Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: Algorithm Development and Validation Study

Rachid Riad (1), Martin Denais (1), Marc de Gennes (1), Adrien Lesage (1) Vincent Oustric (1), Xuan-Nga Cao (1), Stéphane Mouchabac (2,3), Alexis Bourla (2,3,4,5)

(1) Callyope, Paris, France

(2) Department of Psychiatry, Saint-Antoine Hospital, Sorbonne University, AP-HP, Paris F-75012, France

(3) Infrastructure for Clinical Research in Neurosciences (iCRIN), Paris BrainInstitute, Paris, France

(4) Clariane, Medical Strategy and Innovation Department, Paris, France

(5) NeuroStim Psychiatry Practice, Paris, France

Abstract

Background: While speech analysis holds promise for mental health assessment, research often focuses on single symptoms, despite symptom co-occurrences and interactions. In addition, predictive models in Mental Health do not properly assess speech-based systems' limitations, such as uncertainty, or fairness for a safe clinical deployment.

Objective: We investigated the predictive potential of mobile-collected speech data for detecting and estimating depression, anxiety, fatigue, and insomnia, focusing beyond mere accuracy, in the general population.

Methods: We included n=435 healthy adults and recorded their answers concerning their perceived mental and sleep states. We asked them how they felt and if they had slept well lately. Clinically validated questionnaires measured depression, anxiety, insomnia, and fatigue severity. We developed a novel speech and machine learning pipeline involving voice activity detection, feature extraction, and model training. We detected voice activity automatically with a bidirectional neural network and examined participants' speech with a fully ML automatic pipeline to capture speech variability. Then, we modelled speech with a ThinResNet model that was pre-trained on a large open free database. Based on this speech modelling, we evaluated clinical

threshold detection, individual score prediction, model uncertainty estimation, and performance fairness across demographics (age, sex, education). We employed a train-validation-test split for all evaluations: to develop our models, select the best ones and assess the generalizability of held-out data.

Results: Our methods achieved high detection performance for all symptoms, particularly depression (PHQ-9 AP=0.77, BDI AP=0.83), insomnia (AIS AP=0.86), and fatigue (MFI Total Score AP=0.88). These strengths were maintained while ensuring high abstention rates for uncertain cases (Risk-Coverage AUCs < 0.1). Individual symptom scores were predicted with good accuracy (Correlations were all significant, with Pearson strengths between 0.59 and 0.74). Fairness analysis revealed that models were consistent for sex (average Disparity Ratio (DR) = 0.77), to a lesser extent for education level (average Disparity Ratio (DR) = 0.44) and worse for age groups (average Disparity Ratio (DR) = 0.26).

Conclusions: This study demonstrates the potential of speech-based systems for multifaceted mental health assessment in the general population, not only for detecting clinical thresholds but also for estimating their severity. Addressing fairness and incorporating uncertainty estimation with selective classification are key contributions that can enhance the clinical utility and responsible implementation of such systems. This approach offers promise for more accurate and nuanced mental health assessments, potentially benefiting both patients and clinicians.

Keywords: speech; voice; depression; anxiety; insomnia; fatigue; mobile phone; machine learning; prediction

Introduction

Depression and anxiety disorders are recognised as the leading causes of disease burden [1], and their prevalences are high during the entire lifespan, across the sexes and all around the globe [2]. This burden was aggravated by the COVID-19 pandemic [3]. In these disorders, the early identification and evaluation of symptoms severity are of prime importance since the incidence of suicide is associated with a diagnosis of depression more than 50% of the time [4]. Besides, measurement-based care improves the follow-up and treatment of affected individuals with mental health disorders [5]. Yet, the measure of mental health remains a challenge, since manifestations of depression and anxiety are heterogeneous [6], and co-occur with insomnia [7] and fatigue [8]. The exhaustive and objective assessment of these different mental health dimensions through validated questionnaires is long and fastidious for clinical staff, and particularly not adapted to primary care, which is at the forefront of handling mental health disorders[9]. Timely interventions lead to better outcomes in mental health. This proactive approach can ensure early access to treatment and prevent significant

complications. Unfortunately, current methods of follow-up often rely on subjective assessments, which can be time-consuming, susceptible to bias, and limited in their ability to detect early signs of mental health concerns. The development of objective biomarkers, however, has the potential to overcome these limitations. These quantifiable measures, encompassing biological, genetic, or behavioural assessments, could revolutionize early detection, enabling timely and targeted interventions that ultimately improve patient outcomes and well-being. This is particularly significant for screening the general population across diverse mental health dimensions. Biomarkers could provide efficient and objective tools to identify individuals at risk for a variety of mental health conditions, ensuring they receive the necessary support before symptoms escalate.

The study of speech biomarkers in mental health holds great potential, offering a non-invasive and easily accessible avenue to capture significant motor, cognitive and behavioral changes due to mental health disorders such as depression and anxiety [10–13]. Clinical evidence and research studies have increasingly linked specific automated extracted speech features, such as prosody, articulation, and fluency, with various mental health conditions, including depression [10, 14], anxiety [15], suicide-risk assessment [16], fatigue [17, 18], or sleep deprivation [19]. The complexity of human speech extends beyond the intricate motor coordination involved. The speech production system within the brain relies on the synchronization of diverse cognitive, social, and motor processes [20, 21]. This intricate interplay involves hundreds of muscles across the respiratory, phonatory, and supralaryngeal systems, working in concert with critical cognitive skills like attention, memory, and planning. Additionally, social skills such as theory of mind and emotional processing play a vital role. Importantly, disruptions in any of these areas, including those related to mental health dimensions, can introduce perturbations in the resulting speech signal.
Besides, beyond research evidence, clinical practitioners also use voice (un)consciously when evaluating individuals, and these subjective evaluations could be completed and refined with objective measures from automatic speech analysis.

Speech biomarkers emerge also as a promising avenue for mental health assessment due to their unique characteristics. These markers hold potential as non-invasive, cost-effective, and convenient tools. Recent advancements have significantly simplified and reduced the cost of acquiring acoustic data, making it a more accessible option compared to traditional biological, imaging, or cognitive markers. In addition, speech data collection requires minimal effort from both patients and clinicians and can even be conducted remotely, further enhancing its feasibility in various settings.

However, despite its promises, the study of speech biomarkers remains largely fragmented and lacks more evidence to break into clinical practice [22]. Research on speech in mental health in the general population often focuses on one isolated mental health dimension, even though there are proofs supporting the existence of networks of symptoms in mental health that influence each other [23, 24]. Another

limitation of previous speech studies, they are confined to specific populations such as students [25] or senior populations [26]. Besides, machine learning models making predictions at the individual levels in medicine should be 'fair' and have the same quality of prediction no matter the demographics before they can be deployed in a clinical setting [27]. Finally, speech-based systems should be tested beyond the simple classification of binary labels, be tested for the estimation of the severity of symptoms [28] and be capable of abstaining when too uncertain and deferring decisions to health staff in practice[29].

In this study, using mobile-collected speech and mental health data, the main objective was to assess the predictive potential of speech in detecting and estimating the severity of depression, anxiety, fatigue, and insomnia within the general population. Besides, the fairness and uncertainty capabilities of these models were assessed for potential responsible and effective implementation in diverse real-world settings.

Methods

Participants

We recruited adult French healthy participants, above 18 years of age, without any known severe psychiatric or neurological disorder (self-declaration), or speech pathologies such as stuttering or clutter.

After fully understanding the explanation of the study, all participants signed a consent form to participate in the study, in line with the Declaration of Helsinki, current Good Clinical Practice guidelines, and local laws and regulations. All procedures were approved by the French National Institutional Review Board, (identifier 23.00748.OOO2L7#I for the Committee for the Protection of Persons").

Study procedure

The participants completed the protocol on smartphones through the Callyope research mobile application in a home environment. The participants undertook questionnaires that assessed the different mental health dimensions and different speech tasks. In this work, we only focused on one speech task, with spontaneous speech, with a semi-structured question, where participants had to answer: "Describe how you are feeling at the moment and how your nights have been." [30]. The participants were included by speech pathologist interns and recruited through social media platforms. Finally, we examined self-reported symptoms with clinically validated questionnaires (See Figure 1a.). The speech samples were recorded using the Callyope app on smartphones, which provided instructions to the participants. The vocal answers were recorded with the smartphone's microphone. The audio was sampled at 44.1 kHz with a 16-bit resolution.

We refer to the dataset collected in this study as the Callyope-GP dataset. We split randomly the Callyope-GP dataset into three sets: training, validation and testing. Demographic data, such as sex, age, and study level, were collected. We compared groups with adequate tests for their demographics and self-assessments to ensure groups were consistent.

