Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Culturally Anchored Target Construction for Amharic Psychiatric Voice Artificial Intelligence: A Multimodal Data-Centric Study

Patrick Ohiomoba
AYA-Innovation
Yared Alemu
ya@aya-innovation.com

AYA-Innovation
Octavian Balatel
AYA-Innovation
Habib Mohammed
Addis Ababa Science and Technology University
Jale Bikila Abeya
Woosong University

Article

Keywords: data-centric artificial intelligence, speech emotion recognition, Amharic, culturally anchored annotation, inter-rater reliability, psychiatric severity

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10684265/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Culturally Anchored Target Construction for Amharic Psychiatric Voice Artificial Intelligence: A Multimodal Data-Centric Study

Patrick Ohiomoba[1], Yared Alemu[1], Octavian Balatel[1], Habib Mohammed[2], and Jale Bikila Abeya[3]

[1] AYA-Innovation, Bole Sub-City, Woreda 03, Addis Ababa, Ethiopia

[2] Addis Ababa Science and Technology University, Addis Ababa, Ethiopia

[3] Woosong University, 171 Dongdaejeon-ro, Dong-gu, Daejeon 34606, Republic of Korea

Corresponding author: Yared Alemu, AYA-Innovation, Bole Sub-City, Woreda 03, Addis Ababa, Ethiopia.
Email: ya@aya-innovation.com

Abstract

Background: High-stakes clinical voice artificial intelligence requires supervisory targets that are culturally valid in the language and population of deployment and that remain separate from the clinical measures later used for independent validation. In Ethiopian populations with extensive documented trauma, conventional self-report screening systematically understates severity — the defensiveness gap — motivating a voice-based severity channel that does not depend on questionnaire disclosure. This study was designed to answer a specific measurement question: can trained Ethiopian raters, under a culturally adapted listening protocol, judge psychiatric severity from Amharic speech reliably enough—and under what aggregation design—to serve as machine-learning supervisory targets independent of self-report? We report the completed analysis of a culturally anchored voice-labeling snapshot from the Lenegewa Women's Rehabilitation and Skills Development Center that tests this question, together with the pre-specified validity architecture built on it.

Methods: The multimodal resource links five layers: 3,617 supported beneficiaries; 4,163 voice-sample records from 2,600 clients; 19,610 standardized survey records; 111,059 item-level symptom responses; and 67,274 clinical, service, and risk documentation records. Three trained Ethiopian raters scored each Amharic sample on the 0–39 Emotional Distress Index (EDI) and independently rated fear, anger, neutral, sadness, and happiness on 0–4 scales under a culturally adapted listening protocol. Clinical scales, diagnoses, safety indicators, and outcomes were withheld from raters and reserved as independent validation targets. The locked analysis snapshot comprised 2,479 samples, 7,919 rating attempts, 7,086 usable ratings, and 2,267 complete three-rater EDI matrices; four severity bands were derived from sample-level EDI scores rather than separately assigned.

Results: Derived severity was Low in 720 samples (29.0%), Moderate in 1,446 (58.3%), High in 300 (12.1%), and Severe in 13 (0.5%); median EDI was 13.0 (IQR 9.0–17.0). Ordinal Krippendorff's alpha was 0.46 (95% CI 0.43–0.48) for derived severity bands and ranged from 0.28 to 0.51 across the five voice types; mean pairwise quadratic-weighted kappa was 0.47. EDI reliability was substantially higher for the three-rater mean (ICC(2,k) = 0.78, 95% CI 0.76–0.80) than for a single rating (ICC(2,1) = 0.54, 95% CI 0.52–0.57), with a standard error of measurement of 4.80 EDI points; rater means differed by at most 1.12 points, while Bland–Altman limits documented substantial sample-level disagreement.

Conclusions: The research question is answered affirmatively with a design qualification: severity judgment of Amharic speech is reliable enough to supervise machine learning only as a multi-rater aggregate, and a single rating is too noisy a target. The analysis establishes culturally anchored target construction, quantifies annotation uncertainty, preserves non-circular validation, and fixes the machine-learning validity architecture — audio-only inference, client-level data splits, constrained augmentation, calibration, subgroup checks, and teacher–student deployment separation — before any model fitting, without claiming model performance. Full-corpus labeling and subsequent algorithm performance evaluation will be reported separately.

Keywords: *data-centric artificial intelligence; speech emotion recognition; Amharic; culturally anchored annotation; inter-rater reliability; psychiatric severity*

1. Introduction

1.1 Two gaps, one measurement problem

The case for artificial intelligence in African mental health is usually made in workforce terms. Sub-Saharan Africa has among the lowest densities of mental-health professionals in the world, with roughly 1.4 providers per 100,000 population and, in Ethiopia, approximately 0.11 psychiatrists per 100,000 [1, 2, 3]. Even when providers are present, recognition is unreliable: in rural Ethiopian primary care, clinicians detected only a small fraction of depression identified by validated instruments, and a recent cluster-controlled trial confirmed that improving detection in Ethiopian primary healthcare remains an unsolved implementation problem [4, 5]. Conflict, displacement, and gender-based violence further concentrate psychiatric need in populations that health systems reach least [6].

Descriptive analyses of the linked Lenegewa Women's Rehabilitation and Skills Development Center records (Table 3) show that the workforce framing, while correct, is incomplete. In a population of Ethiopian women with extensive documented trauma, substance-use, and social adversity, conventional severity categories made the population appear largely well: only 296 of 2,618 paired baseline clients (11.3%) met conventional PHQ-9 or GAD-7 elevation, while the majority of clients carrying high-risk indicators—elevated ACE burden, substance-use severity, suicide-risk endorsement, or documented sexual abuse—fell below those same thresholds. The linked safety records quantify the clinical stakes: 70.7% of clients endorsing PHQ-9 item 9 (suicidal ideation or self-harm) had PHQ-9 totals below 10; 72.9% of clients with sexual-abuse flags scored below moderate thresholds; and a total-score triage rule of PHQ-9 or GAD-7 ≥ 10 captured only 218 of 840 safety-signal clients (26.0%). We term this second failure the defensiveness gap: a systematic distance between what standardized self-report instruments record and what trauma-affected clients are willing or able to disclose in questionnaire language.

The two gaps interact. The workforce gap means that first contact frequently occurs with non-specialist staff who depend on standardized screeners; the defensiveness gap means that those screeners systematically understate severity in exactly the populations where specialist backup is least available. Adding providers who rely on the same self-report channel does not close the second gap, because the failure is located in the measurement channel itself, not in the number of people administering it. Well-documented mechanisms converge to produce this failure: socially desirable responding and sensitive-question effects [7, 8], trauma-related avoidance, numbing, and dissociative shutdown [9, 10], and culturally patterned idioms of distress that route suffering away from the symptom vocabulary of Western instruments [11, 12, 13]. The instruments themselves are psychometrically sound — the PHQ-9 and GAD-7 are validated, including in East African populations [14, 15, 16, 17] — but validity under candid responding does not guarantee sensitivity under defended responding, and item 9 endorsement is known to be an imperfect proxy for suicide risk even in high-resource systems [18, 19, 20, 21].

1.2 Why voice, and why now

The case for a voice channel begins with arithmetic that self-report cannot fix. Ethiopia has roughly 0.11 psychiatrists per 100,000 population, and sub-Saharan Africa as a whole has approximately 1.4 mental-health professionals per 100,000 — orders of magnitude below need — while fewer than one in ten people with mental illness in the region ever reach professional services [1, 3, 4, 5]. Critically, there is no realistic training path that closes this gap on any policy-relevant timescale: the specialist workforce required does not exist at scale and cannot be trained at the necessary speed, a constraint documented for Ethiopia specifically and for the continent generally [1, 3]. The dominant response — task-shifting structured care to non-specialists under WHO's mhGAP framework — has produced real benefit [2, 22], but the linked cohort evidence summarized in Table 3 identifies what it cannot do: a brief-trained paraprofessional administering a self-report screen inherits exactly the suppressed disclosure channel that hides severity from everyone else. Task-shifting multiplies hands;

it does not create an independent measurement channel. The detection step that stepped care takes for granted is the step most likely to fail in this population.

The companion problem is that the failing channel is the only one conventional screening uses. Standardized instruments assume the respondent can read or hear an item, understand the construct it names, map her experience onto a numeric scale, and be willing to disclose the result — assumptions that break down precisely among women facing stigma, gendered shame, low literacy, and trauma-conditioned self-protection [7, 8]. In the Lenegewa cohort, the result was quantified: 78.6% of high-risk clients scored below conventional PHQ-9/GAD-7 thresholds, and roughly seven in ten clients endorsing suicide-related indicators presented totals a screening program would read as low need. Adding more self-report instruments, more administration time, or more administrators cannot repair a channel the client is — consciously or unconsciously — regulating. What the setting requires is a measurement channel that (i) does not depend on a specialist workforce that cannot be trained in time, and (ii) does not depend on the disclosure behaviors that defensiveness suppresses.

Voice is a candidate for that channel because it carries information about how a person speaks in addition to what she is willing to disclose. Distress is expressed in pitch and pitch variability, intensity, speech rate, pausing, prosodic range, vocal effort, and voice quality — properties that are difficult to fully suppress even when verbal content is guarded [23, 24, 25]. Because these paralinguistic properties do not require the speaker to read an item, name a symptom, or consent to disclose it in words, a voice channel is, in principle, less dependent on the very behaviors the defensiveness gap disrupts [26, 27]. Our prior pediatric work demonstrated that severity-relevant vocal signal can be extracted from brief speech samples collected during routine care [28], and our implementation work showed that voice-derived severity can function inside measurement-based care as a frequent, objective adjunct where clinician time is scarce [27].

Voice also fits the deployment realities of the setting in ways written screeners do not. Speaking is native to an oral communication culture and requires no literacy, so the channel is open to exactly the clients whom text-based instruments exclude. Brief voice samples can be captured on basic smartphones during routine encounters by non-specialist staff, processed offline or under low connectivity, delivered in local languages, and repeated at a marginal cost per assessment that approaches zero as systems scale [22, 26, 27, 29]. Where a psychiatrist-dependent pathway cannot reach a rehabilitation nurse, a community health worker, or a maternal health provider, a voice-based severity signal can — which is why the workforce gap and the defensiveness gap, taken together, point to the same technological response.

The Lenegewa dataset is unusually positioned for developing such a channel because it pairs voice recordings with standardized symptom measures, item-level safety signals, trauma and substance-use screens, clinical and service documentation, and repeated outcome data in the same high-need population, within a functioning rehabilitation program [30]. The goal is not to replace clinicians or to diagnose from speech. The goal is a culturally responsive severity signal that helps frontline teams identify who may need further assessment, safety review, or follow-up when conventional self-report is incomplete — a measurement-based-care function, delivered through a channel that is harder to mask.

1.3 Aims and contribution of this paper

This Research Article reports four linked data-centric AI contributions. First, it identifies the failure modes most consequential in African, low-resource-language voice AI: circular validation, speaker leakage, culturally invalid labels, and unsupported performance claims (Section 2). Second, it characterizes the multimodal Lenegewa resource and separates record-, file-, rating-attempt-, usable-rating-, sample-, and complete-case denominators (Section 3). Third, it reports a completed analysis of a locked labeling snapshot comprising 2,479 samples, 7,086 usable ratings from three Ethiopian raters, and 2,267 complete three-rater EDI matrices (Sections 4–5). Direct 0–39 EDI ratings and five 0–4 voice-type ratings are the annotation outputs; four overall-severity bands are derived from EDI and are not a second, independently assigned target. Fourth, the article fixes the machine-

learning validity architecture before model fitting: audio-only inference, multi-task supervision, client-level splits, constrained augmentation, nested hyperparameter selection, encoder comparison, calibration, subgroup checks, a comparator ladder, and teacher–student deployment separation (Sections 6–7). Full-corpus label completion, algorithm fitting, and model-performance results remain reserved for the follow-up algorithm paper. Section 1.4 states the research question and the testable expectations that structure the completed analysis.

1.4 Research question and hypotheses

The analyses reported in this article were designed to answer one primary research question: can trained Ethiopian raters, listening under a culturally adapted protocol, judge psychiatric severity from short samples of Amharic speech reliably enough—and under what aggregation design—to serve as machine-learning supervisory targets that are independent of self-report? This is a measurement-science question rather than a data-collection exercise: it asks whether culturally anchored human perception of vocal distress in an African language is a sufficiently stable measurand to anchor supervised learning, and it is answerable only with the quantitative reliability evidence reported in Section 5.7.

Three falsifiable hypotheses, fixed in the analysis plan before model development, are tested in this article. H1 (aggregation): the reliability of the mean of three independent raters' EDI scores materially exceeds the reliability of a single rating, such that multi-rater aggregation—not single-rater annotation—is the minimally adequate labeling design; this is tested by comparing ICC(2,k) with ICC(2,1) and by the standard error of measurement. H2 (ordinal coherence): culturally anchored severity judgments show above-chance ordinal inter-rater agreement at the level of derived severity bands and of the five voice types, quantified by ordinal Krippendorff's alpha and quadratic-weighted kappa, with agreement expected to be moderate rather than high given the perceptual difficulty of the constructs. H3 (distributional adequacy): the derived severity distribution of a routine-care rehabilitation population is dominated by low and moderate bands with a small severe minority, implying that any credible later model must be evaluated on minority-class sensitivity rather than aggregate accuracy; this is tested descriptively, and it feeds directly into the pre-specified endpoint families of Section 7.

Answering this question contributes scientific knowledge beyond the dataset itself. If single-rater labels prove unreliable, that finding generalizes as an annotation-design requirement for affective computing in low-resource languages, where single-annotator labeling is the norm. If aggregation restores adequate reliability, the quantified uncertainty—the standard error of measurement in EDI points—becomes a floor that any later model-performance claim must respect, because a model cannot be meaningfully more accurate than its targets are reliable. And the demonstration that a severity target can be constructed without access to the clinical scales later used for validation resolves, by design rather than assertion, the circular-validation weakness that Section 2.6 identifies as recurrent in the voice-biomarker literature.

2. Literature Review: The State of Voice-Based Psychiatric Severity Detection

2.1 Vocal indicators of psychiatric state: converging evidence, persistent caveats

The proposition that psychiatric states can be detected from speech now rests on two decades of accumulating evidence. This literature is often framed in terms of "vocal biomarkers" — a term that, in the speech context, denotes measurable acoustic and paralinguistic characteristics of the voice (prosody, rhythm, pausing, intensity, spectral properties, and voice quality) that covary with clinical states, not biological assays; we accordingly prefer the more precise phrase vocal indicators of psychiatric state, and use "biomarker" only in this qualified sense. Foundational reviews established that depression and suicide risk are accompanied by measurable changes in prosody, source, and spectral features of speech — reduced pitch variability, lowered intensity, slowed speech rate, lengthened pauses, and altered voice quality — and that these changes are detectable by

machine-learning models [23, 31]. Clinical work demonstrated early that vocal acoustic measures track depression severity and treatment response over time, supporting their use as repeated-measures indicators rather than one-time classifiers [32]. Subsequent reviews consolidated the acoustic feature families most consistently associated with depression — fundamental frequency statistics, intensity, temporal and pause structure, spectral energy distribution, and perturbation measures such as jitter and shimmer [33] — and positioned speech as a promising signal source for precision psychiatry more broadly [24].

