Original Paper

Cross-Domain Transfer of Depression Voice Biomarkers Depends on the Outcome Instrument: Leakage-Controlled Cross-Sectional Evaluation Study

Rachel Wiley[1*], MD, MPH; James Schwoebel[2*], BSc; Matias Caccia[3], PhD; Joel Shor[2], PhD; Adolfo M Garcia[4,5,6], PhD; Sheehan D Fisher[7], PhD; Martin Gerbert Frasch[8], MD, PhD

[1]Division of Maternal Fetal Medicine, Department of Obstetrics, Gynecology and Reproductive Sciences, University of California, San Diego, San Diego, CA, United States

[2]Quome, Inc, Los Angeles, CA, United States

[3]TELL, Inc, San Francisco, CA, United States

[4]Cognitive Neuroscience Center, Department of Life and Behavioral Sciences, Universidad de San Andrés, Buenos Aires, Argentina

[5]Global Brain Health Institute (GBHI), University of California, San Francisco, CA, United States

[6]Departamento de Lingüística y Literatura, Facultad de Humanidades, Universidad de Santiago de Chile, Santiago, Chile

[7]Department of Psychiatry and Behavioral Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[8]Department of Obstetrics and Gynecology and Institute of Human Development and Disability, University of Washington, Seattle, WA, United States

*these authors contributed equally

Corresponding Author:

Rachel Wiley, MD, MPH
Division of Maternal Fetal Medicine, Department of Obstetrics, Gynecology and Reproductive Sciences
University of California, San Diego
9300 Campus Point Dr.
San Diego, CA 92037
United States
Phone: +1 214-557-4758
Email: rwiley@health.ucsd.edu

Abstract

Background: Voice-based depression screening is commercializing rapidly, yet whether voice biomarkers generalize across clinical settings is largely untested. Generalization is usually framed as a question about populations, but it is also a question about the outcome instrument a model is scored against, an effect confounded with other between-study differences in the literature.

Objective: We tested whether depression voice biomarkers transfer between perinatal and general-psychiatric speech in both directions under strict leakage control and whether transfer depends on the screening instrument.

Methods: In a cross-sectional, online US study, 390 pregnant participants at a mean of 22 (SD 1) weeks' gestation (analytical sample n=316) participated in 446 sessions; each participant provided 4 voice recordings and completed the modified 8-item Patient Health Questionnaire (PHQ-8) and modified 9-item Edinburgh Postnatal Depression Scale (mEPDS-9). We extracted 375 acoustic descriptors and 330-dimensional Population Dynamics Foundation Model community-health embeddings. Discrimination area under the curve (AUC) was assessed under leakage-controlled cross-validation, with feature and classifier selection inside training folds only and a permutation negative-control harness, across a prespecified grid of 4 speech tasks×4 outcome definitions. The prenatal-trained model was applied to 2 general-psychiatric corpora (Distress Analysis Interview Corpus–Wizard-of-Oz, n=189; Extended Distress Analysis Interview Corpus, n=219) under pure held-out inference, and the open-source Kintsugi Depression/Anxiety Model (DAM), trained on 35,000 individuals, was applied to all 3 cohorts without refitting, with Benjamini-Hochberg multiplicity control.

Results: In the prenatal cohort, the preregistered outcome was at chance (AUC 0.494, 95% CI 0.431-0.560); no cell in the 16-cell grid survived adjustment under either modeling paradigm. The prenatal-trained model did not transfer to either general-psychiatric corpus (AUCs 0.505 and 0.478). Reverse transfer depended on the screening instrument. The DAM reached an AUC of 0.706 to 0.708 on general-psychiatric speech. In the same participants, it performed at chance against the

PHQ-8 (AUC 0.510, 95% CI 0.411-0.609) but discriminated against the mEPDS-9 (AUC 0.645, 95% CI 0.541-0.744), a paired difference of 0.135 (95% CI 0.019-0.248; unadjusted post hoc P=0.02). Exploratory item-level analysis suggests a mechanism: the DAM tracked affective-distress items (crying r=0.26, sadness r=0.21, and overwhelmed r=0.20) but not somatic PHQ-8 items; only crying survived adjustment across 17 tests. Transfer varied by elicitation task (pregnancy narrative 0.653, 95% CI 0.550-0.753; open-ended speech 0.597, 95% CI 0.486-0.699; scripted reading 0.544, 95% CI 0.434-0.651), though these were not formally compared.

Conclusions: Cross-domain comparison showed a compound shift: population, elicitation task, and outcome instrument differed together, while only the instrument could be varied on its own. Whether a depression voice biomarker transfers may depend not only on the population but also on whether the outcome instrument captures what the model can detect. Validation should state the population, elicitation task, and outcome instrument together; a tool validated against one instrument should not be assumed to work against another.

Trial Registration: OSF Registries osf.io/k387t; https://osf.io/k387t/overview

J Med Internet Res 2026;28:e105737; doi: 10.2196/105737

Keywords: perinatal depression; voice biomarkers; digital phenotyping; speech analysis; machine learning; data leakage; external validation; model generalizability; depression screening; Population Dynamics Foundation Model

Introduction

Perinatal depression affects up to 20% of pregnant individuals and is a leading cause of maternal morbidity and pregnancy-related death in the United States, yet only about half of antenatal cases are identified during pregnancy, and fewer than 10% receive adequate treatment [1-3]. Current screening commonly relies on standardized self-report questionnaires such as the 9-item Patient Health Questionnaire (PHQ-9) and the Edinburgh Postnatal Depression Scale (EPDS), administered or reviewed within clinical care and recommended during perinatal care [4], but real-world adherence is variable and even adherent screening can miss interval mood changes. Scalable, low-burden adjuncts that integrate into existing workflows are needed.

Voice carries a biological signal of depressive state: major depression produces psychomotor retardation and altered laryngeal tone, manifesting as reduced fundamental-frequency variability, longer pauses, reduced speech rate, and elevated jitter/shimmer [5-7]. Pregnancy itself produces overlapping but distinct phonatory changes via progesterone-mediated mucosal effects and altered respiratory mechanics [8-10]. This overlay motivates the usual framing of the following questions: Are depression voice biomarkers population specific? Can models trained on general-population speech generalize to perinatal speech, where the phonatory baseline is itself perturbed?

We test these questions in both directions and add one that the literature has left confounded. A systematic review of 25 speech-based depression studies reports that diagnostic performance varied with the diagnostic criteria used, alongside sample size, validation strategy, and language [11]. Because those criteria differ between studies at the same time as the cohorts, the selected outcome measure's contribution cannot be separated from the rest. Each of these models is scored against a self-report instrument, and instruments differ in what they measure. In pregnancy, the PHQ-9 and the EPDS show similar psychometric performance, but the EPDS focuses on psychological manifestations, whereas the PHQ-9 includes more somatic manifestations of depression [12]. Due to the dramatic physical changes during gestation, somatic-weighted instruments may capture symptoms of gestation independent of mood. Generalization may therefore depend not only on who is speaking but on what the outcome measure is asking. Our cohort includes 2 instruments used with the same participants, which allows population and instrument to be separated in one direction.

This study had 3 objectives. First, we tested whether depression voice biomarkers transfer between prenatal and general-psychiatric speech in both directions. We applied a prenatal-trained model to 2 general-psychiatric corpora (Distress Analysis Interview Corpus–Wizard of Oz [DAIC-WOZ] and Extended Distress Analysis Corpus [E-DAIC]), and we applied an open-source model trained on data from approximately 35,000 adults to the prenatal cohort [13]. We hypothesized that transfer would be poor. Second, we tested whether acoustic features fused with community-level embeddings from Google's Population Dynamics Foundation Model (PDFM) [14] discriminate depression within the prenatal cohort, with a score of 10 or higher on the 8-item Patient Health Questionnaire (PHQ-8) as the preregistered outcome. The PDFM summarizes socioeconomic conditions, health care access, and community health by US zip codes, which medical records already hold, so it adds no patient burden. Third, because the cohort completed 2 instruments, we asked whether transfer depends on the outcome instrument. The third question was not prespecified. All analyses followed a leakage-controlled protocol (see Methods section), which we offer as a methodological template. Participants were recruited at a single time point (mean 22, SD 1 weeks' gestation) so the findings concern midgestation antenatal depression.

Methods

Study Design and Participants

This was a cross-sectional, nationwide, online study of pregnant individuals at 22 (SD 1) weeks' gestation conducted in the United States. The study was preregistered on the

Open Science Framework [15]. Participants were recruited via Cint (Cint Group), an online survey recruitment platform that maintains a verified panel of US-based participants [16], and directed to the secure Databoard platform (Chessboard Labs) to consent and complete the study. The study was deemed "not human participants research" by the University of California San Diego Institutional Review Board (IRB#812757) as it collected no identifiable information.

Participants were included if they were female, aged 18 to 45 years, pregnant at 22 (SD 1) weeks, able to respond via a computing device with internet access, and without current suicidal ideation.

