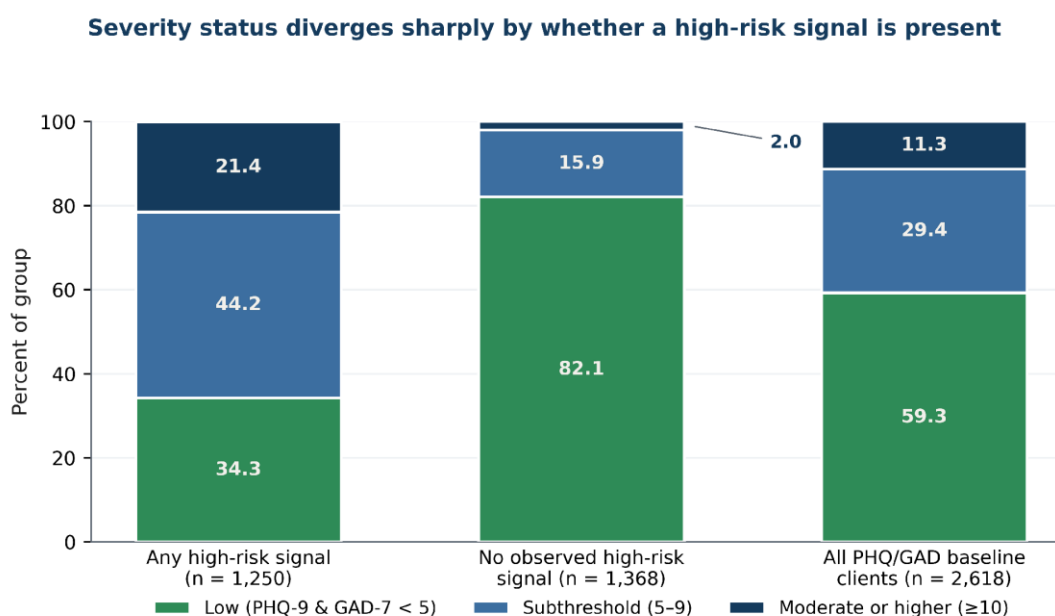


Figure 3. Severity distribution by the presence or absence of a high-risk signal. The high-risk group is disproportionately concentrated in the subthreshold band rather than crossing into the moderate-or-higher category.

4.3 Masked presentation across high-risk domains

The masking pattern is not confined to one risk domain. It holds across high ACE, substance risk, suicide-risk flags, and sexual-abuse flags alike: in every domain, the majority of clients fall below the moderate mood/anxiety cut-off (Figure 4; Table 5). The sexual-abuse domain is especially informative — 72.9% of clients

with a documented sexual-abuse flag fall within the low or subthreshold range — because it places a documented trauma history directly alongside a reassuring symptom score in the same clients.

Across every high-risk domain, most clients fall below the moderate cut-off

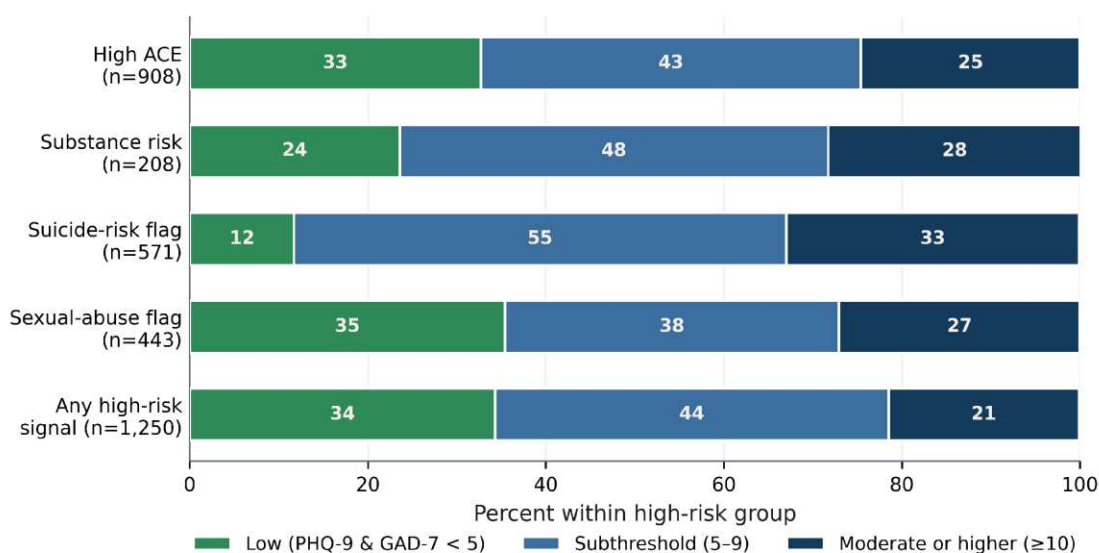


Figure 4. Mood/anxiety severity distribution within each high-risk domain. The combined low and subthreshold categories predominate across all domains. Low and subthreshold bands together dominate every domain.

Risk domain	N	Low	subthreshold	Low + sub	Mod+
High ACE	908	297 (32.7%)	388 (42.7%)	685 (75.4%)	24.6%
High substance risk	208	49 (23.6%)	100 (48.1%)	149 (71.6%)	28.4%
Suicide-risk flag	571	67 (11.7%)	316 (55.3%)	383 (67.1%)	32.9%
Sexual-abuse flag	443	157 (35.4%)	166 (37.5%)	323 (72.9%)	27.1%
Any high-risk signal	1,250	429 (34.3%)	553 (44.2%)	982 (78.6%)	21.4%

Table 5. Mood/anxiety severity distribution within high-risk domains. Low and subthreshold categories indicate potential masked or constrained presentation and should not be interpreted as evidence of intentional minimization.

4.4 Discordant-presentation patterns under multiple operational definitions

Because the masked-presentation construct can be operationalized in several ways, we report its prevalence under multiple definitions (Table 6). The pattern is robust to definition: whether anchored on trauma exposure, substance risk, combined trauma-plus-substance, suicide flags, or sexual-abuse flags, a substantial share of the at-risk group sits below the conventional mood/anxiety threshold. The suicide-flag comparison is the most striking —70.7% of clients with a survey-level suicide flag had a PHQ-9 total score <10 — underscoring that safety-relevant risk is the category most likely to be masked when relying on total-score thresholds.

Pattern	Eligible N	Condition N	Pattern N	%
High ACE with low PHQ-9/GAD-7	2,586	908	297	32.7%
High substance risk with low PHQ-9/GAD-7	1,595	208	49	23.6%
High ACE + substance with low PHQ-9/GAD-7	1,592	234	51	21.8%

Pattern	Eligible N	Condition N	Pattern N	%
Baseline survey suicide flag with PHQ-9 total <10	2,602	564	399	70.7%
Sexual-abuse flag with low PHQ-9/GAD-7	2,586	443	157	35.4%
All-five low endorsement	1,592	1,592	203	12.8%
Multidomain high acuity	1,592	1,592	52	3.3%

Table 6. Discordant screening patterns under multiple operational definitions. Low mood/anxiety is defined as PHQ-9 < 5 and GAD-7 < 5; the suicide-flag comparison uses PHQ-9 <10 to reflect conventional moderate-threshold screening.

4.5 Total-score thresholds conceal safety risk

The clearest demonstration that conventional severity misses clinically urgent information comes from the safety analysis. When safety is evaluated at the item and flag levels rather than by total scores, a large share of at-risk clients falls below the thresholds that would trigger concern (Table 7). Among clients endorsing the PHQ-9 self-harm item, 68.8% had a total PHQ-9 score below 10; among those with a survey-level suicide flag, 70.7% had a PHQ-9 score below 10. A program relying on PHQ-9 scores of moderate or higher as its primary safety screen would therefore fail to identify approximately 7 in 10 clients with active safety indicators.

Safety check	Condition N	Pattern N	% of condition
PHQ-9 item 9 positive with PHQ-9 total <5	590	101	17.1%
PHQ-9 item 9 positive with PHQ-9 total <10	590	406	68.8%
Survey suicide flag with PHQ-9 total <10	564	399	70.7%

Table 7. Safety signals relative to total-score thresholds. Total PHQ-9 cutoffs fail to capture the majority of item-and survey-level safety risks.

Clinical implication. A total-score threshold is not an adequate safety screen in this population. Item-level self-harm endorsement and survey-level suicide flags must be reviewed independently of the PHQ-9 total, or genuine risk will be routinely passed over. This is consistent with evidence that suicide risk is not fully captured by overall depression severity and may be present even at lower PHQ-9 total scores, supporting the need for item-level and dedicated suicide risk assessment (Simon et al., 2013; Nock et al., 2008).

4.6 Exploratory five-profile model and cross-scale correlations

An exploratory five-profile description of clients with complete screens separates a large low-endorsement group from trauma-, substance-, and multidomain-burdened groups (Table 8). Nearly half the complete-screen subset (46.4%) falls into a low-endorsement profile — the subgroup most likely to be misclassified as “low risk” under conventional severity frameworks — while ACE-dominant and subthreshold-plus-ACE profiles together account for a further 37.8% in whom trauma exposure coexists with modest mood/anxiety scores. Only 7.2% fall into the overt multidomain-burden profile that conventional severity would readily detect.