The most recent evidence syntheses strengthen the quantitative case while sharpening the methodological caveats. 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 [34]. A large multi-symptom development-and-validation study demonstrated that automated speech analysis can estimate risk across depression, anxiety, insomnia, and fatigue from short telephone-quality recordings [35], and a scoping review of smartphone-collected voice features supported feasibility for naturalistic longitudinal monitoring [29]. In primary care — the deployment context closest to our own frontline setting — a machine-learning voice-analysis tool identified signals consistent with moderate-to-severe depression in a prospective clinical evaluation [36]. For the safety domain specifically, systematic review evidence indicates that vocal characteristics carry signal relevant to suicide risk [37], and acoustic-plus-language analysis has distinguished suicidal ideation among trauma-exposed veterans, a population whose guarded presentation resembles the defended presentations documented at Lenegewa [38].

Yet every recent synthesis also converges on the same caveats. Reviews of machine-learning vocal-indicator models for depression emphasize small samples, heterogeneous elicitation tasks, inconsistent reference standards, and limited external validation [39, 40]. Current voice-only models do not yet outperform well-administered self-report in general populations and are not standalone diagnostics [40]; their defensible role is as an independent, harder-to-suppress channel used alongside self-report under human review. Most published datasets are English-language, interview-based, and demographically narrow — the widely used Distress Analysis Interview Corpus being the canonical example [41] — which means that reported accuracies may not transfer to African clinical populations at all [42]. These caveats do not weaken the rationale for the present program; they define its requirements. A voice algorithm intended for Ethiopian women with complex trauma must be built on population-matched audio, culturally valid labels, and validation targets that include the hidden-risk presentations the technology exists to detect.

2.2 From hand-engineered acoustics to speech foundation models

Methodologically, the field has moved through two generations. The first generation extracted expert-defined acoustic parameter sets — standardized most influentially in the Geneva Minimalistic Acoustic Parameter Set (GeMAPS), which specifies frequency, energy, spectral, and temporal descriptors selected for their theoretical and empirical relation to affect [43]. Hand-engineered features remain valuable in clinical work precisely because they are interpretable: a model driven by pause duration and reduced pitch range can be audited against the psychomotor-slowing literature in a way that an opaque embedding cannot.

The second generation is built on self-supervised speech foundation models. Architectures such as wav2vec 2.0 and HuBERT learn general-purpose speech representations from large unlabeled corpora by masked prediction and contrastive objectives [44, 45], and weakly supervised multilingual models such as Whisper extend robust representation learning across dozens of languages [46]. These embeddings transfer strongly to paralinguistic and clinical tasks in low-data regimes, which is decisive for a language such as Amharic where labeled clinical audio is scarce. Our development plan therefore treats the two generations as complementary: interpretable acoustic-prosodic features provide clinical auditability and cultural face validity, while foundation-model embeddings provide representational power, with the labeled Lenegewa corpus serving as the supervision layer for both.

2.3 Speech emotion recognition and the problem of emotion ground truth

A parallel literature — speech emotion recognition (SER) — supplies the annotation machinery for the present work. SER research established that vocal emotion is legible to human raters and learnable by machines across pitch, intensity, rhythm, and spectral cues [25], and a recent systematic review documents SER's growing but still indirect use in mental-health applications, spanning depression, suicide risk, and psychotic disorders across highly heterogeneous architectures and datasets [47]. The theoretical basis for cross-linguistic transfer rests on partial universality of emotion expression [48], but the qualification matters more than the universality: meta-analytic evidence shows a consistent in-group advantage in emotion recognition, with accuracy declining across cultural distance [49], and contemporary affective science cautions that emotional expression is variable, context-dependent, and not reducible to fixed signatures [50]. For algorithm development the implication is direct: emotion ground truth is not free. Labels produced by culturally distant raters, or by mapping from English-language emotion norms, import systematic error into the supervision signal itself. This is why the labeling methodology in Section 5 treats cultural anchoring as part of the scientific method rather than an annotation afterthought, and why our raters score five voice types independently rather than forcing a single dominant emotion per sample.

2.4 Cultural and linguistic validity: bias, low-resource languages, and the Amharic context

The risk of culturally biased speech technology is not hypothetical. Commercial automated speech recognition systems exhibit large racial disparities, with roughly doubled word error rates for African American speakers, traceable to unrepresentative training audio [51], and biased clinical algorithms have produced systematic under-service of Black patients at population scale [52]. A voice-severity model trained on culturally naive labels in an Ethiopian population would risk an analogous failure with higher stakes: misclassifying trauma-related shutdown, restraint, or dignity-preserving composure as low severity — reproducing the defensiveness gap in algorithmic form — or, conversely, pathologizing dialect, regional pronunciation, or culturally normative expressive restraint.

Amharic-specific resources are emerging but remain thin. The Amharic Speech Emotion Dataset (ASED) — to our knowledge the first SER dataset for the language — comprises 2,474 acted samples from 65 speakers across four dialects, labeled for five emotions with high inter-judge agreement, and demonstrates that standard SER architectures perform comparably on Amharic and on English and German benchmarks [53]. ASED is an important proof of feasibility, but acted emotion from healthy volunteers is a different construct from clinical distress in trauma-affected women; the transition from acted-emotion benchmarks to clinical-severity labeling on naturalistic patient audio is precisely the gap the Lenegewa corpus and labeling protocol are designed to fill. No existing public resource pairs Amharic clinical audio with standardized psychiatric measures, item-level safety indicators, and longitudinal outcomes; to our knowledge, the resource described in Section 3 is the first of its kind.

2.5 Prior work by this group: from pediatric protocols to African deployment

The present methodology consolidates lessons from three prior publications by members of this team. The first specified a protocol for developing an automated speech-analysis algorithm to detect clinically relevant emotional distress and functional impairment in children and adolescents, establishing the core design commitments carried forward here: clinician-anchored ground truth, ordinal severity targets rather than binary diagnosis, and separation between label assignment and validation instruments [26]. The second reported a pilot study objectively quantifying pediatric psychiatric severity using voice recognition and universal-emotion framing in a predominantly low-income, minority youth population, demonstrating that severity-relevant signal can be extracted from brief speech samples collected in routine care, while also surfacing the limits of universal-emotion assumptions across cultural groups [28]. The third analyzed the implementation of voice AI as measurement-based care (MBC) to address behavioral-health treatment disparities among youth in

economically disadvantaged communities, framing the voice signal not as a diagnostic oracle but as an MBC instrument that increases the frequency and objectivity of severity measurement where clinician time is scarce [27]. A fourth publication documented the AI-enabled, trauma-informed rehabilitation model at Lenegewa itself, within which the present data resource was generated [30].

2.6 The methodological gaps this paper addresses

Across the reviewed literature, four methodological weaknesses recur, and this paper's process model is constructed to answer each. (1) Circular validation: models trained on labels derived from the same instruments later used as validation targets inflate apparent performance and inherit instrument blind spots; we therefore enforce strict separation between voice-label assignment and clinical scales (Sections 5–6). (2) Data leakage: repeated samples from the same speaker distributed across training and test partitions overstate accuracy, a failure now recognized as a leading driver of the reproducibility crisis in machine-learning-based science [54]; we mandate client-level splits. (3) Neglect of data work: annotation quality is the least glamorous and most consequential layer of high-stakes AI, where undocumented labeling shortcuts cascade into downstream harm [55]; we treat the labeling protocol, rater training, reliability rules, and label-lock documentation as first-class scientific artifacts. (4) Accuracy monoculture: discrimination metrics alone are insufficient for clinical prediction — calibration is routinely neglected [56], generalizability is routinely overclaimed [42], and reporting is routinely incomplete [57]; we pre-specify a five-family endpoint matrix including safety sensitivity, cultural false-positive review, clinical usefulness, and frontline viability (Section 7), with staged clinical evaluation to follow DECIDE-AI guidance [58]. These safeguards are not presented as an unexecuted protocol: target construction, cultural annotation, usability gating, locking and analysis of the current snapshot, corpus characterization, and reliability analysis are completed research outputs, while the pre-specified model-validity architecture is the registered standard against which later performance will be evaluated. Together, these four gaps define the research question stated in Section 1.4: the field lacks quantitative evidence on whether culturally anchored severity annotation in an African language is reliable enough to supervise machine learning at all, and, to our knowledge, no prior study reports inter-rater reliability for perceptual psychiatric-severity annotation of Amharic speech.

3. Materials and Methods: The Multimodal Lenegewa Data Resource

3.1 Population and observation window

The data resource derives from routine service delivery at the Lenegewa Women's Rehabilitation and Skills Development Center, a high-need rehabilitation setting serving Ethiopian women, many with complex trauma histories, within the AI-enabled, trauma-informed programmatic model previously described [30]. The analyzed beneficiary-linked dataset covers January 14, 2025, through April 23, 2026. Completed standardized clinical screens span January 14, 2025, through April 8, 2026, and supported voice-sample records span January 15, 2025, through April 8, 2026. After excluding 10 obvious test or dummy records, the analytic dataset identifies 3,617 unique supported beneficiaries as the conservative beneficiary denominator. The population is 100% female.

3.2 Five data layers

The resource comprises five distinct, linkable layers (Table 1; Figure 1). The person layer contains the 3,617 supported beneficiaries. The voice layer contains 4,163 supported voice-sample records from 2,600 clients; repeat samples are counted as usage and measurement records, not additional beneficiaries. The survey layer contains 19,610 standardized survey records across the PHQ-9 [16], GAD-7 [17], ACE questionnaire [59], CAGE [60], and Severity of Dependence Scale [61], instruments with established East African validity evidence where available [14, 15]. The item layer contains 111,059 item-level symptom responses, preserving the fine-grained safety and symptom information — including PHQ-9 item 9 — that total scores discard. The documentation layer contains 67,274 consent, signature, clinical, service, risk-indicator, note, and program records that situate

every measurement in an auditable care context. In aggregate, the resource contains 205,723 beneficiary-linked program documentation records.

Table 1. Multimodal data layers supporting voice-algorithm development.

Data layer	Aggregate scale	Role in algorithm development
Supported beneficiaries	3,617	Person-level denominator after excluding 10 obvious test/dummy records.
Voice samples	4,163 records from 2,600 clients	Primary audio source for future model training and testing.
Standardized survey records	19,610	PHQ-9, GAD-7, ACE, CAGE, SDS, and related survey events.
Item-level symptom responses	111,059	Fine-grained symptom, safety, and response-level validation data.
Clinical/service/risk documentation	67,274	Consent, signature, clinical, service, risk-indicator, note, and program records.
Total beneficiary-linked documentation	205,723	Aggregate evidence base supporting algorithm viability and auditability.

Counts are aggregate and fully de-identified; no record-level data are presented.

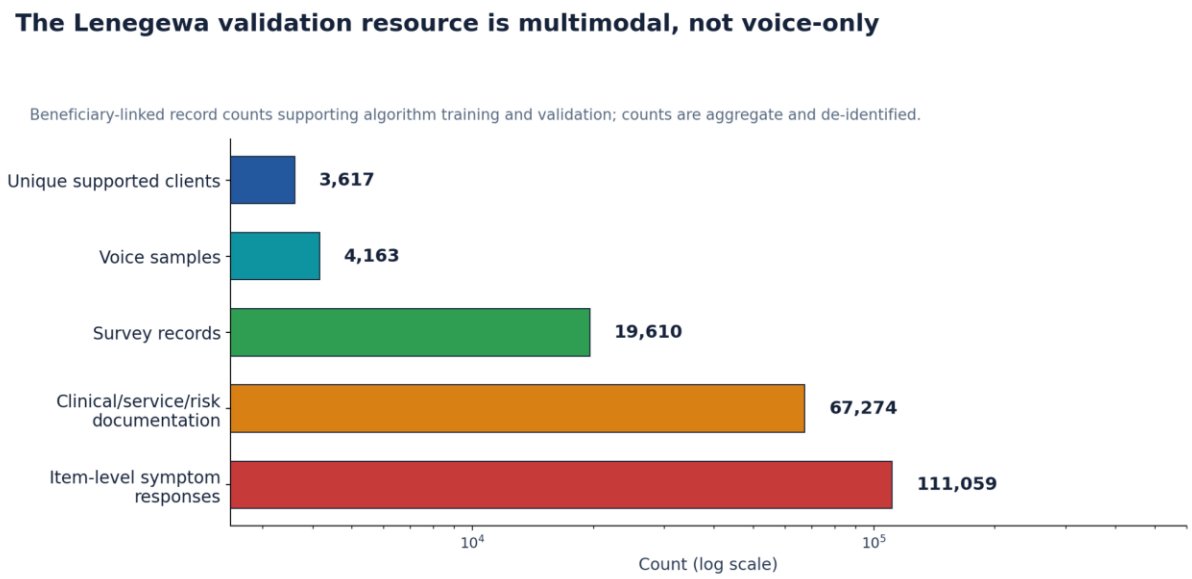


Figure 1. The Lenegewa validation resource is multimodal, not voice-only.

Log-scale chart of beneficiary-linked record counts. The figure intentionally separates people, voice samples, surveys, item responses, and support documentation.

3.3 What multimodality buys

Each layer answers a different question that a voice algorithm must eventually face. The survey layer supports convergent validity against conventional instruments. The item layer supports safety-sensitivity analysis at the level where risk actually lives — individual items and flags — rather than at the level of totals that the linked cohort evidence in Table 3 shows can conceal it. The documentation layer supports viability analysis: whether a signal can be generated, reviewed, and acted upon inside real workflows, the deployment questions emphasized in our MBC implementation work [27] and in the digital-mental-health literature for low- and middle-income countries generally [22]. Repeated surveys and repeated voice samples jointly support change-sensitivity

analysis, echoing the longitudinal use of vocal indicators in treatment monitoring [32]. No single-layer corpus — audio with diagnosis labels, or audio with one questionnaire — can support this evaluation program; the multimodality is the resource's scientific value. Crucially, these layers are not reserved for after-the-fact validation alone: during algorithm development, the locked voice labels are triangulated with survey totals, item-level safety indicators, discordance profiles, repeated outcomes, and documentation, so that the model is built against the full multimodal record rather than the voice layer in isolation (Section 6).

3.4 Voice-sample acquisition protocol

Two task types are used, with recordings of approximately 30–45 seconds per sample. All data — surveys and voice alike — are captured through ClarityConnect, the program's clinical data-collection platform, by AYA clinicians (the clinical guides referenced throughout) using Android tablets or laptops, in Amharic; there is no separate, study-specific voice-collection pathway — every sample was produced inside routine clinical workflow, at one of two occasions: the admission process or a treatment-response progress check. The first task is **structured read speech**: clients read the same standard paragraph, making that subset content-constant by construction — semantic content carries no information about the speaker, and any severity signal recoverable from read samples is vocal by definition. Read-speech collection was discontinued early, after approximately 250 samples, because clients — most of whom have limited formal education — reported feeling judged on their reading ability; the task was withdrawn as a matter of client dignity, and the episode is itself direct evidence for the literacy argument of Section 1.2: even a reading task designed only to capture voice proved to be the wrong channel for this population. The read subset is therefore small and non-random (early-enrollment clients only), and every analysis that uses it is scoped and labeled exploratory. All remaining collection uses the second task; a **semi-structured narrative prompt** administered identically to every client. At admission, the standardized prompt is: "Please tell me your name, your age, the sub-city and woreda where you live, the highest grade you completed in school, the type of work you did previously, and the type of training you would like to receive."