Voice Recording Protocol

Participants completed 4 vocal tasks via internet browser: (1) a 10-second sound check, (2) a 1-minute free speech prompt on pregnancy experiences, (3) a 2-minute structured reading passage, and (4) a 1-minute free speech prompt on a nonpregnancy topic. Tasks 2 and 4 were counterbalanced across participants to control for order effects (Multimedia Appendix 1).

Depression Assessment

Depression was measured at the single mean 22 (SD 1)–week timepoint using the PHQ-8 and the 9-item modified Edinburgh Postnatal Depression Scale (mEPDS-9) [17], both commonly used scales and each with one question removed [18,19]. Both instruments were self-administered online with no clinician involvement; therefore, the suicidality screening question was removed for ethical concerns related to inadequate safety and referral pathways. Binary classification used thresholds of 10 or higher and 13 or higher, respectively, for both instruments, which are common thresholds established in the literature for the PHQ-9 and EPDS with a question on suicidality included. The preregistered primary outcome is the reading passage with PHQ-8 score 10 or higher; all mEPDS-9 analyses are secondary or exploratory, and all 4 thresholds are reported.

A caution about the mEPDS-9 thresholds is required. The instrument as administered used a generic 4-point agreement scale, not the item-specific response anchors of the published EPDS, and omitted item 10. Correcting the scoring restores a monotone symptom score, but it does not establish that the psychometric properties of the published EPDS (including the validation of its ≥10 and ≥13 cutoffs) transfer to a 9-item instrument administered with different response options. We therefore treat mEPDS-9 scores as a reconstructed symptom score, and the values of 10 or greater and 13 or greater as operational thresholds mapped from EPDS convention, not as validated diagnostic cutoffs for this administration.

Rationale for a Single-Timepoint Screening Outcome

The study evaluates voice analysis as a screening adjunct: a single self-report screening instrument at a single clinical touchpoint, which is how perinatal depression screening is delivered in practice [4]. A screen-positive result is not a diagnosis. Individual participant–data meta-analyses report that against a semistructured diagnostic interview, the PHQ-9 at 10 or greater achieves a sensitivity of 0.85 (95% CI 0.79-0.89) and a specificity of 0.85 (95% CI 0.82-0.87) [20], and the EPDS at 10 or greater achieves a sensitivity of 0.85 (0.79-0.90) and a specificity of 0.84 (0.79-0.88) [21]. That is the performance level a screening adjunct would have to approach to be useful, and it is far above anything observed here.

Two caveats attach to the EPDS figure. First, the same meta-analysis found that combined sensitivity and specificity is maximized not at 10 or greater but at 11 or greater; we use 10 or greater because it is the conventional threshold and the one specified in our preregistration, not because it is the optimal operating point in that analysis. Second, and more importantly, these figures apply to the published instruments as validated and cannot be assumed to transfer to the modified administration used here (mEPDS). Throughout, the outcome is described as screen-positive depressive symptomatology rather than depression as a diagnostic entity.

Both instruments were self-administered online, unassisted, in the same session as the voice recordings; no clinician was involved in administration or scoring. The platform presented, for all nine mEPDS-9 items, a single generic 4-point agreement scale (Yes, all of the time/Yes, most of the time/No, not very often/No, not at all) instead of the item-specific anchors of the published EPDS.

Acoustic Feature Extraction (the Privacy Tier)

We extracted 375 audio-only acoustic descriptors per recording, composed of 4 feature blocks. Block sizes and the open-source toolboxes used are stated explicitly.

The 375 descriptors comprise 4 blocks: a 325-feature base-acoustic block (13 Mel-frequency cepstral coefficients [MFCCs], 26 log-Mel-filterbank energies, and 26 spectral subband centroids, each summarized by 5 statistics) adapted from the open-source Allie machine learning framework's tier-1 audio pipeline [22,23], 22 original voice-quality features (fundamental frequency, jitter, shimmer, harmonic-to-noise ratio, and formants) via Parselmouth/Praat [24,25], 12 extended voice-quality distributional moments, and 16 talking-interval/timing features. Per-block extraction detail is mentioned in Multimedia Appendix 2.

These features measure how speech is produced: spectral shape, voice-source characteristics, and timing and require no speech transcription, no language model, and no semantic content. We refer to this block collectively as the "privacy tier" throughout the main-text Results section, the calibration grid, the negative-control harness, and the leakage-controlled Tree-based Pipeline Optimization Tool (TPOT)–cross-validation (CV) sensitivity analysis (Multimedia Appendix 3). Audio loading uses soundfile via a thin wrapper (pipeline_a/audio_loader.py).

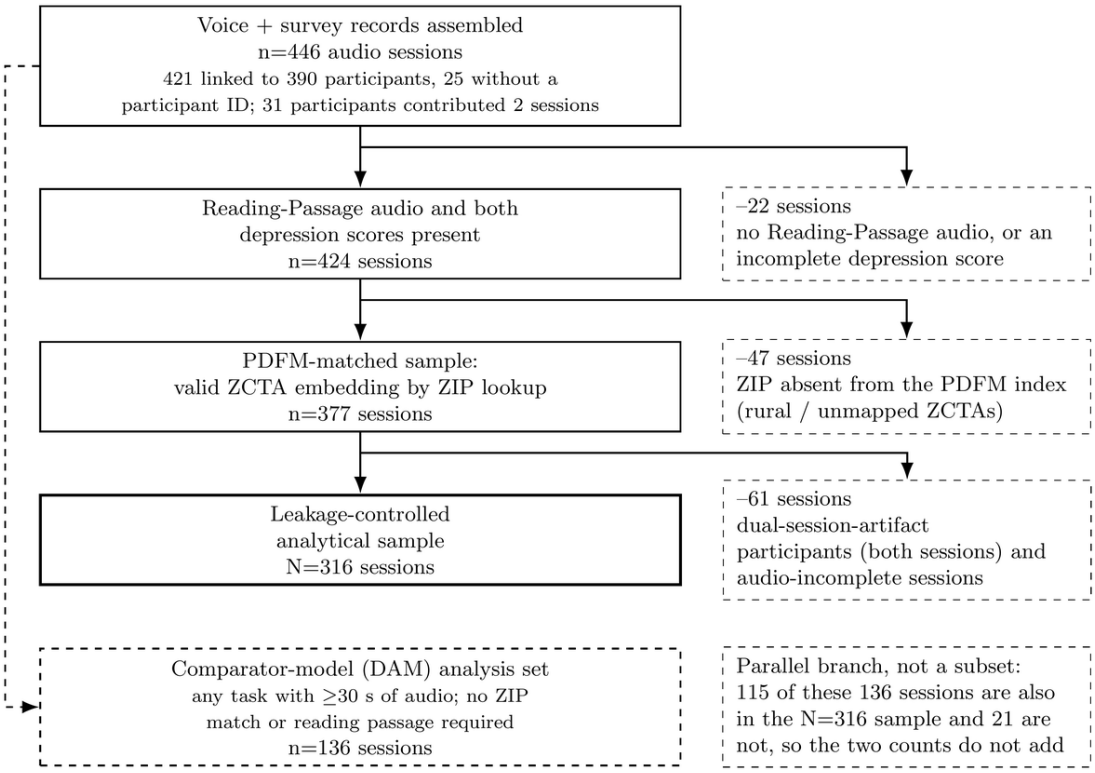
Within each CV training fold, the 375-feature vector was reduced to 20 features by univariate filter feature selection (using SelectKBest with the ANOVA F-statistic

from scikit-learn) with no information from the held-out test fold used in the selection. Technical details of the per-block extraction pipelines, together with documentation of 2 additional feature sources (technology-enhanced language learning linguistic features and Kintsugi Depression/Anxiety Model [DAM] scores [26,27]) that are deliberately excluded from the headline analysis, are mentioned in Multimedia Appendix 2.

Community Health Embeddings

PDFM provides 330-dimensional embeddings per US zip code tabulation area (ZCTA), summarizing community-level socioeconomic conditions, health care access, environmental factors, and health patterns from publicly available data [14]. We merged pretrained embeddings for 32,069 US ZCTAs with participant data; 88.9% (319/390) of participants had a matching ZCTA embedding, and unmatched participants were excluded from the headline analytical sample (Figure 1). As with the acoustic features, the 330-dimensional embedding was reduced to 20 features by SelectKBest fit on the training fold only.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials)–style flow diagram of cohort assembly for the primary (reading passage with a score ≥10 on the 8-item Patient Health Questionnaire) analysis. DAM: Depression/Anxiety Model; PDFM: Population Dynamics Foundation Model; ZCTA: zip code tabulation area.



Critically, PDFM uses only the zip code already present in medical records. It requires no additional demographic questionnaires, no individual-level data collection, and no patient interaction, making it a zero-burden enhancement applicable in any clinical setting with access to a patient's address.

We retained ZCTA-level instead of coarser (eg, urban/rural) granularity; a parallel large-cohort Pregnancy Risk Assessment Monitoring System (PRAMS) analysis indicated that binary urban/rural stratification of PDFM underperforms finer schemes, so coarse aggregation discards signal [28].