Measures of depressive symptoms

To allow broader use of our solutions, we assessed the severity of depression through the Beck Depression Inventory (BDI) [31] and Patient Health Questionnaire-9 (PHQ-9) [32] self-questionnaires. In current practice to monitor depression, it is common that different professionals interacting with a patient use different metrics. While depression assessment through these two measures exhibits a robust correlation at the group level [33], facilitating the development of an equational conversion for research uses, their limited efficacy at the individual level impedes their reliable conversions to predict individual depressive status [34].

The PHQ-9 is a short, self-administered questionnaire mainly used to screen and measure the severity of depression[32] and is sensitive to potential changes [33]. It includes the two cardinal signs of depression: anhedonia and depressed mood. We considered a risk of depression based on the PHQ-9 if the total score was more than 10 (PHQ-9≥10).
The BDI is a self-administered questionnaire with 21 items, each centered around a core theme [31, 35]. Respondents are presented with four ordinal statements for each item, and they are instructed to choose one statement, which is then associated with a score ranging from 0 to 3. The cumulative score for the scale can reach a maximum of 63 points. We considered a risk of depression based on the BDI if the total score was more than 10 (BDI≥10), as it is above the normal range as defined by the authors of the BDI [35].

Measure of anxiety

The Generalized Anxiety Disorder questionnaire (GAD-7) is used to screen for possible generalized anxiety disorder [36]. The GAD-7 is not a diagnostic scale for generalized anxiety disorder nor a severity scale of the disorder. This is a self-administered questionnaire that takes less than 5 minutes to complete. The optimal cutoff for the GAD-7 was found to be a cut-off for the total score of GAD-7≥10 [36].

Measure of insomnia

The Athens Insomnia Scale (AIS) is a self-administered questionnaire to assess the patient's sleep difficulties according to the ICD-10 criteria [37]. The AIS-8 comprises 8 items (5 minutes) and is a good tool for general sleep assessment and insomnia screening, and to measure the intensity of sleep-related problems, but also as a

screening tool in reliably establishing the diagnosis of insomnia. The optimal cutoff for diagnosis to detect sleep-related troubles, for the AIS scale, is 6 [38, 39].

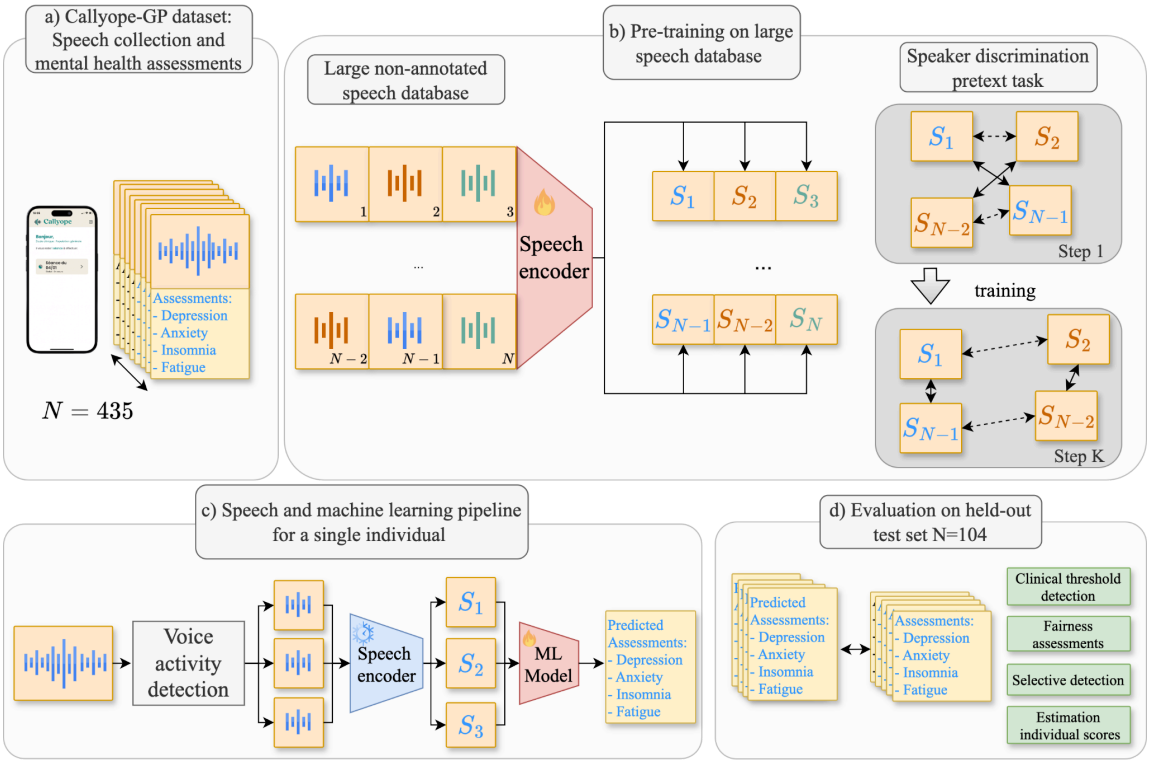
Measures of fatigue

We used the Multidimensional Fatigue Inventory (MFI) to assess the different dimensions of fatigue [40–42]. It is a short self-questionnaire (5-10 minutes) based on 20 questions to determine five dimensions of fatigue: general fatigue, physical fatigue, reduced motivation, reduced activity and mental fatigue. We also reported the total fatigue score as the sum of all sub-components.
We used the normative data from [41, 43] to choose thresholds for each sub-component. Individual sub-components of fatigue in the 75% quantile in the studied populations are all above 10. Therefore, we aimed to predict individuals' scores which are above or equal to 10, for each dimension. As mentioned also in [41], the total score has clinical significance and validity, as it was observed to have the highest correlations with anxiety, depression, and quality of life. There is no consensus cut-off for the total sum fatigue score, yet, based on the Colombian normative data [43], we observed the mean values for each studied sub-group were all above 40, therefore we chose a clinical threshold of 40 for the total sum score.

Machine Learning Analyses

Our machine learning analyses can be decomposed into three main steps: (1) the pre-training of the speech encoder model (See Figure 1b.) (2) the detection of voice speech turn and finetuning machine learning models for each mental health aspect considered in this study (See Figure 1c.) and (3) extensive evaluations of the clinical threshold detection, selective detection, fairness assessments and severity estimations for each clinical scale (See Figure 1d. and Figure 2.).

Figure 1. Overview of this study. (a) Overview of our Callyope-GP dataset, with 435 included participants (b) Flowchart of the pre-training phase of our speech encoder with the VoxCeleb2 dataset, with an illustration of the pretraining speaker embedding process. (c) Graphical illustration of our speech and machine learning pipeline for a single individual. The pre-trained speaker embedding is frozen and only machine learning models on top are trained on training data. (d) Evaluation and comparison between true and predicted assessment on held-out participants (test set N=104). Symbols S_i represent speech turn vector embeddings obtained from the speech encoder model. Colours represent speaker identities for each speech turn and embedding.



Model pretraining and tuning

Audio intensity is normalized per sample, and we established our Speech encoder model with the same architecture ResNet [44] as described in [45, 46]. The architecture is a ThinResNet model with 34 layers, which takes as input speech samples encoded as 40-Mel Spectrograms with a hop length of 10ms and a hamming window. The varying-length speech sample is pooled with a self-attentive pooling method.

We choose speaker recognition as a pretraining task as it has proven great results in mental health and neurology [11, 47]. This model is pre-trained on the VoxCeleb2 dataset [48], which is publicly available and contains over 1 million utterances from 6,112 speakers, from 145 nationalities. The VoxCeleb2 dataset consists of almost only continuous speech. Voice activity detection (VAD) is not used for the pre-training phase. The pre-training learning forces the model to organize speech in terms of speaker characteristics, as we illustrated above in the right panel of Figure 1b. We used an additive-margin softmax loss for this speaker identification task, [49].
In this work, we did not fine-tune the speech encoder model on the Callyope-GP dataset, which we represented in Figure 1c. by a frozen speech encoder model. For each speaker i, we obtained a vector representation, a speech vector embedding denoted S_i, of size 512. Extraction was performed using Python (version 3.9) and

the following packages were used to extract the acoustic features: pytorch 2.0.1, torchaudio 2.0.2, and voxceleb_trainer project [50].

For the finetuning for each task and each clinical score, we compare different machine learning algorithms on the validation set. For each task, once a model is selected, we retrain this final model on the concatenation of the training and validation sets and test on the held-out test, to avoid any inflated results. We used the scikit-learn implementation of each algorithm, splitting and evaluation [51].