At the treatment-response progress check, the prompt asks how she has been doing at Lenegewa. The prompt is read here in English translation. Two properties of this prompt are named at once because later sections depend on them: it elicits direct identifiers—name, age, and sub-city and woreda of residence—so identifying speech is present in most samples by design (Section 5.8); and it elicits education and work history, so transcripts carry socioeconomic information, which the text-only baseline of Section 7 exists to control (Section 5.9). Recording takes place in a standard private clinical room that is not acoustically isolated, so ambient clinic noise is part of the training distribution. Recordings were made within the same ClarityConnect session as the survey battery, at the end of survey collection. The resource contains 4,163 linked supported voice-sample records and 4,182 raw audio files; duration metadata were available for 4,171 files (wall-clock 33.45 h; effective voiced duration 11.31 h; median 26.4 s, IQR 19.4–36.4). These operational denominators describe different layers and are not used interchangeably. The current annotation calculations use the locked 2,479-sample snapshot described in Section 5.7.

Three consequences of this protocol are design facts of the resource, named here rather than left for a reviewer to find. First, the survey battery is a standardized emotional prime, adopted deliberately. Every voice sample was recorded moments after the client worked through childhood-abuse items, dependence items, and PHQ-9 item 9; the voice signal therefore encodes distress as elicited by trauma screening, not distress at rest. A standardized stressor administered identically to every client is what a measurement instrument wants: the screening battery functions as a uniform elicitation context, held constant across all samples. This also strengthens the residual formulation of Section 6.3—because the audio is captured while the client is still in the defended state she occupied while completing the form, voice-predicted and observed PHQ-9 are contemporaneous measurements of the same moment, not sequential ones. Second, the protocol creates a train/deploy distribution shift: deployment contexts without the survey battery remove the prime, and the

acoustic distribution shifts accordingly. This is stated as an explicit limitation in Section 9, and the general-population cohort of Section 7 randomizes a sub-arm to record before the surveys, yielding a direct estimate of the prime's acoustic effect size before any field deployment. Third, guides were not blind to survey scores at the point of recording. The possible leakage path is therefore named plainly: a clinician who has just administered the battery knows the client's scores and can, without intending to, shape the recording moment—tone, patience, permitted pauses, or the space given to continue—in score-correlated ways. Any by-guide analysis will be undertaken only after full-corpus label completion and metadata reconciliation in the follow-up study.

Because the narrative prompts are semi-structured, personal semantic content is present in those samples and raters hear it. This is the sharpest threat to the labeling protocol's organizing rule — label the voice, not the story (Section 5.1) — and the labeling protocol addresses it directly through the cultural rating rules of Section 5.3, which train raters to score vocal quality rather than narrative content. The small early read-speech subset (~250 samples), whose transcripts are identical across clients, offers an exploratory check that recovered severity signal is vocal rather than lexical.

Protocol element	Specification
Task	Structured read speech (same standard paragraph; content-constant), discontinued after ~250 samples early in collection for client-dignity reasons; semi-structured standardized prompt for all remaining samples.
Prompt	Admission (standardized, identical for all clients): 'Please tell me your name, your age, the sub-city and woreda where you live, the highest grade you completed in school, the type of work you did previously, and the type of training you would like to receive.' Progress check: how she has been doing at Lenegewa. The prompt elicits direct identifiers by design (Section 5.8).
Duration	Approximately 30–45 s per sample. Raw inventory: 4,182 files; duration metadata: 4,171 files; wall-clock 33.45 h; effective voiced duration 11.31 h; median 26.4 s (IQR 19.4–36.4).
Administration	AYA clinicians via the ClarityConnect platform, in Amharic, within the same session as the survey battery; occasions: admission or treatment-response progress check.
Order relative to surveys	Recording follows the full battery (PHQ-9, GAD-7, ACE, CAGE, SDS); the battery is a uniform elicitation context.
Guide blinding	Guides were not blind to survey scores at recording. Possible guide-related elicitation effects are pre-specified for later analysis after full-corpus label completion and metadata reconciliation.
Device / format / environment	ClarityConnect on Android tablets or laptops; standard private clinical room, not acoustically isolated.
Minimum voiced duration	A pre-specified minimum voiced-duration threshold will gate feature extraction in the follow-up algorithm study (Section 6.2).

Table 2. Voice-sample acquisition protocol.

3.5 Linkage and analytic attrition

A data resource is only reviewable if each denominator is tied to its operational layer. Figure 2 therefore distinguishes 3,617 supported beneficiaries; 2,600 clients linked to 4,163 supported voice-sample records; 4,182 raw audio files, of which 4,171 had duration metadata; 7,919 rating attempts, including 833 marked unintelligible and excluded; 7,086 usable ratings from three raters; 2,479 samples with a derived EDI severity band; and 2,267 samples with complete three-rater EDI matrices. These counts describe linked but different layers and are not a single subtraction-based attrition sequence. Labeling of the broader corpus is ongoing and will be analyzed only after full-corpus label lock in the follow-up algorithm paper.

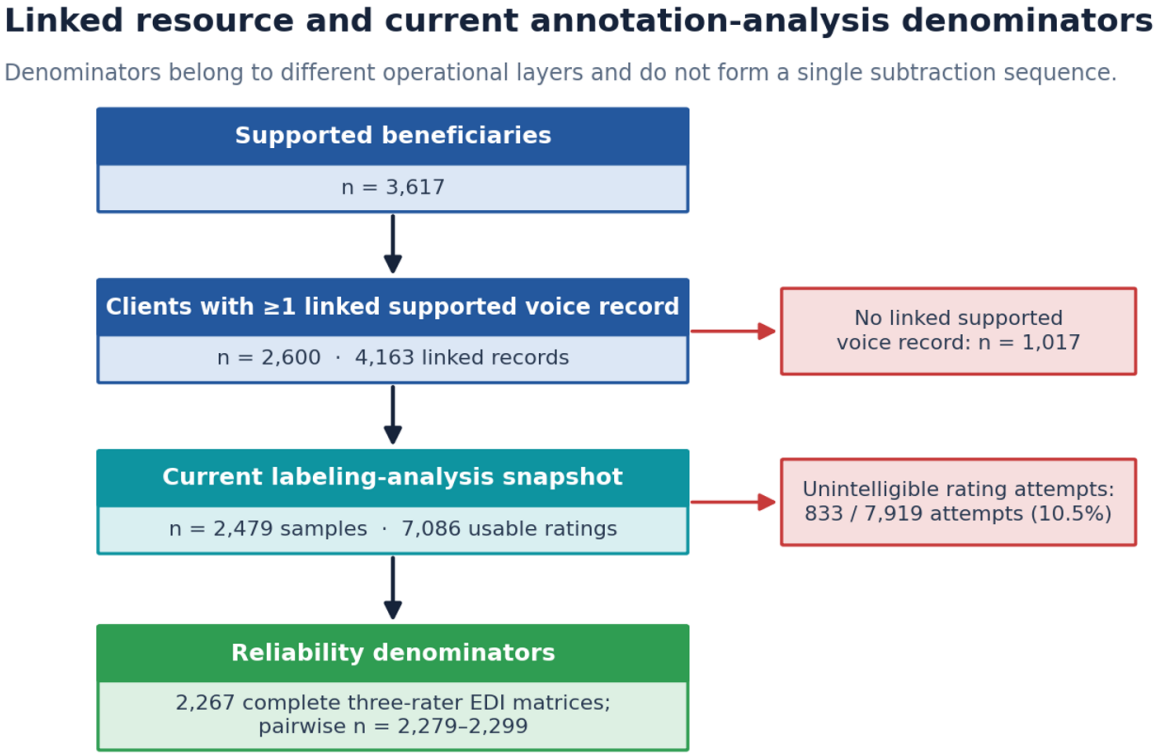


Figure 2. Linked resource and current annotation-analysis denominators.

Final client-level eligibility and endpoint-specific intersections will be enumerated only in the follow-up algorithm paper after full-corpus label completion.

4. Independent Clinical Validation Targets: Discordant Risk

The linked Lenegewa cohort evidence summarized in Table 3 makes the central validation requirement explicit: a voice model must not be evaluated solely by whether it reproduces PHQ-9 or GAD-7 total scores. If the target population contains systematic low disclosure — driven by cultural restraint, shame, trauma shutdown, fear of judgment, or gendered self-protection — then a model that perfectly predicts self-report totals has learned, with high fidelity, the same blind spot the totals contain. In machine-learning terms, the reference standard is corrupted for exactly the minority of cases that matter most; optimizing agreement with it optimizes the error. This is the voice-AI instantiation of a general clinical-AI hazard: a model is only as good as the labels and comparators it is held to, and biased targets produce confidently biased systems [52].

Table 3 summarizes descriptive findings from the linked Lenegewa records as four validation lenses. The general severity lens establishes the base-rate problem: only 296 of 2,618 paired baseline clients met conventional mood/anxiety elevation. The high-risk discordance lens establishes the central use case: the majority of clients with high ACE, substance-use, suicide-risk, or sexual-abuse indicators fell below conventional PHQ/GAD

thresholds. The safety lens establishes that item-level and program-identified risk must be reviewed independently of total-score severity. The outcome lens establishes that repeated screens registered improvement signals, supporting future evaluation of voice-based change sensitivity. A voice algorithm developed against this resource can therefore be asked a sharper question than agreement: does the voice channel add signal precisely where self-report goes quiet?

Table 3. Descriptive cohort evidence motivating independent voice-model validation.

Lens	What it measures	Why it matters for voice validation
General severity lens	PHQ-9 or GAD-7 total ≥ 10	Only 296 of 2,618 paired baseline clients met conventional mood/anxiety elevation.
High-risk discordance lens	High ACE, substance-use, suicide-risk, or sexual-abuse indicator with lower PHQ/GAD totals	The linked cohort records show that most high-risk clients were below conventional PHQ/GAD thresholds.
Safety lens	PHQ-9 item 9, suicide-risk, and sexual-abuse indicators	Item-level and program-identified safety risk must be reviewed independently of total-score severity.
Outcome lens	Repeated PHQ-9, GAD-7, CAGE, and SDS change	Repeated screens showed improvement signals, supporting future evaluation of voice-based change sensitivity.

These descriptive cohort findings motivate the independent validation design; they are not voice-model performance results.

5. Materials and Methods: Culturally Anchored Amharic Voice Labeling

5.1 Unit of labeling, ordinal target, and separation from clinical scores

The labeling methodology converts audio into supervised training data and is designed around a single organizing rule communicated to every rater: label the voice sample—not the diagnosis, trauma history, or clinical score. The unit of labeling is one voice sample or audio segment, never the person's full clinical record. The primary direct output is the 0–39 Emotional Distress Index (EDI), a holistic ordinal-continuous judgment of voice-indicated transdiagnostic distress. For reporting and for the pre-specified ordinal Head A, sample-level EDI is mapped to Low (0–9), Moderate (10–19), High (20–29), or Severe (30–39). These bands are derived from EDI; raters did not assign a separate categorical overall-severity label. The resulting ordinal target matches the algorithm's intended function—indicating severity and prompting provider review, not diagnosing trauma, depression, anxiety, PTSD, or suicide risk.

Validation separation is enforced at the source: clinical scales, diagnoses, and safety flags are excluded from label assignment and stored separately for later model validation. Raters are explicitly instructed not to upgrade severity based on known diagnosis, prior history, or scale scores, and not to infer severity from story content alone — a severe story can be told with regulated affect, and regulated content can be carried by a dysregulated voice. Content may contextualize the voice, but the rating object is vocal quality. This separation-with-linkage design prevents the circularity failure identified in Section 2.6 while preserving the linkage needed for multimodal validation.

5.2 Five voice types, five listening elements

Alongside the direct EDI rating, raters score five voice types—fear, anger, sadness, neutral, and happiness—independently on a 0–4 ordinal intensity scale. These five dimensions are not an arbitrary taxonomy: they operationalize the universal-emotions framework [48] that structured this group's prior published voice work, in which the pediatric voice-analysis protocol and its subsequent pilot were explicitly built around universal

emotion categories as the affective scaffolding for severity detection [26, 28]. Retaining the same five voice types preserves construct continuity and cross-study comparability across the developmental arc from United States pediatric populations to Amharic-speaking Ethiopian adults, while the surrounding protocol deliberately relocates those dimensions inside Amharic-specific cultural rating rules—reflecting the in-group-advantage evidence that emotion recognition is culturally patterned [49] and the constructionist critique that affective expression is mixed and contextual rather than discretely signaled [50]. Raters do not force a single dominant voice type; uncertainty is represented through the recorded confidence field and evidence-based rationale rather than through a default category.

For every voice type, raters evaluate five elements: (1) emotional tone — the feeling the voice conveys; (2) vocal energy and intensity — force, volume, vitality, or depletion; (3) rhythm, pauses, and speech flow — smooth, fragmented, halting, rushed, or silence-laden; (4) organization and ability to continue — whether distress disrupts coherent speech and task completion; and (5) Amharic/Ethiopian cultural expression — restraint, dignity, defensiveness, shame, indirect expression, and social pressure. Elements (1)–(4) map onto the acoustic-prosodic feature families the algorithm will consume [33, 43]; element (5) is the cultural calibration layer that machine features cannot supply.

The annotation instrument produces two complementary direct outputs. First, each rater assigns a 0–39 EDI score as a holistic judgment of voice-indicated distress. Second, each rater scores fear, anger, sadness, neutral, and happiness separately on a 0–4 ordinal scale. The EDI is not a sum of the five voice-type ratings and is not an external clinical instrument. Overall-severity bands—Low 0–9, Moderate 10–19, High 20–29, and Severe 30–39—are derived from the EDI after sample-level aggregation; they are not independently rated. This distinction preserves a direct continuous target for Head B, a derived ordinal target for Head A, and complementary voice-type supervision while keeping every annotation target separate from the clinical scales reserved for validation.

An audio-usability gate precedes analysis. Rating attempts marked unintelligible are excluded; the current tool does not collect separate Clear, Adequate, and Poor tiers. Usability is therefore reported at the rating-attempt level, while confidence remains a separate 1–5 field.

5.3 Ethiopian cultural adaptation is methodological, not cosmetic

The cultural protocol operates in both directions, guarding simultaneously against false negatives and false positives. Against false negatives, raters are trained that trauma presentations include both hyperarousal and hypoarousal: severe distress may appear as panic, anger, or grief, but equally as psychomotor slowing, emotional numbing, dissociation-like detachment, fragmented speech, or marked cognitive slowing — the shutdown presentations documented in the dissociative-subtype literature [10] and in complex trauma generally [9]. Raters therefore must not treat flat, monotone, very slow, hesitant, or detached speech as Low severity when those features represent trauma-related shutdown, shame, grief, or defensive self-protection. In Ethiopian women with complex trauma, High severity may sound quiet, slowed, guarded, or emotionally constricted rather than loud or dramatic. In machine-learning terms, these rules intervene at the point where the supervisory signal is constructed; they are therefore part of the model's measurement methodology, not a post-hoc fairness adjustment.