Model Development Under Leakage-Controlled CV

Acoustic features and PDFM embeddings were combined by late fusion: a separate StandardScaler → SelectKBest → classifier pipeline was refit per CV fold on each modality's features, and the 2 arms' out-of-fold positive-class probabilities were averaged with equal weight to form the fused score (no meta-learner and no fitted fusion weights). The analysis uses a strictly leakage-free protocol (Multimedia Appendix 4): an sklearn Pipeline containing StandardScaler → SelectKBest → classifier was passed to cross_val_predict with 5-fold stratified CV, ensuring scaling and feature selection are refit on each training fold. Class imbalance was handled by setting the classifier's class_weight parameter to "balanced"; no presplit downsampling was applied. Three evaluation protocols are reported: (1) a parsimonious single-classifier model using L2-regularized logistic regression, which provides the headline estimate; (2) a sensitivity analysis using nested CV that selects between logistic regression, gradient boosting, and random forest candidates inside an inner CV loop on training data only; and (3) an automated machine learning (AutoML) sensitivity analysis using TPOT [29], in which the pipeline

itself (preprocessing, feature transformation, and classifier) is searched rather than fixed in advance.

In (3), outer folds are participant grouped (StratifiedGroupKFold); StandardScaler and SelectKBest (k=20) are fitted on outer-training rows only, and TPOT runs its own internal CV on those rows to select a pipeline, leaving the outer-test partition untouched until prediction; the search budget was 3 minutes per outer fold with the linear-light search space. Protocols (2) and (3) are the 2 modeling paradigms across which the prespecified grid is reported; the full AutoML specification is mentioned in Multimedia Appendix 3. Because the analytical sample retains at most one session per participant (dual-session-artifact participants were removed entirely; Figure 1), CV folds contain no within-participant repeated measures; with one session per participant, grouping by participant identifier (GroupKFold) is equivalent to the stratified folds used here, so no participant-level leakage is possible.

Operating-point metrics used the fixed 0.5 threshold and a sensitivity-targeted threshold. No threshold was tuned to improve reported performance. The sensitivity-targeted threshold was selected inside training folds only: the threshold reaching training-fold sensitivity 80% or greater was chosen on training rows and applied unchanged to held-out rows. This is a leakage control, not an optimization step; choosing an operating point on the same predictions that are then reported would inflate every metric derived from it. Discrimination (area under the curve [AUC]) is threshold-independent in any case, so the reported AUCs do not depend on either choice. Positive predictive value (PPV)/negative predictive value (NPV) were Bayes-recalibrated to 12%-30% prevalence (Table 1). Bootstrap 95% CIs (1000 resamples) were computed for all AUC estimates, with pairwise contrasts by DeLong test [30].

Table 1. Operating-point metrics for the preregistered configuration with deployment-prevalence Bayes recalibration.[a]

Configuration	Prevalence (%)	Sensitivity (%)	Specificity (%)	PPV[b] (%)	NPV[c] (%)	NNS[d]	LR+[e]	LR–[f]
RP[g]/privacy+PDFM[h]/PHQ-8[i] ≥10	55.8	0.638	0.377	0.563	0.452	1.8	1.02	0.96
RP/privacy+PDFM/PHQ-8 ≥10	12	0.638	0.377	0.122	0.884	8.2	1.02	0.96
RP/privacy+PDFM/PHQ-8 ≥10	15	0.638	0.377	0.153	0.855	6.5	1.02	0.96
RP/privacy+PDFM/PHQ-8 ≥10	20	0.638	0.377	0.204	0.806	4.9	1.02	0.96
RP/privacy+PDFM/PHQ-8 ≥10	25	0.638	0.377	0.254	0.757	3.9	1.02	0.96
RP/privacy+PDFM/PHQ-8 ≥10	30	0.638	0.377	0.305	0.708	3.3	1.02	0.96

[a]Positive likelihood ratio (LR+)=1.02 and negative likelihood ratio (LR–)=0.96; essentially noninformative, so positive predictive value (PPV) approximates prevalence and negative predictive value (NPV) approximates 1-prevalence at every deployment prevalence, consistent with an area under the curve of 0.494. Sensitivity, specificity, LR+, and LR– are prevalence independent; PPV, NPV, and number needed to screen are recomputed by Bayes theorem across the 12%-30% deployment range (first row shows in-cohort prevalence). Row format is task/feature set/outcome.
[b]PPV: positive predictive value.
[c]NPV: negative predictive value.
[d]NNS: number needed to screen.
[e]LR+: positive likelihood ratio.
[f]LR–: negative likelihood ratio.
[g]RP: reading passage.
[h]Privacy+PDFM: privacy-tier acoustic features fused with Population Dynamics Foundation Model community-health embeddings.
[i]PHQ-8: 8-item Patient Health Questionnaire.

Negative-Control Harness

Before reading the AUC for any pipeline, we ran a 20-permutation negative-control harness (shuffled-label, random-feature, and joint-shuffle controls); a clean pipeline must produce an AUC of approximately 0.5 in all 3 conditions, and all preregistered and exploratory configurations passed. Full harness specification is mentioned in Multimedia Appendix 4.

External Validation

We evaluated prenatal-trained models on 2 independent general-psychiatric datasets: DAIC-WOZ (n=189) and E-DAIC (n=219), both semistructured clinical interviews of mixed-gender adults, and benchmarked against the Kintsugi DAM, an open-source model fine-tuned from Whisper (OpenAI) on approximately 863 hours of calls from approximately 35,000 individuals (Multimedia Appendix 5) [13,31,32]. All cross-domain AUCs reflect pure inference: no refitting, recalibration, or threshold selection was performed on the target cohorts, and cross-domain operating points used the fixed 0.5 threshold. For the prenatal→external direction, the privacy-tier acoustic block with a leakage-controlled logistic regression pipeline (StandardScaler→SelectKBest(20)→class-weighted L2 likelihood ratio [LR]) was fitted on the reading-passage rows of the analytical sample carrying PHQ-8 labels (n=312; Figure 1) and applied to native DAIC-WOZ and E-DAIC features extracted by the identical pipeline (375/375 feature alignment); PDFM is unavailable externally (no zip codes), so the acoustic arm is used (implemented in the external validation scripts; eg, run_prenatal_to_edaic.py).

Statistical Analysis

Overview

Data were tested for normality with appropriate 2-tailed group comparisons at α=.05; demographic confounding was assessed with effect sizes and inverse probability weighting (Multimedia Appendix 6).

Multiplicity

Two families of tests are adjusted for multiple comparisons: the prespecified 16-cell task×outcome grid and the comparator-model cells evaluated on the prenatal cohort. Within each family, one-sided *P* values for H_0:AUC=0.5 were obtained from the Mann-Whitney statistic on out-of-fold predictions and adjusted by the Benjamini-Hochberg procedure [33]. Benjamini-Hochberg controls the false discovery rate instead of the family-wise error rate; at this sample size, a family-wise correction would have little power against effects of the size at issue. The 2 families are adjusted separately because they address distinct analytic questions. Pipeline scripts, the negative-control harness, and per-(task×threshold) result JSONs are available on reasonable request (see Data Availability section).

Ethical Considerations

The University of California San Diego IRB determined that the study did not constitute research on human participants (determination IRB number 812757). The zip code, which this study uses, can be identifying in combination with other attributes. Participants reviewed and agreed to a digital consent statement on the secure Databoard platform (Chessboard Labs) before any data collection.

Current suicidal ideation was an exclusion criterion applied by the recruitment panel at eligibility screening, before enrollment and before any study instrument was administered; the study team did not administer a suicidality item at any point, and the suicidality item (EPDS item 10) was omitted from both administered instruments. Because no suicidality item was administered within the study, no in-study endorsement could occur, and data collection and a deployed system would require a separate suicidality-management pathway. No compensation details, identifiable demographics, or contact information were retained, and the terms of consent did not authorize sharing of the prenatal voice data.

Results

Participants

The full data assembly comprised 446 sessions from 390 unique participants. Two participant-grouping artifacts defined the analytical sample: 31 participants contributed duplicate sessions submitted 6 to 24 minutes apart, treated as platform artifacts, and excluded; and 16 sessions missing protocol-specified audio were excluded. The analytical sample is 316 participants (Figure 1). Participants were aged approximately 30 (SD 7) years with no age difference between groups. Cohort prevalence was elevated relative to general perinatal populations (PHQ-8≥10 55.8%, modified-EPDS-9 ≥10 73.5%; pooled antenatal prevalence ≈21%) [34], consistent with the overrepresentation of depressive symptoms documented in online research panels relative to national samples [35,36]. Both instruments showed good internal consistency (Cronbach α=0.882 for the PHQ-8 and 0.881 for the mEPDS-9; McDonald ω=0.883 and 0.879, respectively) and agreed with one another at Pearson *r*=0.660 (Spearman ϱ=0.639; Cohen $\varkappa$ at ≥10 =0.41), comparable to the median PHQ-9-EPDS correlation of *r*=0.59 reported across perinatal validation studies [37]. One instrument's continuous score predicts the other's 10 or greater binary at AUC values of 0.80 to 0.84. Together, these results indicate that the mEPDS-9, despite its nonstandard response scale, behaved as a coherent measure of the construct the EPDS is designed to capture. Its internal consistency matches that of the PHQ-8, every item correlates positively with the rest of the scale (corrected item–total *r*=0.33-0.77, lowest for the 2 reverse-scored anhedonia items), and its agreement with the PHQ-8 is above the median reported for the published instruments. This supports using its continuous score as an outcome. It does not validate the 10 or greater and 13 or greater cutoffs for this administration (see Methods section).