For the speech collected in the Callyope-GP dataset, we used a Voice Activity Detection (VAD) system trained with pyannote-audio 3.0.0 on open and free datasets as described [52]. We used a bidirectional neural network for the VAD. Then, we applied the frozen speech encoder (the ThinResNet) to each speech turn and propagated mental health assessment labels at the speech turn level, to train and compare the final machine learning model. At inference, for final evaluation, we average predictions at the speaker level, after we obtained varying speech turns from a specific speaker.

We illustrated in Figure 2 each clinical endpoint, translated as a machine learning task based on the aforementioned procedure.

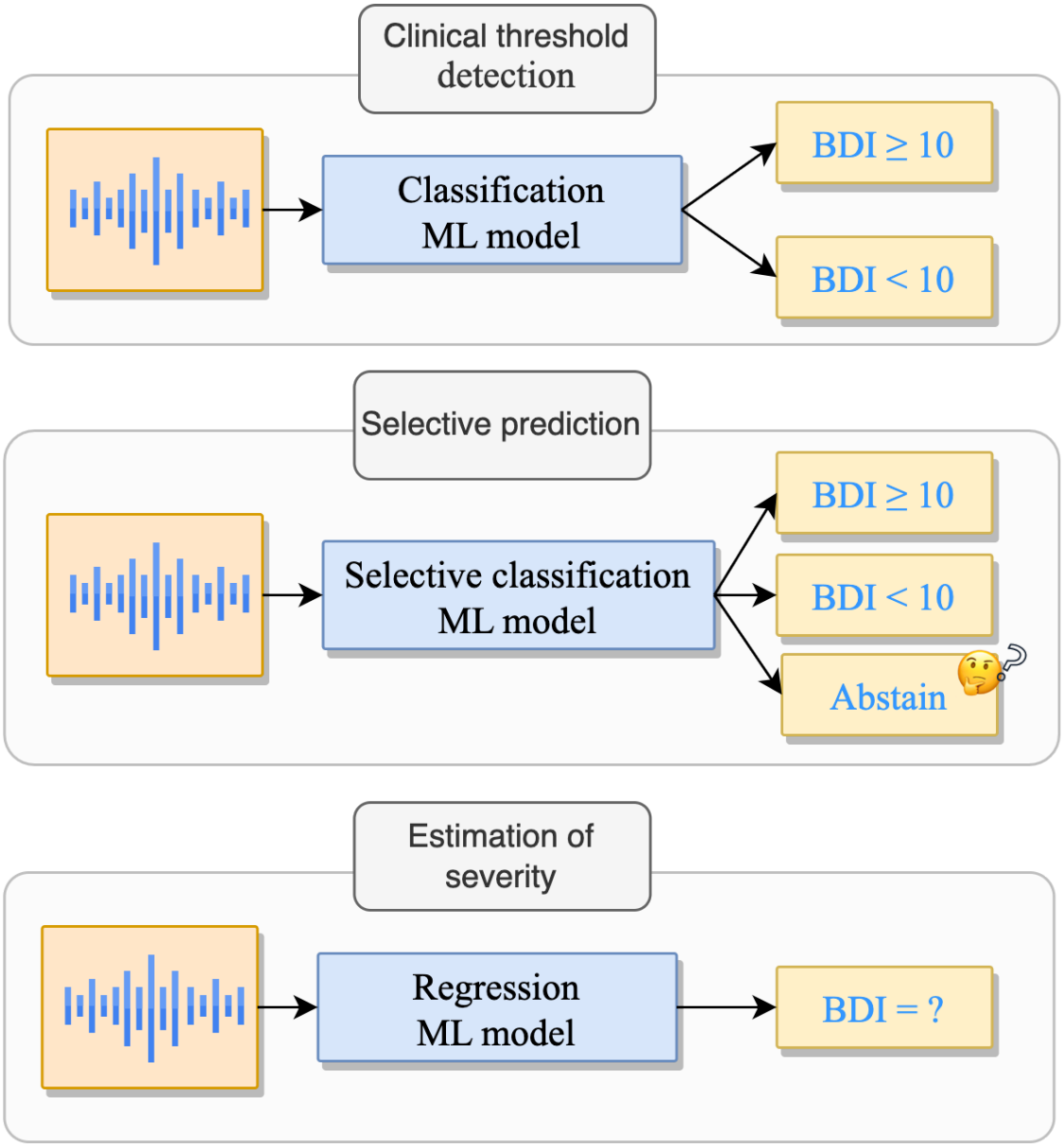


Figure 2. Schematic representation of the clinical tasks that are being assessed in this study. Each task is different in terms of the set and types of outputs. We illustrated the different tasks with the BDI clinical scale and its given threshold of 10.

Clinical threshold detection (Classification)

We first compared the predictive power of the speech encoder to discriminate between the groups to detect individuals who are below or above the threshold for each clinical scale.
The distributions of positive and negative labels vary across clinical dimensions, and to take into account imbalance we reported the performances of the Average

Precision (AP) with the Precision-Recall curve (PR) on the test set, i.e. the area under the curve for the PR at different thresholds. With random predictions of classes, the AP value is the fraction of positive samples. We showed these chance levels with stars for each clinical scale, which represents as mentioned, the fraction of the positive labels in the test set. We also compared the efficiency of our models on different tasks with the ratio $AP_{adjusted} = \frac{AP-Chance}{1-Chance}$, to adjust the performance of a model relative to the baseline performance for imbalanced classes. This ratio quantifies the model's improvement over random guessing as a fraction of the total possible improvement.
We chose to report and focus on the PR curve since it is more sensitive to settings with unbalanced data and prevents inflated results [53].
We compared linear-based models (Logistic Regression for classification with L2 regularization, Ridge regression for regression), Tree-based models (Random forests with 100 estimators) and Gradient-boosting algorithms (Histogram-Based Gradient Boosting). Even though the pre-training phase captured information about the participants' mental health, it is important to build a final model for each mental health dimension to be more specific and more sensitive. In ML terms, the mental health characteristics in speech are not necessarily linearly separable in the last vector space of the speech encoder.

Fairness assessments: Quality-of-services for sex, age, and education level demographics

Machine learning systems can behave unfairly for different reasons and in multiple ways[27]. In medicine, the use of machine learning and predictive models should be carefully evaluated, especially for potential quality-of-service harms, i.e. it can occur when a system is as performant for one specific group of people as it is for another group. We extensively studied the *quality-of-service* of harms for each clinical scale for the final predicted model, for each dimension: sex, age, and education level. We reported the Disparity Ratio (DR) based on clinical threshold detection AP scores [54], to consider both false positives and false negatives and to be more stringent than equality of opportunity. The DR is computed as the fraction of the minimum AP score across sub-groups divided by the average AP score on the full test set:
$DR = \frac{min_{group}(AP_{group})}{max_{group}(AP_{group})}$.
A higher DR is better, as the model performs equally well across groups, the perfect DR is 1, i.e. each group has the same level of performance. We used the fairlearn toolkit to perform our fairness evaluations [55].

Selective clinical threshold detection (Selective prediction)

Machine learning approaches have made great strides in several domains, yet applications to high-stakes settings remain challenging. In our case, in mental health

assessments, communicating the uncertainty associated with the system predictions appropriately is critical [56]. Yet, the communication of probabilities to human users is hard [57], and a pragmatic approach is to determine if an AI system is more likely to make erroneous predictions and defer these cases to clinicians. This setting can be cast as a selective prediction task for an AI system, the capability of a system to abstain from making a prediction when too uncertain (i.e., models saying "I don't know") [58, 59]. In this work, we followed the method from [60], we used the maximum output probabilities of the ML classification system as a way to measure uncertainty.
Based on a threshold we can obtain a specific ML system to choose to abstain when these output probabilities are too low. This specific ML system is evaluated based on the predictions it chooses to make only, thus there is a specific coverage and a specific accuracy, or risk. There is a natural tradeoff between the coverage of the ML system and its accuracy. Therefore, the way to evaluate a selective prediction task is the area under the curve for the risk-coverage curve.

Estimation of severity through predictions of individual scores (Regression)

The conventional approach in mental health assessment through speech analysis typically focuses on the group's statistical analyses, or binary classifications of categorical outcomes, primarily discerning the presence or absence of a specific dimension. Yet, the risks of depression, anxiety, fatigue and insomnia exist along a spectrum of severity levels that exert varying degrees of influence on an individual's well-being.
We go beyond traditional prediction and categorizations by integrating the estimation of severity through predictions of individual scores, employing regression machine learning models. This offers a more nuanced and comprehensive understanding of mental health dynamics, allowing for a more refined and personalized assessment. We evaluated our estimation of the severity of each total score with the Mean Absolute Error (MAE) between actual and predicted scores (Table 3). MAE score directly measures how close the predicted scores are to the actual scores without considering the direction of error. We also reported the Pearson correlation and the P value between the actual test set and the predicted values.