Profile	Clients	%	PHQ-9	GAD-7	ACE/substance
Low endorsement	738	46.4%	1.0	1.1	ACE 1.1; CAGE 0.1; SDS 0.1
ACE-dominant/low mood-anxiety	303	19.0%	2.6	2.3	ACE 4.6; CAGE 0.1; SDS 0.2
Subthreshold mood-anxiety + ACE	300	18.8%	8.2	7.8	ACE 4.1; CAGE 0.1; SDS 0.2

Profile	Clients	%	PHQ-9	GAD-7	ACE/substance
Trauma + alcohol risk/low mood-anxiety	136	8.5%	4.5	3.7	ACE 4.2; CAGE 3.2; SDS 3.0
Multidomain trauma + substance	115	7.2%	9.7	8.0	ACE 6.0; CAGE 3.1; SDS 8.1

Table 8. Exploratory five-profile model (complete-screen subset, n = 1,592). Values represent mean instrument scores; profiles are descriptive, do not constitute a final latent-class solution.

Cross-scale correlations among first-completed baseline scores support a multidomain, rather than a single-distress, interpretation. Depression and anxiety move together (PHQ-9 and GAD-7 are strongly correlated), and the two substance measures move together (CAGE and SDS), but trauma and substance constructs show only moderate association with mood and anxiety. This leaves sufficient discordance across domains to justify a pattern-based, rather than a single-severity-score, account of clinical need. The Spearman correlation matrix is retained in the analysis output as a supporting figure.

4.7 Early symptom change among baseline-elevated clients

Among the smaller group elevated at baseline and re-measured after an adequate interval, repeated screens showed large improvement. Median first-to-last change was –11.5 points on the PHQ-9 and –10.0 on the GAD-7, with most baseline-elevated clients moving below threshold (Table 9). The finding held across seven-day and 30-day windows. We deliberately refrain from causal language: without intervention timing or contact intensity data, these changes are consistent with improvement but not attributable to a specific intervention, and the effect in small, elevated subgroups should be interpreted with potential floor effects in mind. ACE change is not treated as an outcome because ACE measures historical exposure, and any observed variation likely reflects disclosure dynamics or data quality rather than true clinical change.

Scale	Repeat N	Elevated N	Median change	95% interval	Remitted
PHQ-9	902	74	–11.5	–12.0 to –10.0	98.6%
GAD-7	894	53	–10.0	–11.0 to –8.0	100%
CAGE	489	63	–2.0	–2.0 to –1.0	52.4%
SDS	482	43	–5.0	–7.0 to –4.0	93.0%

Table 9. First-to-last change among baseline-elevated clients with ≥7 days between measures (median ~111 days for mood; ~98 for substance). Intervals represent bootstrap 95% intervals.

4.8 Poverty and social vulnerability in the admission and integrated subsets

Although coverage is lower for the admission and integrated psychosocial forms, these subsets provide convergent evidence that the masked-presentation pattern co-occurs with the structural adversity that defines program eligibility (Table 10). Within these subsets, low-income family background, street circumstances, sex work, family non-support, loneliness, self-harm thoughts, and abuse/exploitation each appear alongside substantial proportions of low mood and high ACE. These findings are interpreted as exploratory and context-informing rather than confirmatory, owing to limited coverage of the underlying data.

Source	Indicator	N	Low mood	High ACE	Both
Admission	Low-income family	158	129 (81.6%)	32 (20.3%)	17
Admission	Street	27	18 (66.7%)	10 (37.0%)	5
Admission	Sex work	16	10 (62.5%)	4 (25.0%)	0

Source	Indicator	N	Low mood	High ACE	Both
Admission	Returnee/migration	6	5 (83.3%)	0 (0.0%)	0
Integrated	Family no support	64	38 (59.4%)	24 (37.5%)	9
Integrated	Feels lonely	61	34 (55.7%)	23 (37.7%)	7
Integrated	Self-harm thoughts	40	20 (50.0%)	23 (57.5%)	7
Integrated	Abuse/exploitation	24	15 (62.5%)	13 (54.2%)	6
Integrated	Substance/alcohol use	17	8 (47.1%)	11 (64.7%)	4

Table 10. Poverty and social-vulnerability indicators in the admission and integrated subsets, with co-occurring low mood and high ACE. Exploratory subset evidence only.

5. Discussion

This analysis compares a general severity profile with a masked-presentation profile within the same population and shows that they yield fundamentally different interpretations. The general profile, based on PHQ-9 and GAD-7 thresholds, suggests modest psychiatric burden: only one in ten clients meets a moderate-or-higher mood/anxiety threshold. The masked-presentation profile incorporating trauma exposure, substance risk, and safety indicators shows that the majority of high-risk clients fall below these thresholds and that conventional severity captures only a minority of those with clinically relevant risk. The two profiles are not contradictory because one is incorrect; rather, they address different constructs. The severity profile measures what the instrument can see, whereas the masked-presentation profile reflects the discordance between documented risk and observed symptom severity. In this population, that discordance appears to be a defining clinical feature.

Table 11 formalizes this distinction across the two analytic lenses and names what each lens answers.

Lens	What it answers	Main finding
General severity analysis	Overt baseline symptom burden using standard scale categories.	Mood/anxiety severity appears low: only 11.3% meet moderate-or-higher PHQ-9/GAD-7 criteria.
Masked/subthreshold analysis	Whether high-risk indicators remain present when mood/anxiety endorsement is low or below moderate.	Among 1,250 high-risk clients, 78.6% fall below the moderate PHQ-9/GAD-7 threshold.
The difference	The two analyses answer different questions, not competing ones.	Severity measures visible symptoms; masked-risk analysis measures discordance between risk/safety indicators and symptom endorsement.

Table 11. Two analytic lenses on the same population. General severity is the descriptive baseline; masked/subthreshold presentation is the analytic contribution.

5.1 What conventional severity misses, and why it matters

The cost of relying on conventional severity here is not abstract. It is the 982 high-risk clients who would be read as below threshold; the roughly seven in ten clients with a positive self-harm item or suicide flag whose total scores would not have triggered concern; and the women with documented sexual-abuse histories whose low anxiety scores would, on a naive reading, suggest they were unaffected. In each case, the instrument returns a

number that is reassuring and wrong, and the error runs in the dangerous direction — toward under-detection of risk in the people least able to advocate for re-evaluation.

5.2 Defensive presentation as the better explanation

We have argued that the most defensible interpretation of the dominant low scores, in this specific population, is defensive or constrained presentation rather than a straightforward absence of psychological distress. The argument is not that a low PHQ-9 can never mean wellness; in a general population sample, it often does. The argument is conditional on the population. When a sample is selected for extreme poverty, sexual exploitation, and chronic unemployability, conditions associated with high levels of psychosocial adversity and elevated risk for mental health difficulties, a predominantly low symptom profile is more plausibly read as the joint product of stigma, cultural idiom, normalization, and gendered silence than as genuine psychological health. The prior probability of disorder is high; the measurement is known to be biased downward by identifiable forces; the convergent risk data are abundant. Under these conditions, a masked or constrained presentation interpretation may offer a more parsimonious explanation of the findings than an interpretation based solely on the apparent absence of symptoms.

We are careful not to overclaim. The data establish discordance, not the internal mechanism producing it; they cannot, on their own, distinguish conscious minimization from culturally shaped expression from genuine sub-syndromal states. That distinction requires qualitative work, validity indices, or multimodal corroboration. What the data do establish is that low symptom scores in this population cannot be taken at face value and should be interpreted with caution and in the context of broader clinical and social evidence, and that any clinical or policy conclusion that treats them as proof of low need is unsupported.

5.3 Clinical defensiveness as a system property: bias and diagnostic overshadowing

The second sense of defensiveness lies not in the patient but in the clinical system. A reassuring number is easy to accept, particularly when doing so conserves scarce time and resources and aligns with a prevailing narrative of resilience. There is a risk of diagnostic overshadowing, in which questionnaire findings receive disproportionate weight relative to other clinical and contextual information: the low score becomes the headline finding, the rest of the record recedes, and the clinical question is treated as settled. In a high-volume, under-resourced setting, this is not the failing of any individual clinician but a predictable consequence of relying on a single, imperfect instrument as the arbiter of need. When a measure systematically underestimates distress, overreliance on its results may contribute to under-recognition and under-treatment of clinically significant difficulties. We characterize this as a form of structural bias and potential system-level neglect, and we locate it in the system rather than in the clinician's intent. The negligence is structural: a screening architecture that converts a known measurement limitation into a treatment decision, and a narrative environment that supplies a flattering rationalization — resilience — for the resulting neglect. The remedy is not to blame the frontline staff who are doing demanding work under substantial resource constraints. Rather, it is to strengthen the tools and workflow so that documented social and clinical risk factors are considered alongside symptom scores and cannot be inadvertently overshadowed by a single reassuring total.