Within-Cohort Screening Performance

Overview

Under the leakage-controlled protocol on the analytical sample, the preregistered outcome performed at chance (PHQ-8 ≥10 on the reading-passage task, privacy-tier acoustic features fused with PDFM; AUC 0.494, 95% CI 0.431-0.560). The acoustic arm alone reached 0.452 and the PDFM arm alone 0.541, so the voice increment over community context alone was negative (ΔAUC=−0.047).

At the fixed 0.5 threshold, sensitivity was 0.64 and specificity was 0.38, giving LR+=1.02 and LR−=0.96−. LRs this close to 1 mean a positive or negative result leaves a patient's estimated probability of depression unchanged from baseline. Bayes-recalibrated PPV accordingly approximates prevalence at every deployment prevalence examined (Table 1).

The null was robust to modeling choices. The same cell under the AutoML pipeline-search protocol was also at chance (AUC 0.540, 95% CI 0.477-0.603), so the result does not depend on the choice of classifier or fixing the pipeline in advance. The configuration passed the permutation negative-control harness (real 0.505, shuffled labels 0.508, and random features 0.504), confirming that the null reflects absent signal and not a broken pipeline. The harness applies a single, fixed logistic regression pipeline to the concatenated feature matrix, and not the nested classifier-selection and late-fusion procedure used for the reported estimate, which is why its "real" value (0.505) differs slightly from the headline AUC (0.494); the comparison of interest is between the real run and its own shuffled controls, not against the headline.

The Full Prespecified Grid With Multiplicity Adjustment

Across all 16 task×outcome cells, no cell reached nominal significance under nested CV (maximum AUC 0.550). One cell reached the nominal level under AutoML pipeline search (maximum AUC 0.569, raw P=.02), but did not survive adjustment. After Benjamini-Hochberg adjustment across the grid, no cell survived under either modeling paradigm (Table 2). Under a global null, the expected number of cells clearing the nominal 5% level is 0.8, which is what we observed.

Table 2. Multiplicity across the prespecified 16-cell grid (4 speech tasks×4 outcome definitions) under 2 modeling paradigms.[a]

Paradigm	Max AUC[b] (95% CI)[c]	Cells[b]	Surviving BH[e]
AutoML[f] pipeline search	0.569 (0.502-0.632)	1	0
Nested cross-validation	0.550 (0.488-0.617)	0	0

[a]P is one-sided for H_0 AUC=0.5, from the Mann-Whitney statistic on out-of-fold predictions; adjustment is by the Benjamini-Hochberg procedure across the 16 cells within each row. Under a global null, the expected number of cells clearing the nominal 5% level is 0.8.
[b]AUC: area under the curve.
[c]CIs for the maximum cell are not adjusted for selection across 16 cells.
[d]P<.05.
[e]BH: Benjamini-Hochberg.
[f]AutoML: automated machine learning.

Cell-Level Estimates in This Grid Are Unstable

Repeating the AutoML search under protocol variations reproduces individual cell values only loosely (mean absolute difference 0.036, maximum 0.081, and correlation across the 16 cells r=0.376). The search is stochastic under a wall-clock budget, and cell values additionally shift with the fold-grouping scheme and with whether item-complete subsets are used, so no single cell of this grid should be interpreted on its own. This is why the grid is reported in full with multiplicity adjustment rather than as a single best-performing configuration; selecting a primary outcome by choosing the highest cell would be reading noise.

Sensitivity Analyses

A reduced 325-feature acoustic tier fused with PDFM performed no better (AUC 0.501, 95% CI 0.434-0.565 at the preregistered outcome). Removing the participant-grouping filter did not change the conclusion (n=372, AUC 0.540, 95% CI 0.485-0.600). On the subset with acoustic, DAM, and PDFM features jointly available (n=115), no block among 7 combinations reached significance (maximum 0.576), and adding DAM to voice+PDFM produced no increment in any cell (largest ΔAUC=+0.034, 95% CI –0.012 to +0.083). Adding the individually collected sociodemographic variables (age, education, employment, race, household income, relationship status, and smoking) did not improve performance over voice+PDFM (ΔAUC=+0.002, 95% CI –0.036 to +0.037). The sociodemographic block alone achieved an AUC of 0.508 (95% CI 0.442-0.572), and the addition of semantic transcript features did not improve performance either.

Transfer Depends on the Outcome Instrument, not Only On the Population

Overview

Cross-domain evaluation is a compound shift: population, elicitation task, and outcome instrument all differ at once, so a design that varies them together cannot say which one matters. In one direction, we could hold the population and task constant and vary only the instrument, because our cohort has 2 outcome measures for the same participants. That comparison is the central result of this study.

Prenatal→General-Psychiatric Population: No Transfer

Under pure held-out inference (prenatal-trained leakage-controlled acoustic model; no refitting on target), the model did not separate from chance on either corpus: DAIC-WOZ AUC value was 0.505 (95% CI 0.417-0.593), while the E-DAIC AUC value was 0.478 (95% CI 0.398-0.564). This held across feature-extraction toolkits: our features, OpenSMILE (Open-Source Media Interpretation by Large-Space Feature Extraction) Extended Geneva Minimalistic Acoustic Parameter Set (0.483), and OpenSMILE MFCCs (0.546) were all near chance on E-DAIC (Multimedia Appendix 7), so it is not a toolkit artifact.

General-Psychiatric→Prenatal Population: Transfer Depended on Outcome Instrument

DAM, trained on approximately 35,000 individuals and applied without refitting, reached an AUC of 0.706 to 0.708 on the 2 general-psychiatric corpora. When applied to pregnant women, it reached 0.510 (95% CI 0.411-0.609) against the PHQ-8 and 0.645 (95% CI 0.541-0.744) against the mEPDS-9, in the same participants, on the same recordings, differing only in which questionnaire defined the outcome (Figure 2 and Table 3). The 2 estimates differ from one another on a direct paired test, which the design permits because both are computed on the same participants and recordings (paired bootstrap ΔAUC=+0.135, 95% CI +0.019 to +0.248; unadjusted P=.02; 10,000 resamples; participants were resampled once per iteration and both AUCs were recomputed on that draw so their correlation is preserved). The contrast is significant on all 3 cohort definitions (+0.138, P=.01 at n=149; +0.127, P=.04 at n=115). This contrast was not prespecified, is not part of the 16-cell multiplicity family reported above, and the P value is unadjusted; it requires independent replication. The prespecified 16-cell task×outcome grid and this comparator-model family are

treated as separate multiplicity families because they address distinct analytic questions. The mEPDS-9 estimate survives Benjamini-Hochberg adjustment across the 10 comparator-model-on-prenatal cells examined (4 outcome definitions plus per-task breakdowns) at adjusted P=.005, and it is stable across cohort definitions: 0.651 (95% CI 0.549-0.746) on all 149 participants with DAM-processable audio, 0.645 (95% CI 0.541-0.744) after applying the same participant-grouping exclusions used for the main analysis (n=136, reported here) and 0.642 (95% CI 0.528-0.749) on the 115 participants who additionally meet the reading-passage and zip code–matching requirements of the main analytical sample. The PHQ-8 estimate is 0.510 to 0.515 in all 3 cohort definitions.

Figure 2. Transfer of depression voice biomarkers across cohorts and outcome instruments. Points are area under the curve, bars are 95% bootstrap CIs, and the dashed line marks chance. All cross-domain estimates are pure held-out inference; the prenatal-trained model's 2 prenatal conditions are leakage-controlled cross-validated estimates within its own cohort. The prenatal-trained model (blue circles) does not separate from chance in any condition, including on its own cohort. The general-population model (Depression/Anxiety Model, orange squares) is competent on both general-psychiatric corpora (0.706-0.708), at chance against the 8-Item Patient Health Questionnaire in pregnant women (0.510), and above chance against the modified 9-Item Edinburgh Postnatal Depression Scale in the same women (0.645). The 2 left-hand conditions share participants, recordings, and model and differ only in the questionnaire defining the outcome; that contrast is the study's central finding. As in Table 3, the 2 prenatal conditions for the prenatal-trained model use the acoustic+community-context fusion, while its 2 external conditions use the acoustic block alone, because community-health embeddings do not exist for Distress Analysis Interview Corpus–Wizard-of-Oz or Extended Distress Analysis Interview Corpus. The comparator model is acoustic-only in all 4. CIs are shown instead of a color-coded matrix because several of these estimates are not distinguishable from one another, and a shaded matrix would imply a precision the data do not support. Series are distinguished by hue, marker shape, and direct labels, so the figure remains readable in gray scale. AUC: area under the curve; DAIC-WOZ: Distress Analysis Interview Corpus–Wizard-of-Oz; DAM: Depression/Anxiety Model; E-DAIC: Extended Distress Analysis Interview Corpus; mEPDS-9: modified 9-Item Edinburgh Postnatal Depression Scale; PHQ-8: 8-Item Patient Health Questionnaire.