Against false positives, raters must not interpret soft speech, indirect expression, respectful restraint, limited emotional display, or cultural modesty as pathology, and must not confuse dialect, regional pronunciation, accent, pacing differences, normal reading difficulty, limited educational exposure, or culturally appropriate directness with distress. These rules operationalize the idioms-of-distress literature — suffering in Ethiopian and comparable contexts is routed through culturally specific channels of expression and containment [11, 12, 13] — and they answer the in-group-advantage evidence directly by requiring that Amharic speech be judged against Amharic and Ethiopian norms, not English emotion norms [49].

The protocol further specifies culturally informed guidance for each voice type. Fear may present as constrained breathing, cautious tone, over-controlled phrasing, reluctance to elaborate, or an incongruence between composed words and a constricted, guarded vocal quality — self-protection rather than panic — and may hide behind false pride, exaggerated toughness, or insistence that everything is fine. Anger, which many Ethiopian women have learned to suppress in male-dominated family, social, religious, and institutional settings, may surface as restrained intensity, sharpness, clipped responses, sarcasm, indirect protest, or controlled insistence on dignity rather than raised volume; raters are warned equally against missing polite anger and against over-labeling assertiveness or culturally appropriate directness. Sadness may present as slow rhythm, long pauses, reduced energy, sighing, monotone quality, or a burdened, depleted voice, without tears and in a controlled, dignified register. Neutral is a positive finding, not a residual category: it requires genuine emotional regulation — even, steady, flexible, organized speech — and must be actively distinguished from suppression, shutdown, exhaustion, guardedness, and culturally shaped defensiveness. Happiness requires distinguishing genuine warmth, relief, and connection from defensive positivity — performed strength, gratitude, or hopefulness used to reduce shame, please the interviewer, protect dignity, or signal resilience while distressed; laughter, politeness, gratitude, and saying one is fine are explicitly insufficient grounds for a Happiness rating. Table 4 summarizes the complete labeling architecture described in this section: the direct 0–39 EDI rating, EDI-derived severity bands, the intended three-rater structure, five 0–4 voice types and five listening elements, the unintelligible-audio gate, the cultural-context rules, the recorded confidence and rationale, and the validation separation that keeps clinical scales, diagnoses, and safety flags out of label assignment.

Table 4. Voice-labeling architecture for Amharic psychiatric severity modeling.

Labeling component	Operational definition	Algorithm-development purpose
Primary continuous target	Direct 0–39 EDI judgment by each rater	Supervision for Head B; constructed from voice alone.
Derived severity bands	Low 0–9; Moderate 10–19; High 20–29; Severe 30–39, derived from sample-level EDI	Ordinal target for Head A; not a separately assigned rating.
Rater structure	Up to three independent Ethiopian raters; 2,267 complete three-rater EDI matrices in the current snapshot	Supports complete-case ICC plus pairwise and available-rating reliability analyses.
Voice-type ratings	Fear, anger, sadness, neutral, happiness—each scored independently from 0 to 4	Complementary multi-task supervision without forcing one dominant emotion.
Five listening elements	Emotional tone; vocal energy/intensity; rhythm, pauses, and flow; organization and ability to continue; Amharic/Ethiopian cultural expression	Anchors judgments to auditable vocal evidence.
Audio usability gate	Unintelligible rating attempts are excluded	Separates usability from confidence; no Clear/Adequate/Poor tiers are collected in the current tool.
Cultural context rules	Amharic speech norms, restraint, idioms, gendered expression, regional variation; bidirectional error guards	Reduces culturally naive false-negative and false-positive labels.
Confidence and rationale	Confidence scored 1–5 plus brief evidence-based rationale when recorded	Supports calibration, quality audit, and later explanation-alignment analysis.
Validation separation	Clinical scales, diagnoses, safety indicators, and outcomes excluded from annotation	Prevents circularity and preserves independent multimodal validation.

Labeling component	Operational definition	Algorithm-development purpose
Rater monitoring	Agreement, pairwise differences, leniency, and coverage tracked with explicit denominators	Treats the annotation process as a measured scientific instrument.

Summarized from the program's written Amharic voice-labeling protocol and rater training materials (May 2026).

5.4 Rating rules, decision structure, and adjudication

The EDI decision follows a fixed sequence. First, the audio-usability gate: an unintelligible attempt is excluded. Second, raters listen to the whole sample and assign a holistic EDI score from 0 to 39, using the Low (0–9), Moderate (10–19), High (20–29), and Severe (30–39) ranges as interpretive anchors. Third, disruption informs placement within that continuum: absent or mild transient distress with organized speech supports the lower range; audible disruption of regulation, attention, speech organization, or task completion supports higher values; and profound shutdown, panic-like activation, or inability to continue supports the upper range and urgent human review. Fourth, fear, anger, sadness, neutral, and happiness are each rated independently from 0 to 4. Finally, confidence and a brief evidence-based rationale are recorded. These rules structure judgment without converting clinical scales, diagnoses, or story content into annotation targets.

The protocol aims for three independent Ethiopian raters working blind to one another's scores. Whole-sample listening is required because affective quality early in a sample may differ from later portions. The current snapshot is not fully crossed: usable assignment counts were 2,450 for Rater 3, 2,326 for Rater 2, and 2,310 for Rater 1, and 2,267 samples formed a complete three-rater EDI matrix. Reliability is therefore reported with explicit denominators and with pairwise as well as complete-case statistics [62, 63, 64, 65]. Rater performance is treated as a measured quantity through agreement, leniency, and difference analyses; systematic drift can trigger recalibration or retraining. This treats the label set as a scientific instrument whose reliability must be quantified before model training, rather than assumed [55].

The locked analysis snapshot contains 2,479 samples and 7,086 usable ratings from three raters. The 2,267-sample complete matrix supports ICC estimation; available-rater and pairwise denominators are stated for the other analyses. Each usable assignment can contain a direct 0–39 EDI rating and five 0–4 voice-type ratings, with small item-level denominator differences documented in Table 5. Sample-level EDI supports the derived four-band severity distribution. Per-rater observations are retained alongside sample-level aggregates. These annotation-reliability results do not weight or exclude samples and are not model-performance estimates.

5.5 Rater governance, ethics, and welfare

Governance requirements are integral to the method. Raters work exclusively with access-controlled audio and are instructed to exclude names, phone numbers, precise addresses, and unnecessary identifying narrative details from rationale notes. The label is not a diagnosis and is never returned to participants as a stand-alone clinical determination. Rater teams include Ethiopian clinicians or trained local raters familiar with the language, idioms of distress, dialect, gender norms, and social context of the population—a requirement that follows directly from the in-group-advantage evidence [49] and from the documented failure of speech systems built without population-matched expertise [51]. Because repeated exposure to trauma narratives can produce secondary traumatic stress, rater support is a standing program component. Provenance covariates will be reconciled and analyzed only after full-corpus label completion; no current provenance result is reported.

5.6 Pilot and locked labeling-analysis snapshot

The current labeling-analysis snapshot is locked for the analyses reported here: 2,479 samples, 7,086 usable ratings from three Ethiopian raters, and 2,267 complete three-rater EDI matrices. An earlier pilot exercise (46 individuals; 153 voice-sample rows; 1,980 extracted long-format label rows) validated the schema. Full-corpus

labeling remains ongoing. Table 5 reports only empirical annotation results supported by the current snapshot; Section 3 and Table 2 separately report the raw-file duration inventory. Algorithm fitting and performance evaluation will begin only after full-corpus label completion and lock.

5.7 Results: The locked labeling-analysis snapshot

The present article completes the analysis of the locked 2,479-sample labeling snapshot (Figure 3). Table 5 reports sample-level EDI and derived severity distributions, rating-level 0–4 voice-type marginals, confidence, rater coverage, inter-rater reliability, rater leniency, and pairwise difference analyses. Every denominator is stated because the snapshot contains 7,086 usable ratings but only 2,267 complete three-rater EDI matrices. Clinical scales, safety indicators, diagnoses, discordance profiles, outcomes, and documentation remain outside annotation and are reserved for independent validation. Full-corpus labeling and every algorithm-performance analysis remain outside this article.

Table 5. Descriptive characteristics and reliability of the locked 2,479-sample labeling-analysis snapshot.

Characteristic	Quantity	Value
Analysis coverage	Samples; rating attempts; usable ratings; raters	2,479 samples; 7,919 attempts; 7,086 usable ratings; 3 raters.
EDI-derived severity distribution	n (%) of 2,479 samples	Low 720 (29.0%); Moderate 1,446 (58.3%); High 300 (12.1%); Severe 13 (0.5%). Bands are derived from EDI, not separately rated.
0–39 EDI distribution	Sample-level median [Q1, Q3]	13.0 [9.0, 17.0], n=2,479.
Rating-attempt usability	Unintelligible exclusion flag	Usable 7,086; excluded as unintelligible 833/7,919 attempts (10.5%).
Voice-type marginals	Counts at levels 0/1/2/3/4	Fear 1,499/2,991/2,332/260/3 (n=7,085); anger 6,549/413/108/14/0 (n=7,084); neutral 6,364/719/1/0/0 (n=7,084); sadness 2,970/1,966/1,757/376/15 (n=7,084); happiness 6,984/83/15/2/0 (n=7,084).
Per-rater usable counts	Usable ratings	Rater 3 2,450; Rater 2 2,326; Rater 1 2,310; maximum-minus-minimum difference 140.
Ordinal Krippendorff's alpha	Coefficient [95% CI]	EDI-derived severity bands 0.46 [0.43, 0.48]; fear 0.34 [0.32, 0.37]; anger 0.28 [0.23, 0.33]; neutral 0.36 [0.32, 0.41]; sadness 0.51 [0.48, 0.53]; happiness 0.34 [0.22, 0.43].
Pairwise quadratic-weighted kappa	Coefficient and pairwise n	Mean 0.47; Rater 2–Rater 1 0.49 (n=2,279); Rater 2–Rater 3 0.47 (n=2,299); Rater 1–Rater 3 0.45 (n=2,294).
EDI intraclass correlation	Complete matrix: 2,267 samples × 3 raters	ICC(2,k), three-rater mean: 0.78 [0.76, 0.80]; ICC(2,1), single rating: 0.54 [0.52, 0.57].

Characteristic	Quantity	Value
EDI standard error of measurement	Calculated from unrounded pooled SD and ICC(2,1); displayed inputs round to 7.11 and 0.54	4.80 EDI points.
Rater leniency	EDI mean (SD), rounded; spread calculated from unrounded means	Rater 3 13.5 (7.5); Rater 2 13.5 (6.5); Rater 1 12.4 (7.3); spread of means 1.12.
Pairwise Bland–Altman	Bias; 95% limits of agreement	Rater 2–Rater 1: 1.2, −11.1 to 13.4; Rater 2–Rater 3: −0.2, −13.5 to 13.1; Rater 1–Rater 3: −1.3, −15.3 to 12.6.
Confidence distribution	Levels 1/2/3/4/5; median [IQR]	21/127/663/5,242/524; median 4.0 [4.0, 4.0], n=6,577 ratings.

Table 5 reports the available reliability coefficients as empirical annotation results with explicit denominators. Agreement is modest for EDI-derived severity bands (Krippendorff's alpha 0.46, 95% CI 0.43–0.48) and ranges from 0.28 to 0.51 across the five 0–4 voice types. Mean pairwise quadratic-weighted kappa is 0.47. For the 2,267-sample complete EDI matrix, the three-rater mean is more reliable (ICC(2,k)=0.78, 95% CI 0.76–0.80) than a single rating (ICC(2,1)=0.54, 95% CI 0.52–0.57), with SEM 4.80 points. Rater means differ by 1.12 EDI points, while the Bland–Altman limits show substantial sample-level disagreement. Confidence is concentrated at level 4 (median 4, IQR 4–4). Together, these findings support aggregation while making annotation uncertainty visible; none is a model-performance estimate.

5.8 Data governance, consent scope, and benefit sharing

Voice is a biometric identifier, and short unscripted utterances are re-identifiable; "de-identified audio" is therefore a claim that must be given operational content rather than asserted [66]. The exposure here is not incidental but structural: the standardized prompt itself elicits direct identifiers — the client's name, her age, and the sub-city and woreda where she lives — so identifying speech is present in most samples by design, and "de-identified audio" can only be made true by the identifier-handling controls below. Unscripted speech about work history and education additionally embeds socioeconomic covariates in every transcript. The small read-speech subset, by contrast, carries no personal content, and the exposure analysis distinguishes the two subsets. De-identification here rests on two operational controls: the key linking audio to client identity is held separately from the audio under restricted access, and the audio itself is kept in access-controlled storage. Because the standardized prompt elicits spoken name, age, and sub-city and woreda of residence, direct identifiers are present in the audio signal itself; the compensating controls, not any claim of anonymized content, carry the protective weight, and those spoken identifiers are left in the audio rather than redacted or flagged, so protection rests entirely on the access controls above — the separately held linkage key and access-controlled storage — not on any alteration of the recording itself. Voice is not anonymizable in the strong sense, and the compensating controls, not the label "de-identified," carry the protective weight.

The audio is stored within the ClarityConnect platform on secure cloud infrastructure, with access strictly controlled by role and responsibility. The consent instrument administered within the program's data-governance framework covered the use of de-identified clinical information and placed no restriction on the use of de-identified data for research or product development; it did not, however, explicitly name algorithm development, commercial or otherwise. That gap is stated here rather than smoothed over, and the question is asked deliberately, because three of four authors are affiliated with the company developing a commercial product from this resource, and a conflict-of-interest disclosure alone does not answer it. The position taken is this: development proceeds on de-identified data only, within the unrestricted research-and-product-development scope the consent language permits, under the operational controls of this section; the gap between that scope and an explicit naming of algorithm development is treated as a limitation of the historical consent instrument, not a license.

Consent instruments for all future collection — the general-population calibration cohort of Section 7 and the Phase 3 field test — will name algorithm development, including commercial development, explicitly. No formal benefit-sharing arrangement is currently in place; the authors state this plainly rather than imply a commitment that does not exist. The team is, however, open to arrangements that direct benefit back to the Lenegewa cohort and to similar underserved communities — the populations from whose data the tool is built — and regards such arrangements as the appropriate disposition for a commercial product developed on a vulnerable population's clinical records. The data reported here come from the same center and the same clinical process as the prior manuscripts in this series, governed by the Lenegewa memorandum of understanding and an approved IRB; that approval covers the retrospective analysis of de-identified clinical records on which every result in this paper rests. Prospective activity beyond this corpus — the Phase 3 field test and the general-population calibration cohort of Section 7 — does not proceed under the present approval: consistent with the study's established practice, each new pilot site will be brought on under its own memorandum of understanding, dedicated IRB review, and site-appropriate consent before any prospective data are collected. This is a forward-looking governance process for later stages and does not alter the retrospective target-construction results reported here; no model has been fit or characterized in the present article.

Two data-governance artifacts are designated for supplementary submission: a Datasheet for Datasets and a Data Statement for the speech corpus [67, 68]. For a data-resource paper that cannot release participant audio, these documents provide the auditable description that a public dataset would otherwise supply. The derived annotation layer planned for the public ASED corpus (Section 5.9) is a future releasable artifact; it is not part of the present analysis. All methods were performed in accordance with the relevant guidelines and regulations, including the principles of the Declaration of Helsinki.

