Assessing Depression Severity From Speech: The Role of Clinical Intuition and Artificial Intelligence

By

Ross Langley

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

At

Dalhousie University
Halifax, Nova Scotia
July 2023

Dalhousie University is located in Mi'kma'ki, the ancestral and unceded territory of the Mi'kmaq. We are all Treaty people.

Table of Contents

List of Tables

List of Figures

ABSTRACT

Automated speech analysis methods are used to estimate depression severity. However, it is unclear how these compare to the intuition of expert clinicians using the same information from speech. We distributed 1-minute speech recordings from 72 participants to 12 clinicians, who estimated depression severity after listening to speech recordings. We trained acoustic and text-based AI models to estimate depression severity in the same samples. Clinicians had a higher agreement to MADRS scores (ICC= 0.47, 95%) than the acoustic-based (ICC = 0.35), text-based (ICC = 0.29), and combined acoustic and text (ICC = 0.33) AI model estimations. However, clinicians had larger errors (RMSE = 10.98) than the text-based (RMSE = 10.02), acoustic-based (RMSE = 10.69), and combined (RMSE = 9.71) AI models. Bias analysis showed clinician gender-based differences in depression estimation. These findings provide the first direct comparison of clinical intuition and AI estimation of depression severity from speech.

LIST OF ABBREVIATIONS USED

Artificial Intelligence	AI
Bidirectional Encoder Representations from Transformers	BERT
Concordance Correlation Coefficient	CCC
Deep-neural Network	DNN
Intraclass Correlation Coefficient	ICC
False Discovery Rate	FDR
Generalized End-To-End	GE2E
Long Short-Term Memory	LSTM
Major Depressive Disorder	MDD
Measurement-Based Care	MBC
Mel-Frequency Cepstral Coefficient	MFCC
Montgomery–Åsberg Depression Rating Scale	MADRS
Natural Language Processing	NLP

CHAPTER 1 INTRODUCTION

1.1 Rationale

Depression is among the leading causes of health-related burdens, affecting 11.2% of Canadians over their lifetime (Knoll & MacLennan, 2017). This has been exacerbated by the COVID-19 pandemic, marked by changing social and economic determinants of mental health, increased prevalence of depression, as well as barriers to timely access to care (COVID-19 Mental Disorders Collaborators, 2021).

With this comes the need for improved methods of recognizing and treating depression that can be applied to our changing needs and modes of healthcare delivery. Measurement-based care (MBC) can leverage information from outcome scales and behavioural biomarkers, such as patient speech recordings, to help clinicians to address these challenges. Speech is a viable biomarker of depression, as it is rapid to collect, affordable, and contains information on psychopathology (Kappen et al., 2023). While speech is a promising resource to aid clinicians, ensuring speech measures are valid, reliable, and unbiased is essential to ensure quality, equitable care for patients.

1.2 Measurement-Based Care

1.2.1 Measurement-based Care in Psychiatry

Measurement-based care is the application of standardized, objective measurements to track care and inform treatment plans. The use of MBC offers improvements for psychiatric patients compared to usual care, including higher response and remission rates for depression treatment and better adherence to treatment plans (Hong et al., 2021; Fortney et al., 2017; Scott & Lewis, 2015; Trivedi et al., 2006).

Psychiatric illness is complex, with fewer objective health outcome measurements than other medical fields. Validated outcome scales are the foundation of MBC in psychiatry (Hong et al., 2021). When evaluating clinical scales, some essential characteristics include sensitivity to change, time investment, mode of recording, scoring complexity, inter-rater reliability, and training requirements (Kroenke & Spitzer, 2002).

Clinicians use MBC to improve the delivery of care through several key factors. First, MBC helps clinicians by streamlining the detection and monitoring of depression (Kroenke & Spitzer, 2002). Outcome measures can also help patients understand and quantify their experiences and depressive symptoms (Dowrick et al., 2009). MBC encourages patient involvement in decision-making and treatment plans, which is especially important for those at risk of dropping out or not responding to treatment (Fortney et al., 2017; Scott & Lewis, 2015; Wright et al., 2020). In addition, the process of recording measurements and receiving feedback can be therapeutic. Receiving treatment feedback improved progress for patients compared to those who did not (Hamilton & Bickman, 2008). Recognizing their symptom improvements creates optimism and helps treatment adherence (Fortney et al., 2017).

Despite the effectiveness of MBC in psychiatry, it remains an underused resource (Wright et al., 2020; Zimmerman & McGlinchey, 2008). The primary reported barrier to adopting MBC in busy clinical environments is the time required to conduct, review, and interpret scales (Scott & Lewis, 2015; Zimmerman & McGlinchey, 2008). Cost is another reported barrier, with technology and resources required for large-scale monitoring of patients (Scott & Lewis, 2015). Typical responses for not using MBC also include lacking knowledge or training in appropriate scales (Zimmerman & McGlinchey, 2008). Understanding common barriers to implementing MBC is important when designing measures to increase use by clinicians.

1.2.2 Biases in Measurement-Based Care

While the administration of MBC can aid in quantifying outcomes, responses can be affected by recall biases, cognitive limitations, and social stigma (Low et al., 2020; Dowling et al., 2016; Sung & Wu, 2018; Vetter & Mascha, 2017). Additionally, patients may be aware of the effects of their responses concerning therapeutic options, discrimination, or stigma (Hunt et al., 2003; Logan et al., 2008; Pavlova & Uher, 2020).

MBC can also introduce systematic biases due to sampling or measurement instruments for which responses vary for individuals based on personal characteristics instead of intended measurable symptoms (Vetter & Mascha, 2017; Gonzalez et al., 2021). For example, the mode of measurement can influence how a patient responds to measures. Discrepancies between self-report and clinician scales can be affected by symptom severity or demographic factors, including gender, age, and education (Carter et al., 2010; Hartmann, Fritzche, & Lincoln, 2012; Enns, Larsen, & Cox, 1999). In addition, biases may be introduced through the design of specific items on scales that measure experiences or symptoms that differ across individuals (Yang & Jones, 2007; Salokangas et al., 2002). For example, Dmitrieva and colleagues (2015) found response differences based on patient age, sex, race, ethnicity, and years of education for items on commonly used depression symptom scales, including MADRS, CES-D, and GDS, although these did not result in meaningful differences in overall depression severity scores.

Understanding sources of bias for measurements is essential to ensure equitable treatment for patients.

1.3 Artificial Intelligence

1.3.1 Artificial Intelligence and Healthcare

Artificial Intelligence (AI) is changing healthcare approaches across medical fields, and in many cases, can perform tasks as well or better than clinicians (Brunn et al., 2020). There are increasing strains on healthcare systems, with physician shortages and worsening outcomes. However, the increasing amounts of patient data and improving abilities to apply AI may help healthcare providers meet some of these needs (Lee et al., 2021; Topol, 2019). Integrating AI into healthcare might help rapid, scalable, and affordable disease detection and monitoring, improving the reach and impact of care (Afrose et al., 2022; Timmons et al., 2022; Weissglass, 2022). AI may streamline processes, freeing more time for clinicians to spend with patients (Lee et al., 2021; Weissglass, 2022). In addition, AI may improve clinical judgement by performing complementary tasks. AI models may be able to recognize patterns to help differentiate diagnoses with similar presentations, identify new illness subtypes, and predict the onset of illness (Lee et al., 2021).

However, despite the promise of AI, there are concerns that expectations may exceed the current technological capabilities (Topol, 2019). A recent review of healthcare machine learning models classified over 90% of published studies at low levels of readinesss, finding that they were at the level of basic laboratory research (Monteith et al., 2022). Additionally, it is challenging to incorporate AI into psychiatric medicine because diagnoses are less reliant on objective, specific measures than in other healthcare fields, and there are fewer large, high-quality datasets compared to other health fields (Lee et al., 2021).

1.3.2 Artificial Intelligence Biases

AI models can provide objective assessments of patients and potentially reduce inequities in care; however, clinical models trained on biased data may produce biased outcomes. Health inequities can be perpetuated or even amplified through AI prediction models (Ganel et al., 2022; Timmons et al., 2022). For example, algorithms used to determine risk and treatment program enrollment in US healthcare demonstrated racial biases, with Black patients receiving lower risk scores than White patients at the same level of illness, which can impede enrollment in treatment programs (Obermeyer et al., 2019).

Sampling is one of the primary contributors to bias in AI models. Typically, models optimize performance for groups of people that occur more frequently in datasets, which can reduce generalizability to underrepresented groups of people in research (Zou & Schiebinger, 2018). Research studies are typically performed on White, high-income individuals (Roberts et al., 2020). Sampling-based disparities in AI models are found based on race, gender, age, and socioeconomic status (Akinhanmi et al., 2018; Timmons et al., 2022; Seker et al., 2022; Spencer et al., 2013; Weissglass, 2022; Buolamwini & Gebru, 2018; Larrazabal et al., 2020; Seyyed-Kalantari et al., 2021; Thake & Lowry, 2017).

1.4 **Clinical Intuition**

1.4.1 Dual-Process Decision-Making

Clinicians rely on clinical intuition, or their “gut feelings” to aid decision-making, developing insight without an explicit or conscious process, described as “knowing without knowing how” (Pearson, 2013). However, its use is often underestimated or considered unscientific (Pearson, 2013). The dual-process model explains clinical decision-making as a combination of intuition (System 1) and analytic reasoning (System 2) (Melin-Johansson et al., 2017). System 1 involves rapid, subconscious decision-making built from knowledge and experience. System 1 and 2 decision-making have been proposed to work interactively, where intuition is used to generate hypotheses and acquire cues, which provides a framework and guidance for decision-making (Brush et al., 2017; Price et al., 2017).

There are mixed opinions on the utility of intuition, with some proposing that its reliance on heuristics and “non-scientific” nature leads to increased diagnostic errors. In contrast, others view System 1 thinking as a valuable complement to analytical reasoning, helping guide medical approaches (Vanstone et al., 2019). In some cases, clinicians using rapid, intuitive decision-making had improved accuracy compared to clinicians using slower, more analytical approaches (Norman, 2009; Price et al., 2017). In real-world tests of patient outcomes, measures of clinical intuition have predicted diagnoses, hospital readmission, patient deterioration, and the long-term onset of disease (Haegdorens et al., 2023; Hoops et al., 2020; Pentzek et al., 2019; Zanovello et al., 2020). However, others argue that the strength of clinical intuition alone is insufficient to guide medical decisions (Cabrera et al., 2015).

1.4.2 Clinician Biases

In psychiatry, clinical decisions are often made by subjective interpretation of symptoms, leaving potential opportunities for bias. In addition, decision-making under time pressure, cognitive overload, and fatigue increasingly rely on heuristics or mental shortcuts (Dehon et al., 2017). Heuristics may be necessary in busy clinical environments but can also be influenced by implicit biases (Dehon et al., 2017). Implicit biases result from unconscious stereotypes or beliefs that affect our thoughts and perceptions, often involving automatic associations about group traits (Payne & Hannay, 2021; Schnierle et al., 2019).

Individuals experience differences in quality or access to care based on race or ethnicity, affecting a range of patient outcomes. A review of racial bias in the US healthcare system found that many clinicians have pro-white implicit biases, which can affect patient care (Dehon et al., 2017). Clinicians with higher levels of pro-white implicit bias provide lower-rated care, with less patient satisfaction for Black patients (Chapman et al., 2013).

Implicit biases about gender can also affect the perception of symptoms and diagnoses. In a study on pain treatment, Samulowitz and colleagues (2018) found gender bias in the perception of pain in patients relating to gender stereotypes of pain expression. Clinicians perceived women as more sensitive and likely to report pain (Samulowitz et al., 2018). Due to gender biases, women are less likely to receive psychiatric diagnoses and therapeutic interventions (Garb, 2021; Hamberg, 2008).