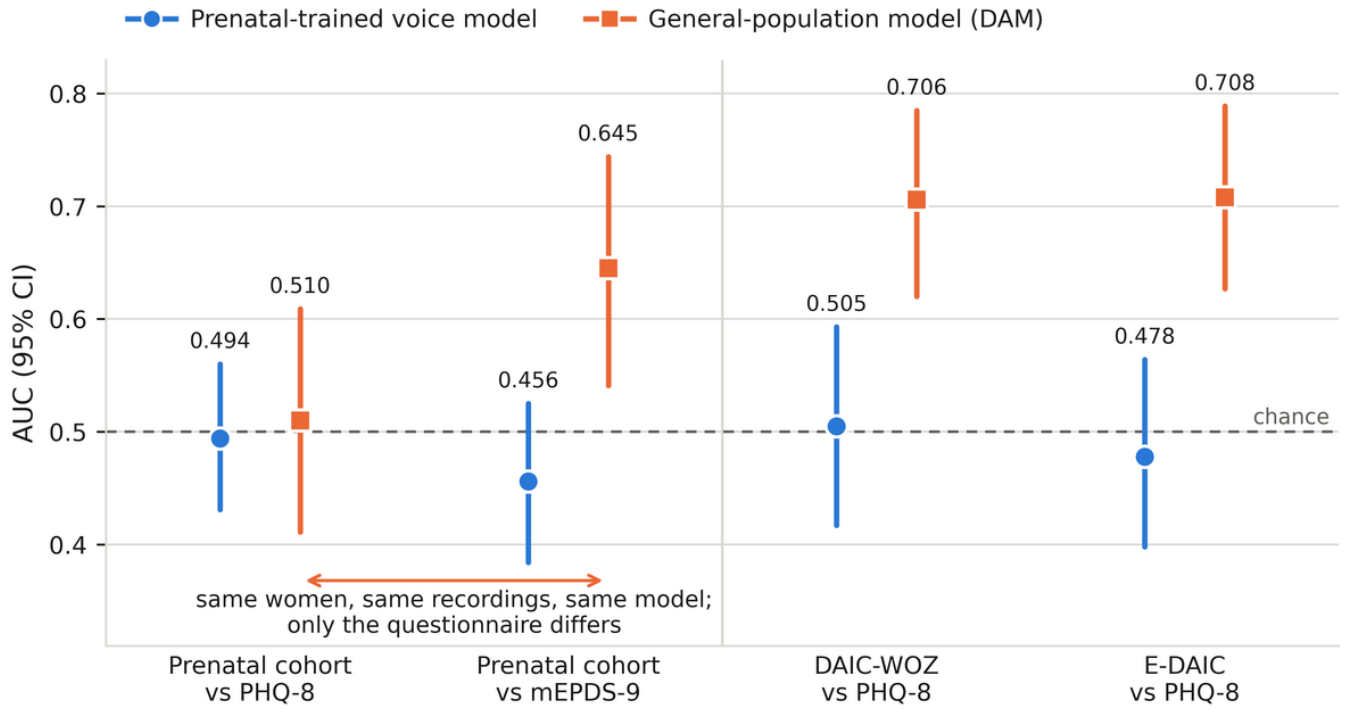


Table 3. Transfer matrix.

Model	Prenatal cohort			
	vs PHQ-8[a]≥10, AUC[b] (95% CI)	vs mEPDS-9[c]≥10, AUC (95% CI)	DAIC-WOZ[d], AUC (95% CI)	E-DAIC[e], AUC (95% CI)
Prenatal, acoustic+PDFM[f,g]	0.494 (0.431-0.560)	0.456 (0.384-0.525)	—	—
Prenatal, acoustic only[h]	0.452 (0.389-0.511)	0.474 (0.402-0.545)	0.505 (0.417-0.593)	0.478 (0.398-0.564)
Kintsugi DAM[i,j]	0.510 (0.411-0.609)	0.645 (0.541-0.744)[k]	0.706 (0.620-0.785)[k]	0.708 (0.627-0.789)[k]

[a]PHQ-8: 8-item Patient Health Questionnaire.
[b]AUC: area under the curve.
[c]mEPDS-9: Modified 9-item Edinburgh Postnatal Depression Scale.
[d]DAIC-WOZ: Distress Analysis Interview Corpus–Wizard-of-Oz.
[e]E-DAIC: Extended Distress Analysis Interview Corpus.

[f]PDFM: Population Dynamics Foundation Model.
[g]Prenatal-trained leakage-controlled model, acoustic features fused with community-health embeddings. Community-health embeddings do not exist for DAIC-WOZ or E-DAIC (no ZIP code), so this row has no external cells.
[h]The same prenatal-trained model using the acoustic block alone (reading passage with PHQ-8 ≥10 training labels). This row is the like-for-like comparison with the Depression/Anxiety Model, which is acoustic-only in every column, and is the only row that can be read across all four columns.
[i]DAM: Depression/Anxiety Model.
[j]DAM applied without refitting; prenatal values are aggregated per participant across tasks. The prenatal DAM set comprises 136 participants: participants with audio meeting DAM's minimum-duration requirement, after the same participant-grouping exclusions applied to the main analysis. It is not nested within the 316-participant analytical sample because DAM requires neither a ZIP-code match nor the reading-passage recording; robustness across 3 cohort definitions is reported in the text.
[k]This CI excludes chance. All cross-domain cells are pure inference, with no refitting, recalibration, or threshold selection on the target cohort. The 2 prenatal cells for the prenatal-trained model are leakage-controlled cross-validated estimates within its own cohort, since a model cannot be applied to its own training cohort without being fitted on it. The 2 prenatal columns differ only in which self-report instrument defines the outcome; the participants, recordings, and models are identical.

In sum, a general-population speech model showed evidence of transfer into pregnancy when evaluated against the mEPDS-9, but not the PHQ-8.

Exploratory Evidence Consistent With Construct Mismatch

The analyses in this paragraph are hypothesis-generating. They were not prespecified, and they rest on correlations of about 0.2 in a subset of 136 participants. DAM's score tracked the part of the mEPDS-9 that the PHQ-8 does not capture: partial r(DAM, mEPDS-9 | PHQ-8)=+0.279 (95% CI 0.11 to 0.43), while partial r(DAM, PHQ-8 | mEPDS-9)=–0.179 (95% CI –0.33 to –0.01). At the item level, across the 17 items examined (mEPDS-9 and PHQ-8), one association survived Benjamini-Hochberg adjustment: crying (r=0.26, adjusted P=.03). The remaining associations were nominal only, but they followed the same pattern. The largest associations were with other affective-distress items: sadness (r=0.21), feeling overwhelmed (r=0.20), and anxiety (r=0.19). Associations with the 2 positively worded anhedonia items were near zero (r≤0.03). We offer this pattern as a candidate explanation that requires replication, not as an established mechanism.

On the PHQ-8, associations were at or near zero for every item except fatigue (r=0.11), with slightly negative associations for appetite change and anhedonia (both r=–0.08); none survived adjustment. The pattern does not appear to reflect anxiety alone: the EPDS anxiety subscale (r=0.185) and depression subscale (r=0.216) correlated with DAM to a similar degree. The PHQ-8 contains substantial somatic and neurovegetative content (sleep, appetite, energy, and psychomotor change), and these symptoms are common in pregnancy for reasons unrelated to mood. If that content contributes disproportionately to the PHQ-8 total in this cohort, a model that responds to affective distress would fail against the PHQ-8 here while succeeding against the same instrument in general-psychiatric adults, where those items are more closely tied to depression.

Transfer Also Varies With Elicitation Task

Within the prenatal cohort and against the same instrument, DAM reached 0.653 (95% CI 0.550-0.753) on spontaneous pregnancy narrative, 0.597 (95% CI 0.486-0.699) on open-ended opinion speech, and 0.544 (95% CI 0.434-0.651) on the scripted reading passage, an ordering that follows similarity to its training distribution of spontaneous speech about personal experience.

Sex Composition Explains None of It

DAM performed comparably on female and male general-psychiatric speech (0.657 vs 0.708 on DAIC-WOZ; 0.662 vs 0.710 on E-DAIC; Δ≈–0.05; P≈0.56), yet dropped to 0.510 against the PHQ-8 in pregnant women. A model that handles female speakers well in one setting but fails on them in another is not failing because of the speakers' sex but because of the model itself. In the forward direction, the prenatal model performed better on female than male external participants (0.587 vs 0.461 on DAIC-WOZ; 0.588 vs 0.417 on E-DAIC). Out of domain, the model score predicts participant sex (AUC 0.62-0.63) better than depression (0.48-0.51). When the forward model was restricted to women, it still did not separate from chance.