5.9 Future-oriented external validation of the labeling instrument

One external-validation study is outlined as future work. It will test the cultural generality of the labeling instrument against an independent public corpus and is executable with the existing rater team. It is described so the validation path is on record; it does not change the pre-specified algorithm-development architecture in Section 6.

ASED as an external check on the labeling instrument. The Amharic Speech Emotion Dataset [53]—acted samples across four dialects and five emotion categories that correspond one-to-one to the five voice types of Section 5.2—is the nearest public linguistic neighbor to this corpus. In the planned validation, the Lenegewa labeling instrument will be applied to a female-speaker subsample of ASED by raters from the Lenegewa pool, blind to the intended emotion, after full-corpus label lock. The analysis will test whether blind raters recover ASED's intended emotions against the five voice types (convergent validity of the universal-emotions inheritance) and whether acted emotion from healthy volunteers is distinguished from psychiatric severity (discriminant validity of the rule that a severe story can be told with regulated affect). The pre-specified hypothesis is that acted emotion will not function as a substitute for clinical-severity labels. Any resulting ASED voice-type annotations will be reported and released with that future validation; no ASED result is claimed here.

6. Pre-Specified Machine-Learning Development and Validity Architecture

6.1 Six stages

Model development proceeds in six pre-specified stages (Figure 3). Stages 1–3 define voice capture, culturally anchored annotation, and label analysis. The present article completes the empirical analysis of the locked 2,479-sample snapshot: 0–39 EDI and 0–4 voice-type ratings are summarized, severity bands are derived from EDI, and reliability is quantified with coverage-specific statistics. Full-corpus labeling and final label lock will precede Stage 4. Stage 4 is the follow-up triangulated algorithm-development study using the complete labeled corpus and the pre-specified voice-severity architecture of Section 6.3, with client-level separation throughout.

Stages 5 and 6 comprise held-out multimodal validation and real-world clinical testing. No model has been fit or evaluated in the present article.

The follow-up development stage will use the fully labeled corpus and the clinical data already collected under the protocol of Section 5; extension to additional general-population cohorts and cross-lingual or cross-cultural comparator models is reserved for future iterations (Section 9). Stage 5 is held-out multimodal validation against the endpoint families of Section 7. Stage 6 is real-world clinical testing, staged under early-clinical-evaluation guidance for AI decision support and reported to prediction-model standards, with a pre-specified correction cycle before deployment [57, 58].

Pre-specification is made verifiable rather than asserted. The full development specification — architecture, label-aggregation rule, loss, splits, endpoint families, decision thresholds, the encoder-selection protocol of Section 6.4, and the sample-size justification of Section 7 — is deposited as a time-stamped registration with OSF Registries; the DOI, registration date, and label-lock date will be listed in the companion algorithm paper. As of the registration date, no model has been fit against the locked labels: the specification recorded here precedes any contact between a model and the label corpus, exploratory or otherwise. Deviations from the registered specification during development will be logged with a date and a reason, and the log will be reported in the companion paper.

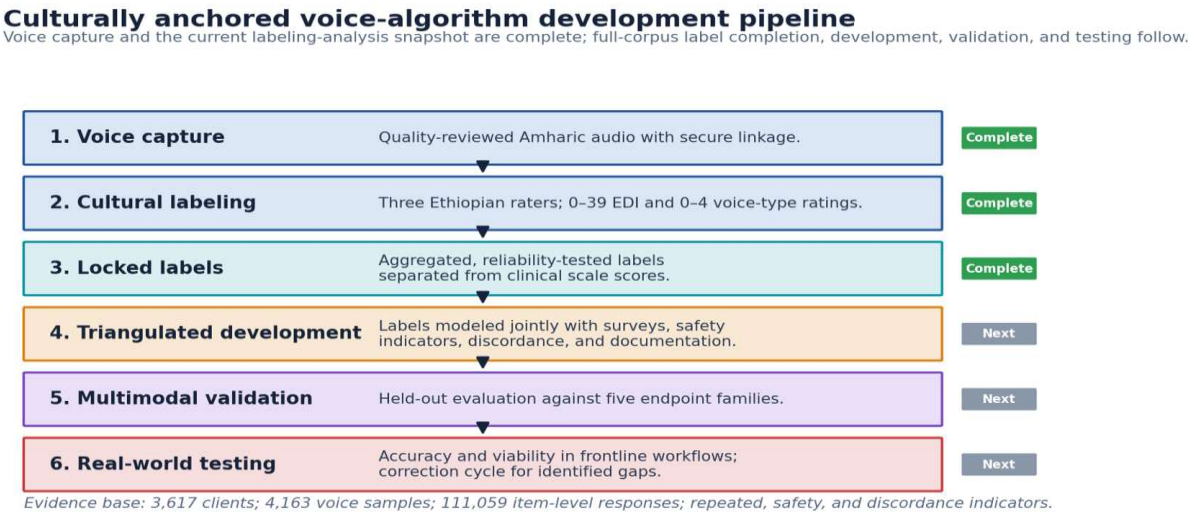


Figure 3. Culturally anchored voice-algorithm development pipeline and the locked labeling-analysis snapshot reported in this article.

The pipeline separates label assignment from clinical-scale validation to protect against circularity and overfitting.

6.2 Pre-specified technical safeguards for machine-learning validity

Table 6 summarizes the design choices that protect model-development validity. The target model is ordinal, respecting the ordered Low–Moderate–High–Severe structure of the label rather than collapsing it to binary classification. Core inputs span both generations of the field reviewed in Section 2.2: interpretable acoustic, prosodic, temporal, and spectral descriptors in the GeMAPS tradition [43] alongside learned embeddings from self-supervised and multilingual speech foundation models [44, 45, 46], so that distress carried in how speech is produced is captured with both auditability and representational power. Required safeguards mandate human review for high-risk, uncertain, discordant, unintelligible, or degraded-audio outputs, preserving the non-diagnostic, provider-mediated use boundary. Fairness checks are pre-specified across cohort, age and sex where available, accent and dialect, audio quality, and literacy or reading difficulty, because subgroup performance gaps in speech systems are the documented default rather than the exception [51, 52].

Three further safeguards are fixed here because each is a documented route to silent invalidation. Augmentation policy: permitted augmentations are channel-only — additive noise, room impulse response, gain, and codec simulation; speed, tempo, and pitch perturbation are excluded, because F0 statistics, speech rate, and pause duration are the label, and the default self-supervised fine-tuning recipe would destroy the target while training without error. Hyperparameter discipline: hyperparameters are selected by nested cross-validation within development folds, or on a fixed inner development fold; selection on the test fold is precisely the leakage failure already cited [54], and the outer test partition is touched exactly once. Minimum voiced duration: at 30–45 seconds of wall-clock audio, effective voiced duration after voice-activity detection can fall below the length at which perturbation measures (jitter, shimmer) and F0-variability statistics are estimable; a pre-specified minimum voiced-duration threshold will gate feature extraction, and the number of samples below it will be reported with the full-corpus algorithm analysis—a second, otherwise silent quality gate.

Table 6. Technical safeguards for model-development validity.

Design choice	Required implementation	Reason
Train/validation/test split	Client-level, not file-level	Prevents repeat-record leakage [54].
Target model	Ordinal severity model	Respects the ordered Low–Moderate–High–Severe target [69].
Core inputs	Acoustic, prosodic, temporal, and spectral features plus learned speech embeddings	Captures distress carried in how speech is produced, with both interpretability and representational power [43, 44, 45].
Required safeguards	Human review for high-risk, uncertain, discordant, unintelligible, or degraded-audio outputs	Maintains non-diagnostic, provider-mediated use [58].
Fairness checks	Cohort; age/sex when available; accent/dialect; audio quality; literacy/reading difficulty	Reduces culturally biased performance claims [51, 52].
Label-lock documentation	Versioned, frozen label corpus with explicit coverage and reliability statistics	Treats annotation as a first-class scientific artifact [55].

Safeguards follow established guidance for validity, leakage prevention, and reporting in clinical machine learning [54, 57, 58].

6.3 Multimodal triangulation with audio-only inference

Figure 4 presents the multimodal voice AI model as it enters development, and its defining commitment is stated first: only audio enters the inference path; surveys never do. A design that feeds PHQ-9, GAD-7, ACE, CAGE, or SDS into the model at inference is rejected because it would put the comparator inside the predictor, break the deployment case, and encourage the model to ignore audio. On the inference path, usability-gated Amharic audio is segmented and encoded through two parallel branches—an interpretable acoustic-prosodic feature set in the GeMAPS tradition [43] and utterance-pooled embeddings from the self-supervised encoder selected under the registered protocol of Section 6.4. The fused representation feeds the unchanged multi-task, audio-only head structure. Head A predicts four-level ordinal severity derived from sample-level EDI, using a rank-consistent ordinal objective (CORN or CORAL) [70, 71]. Head B predicts the aggregate 0–39 EDI. Head C predicts PHQ-9 and GAD-7 totals from voice alone. The five independently scored 0–4 voice types provide

complementary supervision. EDI, its derived severity bands, and voice-type ratings are related but explicitly distinct targets; clinical validation variables never enter label construction or the deployed inference path.

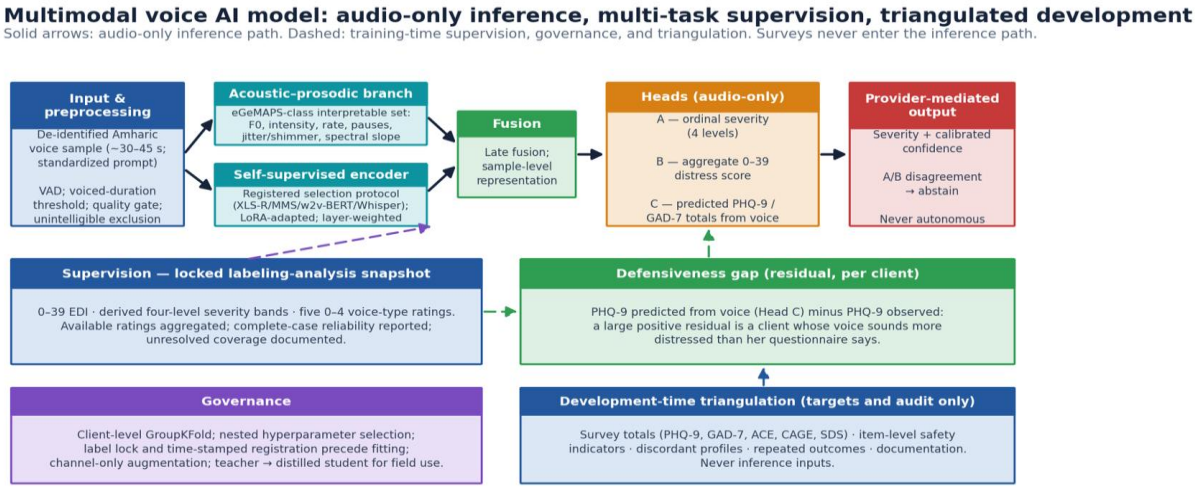


Figure 4. Multimodal voice AI model architecture.

Inference is audio-only (solid arrows): usability-gated Amharic audio is encoded through parallel acoustic-prosodic and self-supervised branches, fused, and mapped by three audio-only heads—A, EDI-derived four-level ordinal severity; B, aggregate 0–39 EDI; and C, PHQ-9/GAD-7 totals predicted from voice, defining the per-client defensiveness gap (predicted minus observed). Five independently scored 0–4 voice types provide complementary supervision. Training-time paths (dashed) use aggregated annotation targets and development-time triangulation with survey, safety, discordance, outcome, and documentation layers; those clinical layers never become inference inputs.

Head C is what converts the program's core construct into a measurable quantity. Per client, the defensiveness gap is defined as PHQ-9 predicted from voice minus PHQ-9 observed: a large positive residual is a client whose voice sounds more distressed than her questionnaire says — measurable, falsifiable, and directly testable against the 70.7% and 72.9% discordant populations summarized in the linked cohort evidence in Table 3. Surveys stay on the target side, never the input side; deployment remains voice-only; and because every sample was recorded moments after the battery was completed (Section 3.4), voice-predicted and observed PHQ-9 are contemporaneous measurements of the same defended state. Figure 5 shows the same commitment from the perspective of the data: the locked 2,479-sample labeling-analysis snapshot sits at the center of a triangulation structure in which every other layer — survey totals, item-level safety indicators, discordant high-risk profiles, repeated outcome measures, and clinical/service documentation — is joined to the labels at client and record level, informing training as targets and auditing model outputs, so that voice is never modeled in isolation from the clinical record it is meant to complement.

Three further training specifications are fixed. Class reweighting, if used, is followed by post-hoc recalibration (temperature scaling or isotonic regression) fit only on development folds, because any reweighting destroys calibration and Section 7 makes calibration a first-class endpoint [56, 72]; calibration is reported both pre- and post-recalibration, with a reliability diagram and a Brier score decomposed into refinement and calibration components. The continuous target is exploited against imbalance: Head B does not suffer the Severe-class starvation of the four-level head, and if the joint loss is weighted toward Head B early in training and annealed toward Head A, that curriculum and its loss weights are registered. If a survey-conditioned model is wanted for development-time auditing, it is built as a clearly labeled secondary, non-deployed concordance model that

never appears on the inference path and is never the model evaluated on the discordance endpoint. The triangulation results and algorithm details will be reported in the companion paper.

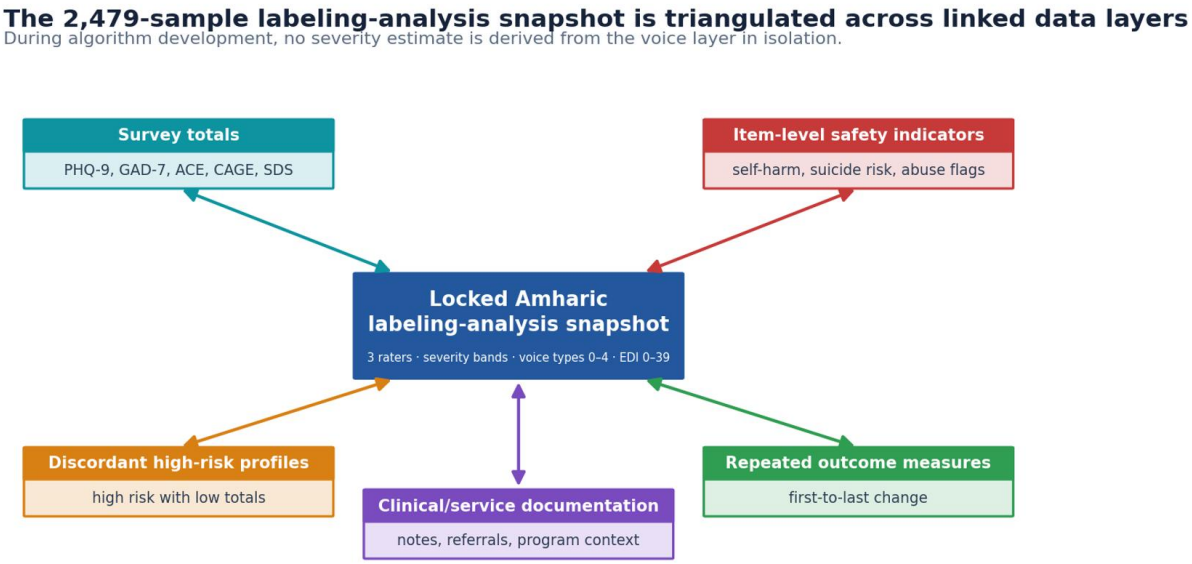


Figure 5. Triangulation of the locked 2,479-sample labeling-analysis snapshot with the other linked data layers.