Results

Data overview and demographics of participants

A total of 563 participants were eligible and accepted to participate in our study, of whom 435 completed the study, giving a recruitment yield of 77.2%.
The reasons for which participants did not complete and therefore were not included in this study, were the following: a missing speech task, missing

demographic information or one missing question in a considered mental health questionnaire.

For our analyses, 231 participants were in the training set, 100 in the validation set, and 104 in the test set, and these groups did not differ in terms of demographics and mental health evaluations (cf Table 1). This yields a dataset sufficient (N>300) to evaluate error bars and predictive algorithms to avoid over-optimistic results [61]. Among the 435 participants, 109 (25.1%) were above the BDI screening threshold, and 52 (12.0%) were above the PHQ9 threshold, 51 (11.7%) above the GAD7 threshold, 165 (37.9%) above the AIS threshold, 222 (51.0%) above the MFI general fatigue threshold, 125 (28.7%) above the MFI physical fatigue threshold, 103 (23.7%) above the MFI reduced activity threshold, 159 (36.6%) above the MFI mental fatigue threshold, 85 (19.5%) above the MFI reduced motivation threshold, and 227 (52.2%) above the threshold of the total score of the MFI. We reported co-occurrences of people at risk for each dimension in the Figure S1 in Multimedia Appendix 1.

Table 1. Demographic characteristics of the training, validation and test groups[a-e]. Categorical variables are compared with Pearson's χ^2 test and one-way ANOVA for continuous variables.
PHQ-9: Patient Health Questionnaire, BDI: Beck Depression Inventory, GAD-7: General Anxiety Disorder 7-item scale, AIS: Athens Insomnia Scale, MFI: Multidimensional Fatigue Inventory

Characteristics[a]	Training	Validation	Test		Group comparisons	
Number of participants	231	100	104		Statistic	P value
Demographics						
Sex (F/M)	140/91	63/37	62/42		χ^2=0.27	0.88
Education level[b]	11/32/16/172	1/15/12/72	6/17/7/74		χ^2=6.23	0.40
Age in years mean (sd) [min-max]	38.7 (17.7) [18.5-84.7]	40.2 (18.0) [18.5-84.5]	40.3 (17.1) [18.5-80.6]		F=0.31	0.73
Clinical evaluation mean (sd) [min-max] (#negative/#positive[c])						
BDI	7.1 (6.4) [0.0-30.0] (170/61)	6.9 (6.6) [0.0-26.0] (75/25)	6.3 (5.8) [0.0-23.0] (81/23)		F=0.58	0.56
PHQ-9	5.0 (4.0) [0.0-21.0] (206/25)	5.0 (4.2) [0.0-21.0] (86/14)	5.2 (4.2) [0.0-21.0] (91/13)		F=0.12	0.88

GAD-7	4.5 (4.0) [0.0-21.0] (206/25)	4.7 (4.8) [0.0-21.0] (86/14)	4.9 (4.0) [0.0-20.0] (92/12)		F=0.98	0.38
AIS	5.0 (3.6) [0.0-19.0] (142/89)	5.0 (3.8) [0.0-19.0] (60/40)	4.9 (3.7) [0.0-18.0] (63/41)		F=0.02.	0.98
MFI general fatigue	10.0 (4.1) [4.0-20.0] (110/121)	10.0 (4.6) [4.0-20.0] (51/49)	9.7 (4.6) [4.0-19.0] (52/52)		F=0.27	0.76
MFI physical fatigue	7.9 (3.3) [4.0-18.0] (166/65)	7.6 (3.5) [4.0-17.0] (73/27)	7.9 (3.1) [4.0-15.0] (71/33)		F=0.31	0.73
MFI reduced activity	7.5 (3.3) [4.0-19.0] (173/58)	7.5 (3.5) [4.0-18.0] (78/22)	7.4 (3.1) [4.0-19.0] (81/23)		F=0.04	0.95
MFI mental fatigue	8.5 (3.8) [4.0-19.0] (147/84)	8.9 (4.3) [4.0-20.0] (64/36)	8.7 (4.2) [4.0-20.0] (65/39)		F=0.27	0.76
MFI reduced motivation	7.1 (2.9) [4.0-18.0] (183/48)	6.9 (2.9) [4.0-16.0] (83/17)	7.3 (2.9) [4.0-18.0] (84/20)		F=0.49	0.61
MFI total score	41.1 (13.0) [13.0-88.0] (106/125)	40.9 (14.4) [20.0-79.0] (50/50)	41.0 (13.7) [20.0-89.0] (52/52)		F=0.01	0.99

[a]Characteristics for the three splits of the dataset to ensure generalization
[b]Highest achieved study level: no diploma/secondary studies/ short post-baccalaureate studies (Bac+2)/Long post-baccalaureate studies (Bac+3 and above)
[c]In addition to statistics for each clinical scale score, we report the number (#) of participants below and above cut-off: cf methods sections for each threshold.

Machine Learning Analyses

We found out that the performance of the Gradient-boosting algorithms always outperformed the linear-based models and the tree-based models on the validation set. Thus, we retrained Gradient-boosting algorithms on the combination of the training and validation sets and reported the final results on the held-out test set in Figure 3, Figure 4, Table 2, Table 3 and Figure S2 in Multimedia Appendix 1.

Clinical threshold detection (Classification)

The clinical threshold detection performed well based on the speech data and our developed system (Figure 3). All systems outperformed the chance levels by a large margin (represented by stars for each clinical score). The chance level is based on

the ratio of the number of above-clinical threshold participants in the held-out test set. The classification results were the highest for the MFI General Fatigue detection (AP=0.91) and the lowest for the anxiety GAD-7 (AP=0.59). The described system obtained strong performances to predict all fatigue measures and insomnia thresholds (AP>0.7).
We observed the largest improvements (cf Table S1 in Multimedia Appendix 1) over the chance for the BDI and MFI General Fatigue component (BDI $AP_{adjusted} = 0.78$, MFI General Fatigue $AP_{adjusted} = 0.82$).

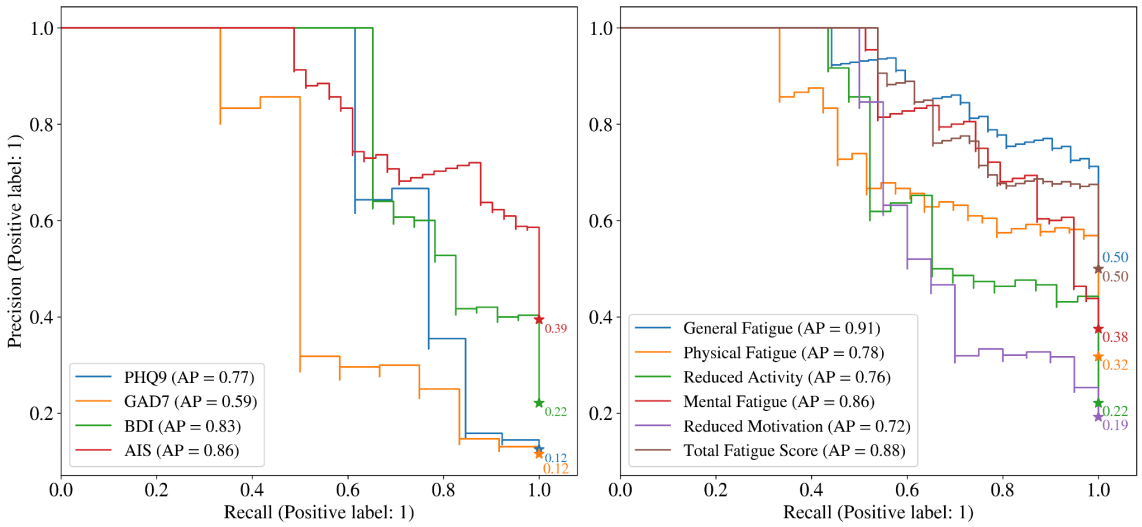


Figure 3. Precision-Recall (PR) curves for the clinical threshold detection task on the held-out test set with Average Precision (AP) scores. Precision-recall curves prioritize the balance between true positives and false positives, making them more informative for imbalanced datasets compared to ROC curves. (Left panel) PR curves to detect clinically relevant thresholds for depression (PHQ-9, BDI), anxiety (GAD-7), and insomnia (AIS). (Right panel): PR curves to detect clinically relevant thresholds for fatigue components (MFI), general fatigue, physical fatigue, reduced activity, mental fatigue, reduced motivation and total fatigue.
Coloured stars represent the chance level for each threshold as defined by the positive ratio in the test set. PHQ-9: Patient Health Questionnaire, BDI: Beck Depression Inventory, GAD-7: General Anxiety Disorder 7-item scale, AIS: Athens Insomnia Scale, MFI: Multidimensional Fatigue Inventory.