How Large Is the Shift the Models Face?

A classifier trained to distinguish prenatal from external recordings, using the same 20 features the transfer model itself selects, reaches AUC values of 0.89 (DAIC-WOZ) and 0.88 (E-DAIC). Domain identity is thus far more recoverable from this feature space than depression, which is why a model fitted in one domain carries so little into another, and why no component of the compound shift can be isolated from these data alone.

Discussion

Principal Findings

Cross-domain evaluation in this study is a compound shift: population, elicitation task, and outcome instrument all differ between cohorts at once. A design that varies them together cannot attribute a failure to any one of them. In this analysis, we were able to isolate one variable, the outcome instrument, in one direction where 2 instruments exist on the same participants. That constraint bounds every cross-domain claim that follows. We therefore describe the external cohort results as evidence of domain specificity, where domain means the population, elicitation task, and outcome

instrument together, and we make no claim about population specificity. The one claim we can make about a single factor is instrument specificity, from the comparison in which participants, recordings, and model are held constant and only the questionnaire differs.

This study yields 4 findings. First, and centrally: whether a depression voice biomarker transfers into a new clinical setting is substantially shaped by the outcome instrument, not by the population alone. A general-population speech model applied without refitting to the same pregnant women, on the same recordings, reached 0.510 against the PHQ-8 and 0.645 against the modified EPDS-9. Nothing about the model, the speakers, or the audio differed: only the questionnaire that defined a screen positive for depression. Second, exploratory item-level associations are consistent with a construct mismatch: the model's score was associated with selected affective-distress items and showed no association with the somatic PHQ-8 items examined in a pregnant population experiencing physical change. Only one of 17 item-level tests survived multiplicity adjustment, so this is a candidate explanation and not a confirmed finding. Third, transfer additionally varies with elicitation task, following similarity to the training distribution (pregnancy narrative>open-ended speech>scripted reading). Fourth, our own prenatal-trained model performed at chance both within our cohort on the preregistered outcome and on 2 independent external databases. No cell of a 16-cell prespecified grid survived multiplicity adjustment.

The practical consequence is that "does this voice biomarker generalize?" is not a well-posed question. It only becomes answerable when population, elicitation task, and outcome instrument are named together. A tool validated against one instrument should not be assumed to work against another, even in the same people.

Neither Voice nor Community Context Discriminates Individuals

For the preregistered outcome, the voice component added nothing to the community context available (ΔAUC=–0.047). The community context was also unable to discriminate between those who would screen positive: the PDFM arm did not clear chance in any of the 16 cells.

A like-for-like comparison sharpens this point. Restricted to the acoustic features alone, the only features on which our model and the general model can be compared directly, the prenatal-trained model reaches 0.452 (PHQ-8) and 0.474 (mEPDS-9), while the comparator reaches 0.510 and 0.645 on the same recordings (Table 3). Both operate on the voice signal alone, without community context, transcripts, or refitting. The difference between them is therefore not the information available in the audio, but what the model learned to extract based on the construct of the outcome measure. This suggests our null result reflects a limitation of the features and fitting procedures, and not evidence that perinatal speech carries no recoverable signal at all.

A graph-based foundation model, pretrained on a large national survey of geospatial context, showed the same absence of individual-level transferability in this cohort (Multimedia Appendix 8). This is an individual-level result and should not be read more broadly. Population-level associations between geospatial context and perinatal depression are well documented [28,38,39], but a context measure can carry such associations and still fail to identify affected individuals. That is the ordinary distinction between ecological and individual inference, and our result speaks only to the latter.

Implications for Validation Practice

Our results argue for 3 changes in how voice-based mental health tools are evaluated. First, validation claims should report population, elicitation task, and outcome instrument together. As our data show, the instrument alone can change a result from chance to discriminatory. Second, a model deployed outside the setting it was validated in gives no internal signal that it has left that setting; DAM's score distribution looks the same whether it is discriminating at 0.71 or at 0.51. This argues for an active detection process for the mismatch, rather than trust in the output. Our data show the information needed for detection exists: domain identity is recoverable at AUC 0.88 from the same features the model uses. Third, the instrument should be chosen to match what the model can plausibly detect. In pregnancy specifically, somatic-weighted instruments are poor targets for acoustic models because gestation alters the somatic items independently of mood.

We do not suggest a deployment scenario for this cohort. With the preregistered outcome at chance and a noninformative operating point (LR+=1.02, LR=–0.96), no configuration examined here supports screening use.

Cross-Domain Nontransfer Is Robust to Audio Format

The cross-domain failure is not an audio-format artifact: format-matching voice-AI audio to DAIC's native 16 kHz PCM and rerunning the comparator model leaves ΔAUC essentially centered on zero, and a symmetric reversal test does not rescue transfer in either direction. Browser-recorded audio in this cohort was, if anything, higher quality than the laboratory reference corpora. Full analyses, including the task-conditional partial-transfer pattern and the pregnancy-specific phonatory substrate, are mentioned in Multimedia Appendix 9.

Comparison With Prior Work

The voice-biomarker literature has repeatedly reported above-chance, sometimes high, within-study discrimination for depression, but external validation across clinical settings is rare, and controls for data leakage are inconsistently applied [5-7,40]. Our results add 2 correctives. First, generalization failure in this literature is routinely attributed to population differences; we show that in the one direction where the two can be separated, the outcome instrument determined whether transfer occurred, with the population held fixed. Reports of population-specific failure should therefore be read with this in mind, since studies

comparing cohorts almost always change instruments at the same time. Second, our own within-cohort estimates collapse to chance under leakage-controlled evaluation with multiplicity adjustment. This is consistent with the view that some previously reported performance reflects analytic flexibility and leakage, rather than genuinely transportable signal.

Limitations

The limitations of this study are as follows. First, the self-report labels are at a single timepoint. Both outcomes are self-administered screening instruments, not clinical diagnoses, collected once. This is the study's most consequential limitation, and it is compounded by the following point. Second, the mEPDS-9 required item-level rescoring. The instrument as administered used a generic response scale in place of the published EPDS anchors, and the responses required rescoring at the item level before analysis. Because the resulting totals are indistinguishable from valid ones at the score-total level, we recommend item-level verification of any self-report label before modeling. Third, there was cell-level instability. Individual cells of the task×instrument grid vary by up to 0.08 across protocol variations, so no single cell should be interpreted; this is why we report the whole grid with multiplicity adjustment and interpret no single best-performing configuration. Fourth, in the post hoc analyses, the instrument-specificity finding, the item-level mechanism, the task ordering, and the sex-stratified analyses were not prespecified. They are hypothesis generating and require preregistered replication. Fifth, the DAM subsample is a different sample; DAM analyses used the 136 participants whose recordings met its minimum-duration requirement and who passed the same participant-grouping exclusions. This set is not nested within the 316-participant analytical sample. The DAM required neither zip code matching nor the reading-passage recording, so the 2 analyses were not drawn from identical participant groups. The instrument contrast is internally consistent because both estimates come from the same participants; the comparison against the within-cohort analysis is not exact. Sixth, there is a compound shift. Population, task, and instrument differ simultaneously between cohorts; we can isolate the instrument only in the direction where 2 instruments were used with the same participants, and we cannot isolate population from task at all. Seventh, we do not have depression history data. Onset timing was not collected, so preexisting and incident antenatal depression cannot be distinguished. Eighth, there are residual demographic differences. Sex is addressed directly, but age could not be matched to the prenatal range within the external corpora without collapsing sample size. Ninth, we used online recruitment selection. Cohort prevalence (53%-74%) far exceeds general perinatal populations; PPV/NPV/number needed to screen are recomputed at 12% to 30% deployment prevalence (Table 1). Tenth, we omitted the suicidality item due to ethical considerations. Additional limitations are listed in Multimedia Appendix 10.

Scope of Inference

The supported claims are narrow, and we state them explicitly. Within this cohort, under leakage-controlled evaluation, neither acoustic features nor community-health context discriminated individuals on the preregistered outcome, and no cell of the prespecified grid survived multiplicity adjustment. Across cohorts, a general-population speech model transferred into pregnancy against one instrument and not another, with an item-level mechanism consistent with that pattern. We do not claim that voice biomarkers cannot work in perinatal populations; we claim that this cohort, with these instruments and these tasks, provides no evidence that they do, and that a substantial part of what looks like population-specific failure in this literature may be instrument-specific.