6. Trauma Has Many Faces: AI Infrastructure for Detection in African Mental Health

The standard argument for artificial intelligence in African mental health is the workforce gap: with roughly 1.4 mental-health professionals per 100,000 population across the continent — far below the WHO benchmark — and approximately 0.11 psychiatrists per 100,000 in Ethiopia, specialist-dependent care cannot reach the people who need it, and AI can extend the reach of every trained provider (Federal Ministry of Health, 2020; Wondimagegn et al., 2023; World Health Organization, 2017). That argument is correct, but the findings in this paper show it to be incomplete. The masked-presentation result identifies a second failure that sits upstream

of the workforce shortage: even where a provider and a validated instrument are present, substantial discordance may exist between symptom scores and broader indicators of vulnerability and risk. Simply increasing the number of providers who rely on the same self-report channel may not, by itself, identify all individuals with significant unmet needs. In our sample, 982 clients with substantial social adversity and multiple documented risk factors nevertheless scored below conventional screening thresholds. Trauma has many faces, and a single self-report total can see only one of them.

6.1 The empirical and research basis for voice-based detection

The proposition that distress can be detected from speech is increasingly supported by empirical evidence. A growing body of research shows that acoustic and prosodic features of voice — pitch, energy, timing, spectral characteristics, and affective coloring, carry a statistically reliable signal about depression, anxiety, and suicide risk, and that machine-learning models can extract that signal at clinically useful accuracy, including for objective quantification of psychiatric severity (Almaghrabi et al., 2023; Caulley et al., 2023; Low et al., 2020; Zhang et al., 2020). A systematic review and meta-analysis of deep-learning speech models for depression reported pooled sensitivity near 0.82 and specificity near 0.85 across studies (Liu et al., 2024), and broader reviews converge on the conclusion that speech is a cost-effective, non-invasive, and remotely collectible biomarker for mood and distress states (Riad et al., 2024; Shin & Bae, 2024). Vocal markers have likewise shown promise for suicide-risk detection specifically (Iyer & Meyer, 2022), the domain where our safety analysis shows conventional totals to be weakest.

Two properties of the voice channel make it well matched to the failure mode documented here. First, because speech can be compared across languages through shared aspects of vocal anatomy and emotion expression, voice-based approaches are extensible to under-resourced languages rather than confined to English (Liu et al., 2024). Second, and more fundamentally, the voice channel does not require the speaker to read an item, name a symptom, or consent to disclose it in words; it registers distress carried in how something is said rather than in whether the speaker is willing to endorse a checklist (Alemu et al., 2023, 2024). This is precisely the property that matters in a population whose low scores are influenced by stigma, gendered shame, idiom, and a strength-schema that discourages verbal disclosure. We note, in the interest of balance, that current voice-only models do not yet outperform well-administered self-report in general populations and are not a standalone diagnostic (Donaghy et al., 2024); their value here lies in serving as an independent, harder-to-suppress channel used alongside self-report, not as a replacement for it.

6.2 From a severity signal to agentic, multilingual care: why detection alone is not enough

Detecting masked severity is necessary but not sufficient to close the workforce gap. A severity indicator that flags a client as high-risk, without describing what is wrong, supporting clinical interpretation, recommending possible next steps, triaging urgency, and tracking response over time, risks relocating rather than resolving the bottleneck: it still requires a scarce trained clinician to interpret the flag, formulate the problem, and plan care. For AI infrastructure to address the workforce gap rather than merely document it, it will likely need to be agentic, able to carry a case from detection through formulation, recommendation, triage, and longitudinal monitoring under human supervision, and do so multilingually, in the languages clients actually speak (World Health Organization, 2019, 2021).

Concretely, this means a system that does more than emit a number. Building on a voice-derived distress estimate, an agentic, multilingual system can perform several connective tasks: it can characterize the presentation, indicating which symptom domains are elevated, whether item-level safety signals are positive, and whether the self-report and voice channels are discordant; it can generate a structured, guideline-aligned set of next steps and intervention options appropriate to the setting; it can triage urgency, escalating discordant or safety-positive cases for human review while routing lower-acuity cases to supported self-help or lay-

delivered intervention; and it can track outcomes across repeated contacts, surfacing non-response so that care can be stepped up. Each of these corresponds to a task that, in a specialist-rich system, a clinician performs, and that, in a specialist-poor system, may not occur consistently.

The components required are becoming increasingly available and capable. Multilingual language models can operate across Amharic, Afan Oromo, and Tigrinya; voice models can extract distress signals without literate disclosure (Liu et al., 2024; Riad et al., 2024); and agentic architectures can chain assessment, recommendation, and follow-up into a supervised workflow (Ouanes & Farhah, 2024; Topol, 2019). We are explicit about the guardrails. Such a system is decision-support, not an autonomous clinician; every safety-relevant action is human-reviewed; and the evidence base for fully agentic mental health care is still emerging, so deployment must be staged, evaluated, and supervised. The claim is not that AI replaces clinical judgment, but that an agentic, multilingual infrastructure can perform the connective work — formulation, recommendation, triage, and monitoring — that turns a detection signal into care, which a bare severity score cannot.

6.3 Discordance as actionable signal, not noise

The architecture our group has developed treats the very discordance documented in this paper as the trigger for action. When an objective voice-derived distress estimate is high, but PHQ-9/GAD-7 endorsement is low, the system raises a structured inconsistency flag rather than averaging the two into a falsely reassuring middle (Alemu et al., 2024). That flag guards against the two errors this analysis exposes: false negatives, where distress is present but unspoken, and false reassurance, where a normal-looking self-report is accepted while underlying severity is missed. In a population where 78.6% of high-risk clients fall below the conventional threshold, a system that escalates on discordance rather than ignoring it is, by construction, designed to identify individuals who may otherwise be missed by questionnaire-based screening alone. The conceptual relationship among defensive-presentation drivers, the suppressed self-report signal, clinician bias, and AI-supported recovery of severity is summarized in Figure 5.

How defensive presentation, low scores, and clinician bias hide true severity, and where AI closes the gap

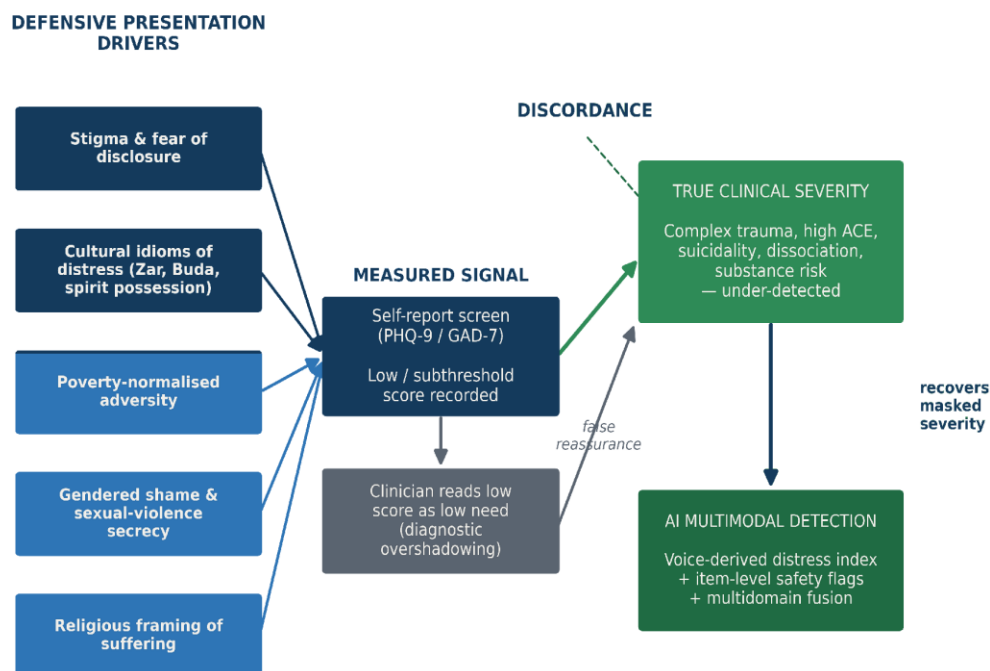


Figure 5. Defensive presentation drivers suppress the self-report signal; clinician bias accepts the low score; true severity remains under-detected. Multimodal AI detection recovers the masked severity.

6.4 Safety detection enforced at the signal level

The safety results in Section 4 show that total-score thresholds miss roughly seven in ten clients with a positive self-harm item or suicide flag. An AI infrastructure can automatically and at scale enforce the item-level safety logic that an overstretched program will not apply by hand to every record: surfacing a positive self-harm item or suicide flag regardless of the total and routing it for human review. This is a low-cost, high-yield function that does not require a specialist at the point of contact, only a system that refuses to let a reassuring total bury an urgent item. Given evidence that vocal markers carry an independent suicide-risk signal (Iyer & Meyer, 2022), a multimodal system can layer a second safety channel on top of item-level logic, potentially reducing the chance that a defended presentation conceals acute risk.