Fairness assessments: Quality-of-services for sex, age, and education level

We computed the sensitive attributes DRs (sex, age, and education level) and assessed the differences in the quality of the predictions made by the speech-based system (Table 2). Overall, the speech-based system had a better quality-of-service for sex (mean 0.77, SD 0.30), and the worst was for age (mean 0.26, SD 0.26). We also identified that the GAD-7 had the worst quality-of-service disparity (Mean of DRs = 0.05), and the best quality-of-service was obtained with the score AIS (Mean

of DRs = 0.77), the BDI score (Mean of DRs = 0.67), and the MFI total score (Mean of DRs = 0.65). We also observed that only the AIS obtained a good performance for age (AIS Age DR=0.77), and, except for PHQ-9 and GAD-7, all DRs were satisfactory or high for sex.

Table 2. Disparity ratios (DRs) based on the AP scores for sex, age and education levels to assess fairness, on the held-out test set, for the clinical threshold detection (Classification) for the different considered dimensions of mental health. Higher DR is better, and 1 is perfect DR. Mean value per score is reported. Mean and Standard Deviation are reported per sensitive dimension.
PHQ-9: Patient Health Questionnaire, BDI: Beck Depression Inventory, GAD-7: General Anxiety Disorder 7-item scale, AIS: Athens Insomnia Scale, MFI: Multidimensional Fatigue Inventory.

	Sex DR	**Age DR**[a]	**Education level DR** [b]		**Mean**
PHQ-9	0.40	0.00	0.51		0.30
GAD-7	0.14	0.00	0.00		0.05
BDI	0.93	0.44	0.62		0.67
AIS	0.97	0.77	0.56		0.77
MFI General Fatigue	0.98	0.23	0.63		0.61
MFI Physical Fatigue	0.83	0.35	0.44		0.54
MFI Reduced Activity	0.62	0.11	0.46		0.40
MFI Mental Fatigue	0.97	0.00	0.65		0.54
MFI Reduced Motivation	0.94	0.33	0.00		0.42

MFI Total Fatigue Score	0.99	0.37	0.58		0.65
Mean (SD)	0.77 (0.30)	0.26 (0.26)	0.44 (0.24)		0.49 (0.21)

[a]Participants are grouped into age categories to allow analysis: [18,30),[30,45),[45,65), >65
[b]Highest achieved study level: no diploma/secondary studies/ short post-baccalaureate studies (Bac+2)/Long post-baccalaureate studies (Bac+3 and above)

Selective clinical threshold detection (Selective prediction)

We also evaluated the capabilities of speech-based models to selectively predict the different clinical thresholds. We achieved great performances for all considered clinical scores (Figure 4). The model that selectively predicts the risk of depression based on the BDI score achieved the best result (BDI risk-coverage AUC=0.02). Besides this model on the BDI score could achieve a perfect prediction for more than 50% coverage of the held-out test set. AIS, BDI, MFI Reduced activity, MFI Total Fatigue Score achieved a zero risk for more than 50% coverage. The scores PHQ-9, GAD-7, MFI Mental Fatigue and MFI-reduced motivation could not achieve such feats of important coverage with no risk.

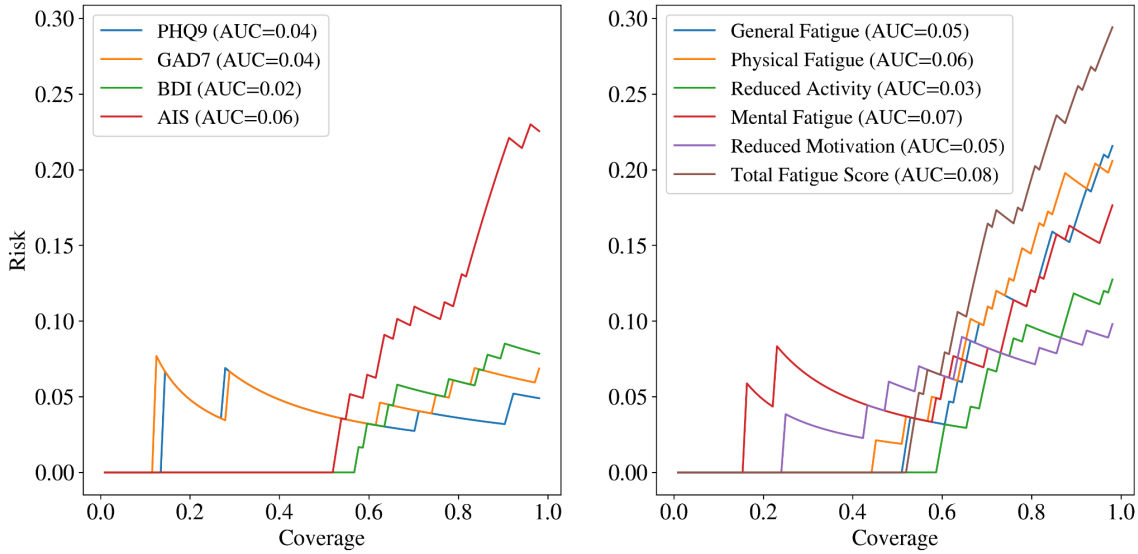


Figure 4. Risk-Coverage curves on the held-out test set for the selective clinical threshold detection task on the held-out test set illustrate the models' selective

screening ability, i.e., risk detection capabilities with the ability to abstain when too uncertain. A lower AUC is better, and 0 is the perfect score. (Left panel) Risk-Coverage curves selectively detect clinically relevant thresholds for depression (PHQ-9, BDI), anxiety (GAD-7), and insomnia (AIS). (Right panel): Risk-Coverage curves to detect clinically relevant thresholds for fatigue components (MFI), general fatigue, physical fatigue, reduced activity, mental fatigue, reduced motivation and total fatigue. PHQ-9: Patient Health Questionnaire, BDI: Beck Depression Inventory, GAD-7: General Anxiety Disorder 7-item scale, AIS: Athens Insomnia Scale, MFI: Multidimensional Fatigue Inventory.

We reported the regression results for the estimation of severity in Table 3. Speech-based models obtained significant results for all clinical variables, based on the evaluation with the Pearson correlations (all $P < 1\times10^{-14}$). The strongest correlations between the prediction and the actual scores on the held-out test were found for MFI general fatigue (r=0.75), MFI Mental Fatigue (r=0.74) and the PHQ-9 (r=0.74). The lowest correlation was found for the MFI Physical fatigue (r=0.59).

Speech-based models also obtained great results in terms of absolute errors. We observed less than 2 points of MAE for PHQ-9 and GAD-7 while we observed that these scales were between 0 and 21. All other scores were predicted on average with less than 3 points, except for the MFI Total Fatigue Score, since its range is [13-88].

Table 3. Estimation of severity results on the held-out test set for the different considered dimensions of mental health (Regression). We reported Mean Absolute Errors (MAE) and Pearson correlations between actual and predicted values. Lower MAE is better, and 0 is perfect. A higher Pearson correlation is better, +1 is perfect. All correlations were significant $P < 1\times10^{-14}$. PHQ-9: Patient Health Questionnaire, BDI: Beck Depression Inventory, GAD-7: General Anxiety Disorder 7-item scale, AIS: Athens Insomnia Scale, MFI: Multidimensional Fatigue Inventory.