Future Directions

Three directions follow: (1) modeling, (2) measurement, and (3) design. First, in terms of modeling, parameter-efficient fine-tuning of foundation speech models on target-domain data would test whether the failures we observe are trainable domain gaps. The task-dependent partial transfer we observed shows that the pretrained representations already carry some target-domain signal. The size of the shift we measured motivates unsupervised domain adaptation, and our domain probe indicates how hard that adaptation will be. Domain-invariant training objectives could target the audio-side axes our analysis identifies, population and elicitation task, and the sex-tracking probe suggests an explicit penalty on speaker-attribute information as well. The instrument axis is a property of the label, not the recording, and is better addressed by choosing, or jointly modeling, the outcome instruments. Second, regarding measurement, confirmatory work should use a structured clinical interview as the primary outcome. Where self-report is used, studies should report item-level psychometrics, not totals alone. In perinatal samples specifically, instruments should be chosen or scored so that pregnancy-related somatic content does not dominate the outcome. Third, when considering design, studies should vary population, task, and instrument separately, not together. Our cohort completed 2 instruments, and that is what made the central finding of this paper visible.

Conclusions

Voice-based depression screening is increasingly being developed and is at risk of being commercialized faster than it can be validated. This study demonstrates why that is a risk: whether a depression voice biomarker transfers to a new clinical setting may depend on the instrument tested, not on the population alone. Applied without retraining to a pregnant population, with participants and recordings held constant, the same model reached an AUC of 0.645 against the modified EPDS-9 and 0.510 against the PHQ-8. Our pregnancy-trained model performed no better than chance on any of the populations tested, suggesting no deployable findings for a screening tool used during pregnancy. This suggests that what the outcome measure is detecting is crucial: the EPDS is more focused on psychological manifestations of depression, while the PHQ-9 includes a somatic component that can be confounded by pregnancy changes. This provides important context for interpreting validation models of vocal biomarkers. Validation should specify population, elicitation task, and

outcome instrument together and should not assume that a tool validated against one instrument works against another in the same people.

Acknowledgments

The teams of Cint, TELL, and Chessboard Labs (Databoard) have gone above and beyond in supporting this project. We thank them. We also thank the Population Dynamics Foundation Model team at Google Research for sharing the model weights used in the community-health analyses.

Generative AI Use

The authors used a generative AI tool (GenAI; Anthropic Claude, accessed through the Claude Code interface, during 2025-2026) during the preparation of this work. Following the GAIDeT taxonomy, the following task categories were delegated to the tool under author direction and supervision: Conceptualization: Limited to the revision. The study's conception, research questions, analysis plan, and preregistration precede any use of GenAI. During revision, the tool proposed the reframing of the central claim from population-specific to instrument-specific transfer failure, which the authors evaluated against the data and adopted. Methodology: Limited to the revision. Selection of analysis methods included the multiplicity adjustment procedure, bootstrap contrasts, participant-grouped cross-validation, domain separability, sex attribution probes, and the partial correlation decomposition. The original study design and protocol were author determined without AI involvement. Software development and automation: Generation and refactoring of the feature extraction, cross-validation, and statistical analysis code, including the leakage control and negative control harnesses. Data management: Data cleaning, statistical computation, and figure and table generation. Writing and editing: Drafting and copyediting of manuscript sections, journal format conversion, and preparation of the response to reviewers. Literature review: Assistance with searching for and locating relevant literature. Assessment of the retrieved sources, and all use made of them in the manuscript, are the authors' own. Supervision: Code and results checking. This surfaced the instrument-scoring error that required the item-level rescoring described under the Limitations section, an inconsistency between cohort definitions in the comparator analyses, and a unit error in the flow diagram, each of which the authors then verified independently. Ethics review was not delegated. All interpretation of results, all decisions about what to report, and the final content of the manuscript are the authors' own. No AI system is listed as an author. Every AI-assisted output was reviewed and verified by the authors, who take full responsibility for the integrity and accuracy of this work, including any content produced with AI assistance. MGF delegated the tasks listed above.

Funding

Nurtur Health, Inc partially funded data collection. Beyond the contributions of the authors affiliated with it, which are itemized under Authors' Contributions, the funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. AMG is partially supported by the National Institute On Aging of the National Institutes of Health (R01AG075775, R01AG083799, 2P01AG019724); Agencia Nacional de Investigación y Desarrollo (ANID; FONDECYT Regular 1250317, 1250091); Agencia Nacional de Promocion Cientifica y Tecnologica (01-PICTE-2022-05-00103); and the Multipartner Consortium to Expand Dementia Research in Latin America (ReDLat), which is supported by the Fogarty International Center and the National Institutes of Health, the National Institute on Aging (R01AG057234, R01AG075775, R01AG21051, and CARDS – NIH), Alzheimer's Association (SG-20-725707), Rainwater Charitable Foundation's Tau Consortium, the Bluefield Project to Cure Frontotemporal Dementia, and the Global Brain Health Institute. The other authors report no specific funding for this project.

Data Availability

The prenatal voice dataset is not publicly available due to the terms of the Institutional Review Board determination and participant consent, which did not authorize data sharing. The DAIC-WOZ (Distress Analysis Interview Corpus–Wizard-of-Oz) and E-DAIC (Extended Distress Analysis Corpus) datasets are available through the University of Southern California's Institute for Creative Technologies upon execution of a data use agreement. The Kintsugi Depression/Anxiety Model is open source (Apache 2.0) and available at the Kintsugi Health GitHub repository. Population Dynamics Foundation Model embeddings are publicly available from Google Research. Analysis code, including the instrument scoring and an independent verification script, together with machine-readable result files, are available from the corresponding author on reasonable request. They are not deposited publicly because the result files derived from this cohort contain participant-level outcome values that participant consent does not authorize us to share.

Authors' Contributions

Conceptualization: J Schwoebel, MGF
Data curation: MGF
Formal analysis: J Schwoebel, MC, J Shor
Investigation: MGF

Methodology: RW, MC, AMG, SDF, MGF
Resources: J Schwoebel, MGF
Software: MGF
Writing – original draft: RW, MGF
Writing – review & editing: RW, J Schwoebel, J Shor, AMG, SDF, MGF

Conflicts of Interest

MGF reports employment at Health Stream Analytics, LLC. MGF and SDF report employment at Nurtur Health, Inc, which partially funded the data collection for this study (see Funding). J Schwoebel and J Shor report employment at Quome Inc. MC and AMG report employment at TELL Inc. RW reports no conflicts of interest.

Multimedia Appendix 1

Study design and voice recording protocol.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Feature extraction pipelines.
[DOCX File (Microsoft Word File), 17 KB-Multimedia Appendix 2]

Multimedia Appendix 3

AutoML sensitivity analysis.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Leakage-controlled evaluation protocol.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Kintsugi DAM details.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Bias-aware deployment assessment.
[DOCX File (Microsoft Word File), 19 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Feature extraction technical validation
[DOCX File (Microsoft Word File), 14 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Foundation-GNN (PRAMS-GAT) transferability.
[DOCX File (Microsoft Word File), 14 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Audio format, task content, and the limits of the cross-domain comparison.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Additional limitations.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 10]

References

1. Bauman BL, Ko JY, Cox S, et al. Vital signs: postpartum depressive symptoms and provider discussions about perinatal depression - United States, 2018. MMWR Morb Mortal Wkly Rep. May 15, 2020;69(19):575-581. [doi: 10.15585/mmwr.mm6919a2] [Medline: 32407302]
2. Cox EQ, Sowa NA, Meltzer-Brody SE, Gaynes BN. The perinatal depression treatment cascade. J Clin Psychiatry. Sep 28, 2016;77(9):1189-1200. [doi: 10.4088/JCP.15r10174]