6.5 Fit with African delivery realities

These capabilities are not generic claims about AI; they are specific to the constraints of African frontline care. A voice-based approach is multilingual by design and can be developed across Amharic, Afan Oromo, and Tigrinya, reaching populations that literacy-dependent and English-centric tools exclude (Alemu et al., 2026). It is literacy- and stigma-bypassing: a brief spoken sample removes the requirement to read, to write, or to first name one's symptoms, the barriers most responsible for the under-detection documented here. It is deployable where specialists are absent, with on-device inference on inexpensive tablets functioning in low-connectivity settings. And it is measurement-based and longitudinal: the same infrastructure that detects masked severity also supports repeated measurement and outcome monitoring, as shown in Section 4, closing the loop from detection to follow-up (Alemu et al., 2024). Existing alternatives do not meet these constraints together, task-shifting models still depend on a trained human intermediary, and digital self-report tools still depend on a literate, willing respondent, reproducing the very bottleneck this population defeats.

7. AI as a Response to Two Gaps: Workforce Scarcity and Clinical Defensiveness

The contribution of this analysis to the case for AI is that it reframes what AI is for. The continent's mental health crisis is usually described as a single gap — insufficient numbers of providers—, and AI is usually justified as a way to stretch a scarce workforce further. This paper shows that there are, in fact, two analytically distinct gaps, that they are independent, and that a multimodal voice-AI infrastructure is potentially among the few interventions capable of addressing both simultaneously. Naming the second gap and showing it with data are the key contributions of this dataset that matter beyond its immediate setting.

7.1 The workforce gap

The first gap is quantitative and familiar. Detection mental disorders at first contact in African primary care is extremely low under routine conditions, particularly in the absence of structured screening approaches (Fekadu et al., 2025; Hanlon et al., 2015); fewer than one in ten people with mental illness in sub-Saharan Africa reach professional services (Komu et al., 2025; Sankoh et al., 2018); and the specialist workforce required to change this does not currently exist at scale and cannot be trained at the necessary speed in the short term (Wondimagegn et al., 2023; World Health Organization, 2017). For this gap, AI functions as a force multiplier: it has the potential to place a clinically meaningful assessment capability in the hands of a community health worker, a rehabilitation nurse, or a maternal health provider who has never had access to a psychiatrist, and it does so offline, in local languages, at a marginal cost per assessment that can approach zero as systems scale (Alemu et al., 2023, 2024, 2026), consistent with broader economic evidence that scaling up depression and anxiety treatment yields a strongly positive return on investment (Chisholm et al., 2016). The standard non-AI

response to this gap is task-shifting to non-specialists and peers; we examine in Section 7.3 why that response, though valuable, cannot, on its own, close the second gap.

7.2 The clinical defensive gap

The second gap is qualitative, and it is the one this analysis brings into focus. Even when a provider and a validated instrument are present, severity is systematically under-detected in this population in part because the measurement channel itself is likely suppressed by stigma, gendered shame, cultural and religious idiom, poverty-driven normalization, and a strength-schema that discourages disclosure (Section 2), and then may be accepted uncritically by a clinical system that may be primed to read a reassuring total as low need (Section 5). The result is the masked-presentation pattern documented across Tables 3–10 and visualized as a cascade in Figure 6: 78.6% of high-risk clients below threshold, roughly seven in ten suicide-flagged clients below the moderate cut-off, and three-quarters of clients with documented sexual-abuse histories scoring as if unaffected. This is a defensive gap, and crucially, it is not closed by adding providers, because additional providers using the same suppressed self-report channel would likely inherit the same blind spot.

The defensiveness gap: near-universal trauma collapses to a small “severe” fraction

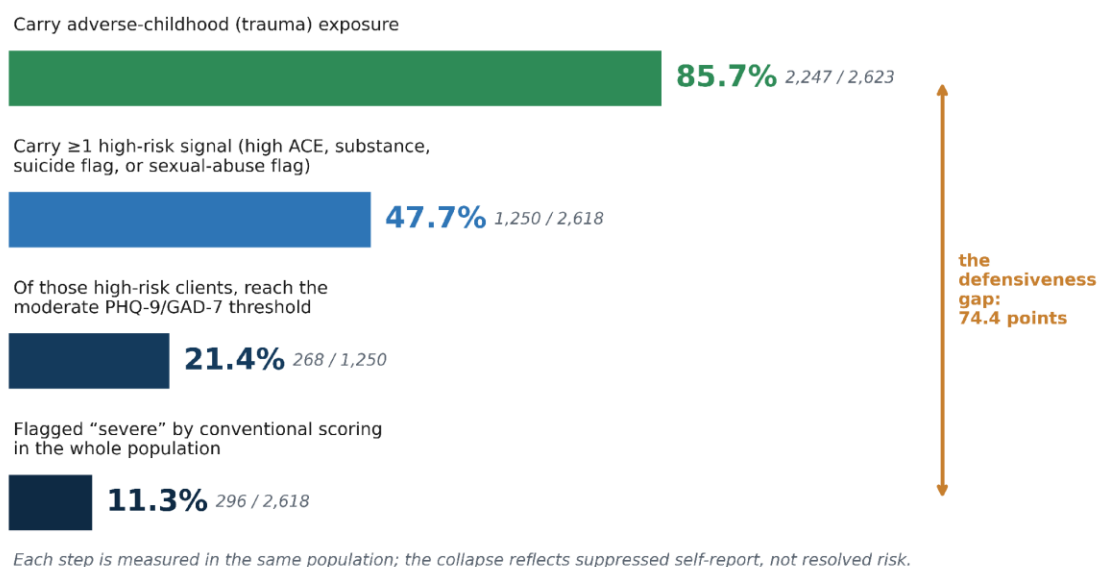


Figure 6. The defensiveness gap is visualized as a cascade. In the same population, near-universal trauma exposure (85.7%) collapses, step by step, to the small fraction of conventional scoring flags as severe (11.3%). The collapse is driven by suppressed self-reporting, not by resolved risk; the 74.4-point gap between true exposure and detected severity is the defensive gap this paper quantifies.

7.3 The paraprofessional paradox: why task-shifting alone cannot close the defensive gap

The dominant global response to the workforce gap is task-shifting: training non-specialists — primary-care workers, lay counselors, and peers with lived experience — to deliver structured, low-intensity psychological interventions, anchored by the World Health Organization’s Mental Health Gap Action Program (mhGAP) and its Intervention Guide (World Health Organization, 2010; Keynejad et al., 2021). The evidence that this approach can work is real and important: lay-delivered interventions for common mental disorders have produced clinically meaningful benefit in low- and middle-income settings (Mutamba et al., 2013; Patel et al., 2010; Singla et al., 2017), ecological, multi-layered frameworks for psychosocial support have been articulated for war-affected populations (Betancourt et al., 2013), and task-shifting is now WHO-endorsed and deployed across more than ninety countries (Keynejad et al., 2021). We do not dispute this evidence; we qualify its scope.

The findings of this paper expose a precondition that task-shifting models frequently assume rather than establish: that severity has been accurately detected before an intervention is matched to it. Effective stepped care is predicated on understanding the problem — its severity, its safety profile, its trauma history — and then selecting and titrating an intervention accordingly. But in this population, the very step that task-shifting takes for granted is the step that may fail or be inadequately performed. A brief-trained paraprofessional administering a self-report screen inherits the same suppressed channel that hides severity from everyone else, and the masked-presentation and safety findings (Sections 4.2–4.5), summarized in Figure 7, show how often that channel is associated with “low need” classifications in the presence of genuine, including suicidal, risk. Lived experience and a short training course do not, by themselves, confer the capacity to detect defended trauma that a validated instrument misses; however, the extent to which they may or may not improve detection is not resolved by the present data. The result can be the appearance of progress — coverage statistics, clients seen, interventions delivered — without the understanding of severity on which genuine benefit depends.

Roughly seven in ten clients flagged for suicide risk fall below the conventional cut-off

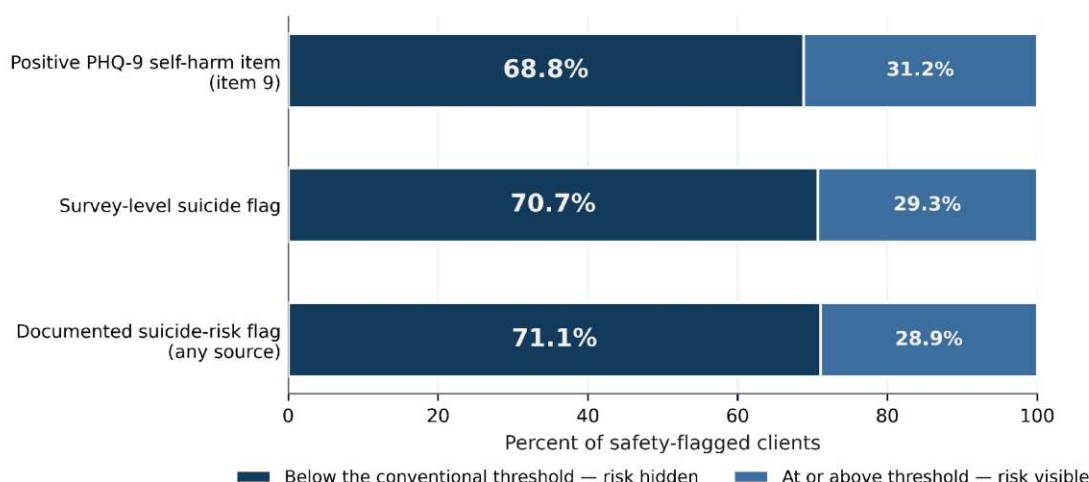


Figure 7. Safety risk hidden below the conventional threshold. Across three ways of identifying suicide-related risk, roughly seven in ten flagged clients score below the total-score cut-off that a screening program would act on, so the risk is present in the record but invisible to the total meant to detect it. This is the safety expression of the defensive gap and the precise failure that brief-trained, self-report-dependent task-shifting cannot remedy.