While other demographic characteristics may remain stable for an individual, ageism is unique because the risk of experiencing ageism changes over all individuals' lifespans if they live long enough (Ayalon & Tesch-Römer, 2017). Age can be used to justify limiting access to resources, paternalizing interactions, or preventing engagement in healthcare decisions (Nemiroff, 2022). Ben-Harush and colleagues (2016) interviewed healthcare professionals across healthcare specialties and found negative attitudes and biases against older individuals. These widespread biases can affect health outcomes, as age discrimination against older patients is associated with new or increased disability (Ayalon & Cohn-Schwartz, 2022; Rogers et al., 2015).

While patients often experience different care based on their personal characteristics, these effects may be moderated or amplified by the characteristics of their clinicians. Tajeu and colleagues (2018) argue that one of the causes of worse outcomes for minority patients is underrepresentation in healthcare professionals, resulting in racial discordance between clinicians and patients. Studies reveal higher levels of pro-white implicit biases in White physicians compared to Black physicians. Race discordance can also affect communication, with higher levels of implicit bias corresponding to changes in the language and communication style used by White clinicians toward Black patients (Hagiwara et al., 2017).

The effect of clinician gender on bias and care has also been investigated. Levels of pro-white bias have also been found to be higher in male physicians (Chapman et al., 2013). Sniezynski & Bkazewicz (2016) tested the effects of therapist gender on clinical decision-making toward male and female patients. They found that male therapists rated different levels of psychological functioning and recommended more involved treatments for female patient vignettes than identical male vignettes. These findings indicate that clinician characteristics may influence the effects of bias toward patients.

1.5 Speech and Depression

1.5.1 Clinical Use of Speech and Language

Clinicians have long recognized differences in the speech of patients with depression. Early research describes the slow, halting speech of patients with depression, with raspy quality and short responses (Kraeplin, 1921; Newman, 1938). Further research explored observable differences in the timing volume, pitch, variation, and quality of speech, as well as changes coinciding with fluctuations in depression severity (Breznitz, 1992; Greden & Carroll, 1980; Sobin & Sackeim, 1997; Widlöcher, 1983).

Interpreting the sound and timing of patient speech is essential to diagnostic interviews and assessments. Psychomotor retardation is listed in the DSM-5 as a diagnostic criterion for major depressive disorder (MDD), which can present through "slowed speech … increased pauses before answering; speech that is decreased in volume, inflection, amount, variety of content, or muteness" (DSM-5).

Clinicians also rely on speech content, asking questions and making decisions on diagnoses, treatment, and progress from the answers provided (Pavlova & Uher, 2020). Patients intentionally and unconsciously communicate information through their descriptions of experiences and internal states, as well as their word choices and language patterns (Smirnova et al., 2018).

1.5.2 Automated Assessment of Depression from Speech

Speech recordings are an attractive biomarker for the objective assessment of the presence or severity of depression. Recording speech is a rapid, non-invasive measure that is easily scalable and contains rich signals in both the sound and content of the speech (Low et al., 2020). In addition, speech can be recorded without explicitly asking about depressive symptoms. As speech processing and prediction models improve, we can better apply the information in speech as an objective biomarker of depression (Cummins et al., 2015).

1.5.3 Speech Acoustics and Depression

Speech production is a complex process, coordinating cognitive planning and neurophysiological actions, which can be affected by cognitive impairments and neurological dysregulation associated with depression. Speech recordings can be broken down into essential acoustic components to independently measure prosodic, source, or spectral speech features (Cummins et al., 2015).

The prosody, or rhythm of speech, is among the earliest objective measure of depressive speech. Individuals with depression typically speak slower, with longer, irregular pausing (Cannizzaro et al., 2004; Mundt et al., 2012; Stassen et al., 1998). Depressive speech also correlates to a smaller range and variation of pitch values, often measured through fundamental frequency (F0). Patients with depression are also observed to speak more quietly, with little variation in the energy of their voice, but studies on the changes in energy in depressed speech have found mixed results (Darby et al., 1984; Cummins et al., 2015).

Source features measure the activity and coordination of vocal production in the glottis. Glottal features of speech measure the timing and amplitude of glottal pulses caused by the vibration of vocal cords. These features are affected by the neuromuscular coordination of the vocal tract during speech and are correlated to the observable changes in voice quality. Jitter, the variation in vocal cord timing, and shimmer, the variation in amplitude of vocal cord vibrations, are increased in depressive speech (Quatieri & Malyska, 2012).

Spectral features measure energy distribution across a voiced sound's frequency spectrum. Some of the most common spectral features are the Mel-frequency cepstral coefficients (MFCC). MFCCs are representations of speech power spectra based on a non-linear

Mel-scale of frequency and have been used to predict diagnoses and severity of depression (Du et al., 2023; Hansen et al., 2022; Taguchi et al., 2018; Zhao et al., 2022).

Recently, speaker embeddings have been proposed to improve the ability of AI networks to estimate depression severity (Dumpala et al., in press; Liu et al., 2023). Speaker embeddings are extracted using acoustic-based speech parameters and were designed initially for speaker recognition systems. Speaker embedding models can be pre-trained on non-clinical datasets, which allows leveraging information from large databases on non-labelled speech recordings (Dumpala et al., in press). The rationale behind speaker embeddings comes from the changes in speech from individuals in a depressive episode, often described as sounding like a different person. If speaker embeddings can use acoustic characteristics to differentiate between speaker identities, they may recognize the identity of depression in speech. Dumpala and colleagues trained LSTM models from commonly used acoustic features from the OpenSMILE Interspeech 2009 feature set, speaker embeddings, and a combination of the inputs. The speaker embedding model outperformed the acoustic feature model, with the best performance from the combined embedding and acoustic feature model (Dumpala et al., in press)

1.5.4 Language and Depression

Information on our mental state can be revealed from the content and form of our speech (Corona Hernández et al., 2023). Observed patterns of language in depressed individuals include increased use of absolutist words, self-reference, increased word count, and negatively valanced words (Al-Mosaiwi & Johnstone, 2018; Rude et al., 2004; Himmelstein et al., 2018). Natural Language Processing (NLP) models leverage the differences in syntax and text of language to measure psychiatric disorders (Corona Hernández et al., 2023). These models can use language from various sources, from social media posts to diagnostic interviews. NLP models can extract text-based information from speech recordings using pre-trained models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). AI networks can be trained using these text features to predict diagnoses of depression or estimate severity levels (Li et al., 2022).

1.5.6 Clinical Intuition and Automated Assessment of Speech

To our knowledge, the ability of clinical intuition to estimate depression severity solely from a short speech recording has not been tested, and the ability of clinical intuition has not been directly compared against the AI-based estimation of depression severity from speech. Clinicians rely on their intuition to subjectively interpret depressive symptoms through the sound and content of speech when speaking with patients. Additionally, language and acoustic-based AI models have the potential to aid clinicians in assessing and monitoring depression. Furthermore, speech is a viable target for incorporating as a biomarker of depression into measurement-based care (Kappen et al., 2023). Given this, measures of estimating depression severity from speech using clinical intuition or AI models should be examined. We look to answer to what extent information on depression severity can be extracted by expert clinicians using short speech samples and compared to language and acoustic-based AI networks.

Through this study, we have several aims:

1. Determine the accuracy and reliability of clinical intuition in estimating depression severity using short speech recordings.

2. Compare clinicians' ability to acoustic-based and text-based AI models to estimate depression severity from speech recordings.

3. Explore potential sources of bias for the clinician and AI-based depression severity estimations.

CHAPTER 2 METHODS

2.1 Participants

We recruited participants for this study from the Canadian Biomarker Integration in Depression (CANBIND) network, the Canadian Depression Research and Intervention Network (CDRIN), and the Families Overcoming Risk and Building Opportunities for Well-being (FORBOW) cohort. All three studies are enriched for participants living with severe mental illness (MDD, bipolar disorder, or schizophreniform disorder). Key inclusion criteria for participants include a diagnosis of MDD, age older than 18, and speaking English as a first language. The exclusion criteria include an inability to give informed consent and a comorbid neurological or language disorder diagnosis.

2.2 Assessments

2.2.1 Depression Assessment

We rated depression severity for participants through the MADRS interview, conducted by graduate students and research staff trained and supervised by clinical psychologists and psychiatrists. The MADRS is a 10-item clinician-rated scale with each item scored from 0 (no symptom present) to 6 (severe or continuous symptom presence) for a total score from 0 to 60, with higher scores reporting increased depression severity. The MADRS is considered a gold standard measure of depression severity and has been found to have good reliability, validity, and diagnostic accuracy (Ntini, 2020).

2.2.2 Speech recording

We record participants' speech samples on the same day as the MADRS interview, in which they provide a prompted autobiographical naturalistic speech sample. Participants sit 2 feet away from a TASCAM DR-05X audio recorder and respond to a neutral, positive, and negative prompt. The assessor will explain the task and allow the participant to ask questions, then will ask the following questions, allowing 3 minutes for each answer, re-prompting the participant if needed:

1. Neutral prompt: *"First, I would like to hear about how your last couple of weeks have been and how you've been spending your time. Tell me how you've been feeling and what you've been up to lately."*

2. Positive prompt: *"Next, I want you to think about a time in the past few weeks when things went well for you. Think about when you had a positive experience or when something good may have happened to you. Take your time to think about it, and you can go ahead whenever you're ready."*

3. Negative prompt: *"Lastly, I now want you to think about a time in the past few weeks when things didn't go well for you. Think about when you had a negative experience or when something bad may have happened to you. Take your time to think about it, and you can go ahead whenever you're ready."*

We processed the speech recordings, removing the interviewer's speech and identifying information. We isolated the first minute of uninterrupted participant speech from the neutral prompt.

2.2.3 Demographic information

Participants recorded demographic information as part of baseline assessments. We asked for information regarding age, biological sex, race and ethnicity, and socioeconomic information, including household income, education level, and location of residence (Table 1).

Table 1: Participant demographics.

	Overall (N=72)
Sex	
Male	19 (26.4%)
Female	53 (73.6%)
Age	
Mean (SD)	41.8 (15.1)
Median [Min, Max]	43.5 [18.0, 81.0]
Race	
White	64 (88.9%)
Black	6 (8.3%)
Asian	1 (1.4%)
Education	
No Post-Secondary	15 (20.8%)
Post-Secondary	52 (72.2%)
Household Income	
<$40,000	40 (55.6%)
>$40,000	30 (41.7%)

2.3 Clinician Estimation of Depression

To test the ability of clinicians to predict depression severity, we distributed speech recordings to 12 clinicians through the CANBIND network. We contacted experienced clinicians specializing in mood disorders with 7 to 40 years of clinical experience (mean = 19.6, sd = 10.3)

(Table 2). The clinicians work as psychiatrists (N = 11) and clinical psychologists (N=1) (Table 2).