	Pearson *r*	**MAE**
PHQ-9	0.74	1.64
GAD-7	0.66	1.76
BDI	0.67	2.61
AIS	0.66	1.72
MFI General Fatigue	0.75	1.37

MFI Physical Fatigue	0.59	1.59
MFI Reduced Activity	0.69	1.31
MFI Mental Fatigue	0.74	1.69
MFI Reduced Motivation	0.71	1.18
MFI Total Fatigue Score	0.69	6.07

Discussions

We aimed to explore the full capabilities and limitations of using speech data extracted from 435 participants in the general population to predict the presence or absence of different mental health self-reported symptoms: depression, anxiety, insomnia and the different dimensions of fatigue.
We built a fully automated speech-based machine learning system that takes as input the audio waveform collected from a simple speech task performed on a smartphone app. Our models were trained and calibrated on training and validation sets of participants, and we demonstrated the system's generalisation on the held-out test set of 104 participants.
Our results indicated that ML-based systems using speech only as input could successfully identify participants above clinical thresholds for depression, insomnia and fatigue components, but to a lesser extent anxiety. All classification results were above chance levels for each clinical threshold.
This result was confirmed with an extensive fairness analysis of quality-of-service for age, sex and levels of education. Depression, insomnia and different dimensions of fatigue clinical threshold detection results were particularly consistent for sex, slightly less for age and a lesser extent for education level. Anxiety risk identification fell behind in accuracy overall and was also unequal per group. The extension of our clinical threshold detection system to be able to abstain, with selective prediction, was conclusive, even for anxiety. All risk-coverage areas under the curves remained low, with all Risk-coverage AUCs below 0.1. Finally, we showed that speech-based models can accurately predict the exact scores, moving beyond binary interpretations of score thresholds. All correlations between the predicted scores and the actual scores given by participants were significant, exhibiting strengths ranging from 0.59 to 0.75.

Our study builds upon existing mental health research on speech analysis and extends the insights for deployment into clinical practice. For risk and anxiety depression in the general population, we found similar strong performances such as [25, 26, 62, 63]. Our recruitment and involvement of participants was in person. For

medium-sized datasets, below 1000 participants, such as ours, it can be observed that web-based crowd-sourced recruitment of participants needs more training data to yield the same level of model accuracy [62, 64], as models trained on speech with face-to-face recruitment. This was also observed in a large study, with over 6000 participants, with online data recruitment for risk detection in the general population in American English [63]. This discrepancy could be explained by the fact that data quality is variable on web crowdsourcing platforming, especially for participants' psychiatric evaluations [65], and also voice recording through laptops.

Our study revealed discrepancies in both clinical threshold detection and estimation of severity through self-reported depression scores between BDI and PHQ-9. This underlines the inherent limitations of score conversion and the crucial role of individual-level assessment in capturing the nuanced and different expressions of depression [34]. This reinforces the necessity of developing assessment tools and interpreting results with meticulous attention to individual variability, particularly by scrutinizing model performance at the individual level, mirroring real-world clinical scenarios.

Our study uniquely addresses the co-occurrence of perceived fatigue and reported insomnia, both prevalent mental health concerns are not detected simultaneously through speech analysis. While prior research, like the work by [19], has explored how sleep deprivation impacts specific vocal features like prosody and voice quality, no previous study has delved into the combined influence of fatigue and insomnia on speech. Addressing this gap is crucial because these conditions often co-occur and significantly impact the symptom trajectory and potential development of other mental health issues [66]. The prevention of sleep recurrent problems can prevent the installation of other mental health troubles or relapses.

Furthermore, to the best of our knowledge, this is the first study to assess the fairness and selective prediction capabilities in speech-based mental health assessments. This is of prime importance since speech signals can be heavily influenced by a multitude of factors such as age, sex, weight, and height [67]. Age factor was the least preserved in our grouped performances, this can be attributed to change in voice changes due to hormones [68], and normal ageing affects the different parts of the vocal production system: larynx, respiratory system, resonators, saliva system and the individual's global emotional status [69]. In addition to lower performances compared to other mental health dimensions, the anxiety risk detection performance collapsed for certain groups of demographics. This could be explained by the heterogeneity and low positive examples in our Callyope-GP dataset. Even though there are limits concerning some groups of individuals, selective classification offers an option to potentially remediate these variable quality-of-services, ensuring a deployment in clinical settings and still bringing overall clinical utility.

Limitations

While valuable, this study has some limitations. The French monolingual, medium sample size dataset (300<N<1000) needs more diverse data for complete generalizability, and the non-longitudinal design misses insights on symptom evolution. Besides, reliance on self-questionnaires introduces potential bias, relying on the insight of participants. Future research with larger samples, longitudinal designs, and the inclusion of severe presentations is crucial for realizing the full potential of voice biomarkers in mental health. Studying a range of speech tasks and prompts also holds the potential for multiple benefits related to user adherence. Tailored tasks can address specific mental health needs, catering to individual preferences and boosting engagement. Varied prompts can prevent user fatigue and maintain interest, leading to more consistent system use and richer data collection.

Conclusions

In conclusion, this study demonstrates the potential of speech-based systems for detecting and predicting various mental health symptoms in the general population. While challenges remain regarding real-world application and ensuring fairness across demographics, our findings pave the way for further development and responsible integration of such tools into clinical settings, potentially advancing personalized mental health assessment and intervention. In future work, we will investigate the extension of this study for longitudinal data, more diverse linguistic and geographic data, and extend this study for severely affected populations, already followed by mental health practitioners. We will also look into fairness and uncertainty mitigation methods to improve the performance of our systems.

Acknowledgements
The authors are thankful to all the participants who volunteered for this research study. Without their active involvement, the present study would not have been possible. We also would like to thank each of the speech pathology interns who helped with the subject recruitment and made sure that the protocol was completed successfully.

Conflicts of Interest

RR, XN, AL, MdG, MD, and AB are shareholders of Callyope and VO was an employee of Callyope.

Abbreviations
JMIR: Journal of Medical Internet Research
RCT: randomized controlled trial
ML: Machine Learning

PHQ-9: Patient Health Questionnaire
BDI: Beck Depression Inventory
GAD-7: General Anxiety Disorder 7-item scale
AIS: Athens Insomnia Scale
MFI: Multidimensional Fatigue Inventory.
DR: Disparity Ratio to measure fair quality-of-service
n.s.:Non significant
MAE: Mean Absolute Error
PR: Precision-Recall curve
ROC: Receiver Operating Characteristic curve
AP: Average Precision (Area under the curve) of the Precision-Recall curves
$AP_{adjusted}$: Average Precision (Area under the curve) of the Precision-Recall curves adjusted by the chance level $AP_{adjusted} = \frac{AP - Chance}{1 - Chance}$
AUC: Area Under the Curve

Appendix

Figure S1. Co-occurrences percentage of people at risk for each dimension, normalization is done per row: first row can be read as 56% of individuals who are at risk for depression based on the PHQ-9 are at risk of GAD-7.