This is not a dismissal of non-specialists; it is a claim about sequence and skill. Building trust and connecting with a traumatized client are not generic acts that any sympathetic listener performs equally well. They depend on understanding the client’s predicament well enough to respond to it, and that understanding is a clinical skill — structured by training and supervision, and measurable (Kohrt et al., 2015). Talking without understanding, engagement that is not anchored in an accurate formulation of severity and risk, has at best a limited and context-dependent benefit and at worst may fail to identify or may obscure deterioration behind activity. The distinction that matters is between complementary or adjunct interventions and primary ones: peer support and lay-delivered low-intensity programs are valuable adjuncts to a system that can detect and formulate severity, but they cannot substitute for that detective and diagnostic function. Treating them as if they could risk confusing the delivery of an intervention with the clinical understanding that should precede it.

In a setting with severe shortages of trained professionals, where most people with mental health problems present at primary and frontline care rather than to specialists (Fekadu et al., 2025; Saraceno et al., 2007), this paradox is not marginal but central: the layer of the system that sees the most patients is the layer least

equipped to detect defended severity, and the instruments it relies on are precisely those that this analysis identifies as failing to adequately capture severity in this context. The problem is therefore beyond the instruments themselves, and beyond the good intentions of frontline workers; it is structural. This is exactly why an agentic, multilingual AI infrastructure is not just another task-shifting tool. By potentially adding an independent, stigma-resistant detection channel and supporting the formulation, triage, and monitoring functions that brief-trained workers cannot reliably perform, it could help supply the missing precondition, an accurate understanding of severity on which the entire task-shifting edifice silently depends.

7.4 Why one infrastructure can close both

The argument that AI addresses both gaps is not an aspiration appended to the analysis; it follows from the data's structure. The masked-presentation finding establishes that the binding problem lies in the measurement channel, not merely in the number of people reading the output. A multimodal voice-AI system addresses both gaps through a single architectural approach. Against the workforce gap, it reduces dependence on a specialist at the point of contact, extending detection to wherever a frontline worker and an inexpensive device can reach, and because it is agentic and multilingual (Section 6.2), it can extend beyond detection into formulation, recommendation, triage, and outcome tracking, performing the connective work a scarce clinician would otherwise have to do.

Against the defensive gap, it introduces an independent detection channel, voice-derived distress estimation that does not depend on literate, willing, construct-aligned disclosure, and it converts discordance between that channel and self-report into an escalation signal rather than a reason to disregard one of the two (Alemu et al., 2024; Liu et al., 2024; Topol, 2019). The same item-level safety logic that the system enforces at scale can address the most dangerous expression of the defensive gap, suicidality masked by low totals. This implies that, the very dataset that exposes the second gap also points toward a potential solution to it: the same infrastructure already justified by the first gap.

The integrating claim. Adding providers addresses scarcity but not masking; better paper screens address neither. A multimodal voice-AI infrastructure is, on the evidence of this analysis, potentially uniquely positioned to address both the severe workforce gap and the clinical defensive gap — detecting more people, and detecting the severity that conventional instruments and clinical bias together may conceal.

8. Limitations

Several limitations bound the claims of this analysis. The most important concern is the distinction between discordance and mechanism. Our data establish, robustly and across multiple operational definitions, that high-risk indicators frequently co-occur with low symptom endorsement; they do not, on their own, prove that the cause of that discordance is defensive minimization rather than culturally shaped symptom expression, genuine sub-syndromal states, or some combination. The defensive-presentation interpretation is an inference, argued from the structural profile of a population selected for extreme adversity, the convergent and abundant risk data, and a well-developed literature on stigma, idioms of distress, and strength-schema self-silencing (Kohrt et al., 2014; Watson & Hunter, 2015; Woods-Giscombé, 2010). It is, in our view, the most parsimonious interpretation of the evidence, but it should be tested directly with qualitative work, validity indices, and multimodal corroboration rather than treated as established by the discordance alone.

The dataset is also uneven in coverage. The standardized mood and trauma screens have strong coverage suitable for full-population analysis, but the substance screens, and especially the admission and integrated psychosocial forms, are sparser; the poverty- and exploitation-related findings are therefore presented as

exploratory subset evidence and theoretical context rather than as full-population inferences. The biological-sex field is too incomplete and imbalanced to support a formal sex or gender comparison, so gender enters our account as interpretive context rather than as a tested variable, notwithstanding its centrality to the argument. Because the data are aggregate and program-derived, we cannot adjust for clustering by site or cohort at the individual level, and recruitment-wave effects cannot be fully ruled out; cohort is best treated as a data-quality and recruitment covariate rather than as a substantive comparison variable.

The repeated-measure findings carry their own cautions. Improvement among baseline-elevated clients is large and consistent across seven-day and 30-day windows, but it cannot be attributed to any specific intervention without exposure timing or contact intensity data. The elevated subgroups are small, so floor effects and regression to the mean are plausible contributors. We deliberately describe these results as improvement-consistent rather than causal. ACE change, in particular, is interpreted as disclosure or data-quality variation rather than clinical improvement, since ACE measures historical exposure. Finally, the voice-AI capabilities discussed in Sections 6 and 7 are presented as a research-grounded rationale rather than as a tested intervention within this dataset; demonstrating that a multimodal system actually recovers the masked severity identified here is the necessary next study, not a claim established by the present analysis.

9. Conclusion

In a population of women selected for extreme poverty, histories of street-based sex work, and chronic unemployability — a population in which complex trauma is structurally near-certain — conventional symptom-severity scoring classifies the majority as low in severity. Only 11.3% met a moderate-or-higher PHQ-9/GAD-7 threshold, compared with 85.7% with adverse childhood exposure, and a high-risk masked group 3.7 times larger than the group conventional severity captures. Put in clinical terms, this means that for every five women carrying a documented high-risk signal, roughly four (78.6%) were scored below the threshold that would prompt care, and that of the women flagged for suicide risk, about seven in ten fell below the conventional cut-off entirely, their risk not detected by the aggregate score that is typically used to guide care decisions.

These are not abstract percentages. Each is a woman with trauma, suicidality, or both, recorded by the instrument as not reaching the threshold for clinical attention, in a system with too few clinicians to catch the error downstream. We have argued, based on the data and a substantial cross-cultural and gender literature, that this is better understood as a limitation of measurement in this context rather than as a direct characterization of the women. The dominant low scores are best understood as defensive or constrained presentation, shaped by stigma, gendered shame, cultural and religious idioms of distress, poverty-driven normalization, and a strength-schema that discourages disclosure, and then accepted uncritically by a clinical system primed to read a reassuring total as low need.

The “resilient African woman” narrative, applied to these scores, is not a description of strength but often functions as a rationalization for non-treatment, and the literature on the Strong Black Woman schema shows it to be a documented hazard to the health of the very women it claims to celebrate. To read low scores as resilience in this population is to convert a detection failure into a flattering attribute of the patient, and to license the diagnostic overshadowing and neglect that the data document. The accurate description is harder and more useful: many of these women are surviving while carrying a severe, under-detected psychiatric burden.

The corrective is to stop treating a single suppressed channel as the measure of severity. This analysis shows that Africa’s mental health crisis is two gaps, not one — a workforce gap and a clinical defensive gap — and that a multimodal, voice-based AI infrastructure is potentially well positioned to address both. It extends scarce expertise to where providers are absent; it adds a detection channel that stigma, literacy, and shame cannot

close; it treats discordance as a signal rather than noise; and it enforces item-level safety logic at scale, in the multilingual, low-connectivity conditions of African delivery. Its value is not merely to substitute for absent specialists but to expand what can be detected beyond the limits of conventional instruments. Trauma has many faces; accurate detection and treatment require infrastructure capable of recognizing them all. Naming what conventional severity misses — and building the systems that can find it — is a precondition for treating these women as the evidence says they should be treated. That is a research and implementation agenda, and this analysis is offered as its empirical foundation.

Declarations

Author contributions. Y. Alemu and P. Ohiomoba conceived the study, designed the analytic framework, led the data analysis, and drafted the manuscript. O. Balatel contributed to the analytic design and technical infrastructure. Y. Kahsay, F. Weldearegay, and L. Hailu contributed to data collection, clinical context, and critical revision. T. Fekadu contributed to initial conceptualization and critical revision. D. Neger contributed to program implementation, data curation, and critical revision. All authors reviewed and approved the final manuscript and agree to be accountable for the work.

Conflict of interest. In the interest of full transparency, the authors disclose that several of them (Y. Alemu, P. Ohiomoba, and O. Balatel) are affiliated with AYA-Innovation, the organization that develops the multimodal, voice-based artificial-intelligence approach discussed in Sections 6 and 7 of this manuscript. The empirical analyses and findings reported here were derived from de-identified clinical records and are independent of that technology; they are presented as obtained, regardless of their implications for it. No author received financial compensation contingent on the conclusions of this manuscript, and the authors declare no other competing financial or personal interests.