Table 2: Clinician characteristics

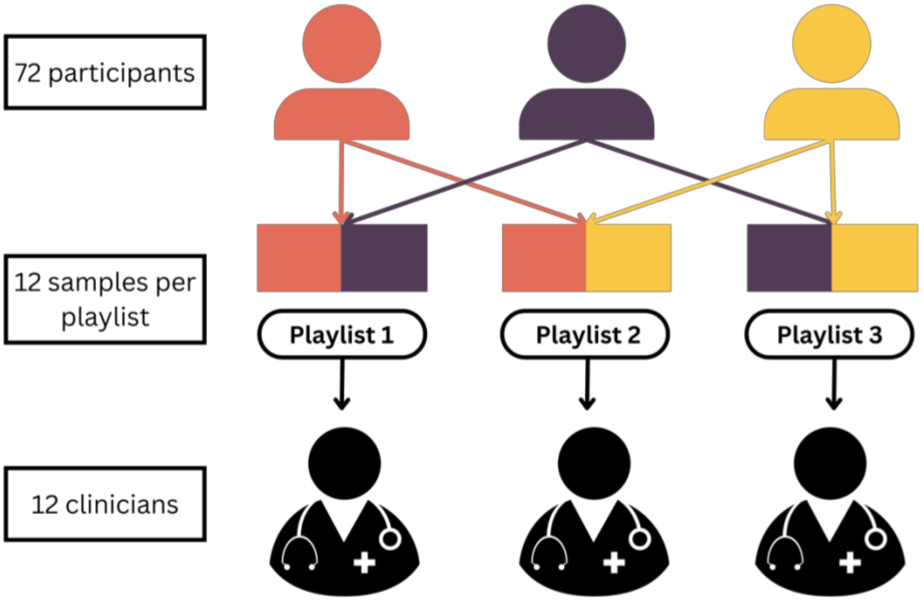
	Overall (N=12)
Gender	
Female	4 (33.3%)
Male	8 (66.7%)
Profession	
Clinical psychologist	1 (8.3%)
Psychiatrist	11 (91.7%)
Age	
Mean (SD)	46.2 (17.8)
Median [Min, Max]	47.0 [0, 64.0]
Years of Experience	
Mean (SD)	20.9 (10.2)
Median [Min, Max]	21.0 [7.00, 40.0]

To obtain rapid, secure ratings of depression severity, we developed a novel clinician estimation measure, which we hosted through Amazon AWS (Cloud Computing Services - Amazon Web Services [AWS], 2023). Clinicians were given a unique link and password to access a playlist of speech recordings. They listened to 1-minute audio recordings, then were prompted to estimate the depression severity of the speaker. Clinicians were provided a scale from 0 - 45, with anchor points corresponding to MADRS mild, moderate, and severe depression severity cut-offs (Hawley et al., 2002; Snaith et al., 1986). After finalizing the depression severity prediction for a speech recording, clinicians proceeded to the following speech sample and could not retroactively change predictions.

A simplified diagram of the speech sample distribution is shown in Figure 1. We selected 72 participants who met the inclusion and exclusion criteria. We aimed to generate a database of speech recordings that covered a range of depression severity, so we sampled participants for a uniform distribution of MADRS scores. We randomly sampled 24 participants with sub-threshold depression severity scores (MADRS < 10), 24 participants with mild depression severity scores (10 ≤ MADRS < 20), and 24 participants with moderate (20 ≤ MADRS < 35) or severe depression scores (MADRS > 35) (Hawley et al., 2002; Snaith et al., 1986).

To test the similarity of ratings between clinicians, each speech recording was rated by two clinicians for a total of 144 depression severity estimations. These recordings were allocated to 12 unique playlists, each containing 12 speech recordings. To minimize the risk of bias across playlists, we allocated 4 speech recordings of participants with subthreshold depression scores, 4 participants with mild depression scores, and 4 participants with moderate to severe depression scores to each playlist.

Figure 1: Speech recording playlist distribution for clinicians



2.4 Artificial Intelligence Estimation of Depression

2.4.1 Training Dataset

To train the acoustic and text-based AI models, we developed a separate training speech recording database of 418 independent speakers. Speakers recorded speech samples following the same procedure as the 72 participants in our testing dataset. However, this training dataset did not have inclusion or exclusion criteria for the diagnosis of mood disorders and was not sampled for a uniform distribution of depression severity. As a result, the training dataset has a low average depression score, with a large proportion of patients with subthreshold depressive severity scores (Table 3). In addition, because of sample size limitations in the training set, we had few participants with moderate or severe depressive symptoms. We trained the AI models using an ensemble approach to resolve training set imbalances. We ran 3 models, under-sampling the training data. We prioritized participants with higher depression severity scores and de-prioritized participants with low depression severity scores. We computed depression severity scores by averaging the model predictions for each participant.

Table 3: Testing and training datasets

	Testing Set (N= 72)	***Training Set (N = 418)***
Age (SD)	41.8 (15.1)	40.2 (12.9)
Sex		
Female (%)	53 (74%)	272 (66%)
Male (%)	19 (26%)	143 (34%)
MADRS Score (SD)	15.1 (11.2)	7.0 (8.2)

2.4.2 Acoustic-based depression estimation model

We developed a combined acoustic feature and speaker embedding-based depression severity prediction model according to the approach outlined by Dumpala and colleagues (in press). First, we pre-trained the generalized end-to-end (GE2E) network for speaker identification on a combined set of speech recordings from the LibriSpeech and VoxCeleb2 datasets to extract d-vector speaker embeddings (Chung et al., 2018; Panayotov et al., 2015).

Before training our model, we segmented speech recordings in the training and test datasets into non-overlapping 5-6 seconds of speech. We extracted segment-level d-vector speaker embeddings using the pre-trained speaker identification network and acoustic features using the OpenSMILE Interspeech 2009 feature set. This acoustic feature set included variables derived from energy, pitch, harmonic-to-noise ratio, and MFCCs (Schuller et al., 2009).

We combined segment-level speaker embeddings and acoustic features using a 2-branch long short-term memory (LSTM) network. To estimate depression severity, we extracted segment-level depression severity predictions and calculated a mean depression severity score. We trained our acoustic-based LSTM network using a mean square error loss function on our training set of speech recordings, and we estimated depression severity using our testing set.

2.4.3 Text-based depression estimation model

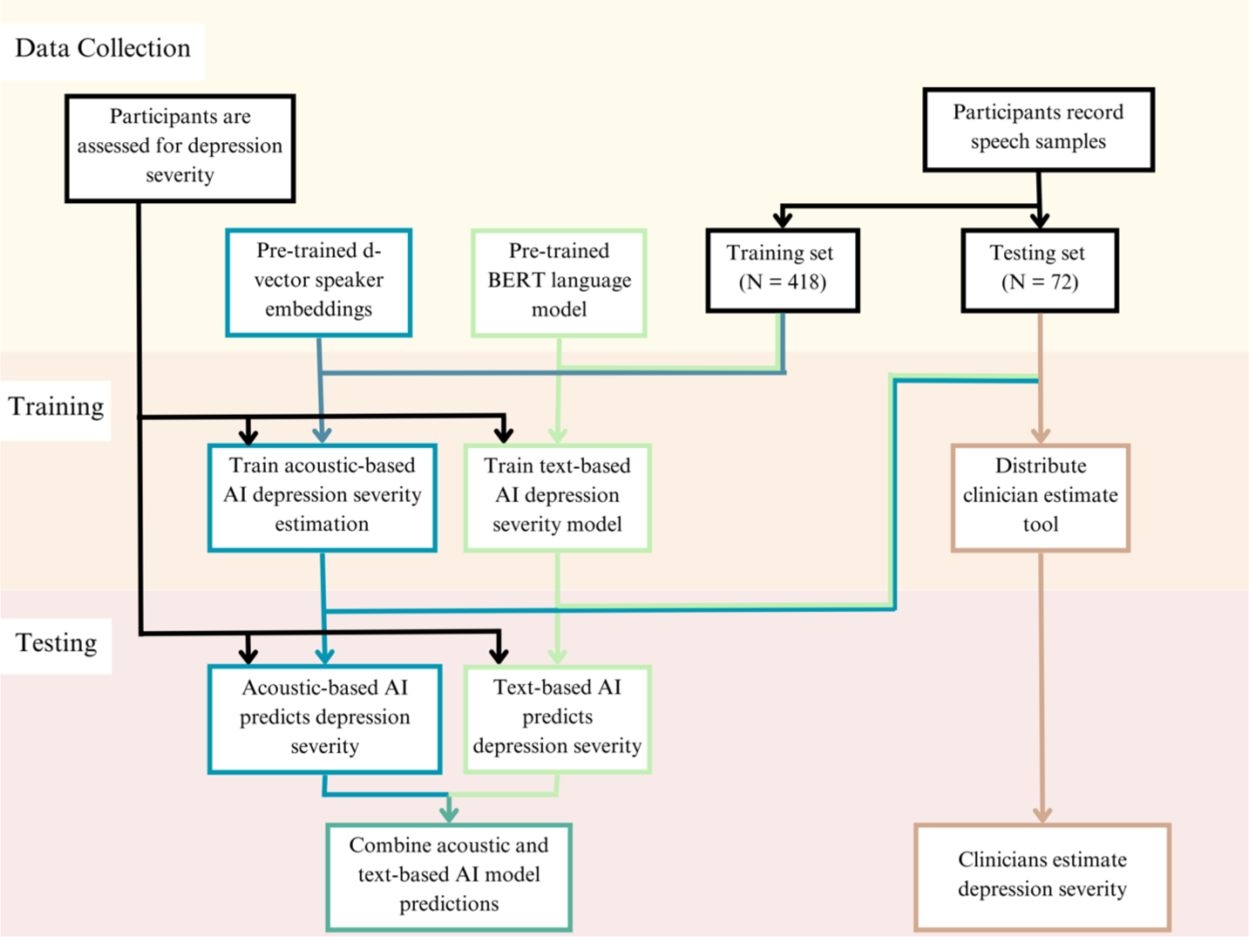
To extract language content from our speech recordings, we used the pre-trained BERT network (Devlin et al., 2019). Initially, the speech recordings were converted to text using the public domain Kaldi Automatic speech recognition engine and the ASPIRE Chain acoustic models (Povey et al., 2011; Peddinti et al., 2015). Without manual editing, this process creates imperfect transcripts with the possibility of errors, but we chose to use automatic transcription for an easily scalable process. We then used this text to fine-tune and test the BERT model.

Using text data and MADRS scores from the training dataset speech sample, we fine-tuned the depression severity estimation task, using a mean square error loss function on a single layer on top of the BERT model. We then used the text-based model to predict depression severity in our dataset.

2.4.3 Combined Acoustic and Text Prediction

We used a post-testing ensemble approach to develop a combined acoustic and text-based AI prediction of depression severity. For regression tasks, the most common method of ensemble predictions post-training is to average the output of the two models, assuming Gaussian distribution of prediction values. We averaged the text-based AI model and acoustic-based AI model estimation of depression severity for a combined AI prediction model of depression severity from speech sound and content.

Figure 2: Flowchart of depression severity estimation pathways


Data Collection
Participants are assessed for depression severity
Participants record speech samples
Pre-trained d-vector speaker embeddings
Pre-trained BERT language model
Training set (N = 418)
Testing set (N = 72)
Training
Train acoustic-based AI depression severity estimation
Train text-based AI depression severity model
Distribute clinician estimate tool
Testing
Acoustic-based AI predicts depression severity
Text-based AI predicts depression severity
Combine acoustic and text-based AI model predictions
Clinicians estimate depression severity

2.5 Analysis

2.5.1 Agreement of Depression Severity Measures

We measured agreement between estimates of depression severity and MADRS scores using intraclass correlation coefficients (ICC). While methods such as Pearson correlation tests the association between scores, and paired t-tests can measure the agreement between scores, ICC measures both the level of agreement and correlation between scores (Chen & Barnhart, 2008). ICC scores can measure the concordance between two measures, ranging from no agreement between scales (ICC = 0) and perfect agreement (ICC = 1). We used STATA 16 software (StataCorp, 2017) to calculate 2-way mixed effects ICC through the absolute agreement of 2 depression severity measures to measure the concordance between MADRS scores and estimates of depression severity within each participant. We calculated concordance between MADRS scores and an ensemble of the combined rating of clinicians for each participant; MADRS scores and an ensemble of the combined acoustic and text-based AI estimations; MADRS scores and acoustic-based AI estimations; and MADRS scores and text-based AI estimations.

We were also interested in measuring the performance of clinicians and AI estimates of depression severity by comparing the magnitude of error through the root mean square error (RMSE) statistic. RMSE is a commonly used machine learning performance metric to measure the error in the estimation of continuous measures. We calculated RMSE from the error between MADRS scores and the estimated values of depression severity.

2.5.2 Inter-rater reliability

We also examined the degree of reliability between clinician estimates of depression severity from the same speech recording. To test this, two clinicians rated each participant's speech sample, and we tested the interrater reliability of the two depression estimates. ICC values were used to calculate interrater reliability, measured by the similarity between two clinician estimates of depression severity for each participant.