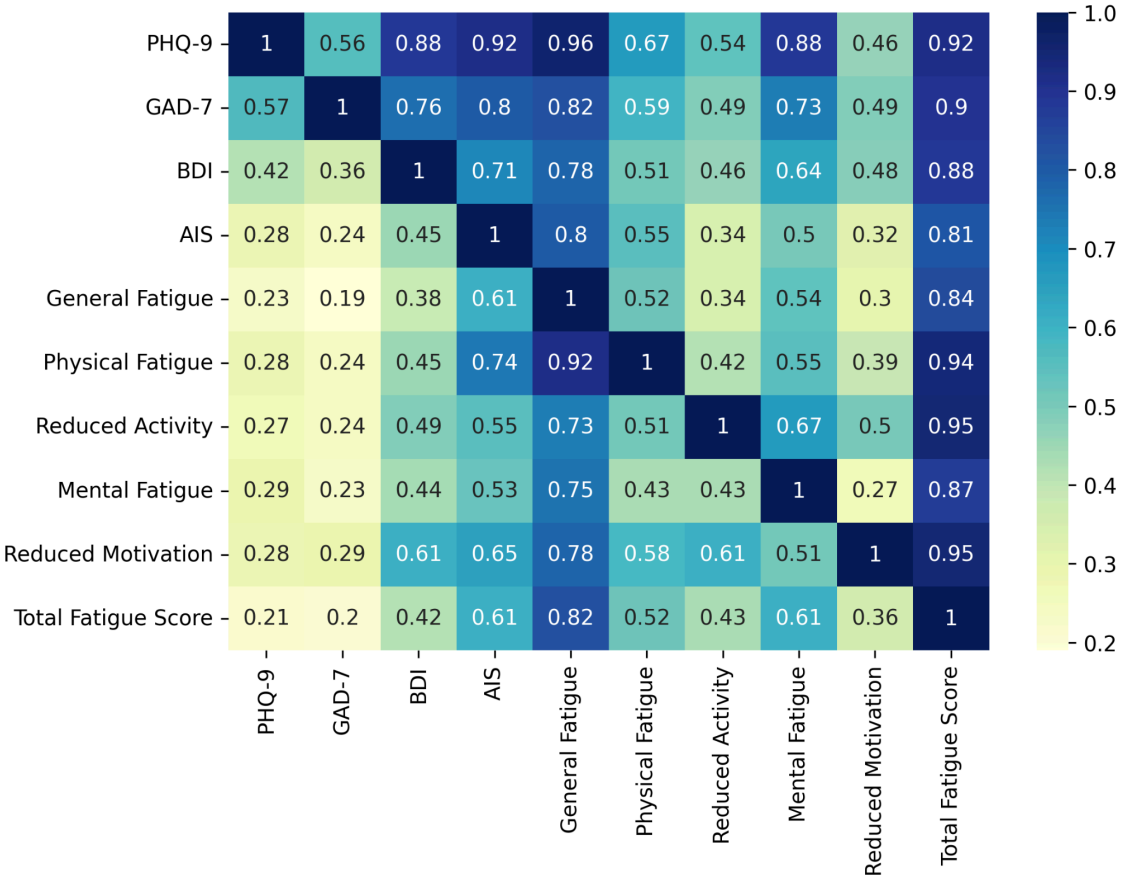


Figure S2. ROC curves for the clinical threshold detection task on the held-out test set.
(Left panel) PR curves to detect clinically relevant thresholds for depression (PHQ-9, BDI), anxiety (GAD-7), and insomnia (AIS). (Right panel): ROC curves to detect clinically relevant thresholds for fatigue components (MFI), general fatigue, physical fatigue, reduced activity, mental fatigue, reduced motivation and total fatigue.
PHQ-9: Patient Health Questionnaire, BDI: Beck Depression Inventory, GAD-7: General Anxiety Disorder 7-item scale, AIS: Athens Insomnia Scale, MFI: Multidimensional Fatigue Inventory.

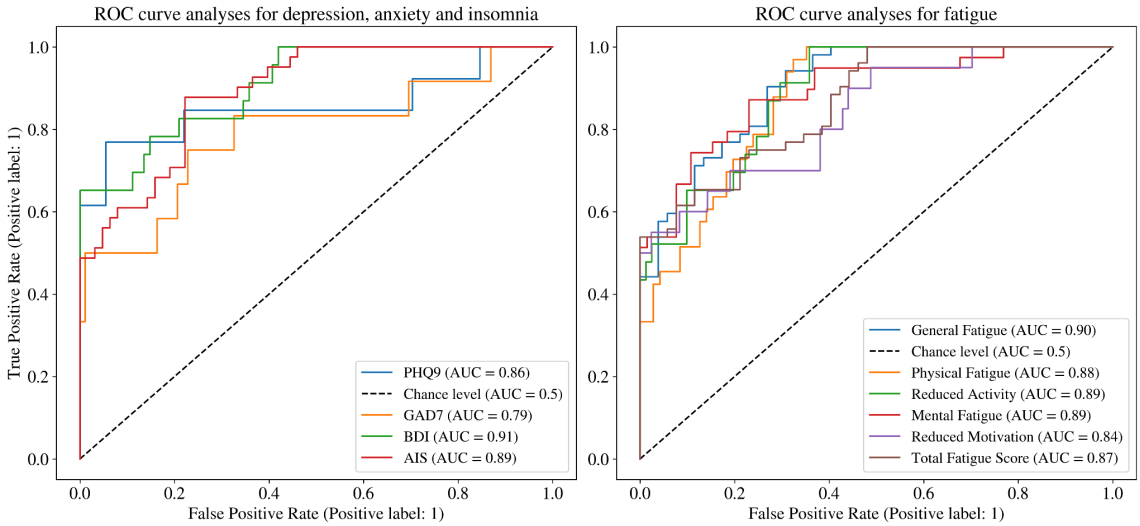


Table S1. Estimation of improvements over chance for clinical threshold detection severity results on the held-out test set for the different considered dimensions of mental health (Classification). We reported adjusted AP values ($AP_{adjusted} = \frac{AP-Chance}{1-Chance}$). Higher $AP_{adjusted}$ is better, and 1 is perfect. PHQ-9: Patient Health Questionnaire, BDI: Beck Depression Inventory, GAD-7: General Anxiety Disorder 7-item scale, AIS: Athens Insomnia Scale, MFI: Multidimensional Fatigue Inventory.

	$AP_{adjusted}$
PHQ-9	0.74
GAD-7	0.51
BDI	0.78
AIS	0.70
MFI General Fatigue	0.82
MFI Physical Fatigue	0.68
MFI Reduced Activity	0.69

MFI Mental Fatigue	0.77
MFI Reduced Motivation	0.65
MFI Total Fatigue Score	0.76

References

1. Vos T, Lim SS, Abbafati C, et al (2020) Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet 396:1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9
2. (2022) Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 9:137–150. https://doi.org/10.1016/S2215-0366(21)00395-3
3. Santomauro DF, Herrera AMM, Shadid J, et al (2021) Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. The Lancet 398:1700–1712. https://doi.org/10.1016/S0140-6736(21)02143-7
4. Henriksson M, Aro H, Marttunen M, et al (1993) Mental Disorders and Comorbidity in Suicide. Am J Psychiatry 150:935–40. https://doi.org/10.1176/ajp.150.6.935
5. Measurement-Based Care Versus Standard Care for Major Depression: A Randomized Controlled Trial With Blind Raters | American Journal of Psychiatry. https://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.2015.14050652. Accessed 13 Feb 2024
6. Zimmerman M, Ellison W, Young D, et al (2015) How many different ways do patients meet the diagnostic criteria for major depressive disorder? Compr Psychiatry 56:29–34. https://doi.org/10.1016/j.comppsych.2014.09.007
7. Soehner AM, Kaplan KA, Harvey AG (2014) Prevalence and clinical correlates of co-occurring insomnia and hypersomnia symptoms in depression. J Affect Disord 167:93–97. https://doi.org/10.1016/j.jad.2014.05.060
8. Corfield EC, Martin NG, Nyholt DR (2016) Co-occurrence and symptomatology of fatigue and depression. Compr Psychiatry 71:1–10. https://doi.org/10.1016/j.comppsych.2016.08.004
9. Park LT, Zarate CA (2019) Depression in the Primary Care Setting. N Engl J Med 380:559–568. https://doi.org/10.1056/NEJMcp1712493
10. Cummins N, Dineley J, Conde P, et al (2023) Multilingual markers of depression in remotely collected speech samples: A preliminary analysis. J Affect Disord

341:128–136. https://doi.org/10.1016/j.jad.2023.08.097
11. Dumpala SH, Dikaios K, Rodriguez S, et al (2023) Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity. Sci Rep 13:11155. https://doi.org/10.1038/s41598-023-35184-7
12. Bone D, Gibson J, Chaspari T, et al (2016) Speech and language processing for mental health research and care. In: 2016 50th Asilomar Conference on Signals, Systems and Computers. pp 831–835
13. Cummins N, Scherer S, Krajewski J, et al (2015) A review of depression and suicide risk assessment using speech analysis. Speech Commun 71:10–49. https://doi.org/10.1016/j.specom.2015.03.004
14. Williamson J, F. Quatieri T, Helfer B, et al (2014) Vocal and Facial Biomarkers of Depression based on Motor Incoordination and Timing. pp 65–72
15. Teferra BG, Borwein S, DeSouza DD, et al (2022) Acoustic and Linguistic Features of Impromptu Speech and Their Association With Anxiety: Validation Study. JMIR Ment Health 9:e36828. https://doi.org/10.2196/36828
16. Stasak B, Epps J, Schatten HT, et al (2021) Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. Speech Commun 132:10–20. https://doi.org/10.1016/j.specom.2021.05.004
17. Bayerl SP, Wagner D, Baumann I, et al (2023) Detecting Vocal Fatigue with Neural Embeddings. J Voice. https://doi.org/10.1016/j.jvoice.2023.01.012
18. Elbéji A, Zhang L, Higa E, et al (2022) Vocal biomarker predicts fatigue in people with COVID-19: results from the prospective Predi-COVID cohort study. BMJ Open 12:e062463. https://doi.org/10.1136/bmjopen-2022-062463
19. Thoret E (2024) Sleep deprivation detected by voice analysis. PLoS Comput Biol 20:e1011849. https://doi.org/10.1371/journal.pcbi.1011849
20. Levelt WJM (1989) Speaking: from intention to articulation. MIT Press, Cambridge, Mass.
21. Guenther FH (2016) Neural Control of Speech. MIT Press
22. Low DM, Bentley KH, Ghosh SS (2020) Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investig Otolaryngol 5:96–116. https://doi.org/10.1002/lio2.354
23. Berta A, Miguel Ángel C, Clara G-S, Rubén H (2022) A bibliometric analysis of 10 years of research on symptom networks in psychopathology and mental health. Psychiatry Res 308:114380. https://doi.org/10.1016/j.psychres.2021.114380
24. Borsboom D (2017) A network theory of mental disorders. World Psychiatry Off J World Psychiatr Assoc WPA 16:5–13. https://doi.org/10.1002/wps.20375
25. König A, Tröger J, Mallick E, et al (2022) Detecting subtle signs of depression with automated speech analysis in a non-clinical sample. BMC Psychiatry 22:830. https://doi.org/10.1186/s12888-022-04475-0
26. Lin D, Nazreen T, Rutowski T, et al (2022) Feasibility of a Machine Learning-Based Smartphone Application in Detecting Depression and Anxiety in a Generally Senior Population. Front Psychol 13:811517. https://doi.org/10.3389/fpsyg.2022.811517
27. Barocas S, Hardt M, Narayanan A (2023) Fairness and machine learning: Limitations and opportunities. MIT Press
28. Martin VP, Rouas J-L (2024) Estimating symptoms and clinical signs instead of