Funding. This manuscript received no external funding, grant support, or commercial sponsorship.

Ethics approval. Ethical approval for this retrospective analysis of de-identified clinical records was obtained from the Institutional Review Board of St. Paul's Hospital Millennium Medical College. The underlying clinical care, case identification, and service delivery activities were conducted as part of routine practice under the memorandum of understanding between AYA-Innovation and the Lenegewa Women's Rehabilitation and Skills Development Center. All records were de-identified before analysis.

Consent and permission for publication. The data were collected as part of routine service delivery, and written informed consent for the use of de-identified clinical information was obtained within the program's data-governance framework. This manuscript reports aggregate findings only; no names, photographs, or individually identifying details are included, and no individual-level case material is reproduced. Permission to analyze and publish the de-identified program data was granted under the memorandum of understanding between AYA-Innovation and the Lenegewa Women's Rehabilitation and Skills Development Center.

Data availability. No public dataset is associated with this manuscript. The underlying client-level records are not publicly available owing to patient confidentiality, ethical restrictions, and the sensitive nature of the population described. The aggregate results supporting the findings are presented in full in the tables and figures of this manuscript.

Reporting guidelines. This observational analysis was prepared in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidance for cross-sectional studies, with available-case denominators reported transparently for every measure and complete-case results used in sensitivity analysis.

Acknowledgments. The authors thank the staff of the Lenegewa Women's Rehabilitation and Skills Development Center for their clinical and program work, and they recognize the women served by the program, whose experiences motivate this analysis.

Generative AI statement. The authors verify and take full responsibility for the use of generative AI in the preparation of this manuscript. Generative AI assisted with language editing, organization, formatting, and clarity, and with structuring author-provided analytic results and tables. All interpretations, analyses, references, and conclusions were reviewed, verified, and approved by the authors. No generative AI tool was used to create, alter, or fabricate patient data, clinical findings, references, or research results.

References

- Abrams, J. A., Hill, A., & Maxwell, M. (2019). Underneath the mask of the Strong Black Woman schema: Disentangling influences of strength and self-silencing on depressive symptoms among U.S. Black women. *Sex Roles*, 80(9–10), 517–526. <https://doi.org/10.1007/s11199-018-0956-y>
- Abreha, G. F., Adhanu, H. H., Aregawi, A. B., Wuneh, A. D., Tesfay, F., Lema, G. K., et al. (2025). Exploring physical, sexual and mental health consequences of gender-based violence among women and girls during conflict in Tigray, Ethiopia. *BMC Public Health*, 25, 2103. <https://doi.org/10.1186/s12889-025-23349-0>
- Alemu, Y., Chen, H., Duan, C., Caulley, D., Arriaga, R. I., & Sezgin, E. (2023). Detecting clinically relevant emotional distress and functional impairment in children and adolescents: Protocol for an automated speech analysis algorithm development study. *JMIR Research Protocols*, 12, e46970. <https://doi.org/10.2196/46970>
- Alemu, Y., Cárdenas Bautista, E., Vinson, S., Ohiomoba, P., Melecio, N., Amare, H., Vivia, J., & Karippara, H. (2024). Implementing and analyzing the advantages of voice AI as measurement-based care to address behavioral health treatment disparities among youth in economically disadvantaged communities. *Telehealth and Medicine Today*, 9(4), Article 506. <https://doi.org/10.30953/thmt.v9.506>
- Alemu, Y., Ohiomoba, P., Kahsay, Y., Weldearegay, F., Amare, H., Negussie, S., Neger, D., Fekadu, T., Balatel, O., Degefu, A., Abdi, S., Ibrahim, R., & Amare, K. (2026). An AI-enabled, trauma-informed rehabilitation model for Ethiopian women with complex trauma: Programmatic implementation, early outcomes, and implications for scalable care in Africa. *International Journal of Psychiatry Research*, 9(1), 1–28. <https://doi.org/10.33425/2641-4317.1237>
- Almaghrabi, S. A., Clark, S. R., & Baumert, M. (2023). Bio-acoustic features of depression: A review. *Biomedical Signal Processing and Control*, 85, 105020. <https://doi.org/10.1016/j.bspc.2023.105020>
- Asefa, E. Y., Haile, A. B., Mohamed, O. Y., & Berhanu, D. (2024). The magnitude of gender-based violence, health consequences, and associated factors among women living in post-war woredas of North Shewa zone, Amhara, Ethiopia, 2022. *Frontiers in Global Women's Health*, 5, 1335254. <https://doi.org/10.3389/fgwh.2024.1335254>
- Baheretibeb, Y., Wondimagegn, D., & Law, S. (2021). Holy water and biomedicine: A descriptive study of active collaboration between religious traditional healers and biomedical psychiatry in Ethiopia. *BJPsych Open*, 7(3), e92. <https://doi.org/10.1192/bjo.2021.56>
- Bauhoff, S. (2011). Systematic self-report bias in health data: Impact on estimating cross-sectional and treatment effects. *Health Services and Outcomes Research Methodology*, 11(1–2), 44–53. <https://doi.org/10.1007/s10742-011-0069-3>
- Beauboeuf-Lafontant, T. (2007). You have to show strength: An exploration of gender, race, and depression. *Gender & Society*, 21(1), 28–51. <https://doi.org/10.1177/0891243206294108>
- Bekeko, S. D., Nimani, T. D., Darcho, S. D., & Bayisa, F. S. (2025). Depression, anxiety and its predictor among internally displaced person in Metekel Ethiopia, 2023: Using a structural equation model. *Frontiers in Psychiatry*, 15, 1458939. <https://doi.org/10.3389/fpsy.2024.1458939>
- Betancourt, T. S., Meyers-Ohki, S. E., Charrow, A. P., & Tol, W. A. (2013). Interventions for children affected by war: An ecological perspective on psychosocial support and mental health care. *Harvard Review of Psychiatry*, 21(2), 70–91. <https://doi.org/10.1097/HRP.0b013e318283bf8f>
- Bryant, R. A. (2019). Post-traumatic stress disorder: A state-of-the-art review of evidence and challenges. *World Psychiatry*, 18(3), 259–269. <https://doi.org/10.1002/wps.20656>

- Castelin, S., & White, G. (2022). "I'm a strong independent Black woman": The Strong Black Woman schema and mental health in college-aged Black women. *Psychology of Women Quarterly*, 46(2), 196–208. <https://doi.org/10.1177/03616843211067501>
- Caulley, D., Alemu, Y., Burson, S., et al. (2023). Objectively quantifying pediatric psychiatric severity using AI and voice recognition technology. *JMIR Research Protocols*, 12, Article e51912. <https://doi.org/10.2196/51912>
- Charlson, F., van Ommeren, M., Flaxman, A., Cornett, J., Whiteford, H., & Saxena, S. (2019). New WHO prevalence estimates of mental disorders in conflict settings: A systematic review and meta-analysis. *The Lancet*, 394(10194), 240–248. [https://doi.org/10.1016/S0140-6736\(19\)30934-1](https://doi.org/10.1016/S0140-6736(19)30934-1)
- Chisholm, D., Sweeny, K., Sheehan, P., Rasmussen, B., Smit, F., Cuijpers, P., & Saxena, S. (2016). Scaling-up treatment of depression and anxiety: A global return on investment analysis. *The Lancet Psychiatry*, 3(5), 415–424. [https://doi.org/10.1016/S2215-0366\(16\)30024-4](https://doi.org/10.1016/S2215-0366(16)30024-4)
- Collins, P. H. (2000). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment* (2nd ed.). Routledge.
- Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6), 1241–1299. <https://doi.org/10.2307/1229039>
- Dellie, E., Salelew, E., Fetene, S. M., et al. (2024). Gender-based violence among women and girls in conflict-affected areas of Northeast Amhara, Ethiopia. *Frontiers in Global Women's Health*, 5, 1453149. <https://doi.org/10.3389/fgwh.2024.1453149>
- Deyessa, N., Getahun, K., Filimona, R., & Tadesse, A. (2025). The impact of wartime conflict on the mental health problems of women in the conflict-hit population in Woldia, Ethiopia. *PLOS ONE*, 20(6), e0322047. <https://doi.org/10.1371/journal.pone.0322047>
- Donaghy, P., Ennis, E., Mulvenna, M., Bond, R., Kennedy, N., McTear, M., O'Connell, H., Blaylock, N., & Brueckner, R. (2024). A review of studies using machine learning to detect voice biomarkers for depression. *Journal of Technology in Behavioral Science*. <https://doi.org/10.1007/s41347-024-00454-2>
- Ewing, J. A. (1984). Detecting alcoholism: The CAGE questionnaire. *JAMA*, 252(14), 1905–1907. <https://doi.org/10.1001/jama.1984.03350140051025>
- Donovan, R. A., & West, L. M. (2015). Stress and mental health: Moderating role of the Strong Black Woman stereotype. *Journal of Black Psychology*, 41(4), 384–396. <https://doi.org/10.1177/0095798414543014>
- Federal Ministry of Health, Ethiopia. (2020). *National mental health strategy 2020–2025*. Ministry of Health. <http://repository.iphce.org/bitstream/handle/123456789/1423/MENTAL-HEALTH%20strategy.pdf>
- Fekadu, A., Medhin, G., Selamu, M., Giorgis, T. W., Lund, C., Alem, A., & Prince, M. (2017). Recognition of depression by primary care clinicians in rural Ethiopia. *BMC Family Practice*, 18(1), 56. <https://doi.org/10.1186/s12875-017-0628-y>
- Fekadu, A., Milkias, B., Birhane, R., Tekola, B., Yared, M., Eshetu, T., Getahun, T., Lempp, H., Tylee, A., Belina, M., Medhin, G., Habtamu, K., Alem, A., Lund, C., Petersen, I., Thornicroft, G., Cleare, A. J., Mayston, R., Assefa, E. A., ... Prince, M. J. (2025). A pilot cluster-controlled trial of interventions to improve detection of depression in primary healthcare in Ethiopia. *BMC Medicine*, 23, 185. <https://doi.org/10.1186/s12916-025-04010-4>
- Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., Koss, M. P., & Marks, J. S. (1998). Relationship of childhood abuse and household dysfunction to many of the leading causes of death