2.5.3 Sources of Bias

We were interested in exploring potential sources of bias as a secondary analysis for clinician and AI estimates of depression severity from speech. To conduct an exploratory analysis of potential sources of bias in estimating depression severity from speech, we tested how the estimation error differed for participant and clinician demographic characteristics. We first calculated an error term through residual scores from the regressions of MADRS scores and depression predictions for the clinician estimation, acoustic-based AI model, and text-based AI model. We looked at the error in the text and acoustic AI prediction models instead of the combined model to separately examine potential differences in the sources of bias in the two models. We tested bias from the least-squares linear regression between the error term and participant or clinician demographic variables. To test the effects of participant characteristics on estimation error, we tested multiple linear regression models, with the error term as a dependent variable, and the sex, age, race, education, or household income as a predictor. To test the effects of clinician characteristics, we tested the correlations between the error term and the years of clinical experience, clinician gender, and the interaction between clinician gender and participant sex. Because each participant was rated by two clinicians, for clinician estimates regression models, we performed mixed effects regression that accounted for clustering within participants.

We used false discovery rate corrections for multiple hypothesis testing (Benjamini & Hochberg, 1995).

CHAPTER 3 RESULTS

3.1 Concordance and Accuracy of Depression Severity Estimates

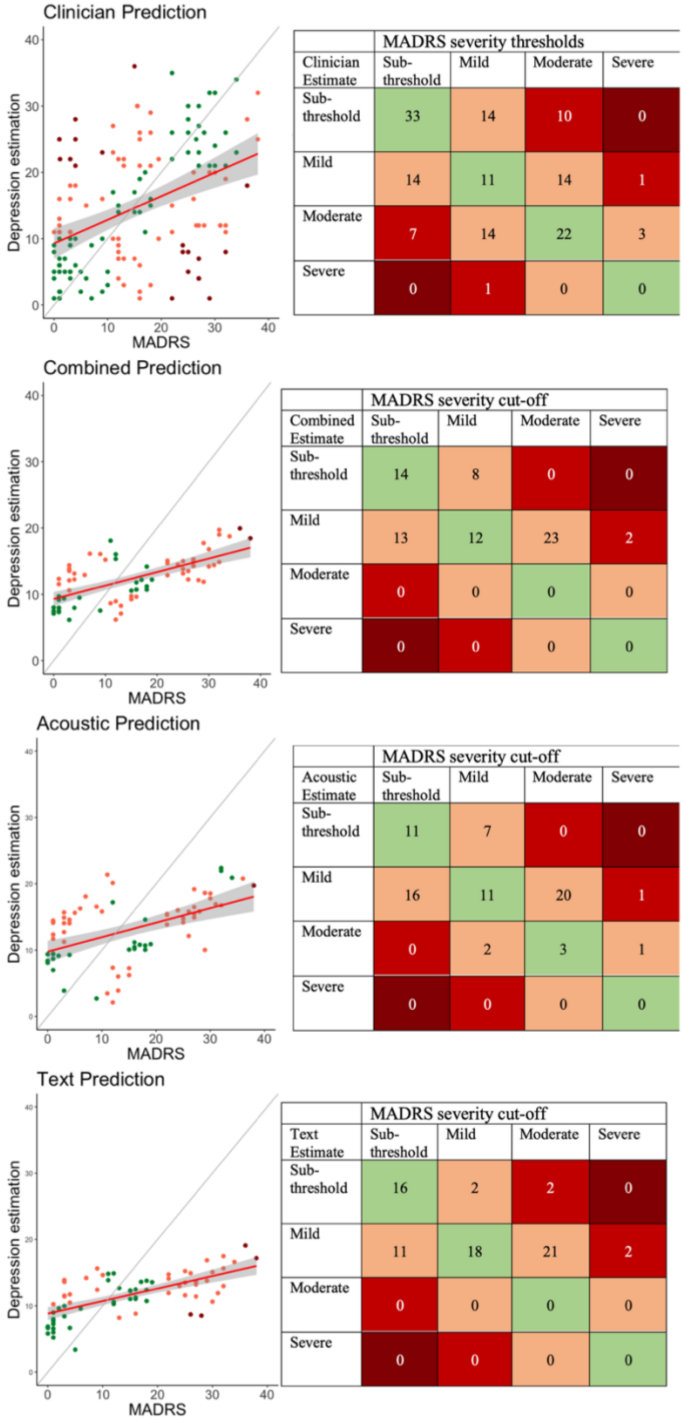
Figure 3 shows the agreement between measures of depression severity. The estimates in Figure 3 are coloured according to their corresponding confusion matrices. Confusion matrices are a commonly used tool in machine learning to highlight the accuracy and biases of prediction models. To visualize patterns in the confusion matrix, we separated depression scores corresponding to recognized MADRS severity thresholds: <10 are subthreshold scores; >10 and <20 are mild scores; >20 and <35 are moderate scores, and > 35 are severe depressive scores (Hawley et al., 2002; Snaith et al., 1986). The confusion matrix shows correct estimations as green, minor misestimations as orange, and large misestimations, resulting in the misclassification of 2 bins or more as red. Comparing the confusion matrices, it can be observed that clinicians had larger misestimations than the AI-based estimations, with more observations in the severe misestimation bins.

Figure 3 shows the agreement between clinician estimates and MADRS scores, compared to a line of perfect concordance in grey. Clinicians estimated the depression severity of participants with moderate levels of agreement to MADRS scores after listening to one minute of speech recordings (ICC= 0.47, 95% CI = 0.30, 0.65). Figure 3 also shows the agreement between AI model estimations of depression severity and MADRS scores. Visually evident clustering of acoustic-based estimations is shown in Figure 3 between participants of subthreshold depression severity symptoms. The acoustic AI-based model (ICC = 0.35, 95% CI = 0.18, 0.57), text-based

AI model (ICC = 0.29, 95% CI = 0.13, 0.53), and combined AI estimation (ICC = 0.33, 95% CI = 0.16, 0.56) had lower agreement scores than the clinician estimates.

We included RMSE to measure the average error for predicting depression severity. The combined acoustic and text-based AI model estimation of depression severity had the smallest average error (RMSE = 9.71). The text-based AI model (RMSE = 10.02) and the acoustic-based AI estimation model (RMSE = 10.69) had larger errors. Clinicians had the largest estimation errors (RMSE = 10.98).

Figure 3: Agreement between MADRS scores and clinician and AI estimates of depression severity from speech



Clinician Prediction

Clinician Estimate	MADRS severity thresholds Sub-threshold	Mild	Moderate	Severe
Sub-threshold	33	14	10	0
Mild	14	11	14	1
Moderate	7	14	22	3
Severe	0	1	0	0

Combined Prediction

Combined Estimate	MADRS severity cut-off Sub-threshold	Mild	Moderate	Severe
Sub-threshold	14	8	0	0
Mild	13	12	23	2
Moderate	0	0	0	0
Severe	0	0	0	0

Acoustic Prediction

Acoustic Estimate	MADRS severity cut-off Sub-threshold	Mild	Moderate	Severe
Sub-threshold	11	7	0	0
Mild	16	11	20	1
Moderate	0	2	3	1
Severe	0	0	0	0

Text Prediction

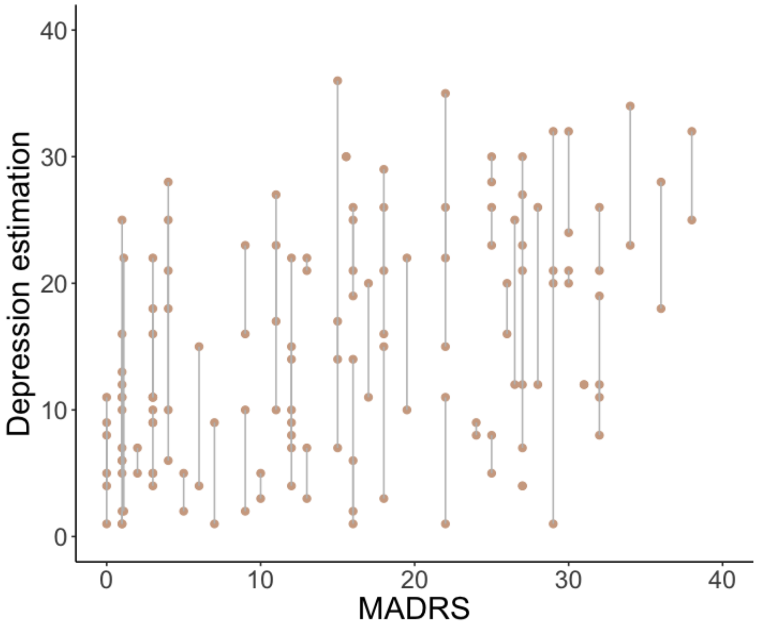
Text Estimate	MADRS severity cut-off Sub-threshold	Mild	Moderate	Severe
Sub-threshold	16	2	2	0
Mild	11	18	21	2
Moderate	0	0	0	0
Severe	0	0	0	0

***Note:** Concordance between MADRS scores and depression severity estimations are shown in descending order: clinician, combined AI, acoustic-based AI, and text-based AI predictions. A line of perfect concordance is shown in grey. Estimations are coloured according to their confusion matrix (right), with correct estimations in green, minor misestimations in orange, and large misestimations in red.

3.2 **Reliability of Clinician Estimates**

We examined the inter-rater reliability between the two clinicians who estimated depression on the same speech sample. We calculated interrater reliability using ICC between the repeat depression severity estimations. Figure 4 shows the similarity of ratings between clinicians, with paired ratings linked by the grey line. We found that clinicians had a moderate level of inter-rater reliability (ICC = 0.46; 95% CI = 0.30, 0.63).

Figure 4: Interrater reliability of clinical intuition depression estimation measure



***Note:** Each observation represents a clinician's estimation of depression severity. Clinician estimates from the same speech recording are connected by a vertical grey line. Observations with no connecting grey line indicate identical estimations for a given speech recording.

3.3 Clinician and Artificial Intelligence Biases

3.3.1 Participant Characteristics and Bias

Figures 5-9 show the effects of participant age, sex, race, income, and education level on estimation errors. The grey line on each figure represents zero bias, where the model would not systematically predict relative over or underestimations of depression severity for those participants. Positive error scores indicate that the clinician or AI model rated the participant relatively higher than other participants of similar depression severity, and negative error scores indicate a relative underestimation of depression severity.

Across the clinician, acoustic-based AI, and text-based AI errors, the sex-based errors were pointed in the same direction, with male participants slightly overestimated and female participants slightly underestimated, but none of the differences were statistically significant (Figure 5). Across the three predictions, there were no significant changes in prediction error across the age of participants (Figure 6). Across clinician and AI model estimations, the depression severity of Black participants was rated numerically higher than White participants, although this effect was non-significant (Figure 7). Across the participants' education level and income, there were no significant differences in error for clinician or AI model predictions (Figures 8-9).