disorders: the path toward the clinical use of voice and speech biomarkers in psychiatry
29. Bondi E, Koster R, Sheahan H, et al (2022) Role of Human-AI Interaction in Selective Prediction. Proc AAAI Conf Artif Intell 36:5286–5294. https://doi.org/10.1609/aaai.v36i5.20465
30. Boschi V, Catricalà E, Consonni M, et al (2017) Connected Speech in Neurodegenerative Language Disorders: A Review. Front Psychol 8:. https://doi.org/10.3389/fpsyg.2017.00269
31. Beck AT, Ward CH, Mendelson M, et al (1961) An inventory for measuring depression. Arch Gen Psychiatry 4:561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004
32. Kroenke K, Spitzer RL, Williams JB (2001) The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 16:606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x
33. Cameron IM, Crawford JR, Lawton K, Reid IC (2008) Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. Br J Gen Pract J R Coll Gen Pract 58:32–36. https://doi.org/10.3399/bjgp08X263794
34. Hawley CJ, Gale TM, Smith PSJ, et al (2013) Equations for converting scores between depression scales (MÅDRS, SRS, PHQ-9 and BDI-II): good statistical, but weak idiographic, validity. Hum Psychopharmacol Clin Exp 28:544–551. https://doi.org/10.1002/hup.2341
35. Beck AT, Steer RA, Garbin MG (1988) Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clin Psychol Rev 8:77–100. https://doi.org/10.1016/0272-7358(88)90050-5
36. Spitzer RL, Kroenke K, Williams JBW, Löwe B (2006) A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 166:1092–1097. https://doi.org/10.1001/archinte.166.10.1092
37. Soldatos CR, Dikeos DG, Paparrigopoulos TJ (2000) Athens Insomnia Scale: validation of an instrument based on ICD-10 criteria. J Psychosom Res 48:555–560. https://doi.org/10.1016/S0022-3999(00)00095-7
38. Soldatos CR, Dikeos DG, Paparrigopoulos TJ (2003) The diagnostic validity of the Athens Insomnia Scale. J Psychosom Res 55:263–267. https://doi.org/10.1016/S0022-3999(02)00604-9
39. Okajima I, Miyamoto T, Ubara A, et al (2020) Evaluation of Severity Levels of the Athens Insomnia Scale Based on the Criterion of Insomnia Severity Index. Int J Environ Res Public Health 17:8789. https://doi.org/10.3390/ijerph17238789
40. Smets EMA, Garssen B, Bonke B, De Haes JCJM (1995) The multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res 39:315–325. https://doi.org/10.1016/0022-3999(94)00125-O
41. Schwarz R, Krauss O, Hinz A (2003) Fatigue in the General Population. Onkologie 26:140–144. https://doi.org/10.1159/000069834
42. Gentile S, DELAROZIèRE JC, Favre F, et al (2003) Validation of the French 'multidimensional fatigue inventory' (MFI 20). Eur J Cancer Care (Engl) 12:58–64. https://doi.org/10.1046/j.1365-2354.2003.00295.x
43. Hinz A, Finck C, Barradas S, et al (2013) Fatigue in the General Population of

Colombia - Normative Values for the Multidimensional Fatigue Inventory MFI-20. Onkologie 36:403–7. https://doi.org/10.1159/000353606
44. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp 770–778
45. Chung JS, Huh J, Mun S (2020) Delving into VoxCeleb: Environment Invariant Speaker Recognition. In: The Speaker and Language Recognition Workshop (Odyssey 2020). ISCA, pp 349–356
46. Chung JS, Huh J, Mun S, et al (2020) In defence of metric learning for speaker recognition. In: Interspeech 2020. pp 2977–2981
47. Arias-Vergara T, Vásquez-Correa JC, Orozco-Arroyave JR, Nöth E (2018) Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. Speech Commun 101:11–25
48. Chung JS, Nagrani A, Zisserman A (2018) VoxCeleb2: Deep Speaker Recognition. In: Interspeech 2018. pp 1086–1090
49. Wang F, Cheng J, Liu W, Liu H (2018) Additive Margin Softmax for Face Verification. IEEE Signal Process Lett 25:926–930. https://doi.org/10.1109/LSP.2018.2822810
50. (2024) clovaai/voxceleb_trainer
51. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
52. Bredin H (2023) pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In: 24th INTERSPEECH Conference (INTERSPEECH 2023). ISCA, Dublin, Ireland, pp 1983–1987
53. Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. Association for Computing Machinery, New York, NY, USA, pp 233–240
54. Agarwal A, Beygelzimer A, Dudik M, et al (2018) A Reductions Approach to Fair Classification. In: Proceedings of the 35th International Conference on Machine Learning. PMLR, pp 60–69
55. Weerts H, Dudík M, Edgar R, et al (2023) Fairlearn: Assessing and Improving Fairness of AI Systems. J Mach Learn Res 24:1–8
56. Kompa B, Snoek J, Beam AL (2021) Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med 4:4. https://doi.org/10.1038/s41746-020-00367-3
57. Zhang H, Maloney LT (2012) Ubiquitous Log Odds: A Common Representation of Probability and Frequency Distortion in Perception, Action, and Cognition. Front Neurosci 6:1. https://doi.org/10.3389/fnins.2012.00001
58. Xin J, Tang R, Yu Y, Lin J (2021) The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing. In: Zong C, Xia F, Li W, Navigli R (eds) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp 1040–1051
59. El-Yaniv R, Wiener Y On the Foundations of Noise-free Selective Classification
60. Hendrycks D, Gimpel K (2016) A Baseline for Detecting Misclassified and

Out-of-Distribution Examples in Neural Networks
61. Varoquaux G (2018) Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 180:68–77
62. Tasnim M, Ehghaghi M, Diep B, Novikova J (2022) DEPAC: a Corpus for Depression and Anxiety Detection from Speech. In: Zirikly A, Atzil-Slonim D, Liakata M, et al (eds) Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology. Association for Computational Linguistics, Seattle, USA, pp 1–16
63. Rutowski T, Harati A, Shriberg E, et al (2022) Toward Corpus Size Requirements for Training and Evaluating Depression Risk Models Using Spoken Language. In: Interspeech 2022. ISCA, pp 3343–3347
64. Teferra BG, Rose J (2023) Predicting Generalized Anxiety Disorder From Impromptu Speech Transcripts Using Context-Aware Transformer-Based Neural Networks: Model Evaluation Study. JMIR Ment Health 10:e44325. https://doi.org/10.2196/44325
65. Kolc KL, Tan YXK, Lo AZY, et al (2023) Measuring psychiatric symptoms online: A systematic review of the use of inventories on Amazon Mechanical Turk (mTurk). J Psychiatr Res 163:118–126. https://doi.org/10.1016/j.jpsychires.2023.05.027
66. McCallum SM, Batterham PJ, Calear AL, et al (2019) Associations of fatigue and sleep disturbance with nine common mental disorders. J Psychosom Res 123:109727. https://doi.org/10.1016/j.jpsychores.2019.05.005
67. Sorokowski P, Pisanski K, Frąckowiak T, et al (2023) Voice-based judgments of sex, height, weight, attractiveness, health and psychological traits based on free speech versus scripted speech. Psychon Bull Rev
68. Amir O, Biron-Shental T (2004) The impact of hormonal fluctuations on female vocal folds. Curr Opin Otolaryngol Head Neck Surg 12:180–184. https://doi.org/10.1097/01.moo.0000120304.58882.94
69. Beton S, Yücel L, Başak H, Çiler Büyükatalay Z (2022) The Elderly Voice: Mechanisms, Disorders and Treatment Methods. Turk Arch Otorhinolaryngol 60:220–226. https://doi.org/10.4274/tao.2022.2022-8-1