- in adults: The Adverse Childhood Experiences (ACE) Study. *American Journal of Preventive Medicine*, 14(4), 245–258. [https://doi.org/10.1016/S0749-3797\(98\)00017-8](https://doi.org/10.1016/S0749-3797(98)00017-8)
- Fisseha, G., Gebregzabher Gebrehiwot, T., Gebremichael, M. W., et al. (2023). *War-related sexual and gender-based violence in Tigray, Northern Ethiopia: A community-based study*. *BMJ Global Health*, 8(7), e010270. <https://doi.org/10.1136/bmjgh-2022-010270>
- Freud, A. (1966). *The ego and the mechanisms of defence* (rev. ed.). International Universities Press. (Original work published 1936)
- Gelaye, B., Williams, M. A., Lemma, S., Deyessa, N., Bahretibeb, Y., Shibre, T., Wondimagegn, D., Lemenhe, A., Fann, J. R., Vander Stoep, A., & Zhou, X.-H. A. (2013). Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. *Psychiatry Research*, 210(2), 653–661. <https://doi.org/10.1016/j.psychres.2013.07.015>
- Gossop, M., Darke, S., Griffiths, P., Hando, J., Powis, B., Hall, W., & Strang, J. (1995). The Severity of Dependence Scale (SDS): Psychometric properties of the SDS in English and Australian samples of heroin, cocaine and amphetamine users. *Addiction*, 90(5), 607–614. <https://doi.org/10.1046/j.1360-0443.1995.9056072.x>
- Grisaru, N., Budowski, D., & Witztum, E. (1997). Possession by the Zar among Ethiopian immigrants to Israel: Psychopathology or culture-bound syndrome? *Psychopathology*, 30(4), 223–233. <https://doi.org/10.1159/000285051>
- Hanlon, C., Medhin, G., Alem, A., Araya, M., Abdulahi, A., Hughes, M., Tesfaye, M., Wondimagegn, D., Patel, V., & Prince, M. (2008). Detecting perinatal common mental disorders in Ethiopia: Validation of the Self-Reporting Questionnaire and Edinburgh Postnatal Depression Scale. *Journal of Affective Disorders*, 108(3), 251–262. <https://doi.org/10.1016/j.jad.2007.10.023>
- Hanlon, C., Medhin, G., Selamu, M., Breuer, E., Worku, B., Hailemariam, M., Lund, C., Prince, M., & Fekadu, A. (2015). Validity of brief screening questionnaires to detect depression in primary care in Ethiopia. *Journal of Affective Disorders*, 186, 32–39. <https://doi.org/10.1016/j.jad.2015.07.015>
- Herman, J. L. (1992). *Trauma and recovery: The aftermath of violence—from domestic abuse to political terror*. Basic Books.
- Hinton, D. E., & Lewis-Fernández, R. (2010). Idioms of distress among trauma survivors: Subtypes and clinical utility. *Culture, Medicine, and Psychiatry*, 34(2), 209–218. <https://doi.org/10.1007/s11013-010-9175-x>
- Iyer, R., & Meyer, D. (2022). Detection of suicide risk using vocal characteristics: Systematic review. *JMIR Biomedical Engineering*, 7(2), Article e42386. <https://doi.org/10.2196/42386>
- Kakuma, R., Minas, H., van Ginneken, N., Dal Poz, M. R., Desiraju, K., Morris, J. E., Saxena, S., & Scheffler, R. M. (2011). Human resources for mental health care: Current situation and strategies for action. *The Lancet*, 378(9803), 1654–1663. [https://doi.org/10.1016/S0140-6736\(11\)61093-3](https://doi.org/10.1016/S0140-6736(11)61093-3)
- Kalu, O. U. (2008). *African Pentecostalism: An introduction*. Oxford University Press.
- Keynejad, R. C., Spagnolo, J., & Thornicroft, G. (2021). WHO mental health gap action programme (mhGAP) intervention guide: Updated systematic review on evidence and impact. *Evidence-Based Mental Health*, 24(3), 124–130. <https://doi.org/10.1136/ebmental-2021-300254>
- Kizilhan, J. I., Steger, F., & Noll-Hussong, M. (2020). Shame, dissociative seizures and their correlation among traumatized female Yazidi with experience of sexual violence. *British Journal of Psychiatry*, 216(3), 138–143. <https://doi.org/10.1192/bjp.2019.197>

- Kleinman, A. (1980). *Patients and healers in the context of culture: An exploration of the borderland between anthropology, medicine, and psychiatry*. University of California Press.
- Kohrt, B. A., Luitel, N. P., Acharya, P., & Jordans, M. J. D. (2016). Detection of depression in low resource settings: Validation of the Patient Health Questionnaire (PHQ-9) and cultural concepts of distress in Nepal. *BMC Psychiatry*, 16, 58. <https://doi.org/10.1186/s12888-016-0768-y>
- Kohrt, B. A., Jordans, M. J. D., Rai, S., Shrestha, P., Luitel, N. P., Ramaiya, M. K., Singla, D. R., & Patel, V. (2015). Therapist competence in global mental health: Development of the ENhancing Assessment of Common Therapeutic factors (ENACT) rating scale. *Behaviour Research and Therapy*, 69, 11–21. <https://doi.org/10.1016/j.brat.2015.03.009>
- Kohrt, B. A., Rasmussen, A., Kaiser, B. N., Haroz, E. E., Maharjan, S. M., Mutamba, B. B., de Jong, J. T. V. M., & Hinton, D. E. (2014). Cultural concepts of distress and psychiatric disorders: Literature review and research recommendations for global mental health epidemiology. *International Journal of Epidemiology*, 43(2), 365–406. <https://doi.org/10.1093/ije/dyt227>
- Komu, C. K., Ngigi, M., & Melson, A. J. (2025). *Barriers and facilitators to accessing mental health services for adults in Sub-Saharan Africa: A systematic review*. *Mental Health Science*, 3(1), e70006. <https://doi.org/10.1002/mhs2.70006>
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
- Lanius, R. A., Vermetten, E., Loewenstein, R. J., Brand, B., Schmahl, C., Bremner, J. D., & Spiegel, D. (2010). Emotion modulation in PTSD: Clinical and neurobiological evidence for a dissociative subtype. *American Journal of Psychiatry*, 167(6), 640–647. <https://doi.org/10.1176/appi.ajp.2009.09081168>
- Liao, K. Y.-H., Wei, M., & Yin, M. (2020). The misunderstood schema of the Strong Black Woman: Exploring its mental health consequences and coping responses among African American women. *Psychology of Women Quarterly*, 44(1), 84–94. <https://doi.org/10.1177/0361684319883198>
- Liu, L., Liu, L., Wafa, H. A., Tydeman, F., Xie, W., & Wang, Y. (2024). Diagnostic accuracy of deep learning using speech samples in depression: A systematic review and meta-analysis. *Journal of the American Medical Informatics Association*, 31(10), 2394–2404. <https://doi.org/10.1093/jamia/ocae189>
- Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1), 96–116. <https://doi.org/10.1002/lio2.354>
- Lund, C., Brooke-Sumner, C., Baingana, F., Baron, E. C., Breuer, E., Chandra, P., Haushofer, J., Herrman, H., Jordans, M., Kieling, C., Medina-Mora, M. E., Morgan, E., Omigbodun, O., Tol, W., Patel, V., & Saxena, S. (2018). Social determinants of mental disorders and the Sustainable Development Goals: A systematic review of reviews. *The Lancet Psychiatry*, 5(4), 357–369. [https://doi.org/10.1016/S2215-0366\(18\)30060-9](https://doi.org/10.1016/S2215-0366(18)30060-9)
- Mutamba, B. B., van Ginneken, N., Smith Paintain, L., Wandiembe, S., & Schellenberg, D. (2013). Roles and effectiveness of lay community health workers in the prevention of mental, neurological and substance use disorders in low- and middle-income countries: A systematic review. *BMC Health Services Research*, 13, 412. <https://doi.org/10.1186/1472-6963-13-412>
- Nelson, T., Cardemil, E. V., & Adeoye, C. T. (2016). Rethinking strength: Black women's perceptions of the "Strong Black Woman" role. *Psychology of Women Quarterly*, 40(4), 551–563. <https://doi.org/10.1177/0361684316646716>