Figure 5: Participant sex biases from clinician and AI estimation of depression severity from speech

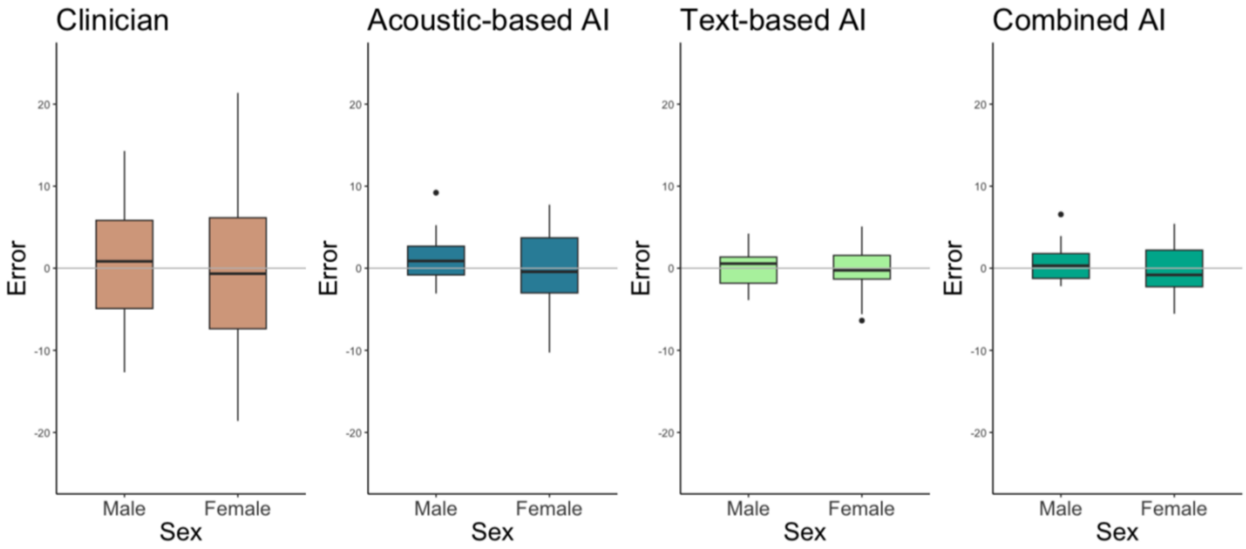


Figure 6: Participant age biases from clinician and AI estimation of depression severity from speech

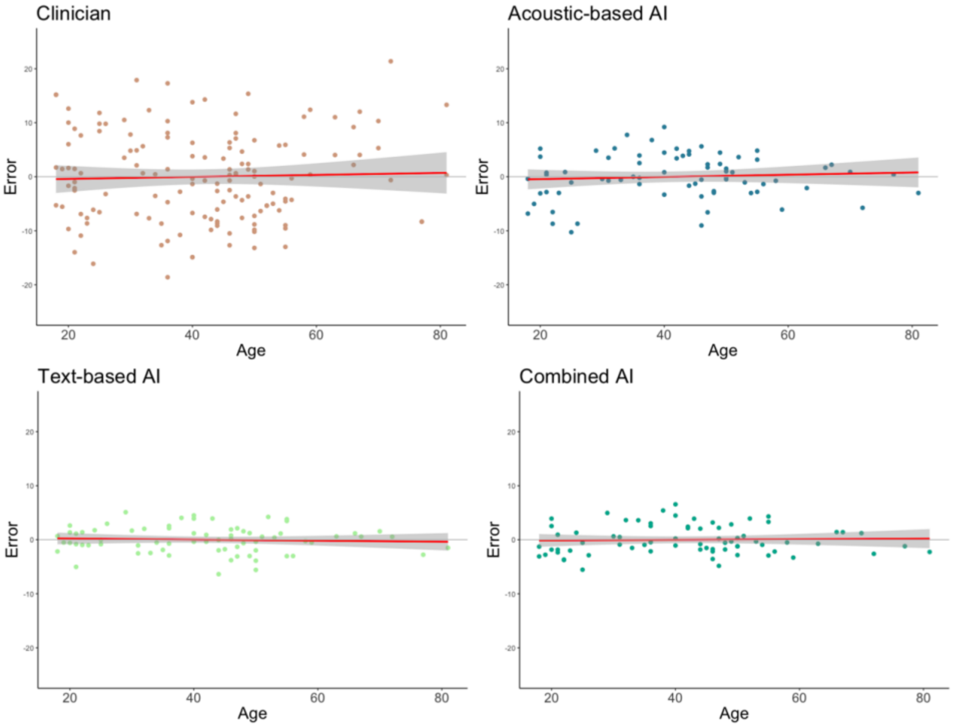


Figure 7: Race biases from clinician and AI estimation of depression severity from speech

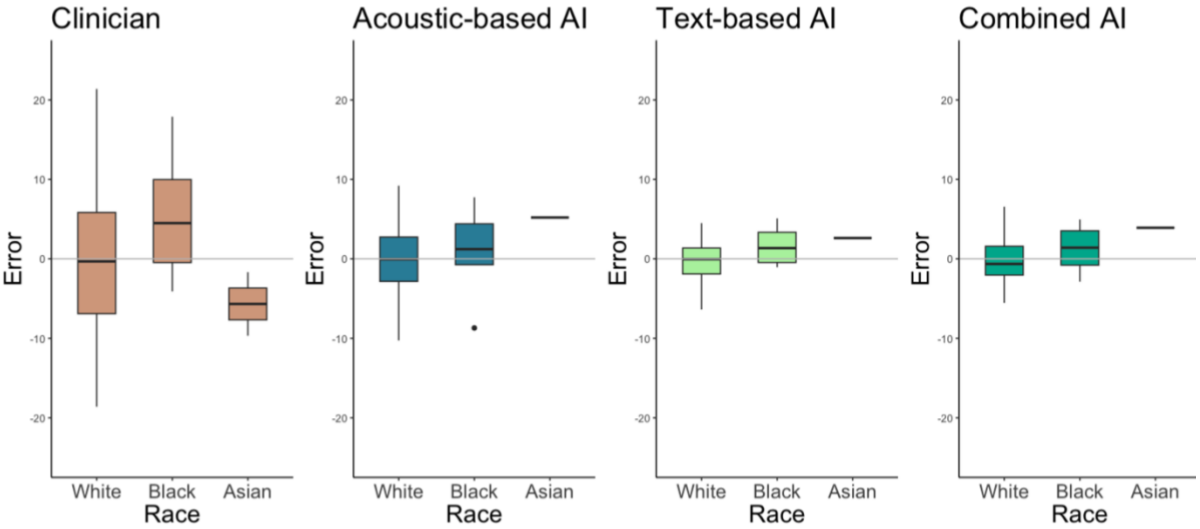


Figure 8: Income biases from clinician and AI estimations of depression from speech

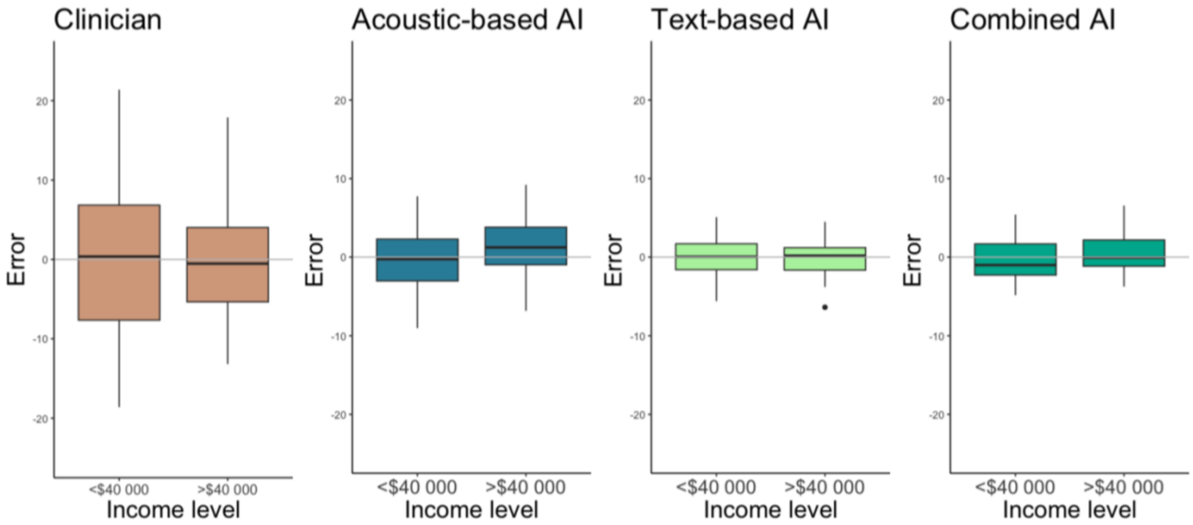
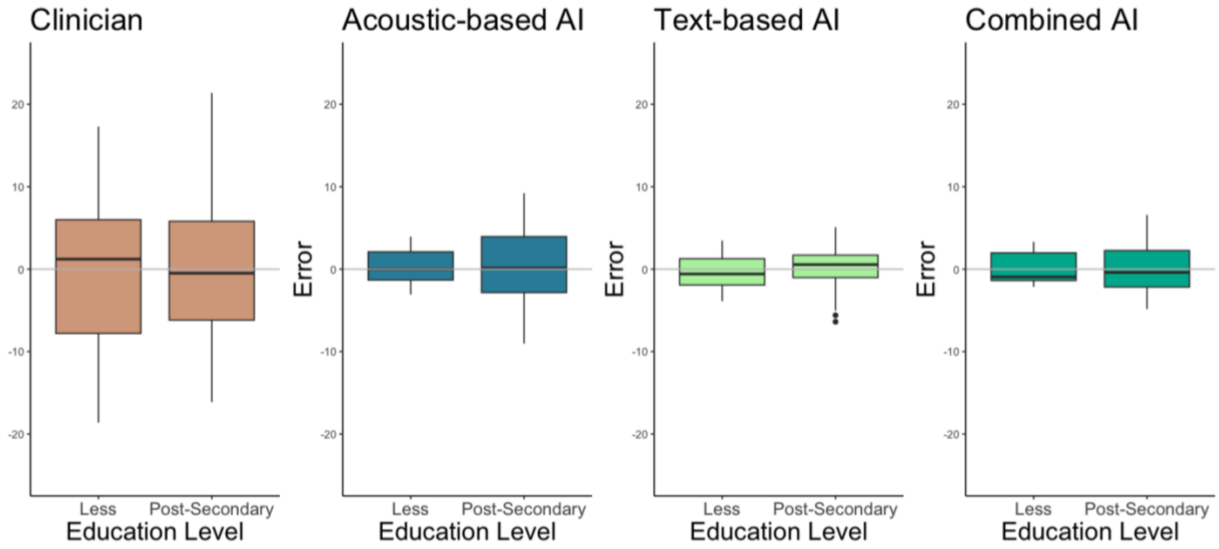


Figure 9: Education biases from clinician and AI estimates of depression severity from speech



3.3.2 Clinician Characteristics and Bias

We examined the effects of clinician gender and experience on their depression prediction error. We found that, after correcting for multiple hypothesis testing, there was a significant difference in depression severity estimation between male and female clinicians (b = 4.13; 95% CI = 1.55, 6.71; p = 0.032). We also tested the interaction between the effect of the clinician's gender and the participant's sex on the prediction error term. Figure 11 shows the average error term for male and female clinicians, split across male and female participants. However, there was no significant interaction between the gender of the clinician and the sex of the participant.