3. Trost SL, Beauregard JL, Smoots AN, et al. Preventing pregnancy-related mental health deaths: insights from 14 US Maternal Mortality Review Committees, 2008-17. Health Aff (Millwood). Oct 2021;40(10):1551-1559. [doi: 10.1377/hlthaff.2021.00615] [Medline: 34606354]
4. ACOG Committee on Clinical Practice Guidelines-Obstetrics. Screening and diagnosis of mental health conditions during pregnancy and postpartum: ACOG Clinical Practice Guideline No. 4. Obstet Gynecol. Jun 1, 2023;141(6):1232-1261. [doi: 10.1097/AOG.0000000000005200] [Medline: 37486660]
5. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol. Feb 2020;5(1):96-116. [doi: 10.1002/lio2.354] [Medline: 32128436]
6. Koops S, Brederoo SG, de Boer JN, Nadema FG, Voppel AE, Sommer IE. Speech as a biomarker for depression. CNS Neurol Disord Drug Targets. 2023;22(2):152-160. [doi: 10.2174/1871527320666211213125847] [Medline: 34961469]
7. Kappen M, Vanderhasselt MA, Slavich GM. Speech as a promising biosignal in precision psychiatry. Neurosci Biobehav Rev. May 2023;148:105121. [doi: 10.1016/j.neubiorev.2023.105121] [Medline: 36914080]
8. Hancock AB, Gross HE. Acoustic and aerodynamic measures of the voice during pregnancy. J Voice. Jan 2015;29(1):53-58. [doi: 10.1016/j.jvoice.2014.04.005] [Medline: 24934324]
9. Ghaemi H, Dehqan A, Mahmoodi-Bakhtiari B, Scherer RC. Voice changes during pregnancy trimesters in Iranian pregnant women. J Voice. May 2020;34(3):358-363. [doi: 10.1016/j.jvoice.2018.09.016] [Medline: 30393048]
10. Rechenberg L, Meurer EM, Melos M, Nienov OH, Corleta H von E, Capp E. Voice, speech, and clinical aspects during pregnancy: a longitudinal study. J Voice. Nov 2024;38(6):1431-1438. [doi: 10.1016/j.jvoice.2022.04.019] [Medline: 35662512]
11. Lu W, Tang X, Huang C, et al. Diagnostic accuracy of traditional and deep learning methods for detecting depression based on speech features: a systematic review and meta-analysis. BMC Psychiatry. Nov 24, 2025;25(1):1190. [doi: 10.1186/s12888-025-07628-z] [Medline: 41286759]
12. Liu W, Li W, Wang Y, et al. Comparison of the EPDS and PHQ-9 in the assessment of depression among pregnant women: similarities and differences. J Affect Disord. Apr 15, 2024;351:774-781. [doi: 10.1016/j.jad.2024.01.219] [Medline: 38290581]
13. KintsugiHealth/dam. Hugging Face. Kintsugi Health; 2024. URL: https://huggingface.co/KintsugiHealth/dam [Accessed 2026-10-29]
14. Agarwal M, Sun M, Kamath C, et al. General geospatial inference with a population dynamics foundation model. arXiv. Preprint posted online on Apr 6, 2026. [doi: 10.48550/arXiv.2411.07207]
15. Wiley RL, Schwoebel J, Caccia M, et al. Cross-domain transfer of depression voice biomarkers depends on the outcome instrument: leakage-controlled cross-sectional evaluation study. arXiv. Preprint posted online on Sep 10, 2026. [doi: 10.48550/arXiv.2505.06412]
16. Cint: The World's Largest Research Marketplace. URL: https://www.cint.com/ [Accessed 2026-10-22]
17. Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. Br J Psychiatry. Jun 1987;150:782-786. [doi: 10.1192/bjp.150.6.782] [Medline: 3651732]
18. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann. Sep 2002;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
19. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord. Apr 2009;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
20. Negeri ZF, Levis B, Sun Y, et al. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. BMJ. Oct 5, 2021;375:n2183. [doi: 10.1136/bmj.n2183] [Medline: 34610915]
21. Levis B, Negeri Z, Sun Y, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) EPDS Group. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. BMJ. Nov 11, 2020;371:m4022. [doi: 10.1136/bmj.m4022] [Medline: 33177069]
22. Schwoebel J. jim-schwoebel/allie: an automated machine learning framework for audio, text, image, video, or.CSV files (50+ featurizers and 15+ model trainers). GitHub. 2021. [Accessed 2026-11-12]
23. Schwoebel JW, Schwartz J, Warrenburg LA, et al. A longitudinal normative dataset and protocol for speech and language biomarker research. medRXiv. Preprint posted online on Aug 24, 2021. [doi: 10.1101/2021.08.16.21262125]
24. Boersma P, Weenink D, Shchupak A. Praat: doing phonetics by computer. Praat. 2021. URL: https://www.fon.hum.uva.nl/praat/ [Accessed 2026-10-22]
25. Jadoul Y, Thompson B, de Boer B. Introducing Parselmouth: a Python interface to Praat. J Phon. Nov 2018;71:1-15. [doi: 10.1016/j.wocn.2018.07.001]

26. García AM, Johann F, Echegoyen R, et al. Toolkit to Examine Lifelike Language (TELL): an app to capture speech and language markers of neurodegeneration. Behav Res Methods. Apr 2024;56(4):2886-2900. [doi: 10.3758/s13428-023-02240-z] [Medline: 37759106]
27. García AM, Ferrante FJ, Pérez G, et al. Toolkit to Examine Lifelike Language v.2.0: optimizing speech biomarkers of neurodegeneration. Dement Geriatr Cogn Disord. 2025;54(2):96-108. [doi: 10.1159/000541581] [Medline: 39348797]
28. Frasch MG. Reducing intersectional disparities in postpartum depression screening through geospatial foundation model embeddings: a cross-sectional study of 209,872 US women. SSRN. Preprint posted online on Mar 2, 2026. [doi: 10.2139/ssrn.6311855]
29. Olson RS, Bartley N, Urbanowicz RJ, Moore JH. Evaluation of a tree-based pipeline optimization tool for automating data science. In: Friedrich T, editor. GECCO '16: Proceedings of the Genetic and Evolutionary Computation Conference 2016. Association for Computing Machinery; 2016:485-492. URL: https://dl.acm.org/doi/proceedings/10.1145/2908812 [Accessed 2026-10-22] [doi: 10.1145/2908812.2908918]
30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. Sep 1988;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]
31. Gratch J, Artstein R, Lucas G, et al. The distress analysis interview corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). European Language Resources Association (ELRA); 2014:3123-3128. [doi: 10.63317/3o7bccg9xequ]
32. Ringeval F, Björn S, Michel V, et al. AVEC 2017: real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. Association for Computing Machinery; 2017:3-9. [doi: 10.1145/3133944.3133953]
33. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc, B: Stat Methodol. Jan 1, 1995;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
34. Yin X, Sun N, Jiang N, et al. Prevalence and associated factors of antenatal depression: systematic reviews and meta-analyses. Clin Psychol Rev. Feb 2021;83:101932. [doi: 10.1016/j.cpr.2020.101932] [Medline: 33176244]
35. Walters K, Christakis DA, Wright DR. Are Mechanical Turk worker samples representative of health status and health behaviors in the US? PLoS One. 2018;13(6):e0198835. [doi: 10.1371/journal.pone.0198835] [Medline: 29879207]
36. Ophir Y, Sisso I, Asterhan CSC, Tikochinski R, Reichart R. The Turker blues: hidden factors behind increased depression rates among Amazon's Mechanical Turkers. Clin Psychol Sci. Jan 2020;8(1):65-83. [doi: 10.1177/2167702619865973]
37. Wang L, Kroenke K, Stump TE, Monahan PO. Screening for perinatal depression with the Patient Health Questionnaire depression scale (PHQ-9): a systematic review and meta-analysis. Gen Hosp Psychiatry. 2021;68:74-82. [doi: 10.1016/j.genhosppsych.2020.12.007] [Medline: 33360526]
38. Lammers S, Lynch CD, Wu J, et al. Neighborhood socioeconomic disadvantage and antenatal depressive symptoms. J Affect Disord. Mar 15, 2026;397:120835. [doi: 10.1016/j.jad.2025.120835] [Medline: 41365450]
39. Onyewuenyi TL, Peterman K, Zaritsky E, et al. Neighborhood disadvantage, race and ethnicity, and postpartum depression. JAMA Netw Open. Nov 1, 2023;6(11):e2342398. [doi: 10.1001/jamanetworkopen.2023.42398] [Medline: 37955900]
40. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). Sep 8, 2023;4(9):100804. [doi: 10.1016/j.patter.2023.100804] [Medline: 37720327]

Abbreviations

AUC: area under the curve
AutoML: automated machine learning
CV: cross-validation
DAIC-WOZ: Distress Analysis Interview Corpus–Wizard-of-Oz
DAM: Depression/Anxiety Model
E-DAIC: Extended Distress Analysis Interview Corpus
EPDS: Edinburgh Postnatal Depression Scale
IRB: Institutional Review Board
LR: likelihood ratio
LR+: positive likelihood ratio
LR−: negative likelihood ratio
mEPDS-9: modified 9-item Edinburgh Postnatal Depression Scale (item 10 omitted)
MFCC: Mel-frequency cepstral coefficient
NPV: negative predictive value
OpenSMILE: Open-Source Media Interpretation by Large-Space Feature Extraction

PDFM: Population Dynamics Foundation Model
PHQ-8: 8-item Patient Health Questionnaire
PPV: positive predictive value
PRAMS: Pregnancy Risk Assessment Monitoring System
TPOT: Tree-based Pipeline Optimization Tool
ZCTA: zip code tabulation area

Edited by Matthew Balcarras; peer-reviewed by Andy Tai, Sijia He; submitted 28.Jun.2026; final revised version received 24.Sep.2026; accepted 25.Sep.2026; published 16.Nov.2026

Please cite as:
Wiley R, Schwoebel J, Caccia M, Shor J, Garcia AM, Fisher SD, Frasch MG
Cross-Domain Transfer of Depression Voice Biomarkers Depends on the Outcome Instrument: Leakage-Controlled Cross-Sectional Evaluation Study
J Med Internet Res 2026;28:e105737
URL: https://www.jmir.org/2026/1/e105737
doi: 10.2196/105737

© Rachel Wiley, James Schwoebel, Matias Caccia, Joel Shor, Adolfo M Garcia, Sheehan D Fisher, Martin Gerbert Frasch. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 16.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.