- Nicolaidis, C., Timmons, V., Thomas, M. J., Waters, A. S., Wahab, S., Mejia, A., & Mitchell, S. R. (2010). "You don't go tell white people nothing": African American women's perspectives on the influence of violence and race on depression and depression care. *American Journal of Public Health*, 100(8), 1470–1476. <https://doi.org/10.2105/AJPH.2009.161950>
- Nock, M. K., Borges, G., Bromet, E. J., Alonso, J., Angermeyer, M., Beautrais, A., Bruffaerts, R., et al. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry*, 192(2), 98–105. <https://doi.org/10.1192/bjp.bp.107.040113>
- Ouanes, K., & Farhah, N. (2024). Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. *Journal of Medical Systems*, 48(1), 74. <https://doi.org/10.1007/s10916-024-02098-4>
- Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., et al. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, 392(10157), 1553–1598. [https://doi.org/10.1016/S0140-6736\(18\)31612-X](https://doi.org/10.1016/S0140-6736(18)31612-X)
- Patel, V., Weiss, H. A., Chowdhary, N., Naik, S., Pednekar, S., Chatterjee, S., De Silva, M. J., Bhat, B., Araya, R., King, M., Simon, G., Verdeli, H., & Kirkwood, B. R. (2010). Effectiveness of an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in Goa, India (MANAS): A cluster randomised controlled trial. *The Lancet*, 376(9758), 2086–2095. [https://doi.org/10.1016/S0140-6736\(10\)61508-5](https://doi.org/10.1016/S0140-6736(10)61508-5)
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46(3), 598–609. <https://doi.org/10.1037/0022-3514.46.3.598>
- Riad, R., Denais, M., de Gennes, M., Lesage, A., Oustric, V., Cao, X. N., Mouchabac, S., & Bourla, A. (2024). Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: Algorithm development and validation study. *Journal of Medical Internet Research*, 26, Article e58572. <https://doi.org/10.2196/58572>
- Sankoh, O., Sevalie, S., & Weston, M. (2018). Mental health in Africa. *The Lancet Global Health*, 6(9), e954–e955. [https://doi.org/10.1016/S2214-109X\(18\)30303-6](https://doi.org/10.1016/S2214-109X(18)30303-6)
- Saraceno, B., van Ommeren, M., Batniji, R., Cohen, A., Gureje, O., Mahoney, J., Sridhar, D., & Underhill, C. (2007). Barriers to improvement of mental health services in low-income and middle-income countries. *The Lancet*, 370(9593), 1164–1174. [https://doi.org/10.1016/S0140-6736\(07\)61263-X](https://doi.org/10.1016/S0140-6736(07)61263-X)
- Shedler, J. (2010). The efficacy of psychodynamic psychotherapy. *American Psychologist*, 65(2), 98–109. <https://doi.org/10.1037/a0018378>
- Shedler, J., Mayman, M., & Manis, M. (1993). The illusion of mental health. *American Psychologist*, 48(11), 1117–1131. <https://doi.org/10.1037/0003-066X.48.11.1117>
- Shin, J., & Bae, S. M. (2024). Use of voice features from smartphones for monitoring depressive disorders: Scoping review. *Digital Health*, 10, 20552076241261920. <https://doi.org/10.1177/20552076241261920>
- Sifneos, P. E. (1973). The prevalence of "alexithymic" characteristics in psychosomatic patients. *Psychotherapy and Psychosomatics*, 22(2–6), 255–262. <https://doi.org/10.1159/000286529>
- Simie, D., Azale, T., Gashaw, F., Wondie, M., Gebeyehu, D. A., & Mekuriaw, B. Y. (2025). *Post-traumatic stress disorder symptoms among youths in war-affected areas of Northeast Ethiopia*. *BMC Psychiatry*, 25, 871. <https://doi.org/10.1186/s12888-025-07356-4>
- Simon, G. E., Rutter, C. M., Peterson, D., Oliver, M., Whiteside, U., Operskalski, B., & Ludman, E. J. (2013). Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? *Psychiatric Services*, 64(12), 1195–1202

- Singla, D. R., Kohrt, B. A., Murray, L. K., Anand, A., Chorpita, B. F., & Patel, V. (2017). Psychological treatments for the world: Lessons from low- and middle-income countries. *Annual Review of Clinical Psychology*, 13, 149–181. <https://doi.org/10.1146/annurev-clinpsy-032816-045217>
- Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092–1097. <https://doi.org/10.1001/archinte.166.10.1092>
- Starrs, A. M., Ezeh, A. C., Barker, G., Basu, A., Bertrand, J. T., Blum, R., et al. (2018). Accelerate progress—Sexual and reproductive health and rights for all: Report of the Guttmacher–Lancet Commission. *The Lancet*, 391(10140), 2642–2692. [https://doi.org/10.1016/S0140-6736\(18\)30293-9](https://doi.org/10.1016/S0140-6736(18)30293-9)
- Tadesse, G. M. (2025). Sacred suffering and stigma: Cultural pathways to mental health in Ethiopia. *Explore*, 22(1), 103289. <https://doi.org/10.1016/j.explore.2025.103289>
- Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., et al. (2017). Undertreatment of people with major depressive disorder in 21 countries. *British Journal of Psychiatry*, 210(2), 119–124. <https://doi.org/10.1192/bjp.bp.116.188078>
- Tinsae, T., Shumet, S., Tadesse, G., Takelle, G. M., Rtbe, G., Melkam, M., Andualem, F., Nakie, G., Segon, T., Koye, S., Fentahun, S., & Alemu, W. G. (2024). Post-traumatic stress disorder in the Ethiopian population dwelling in war-affected communities: A systematic review and meta-analysis. *Frontiers in Psychiatry*, 15, 1399013. <https://doi.org/10.3389/fpsy.2024.1399013>
- Tomlinson, M., Grimsrud, A. T., Stein, D. J., Williams, D. R., & Myer, L. (2009). The epidemiology of major depression in South Africa: Results from the South African Stress and Health study. *South African Medical Journal*, 99(5), 368–373.
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- Vaillant, G. E. (1992). Ego mechanisms of defense: A guide for clinicians and researchers. American Psychiatric Press.
- Van de Mortel, T. F. (2008). Faking it: Social desirability response bias in self-report research. *Australian Journal of Advanced Nursing*, 25(4), 40–48.
- van der Kolk, B. A. (2014). The body keeps the score: Brain, mind, and body in the healing of trauma. Viking.
- Ventevogel, P., Jordans, M., Reis, R., & de Jong, J. (2013). Madness or sadness? Local concepts of mental illness in four conflict-affected African communities. *Conflict and Health*, 7, 3. <https://doi.org/10.1186/1752-1505-7-3>
- Watson, N. N., & Hunter, C. D. (2015). Anxiety and depression among African American women: The costs of strength and negative attitudes toward psychological help-seeking. *Cultural Diversity and Ethnic Minority Psychology*, 21(4), 604–612. <https://doi.org/10.1037/cdp0000015>
- Weinberger, D. A., Schwartz, G. E., & Davidson, R. J. (1979). Low-anxious, high-anxious, and repressive coping styles: Psychometric patterns and behavioral and physiological responses to stress. *Journal of Abnormal Psychology*, 88(4), 369–380. <https://doi.org/10.1037/0021-843X.88.4.369>
- Wondimagegn, D., Pain, C., Seifu, N., Cartmill, C., Alemu, A. A., & Whitehead, C. R. (2023). Reimagining global mental health in Africa. *BMJ Global Health*, 8(9), e013232. <https://doi.org/10.1136/bmjgh-2023-013232>
- Woods-Giscombé, C. L. (2010). Superwoman schema: African American women’s views on stress, strength, and health. *Qualitative Health Research*, 20(5), 668–683. <https://doi.org/10.1177/1049732310361892>

- Woods-Giscombé, C. L., Robinson, M. N., Carthon, D., Devane-Johnson, S., & Corbie-Smith, G. (2019). Superwoman schema, stigma, spirituality, and culturally sensitive providers: Factors influencing African American women's use of mental health services. *Journal of Best Practices in Health Professions Diversity*, 12(1), 51–72.
- World Health Organization. (2010). *mhGAP intervention guide for mental, neurological and substance use disorders in non-specialized health settings*. World Health Organization
- World Health Organization. (2017). *Mental health atlas 2017*. World Health Organization. <https://www.who.int/publications/i/item/9789241514019>
- World Health Organization. (2019). *WHO guideline: Recommendations on digital interventions for health system strengthening*. World Health Organization. <https://www.who.int/publications/i/item/9789241550505>
- World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization. <https://www.who.int/publications/i/item/9789240029200>
- World Health Organization Regional Office for Africa. (2025, October 10). Mental health conditions affect 150 million in Africa amid insufficient care services. WHO Regional Office for Africa.
- Zhang, L., Duvvuri, R., Chandra, K. K. L., Nguyen, T., & Ghomi, R. H. (2020). Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative. *Depression and Anxiety*, 37(7), 657–669. <https://doi.org/10.1002/da.23020>