Figure 10: Depression severity estimation biases by clinician gender

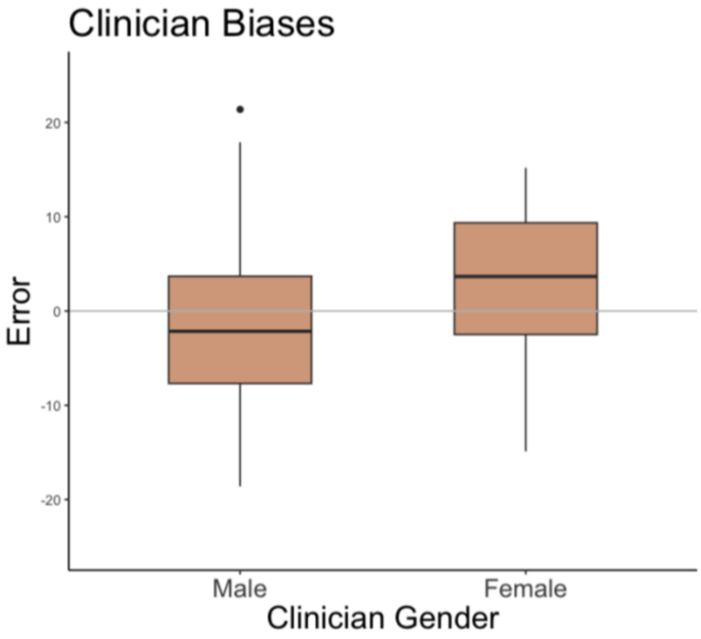
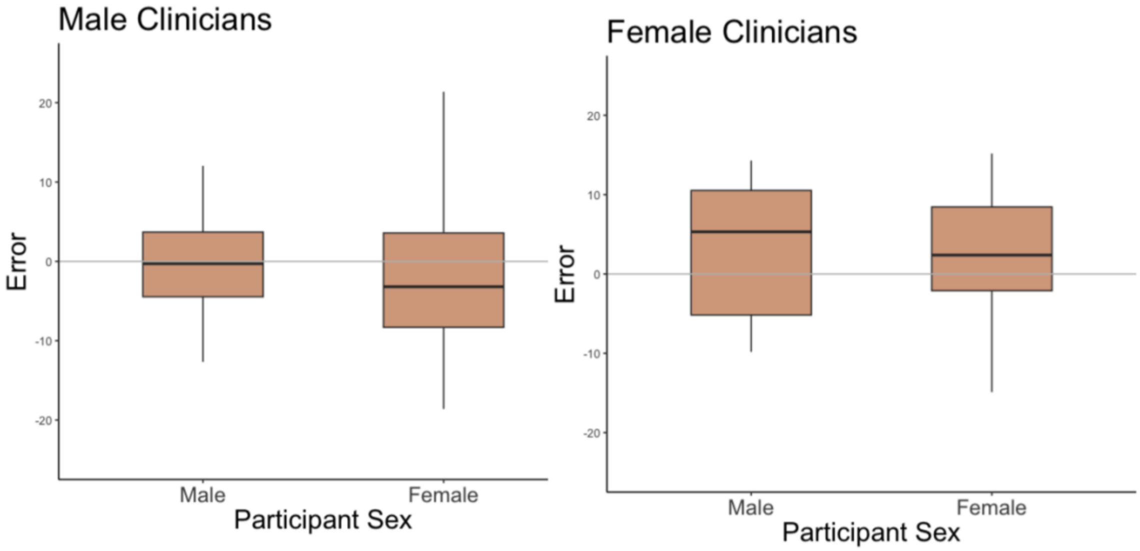


Figure 11: Estimation error by gender of participants and clinicians

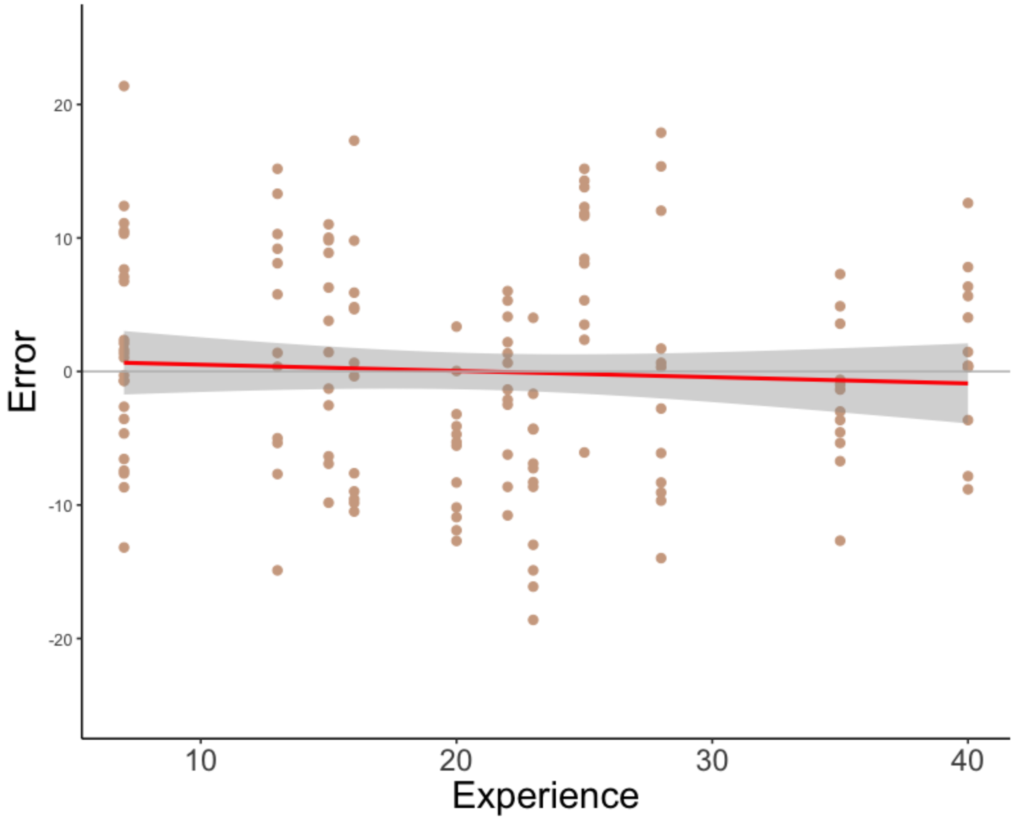


Male Clinicians

Female Clinicians

We recorded depression severity estimations from clinicians with varying levels of clinical experience. All clinicians had experience working in mood disorders, from 7 to 40 years of clinical experience. However, there was no relationship between the error and the number of years practicing medicine (Figure 12)

Figure 12: Estimation error across years of clinical experience



CHAPTER 4 DISCUSSION

4.1 Impact

Our study provides the first measure of clinical intuition from speech and the first direct comparison between clinician intuition and AI estimation of speech-based measures of depression severity. We developed a novel measure of clinical intuition from speech for estimating depression severity that can be performed rapidly, with moderate reliability and agreement with a validated depression scale. Comparison to acoustic and text-based AI network predictions showed better clinician concordance to MADRS scores, although the average error of the AI models, as measured by RMSE, was smaller than the average clinician error. Additionally, we found a potential source of bias in the clinical intuition-based depression estimation from clinician gender. These findings provide a better understanding of how clinicians and AI networks form impressions of depression severity and highlight a potential new measure with the potential to aid clinicians in the assessment and monitoring of depression.

Using the limited information provided by one minute of speech, clinicians were able to approximately estimate depression severity (ICC= 0.47, 95% CI = 0.30, 0.65) with moderate reliability. These findings support the validity of intuitive decision-making for clinicians. The clinicians were provided with no information outside of listening to the sounds and content of a short speech recording, drawing upon their “gut feeling” to estimate depression severity. These results support findings on the utility of clinical intuition measures for measuring symptoms or predicting outcomes, where clinical outcomes were successfully predicted by measures of clinician intuition (Haegdorens et al., 2023; Hoops et al., 2020; Pentzek et al., 2019; Zanovello et al., 2020).

We found unexpected results when comparing the agreement between MADRS scores and estimates of depression severity. Despite listening to only 1-minute of speech without prompting for depressive symptoms, clinicians achieved higher levels of agreement to MADRS scores than AI models using the same information. Given the body of research and interest in automated speech analysis as a biomarker for depression, these findings present an interesting discussion on the potential of measures of clinical intuition through speech.

We also compared the estimation of clinicians and AI models through RMSE values. RMSE is a performance metric commonly used in machine learning regression tasks, computed through the average model estimation error. It would be logical that two performance metrics for the same task would be related, with a stronger agreement resulting in smaller errors. However, the measure of agreement and the measure of error of the estimations provided contradictory results. We found that the combined acoustic and text-based AI model had the smallest errors (RMSE = 9.71), compared to larger clinician errors (RMSE = 10.98). Different priorities between clinicians and AI model estimations could explain this. We trained our AI models using a mean square error loss function, which incentivized reducing the estimation error. In addition, the models were trained on a dataset of relatively lower depression severity scores, which would further incentivize lower, conservative estimations of depression severity. In contrast, clinicians had a larger range of depression severity estimations, which resulted in larger errors. However, the willingness to estimate the extremes of the MADRS scale may provide better agreement to MADRS scores than the AI model estimates. There has been some debate that typical machine learning classification metrics may not best capture the performance required for clinical uses

(Cho, 2021; Wiens et al., 2019). These findings support the importance of selecting models that best measure the desired outcomes. In our study, the AI models that estimate close to the average may provide less discrimination of depression severity for participants with more severe depression, but the clinician estimation model may be more likely to have larger errors. Understanding the risk and rewards of the performance metrics may help inform the best option for a given task.

The potential clinical implications of these findings should be considered. One benefit of the clinical intuition estimate measure was the speed and ease of application. The clinicians had no prior experience using the clinical intuition estimation measure. Despite this, they used the tool to estimate depression severity quickly and with moderate accuracy. Comparatively, clinician-rated scales like the MADRS require training and time investment. The typical MADRS interview can be completed within 20-30 minutes, while the clinicians could estimate depression severity after listening to 1 minute of speech. A comparable time requirement for assessing depression is self-report scales, typically administered in 1-10 minutes. However, while self-report scales require a similar time investment to the clinician intuition tool, they typically have a stronger agreement with clinician-rated depression severity scales. For example, Cunningham and colleagues (2011) found better agreement between self-report and clinician report of the MADRS (ICC = 0.47-0.75). These findings indicate that while clinical intuition can rapidly extract information about depression severity from speech, alternative rapid measures, such as self-report scales, may have better agreement to gold-standard clinical scales, such as the MADRS.

The usefulness of depression severity measures also relies on the validity of observations for all individuals. To examine the effects of personal characteristics on the prediction error, we conducted an exploratory examination of potential sources of bias. Implicit biases from clinicians and methodological biases in AI prediction models affect patient outcomes based on personal characteristics, such as gender, race, age, sex, and socioeconomic status. We did not find any statistically significant differences in the prediction of depression severity across demographic characteristics. When examining the differences in prediction errors across clinician genders, we found a significant difference between the estimations of male and female clinicians. While there is a focus on clinical biases that arise from participant gender, there are relatively few studies that examine the effects of the clinician's gender (Champagne-Langabeer & Hedges, 2021). In our sample, the male clinicians tended to estimate depression severity lower relative to female clinicians. Given the relatively few male (N = 8) and female (N = 4) clinicians that provided depression severity estimates, these findings may benefit from follow-up after obtaining additional ratings from clinicians.

4.2 Limitations

When designing our study, we looked to develop the best dataset to test clinicians and replicate assessments of patients with depression in the clinical environment. To accomplish this, we non-randomly sampled from our database of speech recordings, with the understanding that this would create imbalances between our testing dataset of participants and our available training data for our acoustic-based and text-based models. As a result, we had several key differences between our testing and training sets.

First, we anticipated lower depression severity scores in our training set, as we selected our testing set with a uniform distribution of depression severity. In contrast, the training dataset was sampled from our studies. These studies are enriched for participants with depression, but participants provide many speech samples with low depressive symptoms. Ensemble models can be designed to mitigate some of these training and testing imbalances, using multiple iterations of reduced-size training sets, combining under-sampling of low scores and over-sampling of high scores. These techniques can mitigate some of the effects. However, the relative scarcity of moderate and severe depression severity estimations of the acoustic-based and text-based AI models show the effect of the training and testing set imbalances.

Another difference between the training and testing datasets for the AI models is the inclusion and exclusion criteria. For the testing dataset, our participants all had a diagnosis of depression and did not have a comorbid mood disorder. However, the training set included both individuals without a diagnosis of MDD and participants with diagnoses of other mood disorders, such as bipolar disorder. Classification tasks have shown differences between healthy speakers and speakers with depression, as well as differences between speakers with unipolar and bipolar depression (Dikaios et al., 2023). These diagnosis-based differences in speech and language may have introduced noise that may reduce the accuracy of the AI depression severity estimates.

Combined, these differences between the training and testing set may affect the predictive abilities of the acoustic-based and text-based AI models and limit the conclusions drawn from the comparison between clinical intuition and AI-based estimation of depression severity from speech.