During algorithm development, no severity estimate is derived from the voice layer in isolation; every layer is joined to the labels at client and record level.

6.4 Encoder selection protocol, teacher–student separation, and deployment

The self-supervised encoder is selected by a registered protocol, not assumed. The tempting default — committing to a single multilingual encoder on the grounds that its pretraining includes Amharic — is weaker than it appears: XLS-R's Amharic exposure enters only through BABEL and VoxLingua107, plausibly tens of hours of telephone-band speech [73], which does not underwrite a claim of native prosodic representation; MMS, pretrained across more than a thousand languages, reports better low-resource performance on Amharic specifically [74]; and at roughly 40 hours of audio, well under half of it voiced, this corpus sits firmly in the very-low-resource regime where the largest multilingual encoders are not the documented best choice. The registered shortlist is four arms in five runs: XLS-R-300M; MMS-1B in two sub-arms (pretrained only, and with the frozen Amharic ASR adapter, with the registered prediction that the adapter arm loses — ASR objectives reward invariance to speaker, affect, and prosody, the precise information the severity label depends on); w2v-BERT 2.0, the a-priori favorite on pretraining scale and on the conformer's inductive bias for the local temporal structure that listening element 3 asks raters to attend to [75]; and the Whisper-large-v3 encoder, the most robust arm to noise and channel variation and the registered favorite for the as-collected robustness analysis, though predicted to lose the primary endpoint because transcription training discards paralinguistic information [46]. The interpretable eGeMAPS branch is retained in every configuration and is a comparator, not an arm.

A fifth arm is retained from this group's prior, reduced-to-practice architecture [76], deployed in earlier pediatric and adult populations [26, 27, 28]: a convolutional front end whose first layer applies 512 filters of kernel size 320 with stride 160, matched to the ~20 ms quasi-stationarity of speech, followed by DenseNet-style skip connections in which each layer receives the concatenated output of all preceding layers, and a transformer encoder applying multi-head self-attention across temporal segments; the encoder is pretrained by self-supervised triplet loss over temporal proximity on population-matched speech, drawing temporally adjacent segments together in representation space without requiring labels. Its inclusion answers a question none of the public arms can: whether population-matched in-house pretraining at modest scale outperforms public

multilingual scale. It competes under the identical frozen downstream stack, adaptation budget, and selection criterion as every other arm; its earlier per-emotion classification heads and derived severity composite are superseded by the multi-task head structure of Section 6.3, consistent with the corrected description of the independently rated 0–39 score (Section 5.2). Because this arm is the architecture of authors affiliated with the developing organization, its inclusion in an otherwise arm-agnostic comparison is governed explicitly to prevent bias in its favor: its predicted rank is registered a priori alongside those of the public encoders; it competes under the identical frozen downstream stack, adaptation budget, augmentation policy, and selection metric; fold assignment is fixed and blinded before any arm is fit; and the selection criterion is computed on development folds only, with the outer test partition untouched. A win under these constraints is interpretable precisely because the constraints were fixed in the registration before results were seen.

The comparison is interpretable only if the encoder is the sole varying factor: chunking, layer aggregation (a learnable weighted sum over all hidden states, whose learned weights are themselves reportable — at what depth does an encoder carry Amharic vocal distress?), voiced-frame-masked attentive statistics pooling, projection width, heads, augmentation, optimizer, and the trainable-parameter budget are identical across arms and fixed before the first fit. Adaptation is parameter-efficient by design — the front end and lower half of each context network are frozen and the upper half adapted via low-rank adaptation, with trainable parameters held to 1–2% of encoder size in every arm [77] — because a 300M–1B-parameter encoder against 2,600 clients is a memorization machine. The selection criterion is registered before selection: highest mean quadratic-weighted kappa on development folds only, across five seeds and all folds under client-level GroupKFold, with the outer test partition untouched; tie-breaks are expected calibration error, then measured student latency. The actual hours of Amharic in the selected checkpoint's pretraining mix are reported, or the inability to determine them is reported as a finding about the low-resource regime. A-priori predictions — including the ranking across arms and the deliberately uncomfortable prediction that no arm exceeds the eGeMAPS ordinal-logistic comparator by more than 0.10 quadratic-weighted kappa at this data scale — are registered so that the selection is an experiment rather than a search; full per-arm architecture specifications, the fixed downstream stack, and the compute budget are given in the registered protocol.

None of the shortlisted encoders runs on a mid-range Android device, and the Viability endpoint of Section 7 therefore measures a different object than the selection protocol evaluates. The two objects are named and separated. The teacher is the selected encoder, evaluated on the Validity, Safety-sensitivity, Specificity, and Discordance families. The student is a distilled, int8-quantized model — target roughly 30–95M parameters — trained against the teacher's outputs and the locked labels [78]; the student is what the Viability endpoint measures and what Phase 3 deploys. The acceptable teacher-to-student degradation is pre-specified before distillation — a bounded loss of quadratic-weighted kappa and no loss of safety-sensitivity recall at the operating point — and the fallback is named: server-side inference with offline queueing and store-and-forward, often the correct architecture for a provider-mediated tool whose output is reviewed rather than acted on in real time. The eGeMAPS ordinal-logistic comparator runs on any device; if it trails the teacher by less than the distillation gap, it is the deployable model, and that is a legitimate, publishable outcome.

7. Pre-Specified Validation Strategy: Accuracy Is Not Enough

Development-time triangulation (Section 6.3) builds the model against the full record; held-out validation then asks whether the result generalizes. A final voice model should be evaluated in multiple ways because no single comparator is sufficient, and because the comparator most readily available — conventional screen totals — is precisely the channel whose blind spot motivated the program (Table 8). Five held-out performance endpoint families are pre-specified. Validity endpoints assess agreement with standardized instruments and clinician-reviewed labels [65], using correlation, ordinal accuracy, weighted kappa [69], and calibration — the last routinely neglected in clinical prediction despite being decisive for whether risk estimates can be trusted at the point of care [56]. Safety-sensitivity endpoints assess detection of suicide-risk indicators, severe distress, and

discordant high-risk presentations, prioritizing sensitivity and structured false-negative review, since item 9 and related indicators carry real predictive weight for subsequent suicidal behavior [19, 20, 21] and false negatives here are the costliest error class.

Specificity endpoints protect against over-pathologizing culturally normative speech, with false-positive review conducted by cultural and clinical raters — the mirror-image error the cultural protocol targets. Clinical-usefulness endpoints ask whether the signal changes triage, referral, follow-up, or care planning, measured through workflow adoption, provider trust, action rates, and referral completion, consistent with the measurement-based-care framing of our prior implementation work [27] and with DECIDE-AI's emphasis on human-factors evaluation [58]. Viability endpoints ask whether the system functions offline or under low connectivity with non-specialist staff — completion rate, audio quality yield, runtime, training burden, and cost per screen — the deployment realities emphasized across the LMIC digital mental health literature [22].

The discordance family deserves emphasis because it operationalizes the program's core claim. For clients whose item-level or program-identified risk is high while PHQ/GAD totals are low — the 70.7% and 72.9% populations summarized in Table 3 — the evaluation asks whether the voice-severity estimate is elevated where the total is not. A model that succeeds only on concordant cases has learned self-report by acoustic proxy; a model that adds signal on discordant cases has closed part of the defensiveness gap. All performance statistics will be reported only after the model and locked validation set are complete, against TRIPOD+AI reporting requirements [57], with explicit subgroup reporting and no generalizability claims beyond the validated population [42].

Comparator ladder. The five held-out performance families state how the model is evaluated; the comparator ladder states what it must beat, and it is registered in advance because an unregistered baseline makes the companion paper's central claim unfalsifiable (Table 7). The survey-only baseline is the critical rung: it answers whether voice adds anything over the channel it exists to replace. The text-only baseline is the newly critical rung, a registered comparator: it tests whether voice adds over transcript content alone, holding the model to the standard that its signal is vocal rather than lexical. The duration-and-quality-only baseline answers whether the model is reading recording conditions.

Comparator	Question it answers
Majority class / prevalence	Floor.
Survey-only (PHQ-9 + GAD-7 totals → severity)	Does voice add anything over the channel it exists to replace? The critical baseline.
Text-only (transcript → severity)	Does voice add over transcript content alone? A registered comparator.
eGeMAPS + ordinal logistic regression	Does the foundation model earn its complexity and its loss of auditability?
SSL embeddings, frozen + linear probe	Does adaptation earn its overfitting risk at n ≈ 2,600 clients?
Duration + audio-quality only	Is the model reading recording conditions?
Full pre-specified model	—

Table 7. Registered comparator ladder for the companion paper.

All comparators are evaluated under identical client-level splits; the survey-only and text-only baselines are registered before any model is fit.

Sample size and precision. The present calculations describe annotation rather than future algorithm performance: 2,479 samples, 7,086 usable ratings, and 2,267 complete three-rater EDI matrices. Derived severity is strongly concentrated in Low and Moderate, with only 13 Severe samples (0.5%). No Severe-specific performance or calibration claim is possible because no model has been fit and final client-level eligibility, split sizes, held-out class counts, and precision calculations require full-corpus label completion under prediction-model sample-size guidance [79]. Those quantities will be locked and reported in the follow-up algorithm paper; no current client-level or subgroup estimate is reported here.

Viability, as-collected performance, and review burden. Unintelligible attempts are excluded from the current annotation analysis, but a community health worker on a mid-range device in a noisy room will still produce unusable or degraded audio, so the Viability family closes the deployment/evaluation mismatch explicitly. Pre-specified: an as-collected analysis including degraded audio with an abstain/defer option; yield, the fraction of real-world attempts producing an actionable output; and the operating point with its resulting review burden — flags per 100 screens at threshold t — which is the number the frontline clinical partner actually needs. The minimum voiced-duration threshold of Section 6.2 will be evaluated in the follow-up algorithm study after the duration analysis is finalized. The Viability family evaluates the distilled student of Section 6.4, not the teacher. The Specificity family gains a base-rate anchor: a general-population calibration cohort of 200–500 Amharic-speaking women recruited from routine maternal-health visits under an IRB amendment, frequency-matched on age band and region with residual imbalance reported, assessed under the identical two-task elicitation protocol, rated blind in a single interleaved queue with Lenegewa samples (raters are never told which cohort a sample belongs to — without blinding the arm is worthless as a specificity test), held out as a test set only and siloed from every training partition; a randomized sub-arm records before the survey battery, yielding the direct estimate of the elicitation prime's acoustic effect size that the deployment argument requires (Section 3.4). The pre-specified expectation is that the model outputs Low severity on this cohort at a registered rate, demonstrating that it has not lost its baseline reference.

Explanation alignment. Attention weights are not necessarily faithful explanations, and the point remains contested [80, 81]. The available annotation records include human-authored, culturally grounded rationales aligned with ratings. The follow-up paper will evaluate model attributions from the interpretable GeMAPS branch against those rationales—for example, whether pause-duration features are emphasized where raters document long pauses or depleted, monotone speech. This is a cross-cutting audit family distinct from the five held-out performance families and will be implemented only after full-corpus label completion.

Table 8. Recommended endpoint families for final algorithm evaluation.

Endpoint family	Question	Example metrics
Validity	Agreement with standardized instruments and clinician-reviewed labels	Correlation, ordinal accuracy, weighted kappa, calibration.
Safety sensitivity	Detection of suicide-risk, severe distress, and discordant high-risk presentations	Sensitivity, false-negative review, high-risk recall.
Specificity	Avoidance of over-pathologizing culturally normative speech	Specificity and false-positive review by cultural/clinical raters.
Clinical usefulness	Does the signal change triage, referral, follow-up, or care planning?	Workflow adoption, provider trust, action rate, referral completion.
Viability	Can it work offline or under low connectivity with non-specialist staff?	Completion rate, audio quality, runtime, training burden, cost per screen.

Endpoint family	Question	Example metrics
Explanation alignment	Do model feature attributions correspond to the raters' evidence-based rationales?	Rationale–attribution concordance, feature-level overlap, and blinded qualitative review.

These are pre-specified evaluation endpoints, not results attributed to an untrained model. Final model-performance statistics are reported only after fitting and locked-set evaluation are complete.

8. Discussion

8.1 The scientific bridge

This Research Article reports a completed empirical data-centric bridge between the linked Lenegewa discordance findings and later Amharic psychiatric voice-algorithm development. The completed outputs are the culturally anchored annotation method, the locked 2,479-sample snapshot and its distributional and reliability findings, strict separation of annotation from clinical validation variables, and the pre-specified validity architecture. The linked cohort findings in Table 3 establish why self-report alone can miss clinically important severity; prior voice and measurement-based-care work established feasibility [26, 27, 28]; and current literature establishes both the promise of voice [34, 35, 39] and the failures that can invalidate it [40, 42, 54]. The present answer is methodological and empirical: construct culturally valid targets, quantify their uncertainty, link rather than collapse multimodal evidence, and pre-specify how later model claims will be tested. Stated as an answer to the research question of Section 1.4: culturally anchored severity judgment of Amharic speech is reliable enough to supervise machine learning only under multi-rater aggregation. A single rating is not an adequate target—ICC(2,1) of 0.54 with a standard error of measurement of 4.80 EDI points, nearly half a severity band—whereas the three-rater mean reaches ICC(2,k) of 0.78. Ordinal agreement is moderate (Krippendorff's alpha 0.46 for derived bands; 0.28–0.51 across voice types) and is reported as a property of the construct to be carried into model evaluation rather than smoothed away. These quantitative findings are the constraints the follow-up algorithm paper inherits.

8.2 Risk one: circular validation

If voice-derived supervisory labels were assigned from PHQ-9, GAD-7, clinician diagnosis, or safety flags, later validation against those same measures would be circular and the model would inherit the instruments' documented insensitivity to defended presentation. Conversely, labels assigned with no clinical linkage risk becoming acoustically consistent but clinically unhelpful. The completed compromise is separation with linkage: raters construct supervisory labels from audio under a culturally specific written guide, while clinical scales, item-level safety flags, discordance indicators, and outcomes remain independent validation targets. The same clients contribute both channels, enabling multimodal triangulation without contaminating label construction or the audio-only inference path.

8.3 Risk two: cultural misclassification

In Amharic-speaking Ethiopian women with complex trauma, distress may sound quiet, guarded, slowed, restrained, deferential, firm, or emotionally constricted; it may also hide beneath performed strength, gratitude, politeness, or controlled optimism. A model trained on culturally naive labels would misclassify these patterns as neutral or low severity — an algorithmic reproduction of the defensiveness gap — while simultaneously risking the opposite error of pathologizing dialect, restraint, and reading difficulty. The bidirectional cultural rules of Section 5.3, the requirement for Ethiopian raters, the treatment of Neutral as a positive finding rather than a default, and the pre-specified accent, dialect, and literacy fairness checks are the method's answer. The broader lesson generalizes beyond this project: in affective computing for global mental health, the labeling protocol is where cultural validity is won or lost [47, 49, 50], and label quality is a scientific deliverable, not a

procurement detail [55]. This is a machine-learning design decision: cultural knowledge changes the supervisory target at construction time, before any feature extractor or classifier can learn and amplify a labeling error.