We were also limited in our examination of biases in clinicians and AI estimates of depression severity. When exploring sources of bias, we relied upon the MADRS score as a "true depression" score. However, MADRS item scoring can be affected by clinician and scale biases, which may differ due to participant demographic characteristics (Dmitrieva et al., 2015). In addition, we did not measure whether clinicians or AI models could predict demographic characteristics solely from speech, which could influence the presence of bias in their estimations. Our bias analysis may have also introduced a potential limitation of our study. We structured our bias analysis to account for the clustering of similar scores for repeat observations within participants. However, we were unable to account for the effect of clinician identity on the prediction error term, which may result in clustering effects of similar errors across multiple estimations of each clinician.

CHAPTER 5 CONCLUSION

This study revealed previously unknown comparisons between the abilities of clinical intuition and AI-based prediction of depression severity. We aimed to determine the extent of information that expert clinicians can extract about depression severity from a short, one-minute speech sample. We also compared the ability of clinicians to AI-based estimates of depression severity from the same speech recordings. Lastly, we explored potential sources of bias from clinicians and AI model estimations of depression severity.

We continue to record speech samples and aim to obtain more clinician estimates of depression severity from speech. We aim to further develop the clinical intuition estimate database to increase our sample's power. In addition, we examine other characteristics of clinical intuition in predicting depression severity from speech. Given the importance of monitoring progress in MBC, the sensitivity of clinical intuition to changes in depressive symptoms should be tested. Future studies should track longitudinal speech samples to test the sensitivity of clinical intuition to changes in depression severity over time.

This project provided the first direct comparison between a measure of clinical intuition and AI models to estimate depression severity from speech. These findings reveal information on how clinicians and AI models interpret depression in speech and support the validity of clinical intuition as a potential measure of depression severity from speech recordings.

References

Afrose, S., Song, W., Nemeroff, C. B., Lu, C., & Yao, D. (Daphne). (2022). Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. *Communications Medicine*, *2*(1), Article 1. https://doi.org/10.1038/s43856-022-00165-w

Akinhanmi, M. O., Biernacka, J. M., Strakowski, S. M., McElroy, S. L., Balls Berry, J. E., Merikangas, K. R., Assari, S., McInnis, M. G., Schulze, T. G., LeBoyer, M., Tamminga, C., Patten, C., & Frye, M. A. (2018). Racial disparities in bipolar disorder treatment and research: A call to action. *Bipolar Disorders*, *20*(6), 506–514. https://doi.org/10.1111/bdi.12638

Al-Mosaiwi, M., & Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical Psychological Science*, *6*(4), 529–542. https://doi.org/10.1177/2167702617747074

Ayalon, L., & Cohn-Schwartz, E. (2022). The Relationship Between Perceived Age Discrimination in the Healthcare System and Health: An Examination of a Multi-Path Model in a National Sample of Israelis Over the Age of 50. *Journal of Aging and Health*, *34*(4–5), 684–692. https://doi.org/10.1177/08982643211058025

Ayalon, L., & Tesch-Römer, C. (2017). Taking a closer look at ageism: Self- and other-directed ageist attitudes and discrimination. *European Journal of Ageing*, *14*(1), 1–4. https://doi.org/10.1007/s10433-016-0409-9

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (arXiv:1607.06520). arXiv. http://arxiv.org/abs/1607.06520

Breznitz, Z. (1992). Verbal indicators of depression. *The Journal of General Psychology*, *119*(4), 351–363. https://doi.org/10.1080/00221309.1992.9921178

Brunn, M., Diefenbacher, A., Courtet, P., & Genieys, W. (2020). The Future is Knocking: How Artificial Intelligence Will Fundamentally Change Psychiatry. *Academic Psychiatry: The Journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*, *44*(4), 461–466. https://doi.org/10.1007/s40596-020-01243-8

Brush, J. E., Sherbino, J., & Norman, G. R. (2017). How Expert Clinicians Intuitively Recognize a Medical Diagnosis. *The American Journal of Medicine*, *130*(6), 629–634. https://doi.org/10.1016/j.amjmed.2017.01.045

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Cabrera, D., Thomas, J. F., Wiswell, J. L., Walston, J. M., Anderson, J. R., Hess, E. P., & Bellolio, M. F. (2015). Accuracy of 'My Gut Feeling:' Comparing System 1 to System 2 Decision-Making for Acuity Prediction, Disposition and Diagnosis in an Academic Emergency Department. *Western Journal of Emergency Medicine*, *16*(5), 653–657. https://doi.org/10.5811/westjem.2015.5.25301

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., & Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain and Cognition*, *56*(1), 30–35. https://doi.org/10.1016/j.bandc.2004.05.003

Champagne-Langabeer, T., & Hedges, A. L. (2021). Physician gender as a source of implicit bias affecting clinical decision-making processes: A scoping review. *BMC Medical Education*, *21*, 171. https://doi.org/10.1186/s12909-021-02601-2

Chapman, E. N., Kaatz, A., & Carnes, M. (2013). Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Health Care Disparities. *Journal of General Internal Medicine*, *28*(11), 1504–1510. https://doi.org/10.1007/s11606-013-2441-1

Cloud Computing Services—Amazon Web Services (AWS). (n.d.). Amazon Web Services, Inc. Retrieved May 17, 2023, from https://aws.amazon.com/

Corona Hernández, H., Corcoran, C., Achim, A. M., de Boer, J. N., Boerma, T., Brederoo, S. G., Cecchi, G. A., Ciampelli, S., Elvevåg, B., Fusaroli, R., Giordano, S., Hauglid, M., van Hessen, A., Hinzen, W., Homan, P., de Kloet, S. F., Koops, S., Kuperberg, G. R., Maheshwari, K., … Palaniyappan, L. (2023). Natural Language Processing Markers for Psychosis and Other Psychiatric Disorders: Emerging Themes and Research Agenda From a Cross-Linguistic Workshop. *Schizophrenia Bulletin*, *49*(Supplement_2), S86–S92. https://doi.org/10.1093/schbul/sbac215

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

Dehon, E., Weiss, N., Jones, J., Faulconer, W., Hinton, E., & Sterling, S. (2017). A Systematic Review of the Impact of Physician Implicit Racial Bias on Clinical Decision Making. *Academic Emergency Medicine*, *24*(8), 895–904. https://doi.org/10.1111/acem.13214

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dikaios, K., Rempel, S., Dumpala, S. H., Oore, S., Kiefte, M., & Uher, R. (2023). Applications of Speech Analysis in Psychiatry. *Harvard Review of Psychiatry*, *31*(1), 1–13. https://doi.org/10.1097/HRP.0000000000000356

Dmitrieva, N. O., Fyffe, D., Mukherjee, S., Fieo, R., Zahodne, L. B., Hamilton, J., Potter, G. G., Manly, J. J., Romero, H. R., Mungas, D., & Gibbons, L. E. (2015). Demographic characteristics do not decrease the utility of depressive symptoms assessments: Examining the practical impact of item bias in four heterogeneous samples of older adults. *International Journal of Geriatric Psychiatry*, *30*(1), 88–96. https://doi.org/10.1002/gps.4121

Dowling, N. M., Bolt, D. M., Deng, S., & Li, C. (2016). Measurement and control of bias in patient reported outcomes using multidimensional item response theory. *BMC Medical Research Methodology*, *16*, 63. https://doi.org/10.1186/s12874-016-0161-z

Dowrick, C., Leydon, G. M., McBride, A., Howe, A., Burgess, H., Clarke, P., Maisey, S., & Kendrick, T. (2009). Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: Qualitative study. *BMJ*, *338*, b663. https://doi.org/10.1136/bmj.b663

Du, M., Liu, S., Wang, T., Zhang, W., Ke, Y., Chen, L., & Ming, D. (2023). Depression recognition using a proposed speech chain model fusing speech production and perception features. *Journal of Affective Disorders*, *323*, 299–308. https://doi.org/10.1016/j.jad.2022.11.060

Fortney, J. C., Unützer, J., Wrenn, G., Pyne, J. M., Smith, G. R., Schoenbaum, M., & Harbin, H. T. (2017). A Tipping Point for Measurement-Based Care. *Psychiatric Services (Washington, D.C.)*, *68*(2), 179–188. https://doi.org/10.1176/appi.ps.201500439

Ganel, T., Sofer, C., & Goodale, M. A. (2022). Biases in human perception of facial age are present and more exaggerated in current AI technology. *Scientific Reports*, *12*, 22519. https://doi.org/10.1038/s41598-022-27009-w

Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. (2021). *Lancet (London, England)*, *398*(10312), 1700–1712. https://doi.org/10.1016/S0140-6736(21)02143-7

Gonzalez, O., Georgeson, A. R., Pelham, W. E., & Fouladi, R. T. (2021). Estimating Classification Consistency of Screening Measures and Quantifying the Impact of Measurement Bias. *Psychological Assessment*, *33*(7), 596–609. https://doi.org/10.1037/pas0000938

Greden, J. F., & Carroll, B. J. (1980). Decrease in speech pause times with treatment of endogenous depression. *Biological Psychiatry*, *15*(4), 575–587.

Haegdorens, F., Wils, C., & Franck, E. (2023). Predicting patient deterioration by nurse intuition: The development and validation of the nurse intuition patient deterioration scale. *International Journal of Nursing Studies*, *142*, 104467. https://doi.org/10.1016/j.ijnurstu.2023.104467

Hagiwara, N., Slatcher, R. B., Eggly, S., & Penner, L. A. (2017). Physician Racial Bias and Word Use during Racially Discordant Medical Interactions. *Health Communication*, *32*(4), 401–408. https://doi.org/10.1080/10410236.2016.1138389

Hamilton, J. D., & Bickman, L. (2008). A Measurement Feedback System (MFS) Is Necessary to Improve Mental Health Outcomes. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(10), 1114–1119. https://doi.org/10.1097/CHI.0b013e3181825af8

Hansen, L., Zhang, Y.-P., Wolf, D., Sechidis, K., Ladegaard, N., & Fusaroli, R. (2022). A generalizable speech emotion recognition model reveals depression and remission. *Acta Psychiatrica Scandinavica*, *145*(2), 186–199. https://doi.org/10.1111/acps.13388

Hawley, C. J., Gale, T. M., & Sivakumaran, T. (2002). Defining remission by cut-off score on the MADRS: Selecting the optimal value. *Journal of Affective Disorders*, *72*(2), 177–184. https://doi.org/10.1016/S0165-0327(01)00451-7

Himmelstein, P., Barb, S., Finlayson, M. A., & Young, K. D. (2018). Linguistic analysis of the autobiographical memories of individuals with major depressive disorder. *PloS One*, *13*(11), e0207814. https://doi.org/10.1371/journal.pone.0207814

Hong, R. H., Murphy, J. K., Michalak, E. E., Chakrabarty, T., Wang, Z., Parikh, S. V., Culpepper, L., Yatham, L. N., Lam, R. W., & Chen, J. (2021). Implementing Measurement-Based Care for Depression: Practical Solutions for Psychiatrists and Primary Care Physicians. *Neuropsychiatric Disease and Treatment*, *17*, 79–90. https://doi.org/10.2147/NDT.S283731

Hoops, K. E. M., Fackler, J. C., King, A., Colantuoni, E., Milstone, A. M., & Woods-Hill, C. (2020). How good is our diagnostic intuition? Clinician prediction of bacteremia in critically ill children. *BMC Medical Informatics and Decision Making*, *20*, 144. https://doi.org/10.1186/s12911-020-01165-3

Hunt, M., Auriemma, J., & Cashaw, A. C. A. (2003). Self-report bias and underreporting of depression on the BDI-II. *Journal of Personality Assessment*, *80*(1), 26–30. https://doi.org/10.1207/S15327752JPA8001_10

Kales, H. C., Neighbors, H. W., Blow, F. C., Taylor, K. K. K., Gillon, L., Welsh, D. E., Maixner, S. M., & Mellow, A. M. (2005). Race, Gender, and Psychiatrists' Diagnosis and Treatment of Major Depression Among Elderly Patients. *Psychiatric Services*, *56*(6), 721–728. https://doi.org/10.1176/appi.ps.56.6.721

Knoll, A. D., & MacLennan, R. N. (2017). Prevalence and correlates of depression in Canada: Findings from the Canadian Community Health Survey. *Canadian Psychology/Psychologie Canadienne*, *58*(2), 116–123. https://doi.org/10.1037/cap0000103

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*. https://doi.org/10.3928/0048-5713-20020901-06

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, *117*(23), 12592–12594. https://doi.org/10.1073/pnas.1919012117

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H.-C., Paulus, M. P., Krystal, J. H., & Jeste, D. V. (2021). Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(9), 856–864. https://doi.org/10.1016/j.bpsc.2021.02.001

Li, M., Xu, H., Liu, W., & Liu, J. (2022). Bidirectional LSTM and Attention for Depression Detection on Clinical Interview Transcripts. *2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN)*, 638–643. https://doi.org/10.1109/ICICN56848.2022.10006532

Liu, Z., Yu, H., Li, G., Chen, Q., Ding, Z., Feng, L., Yao, Z., & Hu, B. (2023). Ensemble learning with speaker embeddings in multiple speech task stimuli for depression detection. *Frontiers in Neuroscience*, *17*, 1141621. https://doi.org/10.3389/fnins.2023.1141621

Logan, D. E., Claar, R. L., & Scharff, L. (2008). Social desirability response bias and self-report of psychological distress in pediatric chronic pain patients. *Pain*, *136*(3), 366–372. https://doi.org/10.1016/j.pain.2007.07.015

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96–116. https://doi.org/10.1002/lio2.354

Melin-Johansson, C., Palmqvist, R., & Rönnberg, L. (2017). Clinical intuition in the nursing process and decision-making—A mixed-studies review. *Journal of Clinical Nursing*, *26*(23–24), 3936–3949. https://doi.org/10.1111/jocn.13814

Monteith, S., Glenn, T., Geddes, J., Whybrow, P. C., Achtyes, E., & Bauer, M. (2022). Expectations for Artificial Intelligence (AI) in Psychiatry. *Current Psychiatry Reports*, *24*(11), 709–721. https://doi.org/10.1007/s11920-022-01378-5

Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R. (2012). Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biological Psychiatry*, *72*(7), 580–587. https://doi.org/10.1016/j.biopsych.2012.03.015

Nemiroff, L. (2022). We can do better: Addressing ageism against older adults in healthcare. *Healthcare Management Forum*, *35*(2), 118–122. https://doi.org/10.1177/08404704221080882

Norman, G. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education*, *14*(1), 37–49. https://doi.org/10.1007/s10459-009-9179-x

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science (New York, N.Y.)*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Pavlova, B., & Uher, R. (2020). Assessment of Psychopathology: Is Asking Questions Good Enough? *JAMA Psychiatry*, *77*(6), 557–558. https://doi.org/10.1001/jamapsychiatry.2020.0108

Payne, B. K., & Hannay, J. W. (2021). Implicit bias reflects systemic racism. *Trends in Cognitive Sciences*, *25*(11), 927–936. https://doi.org/10.1016/j.tics.2021.08.001

Pearson, H. (2013). Science and intuition: Do both have a place in clinical decision making? *British Journal of Nursing*, *22*(4), 212–215.

Pentzek, M., Wagner, M., Abholz, H.-H., Bickel, H., Kaduszkiewicz, H., Wiese, B., Weyerer, S., König, H.-H., Scherer, M., Riedel-Heller, S. G., Maier, W., & Koppara, A. (2019). The value of the GP's clinical judgement in predicting dementia: A multicentre prospective cohort study among patients in general practice. *The British Journal of General Practice*, *69*(688), e786–e793. https://doi.org/10.3399/bjgp19X706037

Price, A., Zulkosky, K., White, K., & Pretz, J. (2017). Accuracy of intuition in clinical decision-making among novice clinicians. *Journal of Advanced Nursing*, *73*(5), 1147–1157. https://doi.org/10.1111/jan.13202

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial Inequality in Psychological Research: Trends of the Past and Recommendations for the Future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. https://doi.org/10.1177/1745691620927709

Rogers, S. E., Thrasher, A. D., Miao, Y., Boscardin, W. J., & Smith, A. K. (2015). Discrimination in Healthcare Settings is Associated with Disability in Older Adults: Health and Retirement Study, 2008-2012. *Journal of General Internal Medicine*, *30*(10), 1413–1420. https://doi.org/10.1007/s11606-015-3233-6

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Salokangas, R. K. R., Vaahtera, K., Pacriev, S., Sohlman, B., & Lehtinen, V. (2002). Gender differences in depressive symptoms: An artefact caused by measurement instruments? *Journal of Affective Disorders*, *68*(2), 215–220. https://doi.org/10.1016/S0165-0327(00)00315-3

Samulowitz, A., Gremyr, I., Eriksson, E., & Hensing, G. (2018). “Brave Men” and “Emotional Women”: A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain. *Pain Research & Management*, *2018*, 6358624. https://doi.org/10.1155/2018/6358624

Schnierle, J., Christian-Brathwaite, N., & Louisias, M. (2019). Implicit Bias: What Every Pediatrician Should Know About the Effect of Bias on Health and Future Directions. *Current Problems in Pediatric and Adolescent Health Care*, *49*(2), 34–44. https://doi.org/10.1016/j.cppeds.2019.01.003

Scott, K., & Lewis, C. C. (2015). Using Measurement-Based Care to Enhance Any Treatment. *Cognitive and Behavioral Practice*, *22*(1), 49–59. https://doi.org/10.1016/j.cbpra.2014.01.010

Seker, E., Talburt, J. R., & Greer, M. L. (2022). Preprocessing to Address Bias in Healthcare Data. *Challenges of Trustable AI and Added-Value on Health*, 327–331. https://doi.org/10.3233/SHTI220468

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, *27*(12), Article 12. https://doi.org/10.1038/s41591-021-01595-0

Smirnova, D., Cumming, P., Sloeva, E., Kuvshinova, N., Romanov, D., & Nosachev, G. (2018). Language Patterns Discriminate Mild Depression From Normal Sadness and Euthymic State. *Frontiers in Psychiatry*, *9*, 105. https://doi.org/10.3389/fpsyt.2018.00105

Snaith, R. P., Harrop, F. M., Newby, D. A., & Teale, C. (1986). Grade Scores of the Montgomery—Åsberg Depression and the Clinical Anxiety Scales. *The British Journal of Psychiatry*, *148*(5), 599–601. https://doi.org/10.1192/bjp.148.5.599

Sobin, C., & Sackeim, H. A. (1997). Psychomotor symptoms of depression. *The American Journal of Psychiatry*, *154*(1), 4–17. https://doi.org/10.1176/ajp.154.1.4

Spencer, C. S., Gaskin, D. J., & Roberts, E. T. (2013). The Quality Of Care Delivered To Patients Within The Same Hospital Varies By Insurance Type. *Health Affairs*, *32*(10), 1731–1739. https://doi.org/10.1377/hlthaff.2012.1400

Straw, I., & Callison-Burch, C. (2020). Artificial Intelligence in mental health and the biases of language based models. *PLoS ONE*, *15*(12), e0240376. https://doi.org/10.1371/journal.pone.0240376

Sung, Y.-T., & Wu, J.-S. (2018). The Visual Analogue Scale for Rating, Ranking and Paired-Comparison (VAS-RRP): A new technique for psychological measurement. *Behavior Research Methods*, *50*(4), 1694–1715. https://doi.org/10.3758/s13428-018-1041-8

Taguchi, T., Tachikawa, H., Nemoto, K., Suzuki, M., Nagano, T., Tachibana, R., Nishimura, M., & Arai, T. (2018). Major depressive disorder discrimination using vocal acoustic features. *Journal of Affective Disorders*, *225*, 214–220. https://doi.org/10.1016/j.jad.2017.08.038

Thake, M., & Lowry, A. (2017). A systematic review of trends in the selective exclusion of older participant from randomised clinical trials. *Archives of Gerontology and Geriatrics*, *72*, 99–102. https://doi.org/10.1016/j.archger.2017.05.017

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2022). A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health. *Perspectives on Psychological Science*, 17456916221134490. https://doi.org/10.1177/17456916221134490

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), Article 1. https://doi.org/10.1038/s41591-018-0300-7

Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Nierenberg, A. A., Warden, D., Ritz, L., Norquist, G., Howland, R. H., Lebowitz, B., McGrath, P. J., Shores-Wilson, K., Biggs, M. M., Balasubramani, G. K., & Fava, M. (2006). Evaluation of Outcomes With Citalopram for Depression Using Measurement-Based Care in STAR*D: Implications for Clinical Practice. *American Journal of Psychiatry*, *163*(1), 28–40. https://doi.org/10.1176/appi.ajp.163.1.28

Vanstone, M., Monteiro, S., Colvin, E., Norman, G., Sherbino, J., Sibbald, M., Dore, K., & Peters, A. (2019). Experienced physician descriptions of intuition in clinical reasoning: A typology. *Diagnosis (Berlin, Germany)*, *6*(3), 259–268. https://doi.org/10.1515/dx-2018-0069

Vetter, T. R., & Mascha, E. J. (2017). Bias, Confounding, and Interaction: Lions and Tigers, and Bears, Oh My! *Anesthesia & Analgesia*, *125*(3), 1042. https://doi.org/10.1213/ANE.0000000000002332

Weissglass, D. E. (2022). Contextual bias, the democratization of healthcare, and medical artificial intelligence in low- and middle-income countries. *Bioethics*, *36*(2), 201–209. https://doi.org/10.1111/bioe.12927

Widlöcher, D. J. (1983). Psychomotor retardation: Clinical, theoretical, and psychometric aspects. *The Psychiatric Clinics of North America*, *6*(1), 27–40.

Wright, C. V., Goodheart, C., Bard, D., Bobbitt, B. L., Butt, Z., Lysell, K., McKay, D., & Stephens, K. (2020). Promoting measurement-based care and quality measure development: The APA mental and behavioral health registry initiative. *Psychological Services*, *17*(3), 262–270. https://doi.org/10.1037/ser0000347

Yang, F. M., & Jones, R. N. (2007). Center for Epidemiologic Studies—Depression Scale (CES-D) Item Response Bias Found with Mantel-Haenszel Method Successfully Replicated Using Latent Variable Modeling. *Journal of Clinical Epidemiology*, *60*(11), 1195–1200. https://doi.org/10.1016/j.jclinepi.2007.02.008

Zanovello, S., Donisi, V., Tedeschi, F., Ruggeri, M., Moretti, F., Rimondini, M., & Amaddeo, F. (2020). Predicting Patients' Readmission: Do Clinicians Outperform a Statistical Model? An Exploratory Study on Clinical Risk Judgment in Mental Health. *Journal of Nervous & Mental Disease*, *208*(5), 353–361. https://doi.org/10.1097/NMD.0000000000001140

Zhao, Q., Fan, H.-Z., Li, Y.-L., Liu, L., Wu, Y.-X., Zhao, Y.-L., Tian, Z.-X., Wang, Z.-R., Tan, Y.-L., & Tan, S.-P. (2022). Vocal Acoustic Features as Potential Biomarkers for Identifying/Diagnosing Depression: A Cross-Sectional Study. *Frontiers in Psychiatry*, *13*, 815678. https://doi.org/10.3389/fpsyt.2022.815678

Zimmerman, M., & McGlinchey, J. B. (2008). Why Don't Psychiatrists Use Scales to Measure Outcome When Treating Depressed Patients? *The Journal of Clinical Psychiatry*, *69*(12), 1916–1919. https://doi.org/10.4088/JCP.v69n1209

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—It is time to make it fair. *Nature*, *559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8