8.4 Risk three: overclaiming

The completed result reported here is not model accuracy; it is the analysis of a locked annotation snapshot and the target-construction method required for a valid later accuracy claim. The present evidence comprises 2,479 samples, 7,086 usable ratings, direct EDI and 0–4 voice-type outputs, EDI-derived severity bands, explicit coverage denominators, and a reliability profile that includes confidence intervals, pairwise agreement, ICC, SEM, leniency, and Bland–Altman analyses. Model accuracy, sensitivity, calibration, subgroup performance, and clinical effectiveness remain unreported because full-corpus label completion and algorithm fitting are reserved for the follow-up paper.

Reporting this completed data-centric stage — and depositing the model-development specification as a time-stamped registration before any model is fit — is a commitment device against future overclaiming: the target-construction results can be reviewed now, and later model claims can be compared with a verifiable specification. One overclaiming risk is specific to this design and is flagged here rather than deferred: because every training sample was elicited immediately after the standardized screening battery (Section 3.4), the corpus is drawn from a single, primed elicitation distribution that routine deployment, where no battery precedes the recording, may not reproduce. The magnitude of that distribution shift is therefore a gating result for the companion paper: if the pre-survey sub-arm of Section 7 shows a large acoustic effect, performance obtained under the primed condition cannot be claimed to transfer to unprimed deployment, and accuracy must be reported under both conditions or restricted to the primed setting.

8.5 Positioning: a data-centric, clinically governed contribution

The contribution of this paper is data-centric AI applied to global mental health. What is scarce is not model architecture — foundation models and ordinal classifiers are broadly available [44, 45, 46] — but clinically linked, culturally labeled, population-matched audio with independent validation layers and quantified label uncertainty. To our knowledge, no existing resource pairs Amharic clinical audio at this scale with standardized instruments, item-level safety indicators, longitudinal outcomes, and service documentation; ASED, the nearest linguistic neighbor, is an acted-emotion corpus from healthy volunteers [53]. The reliability results in Table 5 reinforce the data-centric premise: culturally informed psychiatric voice annotation is a noisy measurement problem, and the improvement from single-rater to aggregate reliability is evidence for the multi-rater design, not a reason to conceal disagreement. The workforce arithmetic that motivates AI-supported detection in Africa [1, 2, 3, 4, 5], the suicide burden that gives safety sensitivity its urgency [82], and the disclosure dynamics that defeat questionnaire-only measurement [7, 8] jointly define the specification this resource and process were built to meet.

9. Limitations and Next Steps

The limitations are framed to prevent premature algorithm-performance or causal claims (Table 9). The locked snapshot includes 2,479 samples and 7,086 usable ratings, but coverage is incomplete: 2,267 samples form a complete three-rater EDI matrix, confidence is available for 6,577 ratings, and voice-type denominators vary slightly because of item-level missingness. Agreement is modest, with Krippendorff's alpha 0.46 for EDI-derived severity bands and 0.28–0.51 across voice types; ICC for the three-rater EDI mean is 0.78, compared with 0.54 for a single rating, and SEM is 4.80. These values document annotation uncertainty rather than model performance. They also bound inference in three further ways: the reliability estimates are complete-case and could differ in samples without full three-rater coverage; the restricted severity range of a routine-care population (58.3% Moderate) attenuates chance-corrected agreement; and the cohort is a single-site population of Ethiopian women in residential rehabilitation, so reliabilities and distributions may not transfer to

men, community settings, or other Amharic-speaking populations. The hypotheses in Section 1.4 were fixed in the analysis plan before model development but were not lodged in a public registry; the model-development specification itself is committed to time-stamped registration before any model is fit (Section 8.4). Full-corpus labeling remains ongoing. No accuracy, sensitivity, specificity, calibration, client-level, subgroup, provenance, adjudication-rate, or model-gating result is claimed here.

Clinical exposure data — structured intervention dates, dose, medication and detoxification exposure, group attendance, and follow-up — are not complete at person level and must be added before treatment-response modeling. Safety outcomes beyond flag recurrence — clinician-rated safety resolution, crisis referral, and follow-up status — are not fully captured and are required before safety-utility claims. Finally, Lenegewa is a single high-need Ethiopian rehabilitation population; validation in additional Amharic-speaking clinical and community settings is required before broader deployment, and no generalizability is asserted beyond populations in which the model has been evaluated [42]. Extension to Afan Oromo and Tigrinya, planned within our multilingual program, will require language-specific labeling protocols developed by the same method, not translated shortcuts.

Two limitations and two commitments are added to this frame. Elicitation-context distribution shift: every training sample was collected downstream of a completed, literate-mediated survey battery (Section 3.4), and deployment contexts without that battery remove the standardized prime, shifting the acoustic distribution; the size of that shift is currently unknown, and the general-population sub-arm that records before the surveys (Section 7) is the pre-specified direct estimate of it — obtained before Phase 3, not after. Field-test safety protocol: a real-world test of a suicide-risk-adjacent tool requires a written safety protocol that does not yet exist and is committed here: the handling rule when the model outputs Severe and the clinician disagrees, on-call clinical coverage, the adverse-event definition and reporting path, and the stopping rule, finalized and prospectively registered — with DECIDE-AI's human-factors instruments named in advance — before the field test begins [58]. Shadow deployment precedes Phase 3: the model runs silently on live encounters, outputs logged and reviewed but never shown to clinicians, surfacing audio-quality yield, inference latency, and drift before any clinical decision is influenced; it is also where the deployment-side consequence of removing the survey prime is measured. External validation: Lenegewa is a single site, and the validation in additional Amharic-speaking clinical and community settings noted above is out of scope for the companion paper; it will constitute a third study, and no second site is named here.

Table 9. Limits of the current evidence and required next steps.

Area	Current status	Required next step
Voice-label corpus	Locked analysis snapshot: 2,479 samples, 7,086 usable ratings, and 2,267 complete three-rater EDI matrices; full-corpus labeling is ongoing.	Complete remaining ratings, reconcile coverage, lock the full corpus, and conduct algorithm analyses only afterward.
Algorithm performance	No model has been fit; no accuracy, sensitivity, specificity, calibration, or gating result is reported.	Report triangulation and algorithm results in the follow-up paper, followed by real-world clinical testing with a correction cycle.
Clinical exposure data	Treatment components and dose are not complete at person level.	Add structured intervention dates, dose, medication/detox exposure, group attendance, and follow-up.

Area	Current status	Required next step
Safety outcomes	Safety flags can be compared over time, but safety-plan completion and adverse events are not fully captured.	Add clinician-rated safety resolution, crisis referral, and follow-up status.
Generalizability	Lenegewa is a high-need Ethiopian rehabilitation population.	Validate in additional Amharic-speaking clinical/community settings before broader deployment.
Cross-cultural and cross lingual comparison	Not yet conducted	Planned as an iterative validation step (Amharic-specific vs. general-purpose model comparison; ASED as general-population comparator; rater cross-cultural ablation), contingent on future data-collection and labeling resources.

Limitations are framed to prevent premature algorithm-performance or causal claims.

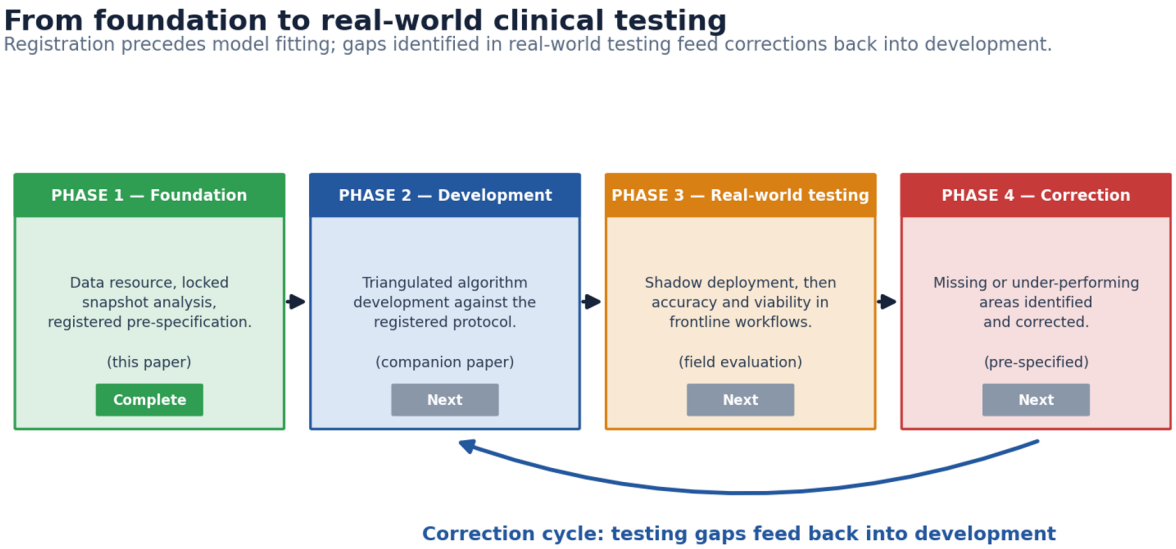


Figure 6. From foundation to real-world clinical testing.

The present analysis is complete for the locked 2,479-sample labeling snapshot; full-corpus label completion and Phase 2 algorithm development follow under the time-stamped specification. Phases 3–4 will test the resulting algorithm in real-world clinical environments and correct identified gaps before deployment.

10. Conclusions

The Lenegewa resource provides a rare foundation for Amharic voice-based psychiatric-severity research: population-matched audio, standardized instruments, item-level safety indicators, discordant high-risk profiles, repeated outcomes, and clinical documentation linked at the client level. Within that resource, this article completes and reports analysis of a locked snapshot containing 2,479 samples and 7,086 usable ratings from three Ethiopian raters. Direct 0–39 EDI ratings and five 0–4 voice-type ratings were constructed from voice alone; overall-severity bands were derived from EDI, while clinical scales and safety variables remained independent validation targets. The observed distributions and agreement statistics make annotation uncertainty explicit without inventing algorithm performance. Stated as the answer to the research question posed in Section 1.4: culturally anchored severity judgment of Amharic speech can supervise machine learning,

but only as a multi-rater aggregate—a single rating is too noisy a target—and the quantified residual uncertainty travels forward as an explicit constraint on any later performance claim.

This Research Article contributes four linked artifacts: a culturally anchored methodology for constructing machine-learning supervisory targets; a completed empirical analysis of a locked 2,479-sample, 7,086-rating snapshot; strict separation of annotation targets from clinical variables reserved for validation; and a pre-specified validity architecture covering audio-only inference, client-level splits, ordinal modeling, constrained augmentation, encoder comparison, calibration, subgroup checks, comparator baselines, and provider-mediated deployment. Full-corpus label completion, algorithm fitting, and model performance will be reported in the follow-up paper. The staged development process and fundamental model architecture are unchanged.

Declarations

Author contributions. P. Ohiomoba conceived the study with Y. Alemu, designed the technical and data architecture, led the multimodal algorithm-development framework, and drafted the manuscript. Y. Alemu designed the clinical and cultural methodology, supervised the voice-labeling program, and co-drafted the manuscript. O. Balatel contributed to the technical infrastructure, data engineering, and analytic design. H. Mohammed contributed to the design of the validation framework, methodological review, and critical revision of the manuscript. J. B. Abeya contributed to the analytic framework, methodological review, 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 (P. Ohiomoba, Y. Alemu, and O. Balatel) are affiliated with AYA-Innovation, the organization developing the multimodal voice-based artificial-intelligence approach whose development methodology is described in this manuscript. This paper reports a methodology and data resource, not model performance; no accuracy, sensitivity, or effectiveness claims are made for any commercial product. 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 the retrospective analysis of de-identified clinical records underpinning this data resource was obtained from the Institutional Review Board of St. Paul's Hospital Millennium Medical College. The underlying clinical care, screening, and voice-sample collection 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, and voice samples are de-identified before labeling; the operational meaning of de-identification for audio, consent scope, data residency, and benefit sharing are addressed in Section 5.8. All methods were performed in accordance with the relevant guidelines and regulations.

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 counts and methodology only; no names, recordings, transcripts, or individually identifying details are included.

Data availability. No public dataset is associated with this manuscript. The underlying client-level records and audio recordings are not publicly available owing to patient confidentiality, ethical restrictions, and the sensitive nature of the population described. The aggregate counts supporting the data-resource and labeling-subset description are presented in the tables and figures. A Datasheet for Datasets and a Data Statement for the speech corpus are designated for supplementary submission [67, 68]. The future ASED annotation layer described in Section 5.9 is not part of the present results. Requests for further information about the data, or for access to de-identified data where ethical and legal restrictions permit, should be directed to the corresponding author, Dr. Yared Alemu (ya@aya-innovation.com).

Code availability. No new algorithm, software package, or computational pipeline is released with this manuscript. The reliability and agreement analyses use standard, previously published statistical estimators—ordinal Krippendorff's alpha, quadratic-weighted kappa, two-way intraclass correlation, the standard error of measurement, and Bland–Altman limits of agreement—computed with standard open-source statistical software. The analysis scripts that reproduce the descriptive and reliability results of Section 5.7 are available from the corresponding author upon reasonable request. The pre-specified model-development pipeline (Sections 6–7) is a registered specification rather than released software; its implementation will be deposited in a DOI-assigning repository with the follow-up algorithm paper.

Reporting guidelines. This is a completed data-centric AI Research Article reporting analysis of a locked 2,479-sample labeling snapshot and the target-construction methodology; it does not report algorithm-performance results. The follow-up development and validation program is designed for prospective conformity with TRIPOD+AI and DECIDE-AI and for risk-of-bias assessment under PROBAST+AI [57, 58, 83].

Generative AI statement. A generative AI assistant supported literature organization, drafting, and formatting under author direction. The authors verify and take full responsibility for the content, methodology, data counts, and conclusions of this manuscript.

Electronic Supplementary Material: A Datasheet for Datasets and a Data Statement for the speech corpus are designated for supplementary submission. No participant-level data, audio, algorithm results, or future ASED-derived annotations are included.

References

[1] Kakuma, R., Minas, H., van Ginneken, N., Dal Poz, M. R., Desiraju, K., Morris, J. E., Saxena, S., & Scheffler, R. M. Human resources for mental health care: Current situation and strategies for action. The Lancet, 378(9803), 1654–1663, 2011. https://doi.org/10.1016/S0140-6736(11)61093-3

[2] Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P. Y., Cooper, J. L., Eaton, J., Herrman, H., Herzallah, M. M., Huang, Y., Jordans, M. J. D., Kleinman, A., Medina-Mora, M. E., Morgan, E., Niaz, U., Omigbodun, O., … UnÜtzer, J. The Lancet Commission on global mental health and sustainable development. The Lancet, 392(10157), 1553–1598, 2018. https://doi.org/10.1016/S0140-6736(18)31612-X

[3] Wondimagegn, D., Pain, C., Seifu, N., Cartmill, C., Alemu, A. A., & Whitehead, C. R. Reimagining global mental health in Africa. BMJ Global Health, 8(9), e013232, 2023. https://doi.org/10.1136/bmjgh-2023-013232

[4] Fekadu, A., Medhin, G., Selamu, M., Giorgis, T. W., Lund, C., Alem, A., & Prince, M. Recognition of depression by primary care clinicians in rural Ethiopia. BMC Family Practice, 18, 56, 2017. https://doi.org/10.1186/s12875-017-0628-y

[5] 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. A pilot cluster-controlled trial of interventions to improve detection of depression in primary healthcare in Ethiopia. BMC Medicine, 23, 185, 2025. https://doi.org/10.1186/s12916-025-04010-4

[6] Charlson, F., van Ommeren, M., Flaxman, A., Cornett, J., Whiteford, H., & Saxena, S. New WHO prevalence estimates of mental disorders in conflict settings: A systematic review and meta-analysis. The Lancet, 394(10194), 240–248, 2019. https://doi.org/10.1016/S0140-6736(19)30934-1

[7] Paulhus, D. L. Two-component models of socially desirable responding. Journal of Personality and Social Psychology, 46(3), 598–609, 1984. https://doi.org/10.1037/0022-3514.46.3.598

[8] Tourangeau, R., & Yan, T. Sensitive questions in surveys. Psychological Bulletin, 133(5), 859–883, 2007. https://doi.org/10.1037/0033-2909.133.5.859

[9] Herman, J. L. Trauma and recovery: The aftermath of violence—from domestic abuse to political terror. Basic Books, 1992.

[10] Lanius, R. A., Vermetten, E., Loewenstein, R. J., Brand, B., Schmahl, C., Bremner, J. D., & Spiegel, D. Emotion modulation in PTSD: Clinical and neurobiological evidence for a dissociative subtype. American Journal of Psychiatry, 167(6), 640–647, 2010. https://doi.org/10.1176/appi.ajp.2009.09081168

[11] Hinton, D. E., & Lewis-Fernández, R. Idioms of distress among trauma survivors: Subtypes and clinical utility. Culture, Medicine, and Psychiatry, 34(2), 209–218, 2010. https://doi.org/10.1007/s11013-010-9175-x

[12] 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. 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, 2014. https://doi.org/10.1093/ije/dyt227

[13] Nichter, M. Idioms of distress revisited. Culture, Medicine, and Psychiatry, 34(2), 401–416, 2010. https://doi.org/10.1007/s11013-010-9179-6

[14] 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. Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. Psychiatry Research, 210(2), 653–661, 2013. https://doi.org/10.1016/j.psychres.2013.07.015

[15] Hanlon, C., Medhin, G., Selamu, M., Breuer, E., Worku, B., Hailemariam, M., Lund, C., Prince, M., & Fekadu, A. Validity of brief screening questionnaires to detect depression in primary care in Ethiopia. Journal of Affective Disorders, 186, 32–39, 2015. https://doi.org/10.1016/j.jad.2015.07.015

[16] Kroenke, K., Spitzer, R. L., & Williams, J. B. W. The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606–613, 2001. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[17] Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine, 166(10), 1092–1097, 2006. https://doi.org/10.1001/archinte.166.10.1092

[18] Levis, B., Benedetti, A., & Thombs, B. D. Accuracy of the Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. BMJ, 365, l1476, 2019. https://doi.org/10.1136/bmj.l1476

[19] Louzon, S. A., Bossarte, R., McCarthy, J. F., & Katz, I. R. Does suicidal ideation as measured by the PHQ-9 predict suicide among VA patients? Psychiatric Services, 67(5), 517–522, 2016. https://doi.org/10.1176/appi.ps.201500149

[20] Na, P. J., Yaramala, S. R., Kim, J. A., Kim, H., Goes, F. S., Zandi, P. P., Vande Voort, J. L., Sutor, B., Croarkin, P., & Bobo, W. V. The PHQ-9 item 9 based screening for suicide risk: A validation study of the Patient Health Questionnaire (PHQ)-9 item 9 with the Columbia Suicide Severity Rating Scale (C-SSRS). Journal of Affective Disorders, 232, 34–40, 2018. https://doi.org/10.1016/j.jad.2018.02.045

[21] Simon, G. E., Rutter, C. M., Peterson, D., Oliver, M., Whiteside, U., Operskalski, B., & Ludman, E. J. Does response on the PHQ-9 Depression Questionnaire predict subsequent suicide attempt or suicide death? Psychiatric Services, 64(12), 1195–1202, 2013. https://doi.org/10.1176/appi.ps.201200587

[22] Naslund, J. A., Aschbrenner, K. A., Araya, R., Marsch, L. A., Unützer, J., Patel, V., & Bartels, S. J. Digital technology for treating and preventing mental disorders in low-income and middle-income countries: A narrative review. The Lancet Psychiatry, 4(6), 486–500, 2017. https://doi.org/10.1016/S2215-0366(17)30096-2

[23] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. A review of depression and suicide risk assessment using speech analysis. Speech Communication, 71, 10–49, 2015. https://doi.org/10.1016/j.specom.2015.03.004

[24] Kappen, M., Vanderhasselt, M.-A., & Slavich, G. M. Speech as a promising biosignal in precision psychiatry. Neuroscience & Biobehavioral Reviews, 148, 105121, 2023. https://doi.org/10.1016/j.neubiorev.2023.105121
[25] Scherer, K. R. Vocal communication of emotion: A review of research paradigms. Speech Communication, 40(1–2), 227–256, 2003. https://doi.org/10.1016/S0167-6393(02)00084-5
[26] Alemu, Y., Chen, H., Duan, C., Caulley, D., Arriaga, R. I., & Sezgin, E. 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, 2023. https://doi.org/10.2196/46970
[27] Alemu, Y., Cárdenas Bautista, E., Vinson, S., Ohiomoba, P., Melecio, N., Amare, H., Vivian, J., & Karippara, H. Implementing and analyzing the advantages of Voice AI as measurement-based care (MBC) to address behavioral health treatment disparities among youth in economically disadvantaged communities. Telehealth and Medicine Today, 9(4), Article 506, 2024. https://doi.org/10.30953/thmt.v9.506
[28] Caulley, D., Alemu, Y., Burson, S., Cárdenas Bautista, E., Tadesse, G. A., Kottmyer, C., Aeschbach, L., Cheungvivatpant, B., & Sezgin, E. Objectively quantifying pediatric psychiatric severity using artificial intelligence, voice recognition technology, and universal emotions: Pilot study for artificial intelligence-enabled innovation to address youth mental health crisis. JMIR Research Protocols, 12, e51912, 2023. https://doi.org/10.2196/51912
[29] Shin, J., & Bae, S. M. Use of voice features from smartphones for monitoring depressive disorders: Scoping review. Digital Health, 10, 20552076241261920, 2024. https://doi.org/10.1177/20552076241261920
[30] 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. 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, 2026. https://doi.org/10.33425/2641-4317.1237
[31] Low, D. M., Bentley, K. H., & Ghosh, S. S. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology, 5(1), 96–116, 2020. https://doi.org/10.1002/lio2.354
[32] Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R. Vocal acoustic biomarkers of depression severity and treatment response. Biological Psychiatry, 72(7), 580–587, 2012. https://doi.org/10.1016/j.biopsych.2012.03.015
[33] Almaghrabi, S. A., Clark, S. R., & Baumert, M. Bio-acoustic features of depression: A review. Biomedical Signal Processing and Control, 85, 105020, 2023. https://doi.org/10.1016/j.bspc.2023.105020
[34] Liu, L., Liu, L., Wafa, H. A., Tydeman, F., Xie, W., & Wang, Y. 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, 2024. https://doi.org/10.1093/jamia/ocae189
[35] Riad, R., Denais, M., de Gennes, M., Lesage, A., Oustric, V., Cao, X. N., Mouchabac, S., & Bourla, A. Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: Algorithm development and validation study. Journal of Medical Internet Research, 26, e58572, 2024. https://doi.org/10.2196/58572
[36] Mazur, A., Costantino, H., Tom, P., Wilson, M. P., & Thompson, R. G. Evaluation of an AI-based voice biomarker tool to detect signals consistent with moderate to severe depression. Annals of Family Medicine, 23(1), 60–65, 2025. https://doi.org/10.1370/afm.240091
[37] Iyer, R., & Meyer, D. Detection of suicide risk using vocal characteristics: Systematic review. JMIR Biomedical Engineering, 7(2), e42386, 2022. https://doi.org/10.2196/42386
[38] Belouali, A., Gupta, S., Sourirajan, V., Yu, J., Allen, N., Alaoui, A., Dutton, M. A., & Reinhard, M. J. Acoustic and language analysis of speech for suicidal ideation among US veterans. BioData Mining, 14, Article 11, 2021. https://doi.org/10.1186/s13040-021-00245-y
[39] Briganti, G., & Lechien, J. R. Speech and voice quality as digital biomarkers in depression: A systematic review. Journal of Voice. Advance online publication, 2025. https://doi.org/10.1016/j.jvoice.2025.05.002
[40] Donaghy, P., Ennis, E., Mulvenna, M., Bond, R., Kennedy, N., McTear, M., O'Connell, H., Blaylock, N., & Brueckner, R. A review of studies using machine learning to detect voice biomarkers for depression. Journal of Technology in Behavioral Science. Advance online publication, 2024. https://doi.org/10.1007/s41347-024-00454-2
[41] Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, A., & Morency, L.-P. The Distress Analysis Interview Corpus of human and computer interviews. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), 3123–3128, 2014.
[42] Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health, 2(9), e489–e492, 2020. https://doi.org/10.1016/S2589-7500(20)30186-2
[43] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190–202, 2016. https://doi.org/10.1109/TAFFC.2015.2457417
[44] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460, 2020.
[45] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460, 2021. https://doi.org/10.1109/TASLP.2021.3122291
[46] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 28492–28518, 2023.

[47] Jordan, E., Terrisse, R., Lucarini, V., Alrahabi, M., Krebs, M.-O., & Desclés, J. Speech emotion recognition in mental health: Systematic review of voice-based applications. JMIR Mental Health, 12, e74260, 2025. https://doi.org/10.2196/74260
[48] Ekman, P. An argument for basic emotions. Cognition and Emotion, 6(3–4), 169–200, 1992. https://doi.org/10.1080/02699939208411068
[49] Elfenbein, H. A., & Ambady, N. On the universality and cultural specificity of emotion recognition: A meta-analysis. Psychological Bulletin, 128(2), 203–235, 2002. https://doi.org/10.1037/0033-2909.128.2.203
[50] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. Psychological Science in the Public Interest, 20(1), 1–68, 2019. https://doi.org/10.1177/1529100619832930
[51] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684–7689, 2020. https://doi.org/10.1073/pnas.1915768117
[52] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453, 2019. https://doi.org/10.1126/science.aax2342
[53] Retta, E. A., Sutcliffe, R., Mahmood, J., Berwo, M. A., Almekhlafi, E., Khan, S. A., Chaudhry, S. A., Mhamed, M., & Feng, J. A new Amharic speech emotion dataset and classification benchmark. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(1), Article 14, 2023. https://doi.org/10.1145/3529759
[54] Kapoor, S., & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804, 2023. https://doi.org/10.1016/j.patter.2023.100804
[55] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Article 39, 2021. https://doi.org/10.1145/3411764.3445518
[56] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230, 2019. https://doi.org/10.1186/s12916-019-1466-7
[57] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378, 2024. https://doi.org/10.1136/bmj-2023-078378
[58] Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., … McCulloch, P. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 28(5), 924–933, 2022. https://doi.org/10.1038/s41591-022-01772-9
[59] Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., Koss, M. P., & Marks, J. S. 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, 1998. https://doi.org/10.1016/S0749-3797(98)00017-8
[60] Ewing, J. A. Detecting alcoholism: The CAGE questionnaire. JAMA, 252(14), 1905–1907, 1984. https://doi.org/10.1001/jama.1984.03350140051025
[61] Gossop, M., Darke, S., Griffiths, P., Hando, J., Powis, B., Hall, W., & Strang, J. 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, 1995. https://doi.org/10.1046/j.1360-0443.1995.9056072.x
[62] Krippendorff, K. Content analysis: An introduction to its methodology (4th ed.). SAGE Publications, 2018.
[63] Shrout, P. E., & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428, 1979.
[64] Koo, T. K., & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163, 2016.
[65] Bland, J. M., & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307–310, 1986.
[66] Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., & Evans, N. The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding. Interspeech 2019, 3695–3699, 2019.
[67] Bender, E. M., & Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587–604, 2018.
[68] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. Datasheets for datasets. Communications of the ACM, 64(12), 86–92, 2021.
[69] Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213–220, 1968. https://doi.org/10.1037/h0026256
[70] Cao, W., Mirjalili, V., & Raschka, S. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters, 140, 325–331, 2020.
[71] Shi, X., Cao, W., & Raschka, S. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. Pattern Analysis and Applications, 26, 941–955, 2023.
[72] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 1321–1330, 2017.

[73] Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv:2111.09296, 2021.
[74] Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. Scaling speech technology to 1,000+ languages. arXiv:2305.13516, 2023.
[75] Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., & Wu, Y. w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 244–250, 2021.
[76] Alemu, Y., Caulley, D., & Joshi, A. A. Acquiring speech features for predicting emotional severity of adverse events on individuals (U.S. Patent Application Publication No. US 2022/0189502 A1). U.S. Patent and Trademark Office, 2022.
[77] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR), 2022.
[78] Hinton, G., Vinyals, O., & Dean, J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
[79] Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G. S., & van Smeden, M. Calculating the sample size required for developing a clinical prediction model. BMJ, 368, m441, 2020. https://doi.org/10.1136/bmj.m441
[80] Jain, S., & Wallace, B. C. Attention is not explanation. Proceedings of NAACL-HLT 2019, 3543–3556, 2019.
[81] Wiegreffe, S., & Pinter, Y. Attention is not not explanation. Proceedings of EMNLP-IJCNLP 2019, 11–20, 2019.
[82] World Health Organization. Suicide worldwide in 2019: Global health estimates. World Health Organization, 2021.
[83] Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., Maier-Hein, L., Liu, X., Logullo, P., McCradden, M. D., Liu, N., … van Smeden, M. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 388, e082505, 2025. https://doi.org/10.1136/bmj-2024-082505

Author Biographies

Patrick Ohiomoba, MS. Co-Founder and Chief Technology Officer of AYA-Innovation. He leads artificial-intelligence product strategy, technical architecture, and scalable implementation of voice-based mental-health tools. His research interests include speech artificial intelligence, multimodal machine learning, data-centric target construction, and responsible clinical deployment.

Yared Alemu, PhD. Co-Founder and Chief Executive Officer of AYA-Innovation and corresponding author. He leads clinical methodology and implementation of artificial-intelligence-enabled mental-health systems in underserved settings. His research interests include voice-based psychiatric-severity measurement, measurement-based care, global mental health, and trauma-informed digital health.

Octavian Balatel. Affiliated with AYA-Innovation. His contributions to this study include technical infrastructure, data engineering, and analytic design. His research interests include clinical data systems, multimodal artificial intelligence, and reliable infrastructure for voice-based mental-health applications.

Habib Mohammed. Affiliated with Addis Ababa Science and Technology University. Contributions to this study include the design of the validation framework, methodological review, and critical revision. Research interests include clinical validation of artificial-intelligence systems, algorithmic fairness, and digital health in low-resource settings.

Jale Bikila Abeya. Affiliated with Woosong University. Contributions to this study include development of the analytic framework, methodological review, and critical revision. Research interests include applied analytics, artificial intelligence methodology, and the evaluation of digital-health systems.