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 on other factors than mere accuracy, in the general population.

Methods: We included n=865 healthy adults and recorded their answers regarding their perceived mental and sleep states. We asked how they felt and if they had slept well lately. Clinically validated questionnaires measuring depression, anxiety, insomnia, and fatigue severity were also used. We developed a novel speech and machine learning pipeline involving voice activity detection, feature extraction, and model training. We automatically analyzed participants' speech with a fully ML automatic pipeline to capture speech variability. Then, we modelled speech with pretrained deep learning models that were pre-trained on a large open free database and we selected the best one on the validation set. Based on the best

speech modelling approach, 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: The best model was WhisperM with a max pooling, and oversampling method. Our methods achieved good detection performance for all symptoms, depression (PHQ-9 AUC= 0.76F1=0.49, BDI AUC=0.78, F1=0,65), anxiety (GAD-7 F1=0.50, AUC=0.77) insomnia (AIS AUC=0.73, F1=0.62), and fatigue (MFI Total Score F1=0.88, AUC=0.68). These strengths were maintained for depression detection with BDI and Fatigue for abstention rates for uncertain cases (Risk-Coverage AUCs < 0.4). Individual symptom scores were predicted with good accuracy (Correlations were all significant, with Pearson strengths between 0.31 and 0.49). Fairness analysis revealed that models were consistent for sex (average Disparity Ratio (DR) = 0.86), to a lesser extent for education level (average Disparity Ratio (DR) = 0.47) and worse for age groups (average Disparity Ratio (DR) = 0.33).

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, 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, early identification and evaluation of severity of the symptoms 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, via the use of clinically valid scales, improves the follow-up and treatment of affected individuals with mental health disorders [5]. Timely interventions lead to better outcomes in mental health. This proactive approach can ensure early access to treatment and prevent significant complications. Yet, measuring 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 assessment scales 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]. The development of objective biomarkers, easy to collect without synchronization of clinicians and patients, 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 the patient's outcomes and well-being. This is particularly significant for screening the general population across diverse mental health dimensions. Indeed, screening the first signs of mental health problems/symptoms could help to avoid escalation of symptoms, as different dimensions interact in time [7, 10].

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 [11–14]. Clinical evidence and research studies have increasingly linked specific automated extracted speech features, such as prosody, articulation, and fluency, with various mental health conditions, such as depression [11, 15], anxiety [16], suicide-risk assessment [17], fatigue [18, 19], or sleep deprivation [20]. 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 [21, 22]. 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 the aforementionned motor, cognitive or social skills, as well as mental health states, 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 complemented 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: they are non-invasive, cost-effective, and convenient tools. Recent hardware and software 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, in laboratory settings, or not evaluated for deployment into clinical practice. This gap calls for more evidence to be integrated into clinical practice [23].

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 and syndromes in mental health that influence each other [24, 25]. Also, previous speech studies were limited to specific populations such as students [26] or the elderly[27]. However, 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 [28]. Finally, speech-based systems should not be tested only for the simple classification of binary labels (ex: depressed or not depressed), but rather for the estimation of the severity of symptoms [29] and their ability to refrain from giving an output when uncertainty is too high, therefore deferring decisions to the health staff in practice[30].

In this study, the main objective was to assess the predictive potential of speech models in detecting and estimating the severity of depression, anxiety, fatigue, and insomnia within the general population, using mobile-collected speech and mental health data. Besides, to prove that these models could be effectively implemented in diverse real-world settings, they are assessed for their fairness and uncertainty capabilities.

Methods

Participants

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

All participants signed an informed 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 completed self-assessment scales for different mental health dimensions and recorded different speech tasks. In this work, we only focused on one spontaneous and semi-structured speech task, where participants had to answer: "Describe how you are feeling at the moment and how your nights' sleep have been lately." [31]. 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.). Participants followed the instructions

displayed on the Callyope application and their vocal answers were recorded with the smartphone's microphone. The audio was sampled at 44.1 kHz with a 16-bit resolution. Each participant was asked to place his phone on a flat surface (for example a table) with the microphone pointing towards the speaker i.e. himself. The session should take place in a quiet environment, whenever possible.

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 education level, were collected. We compared groups with adequate tests for their demographics and self-assessments to ensure that 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) [32] and Patient Health Questionnaire-9 (PHQ-9) [33] self-report questionnaires. In current clinical practices, it is common that different professionals interacting with a patient use different metrics to monitor depression. While depression assessment through these two measures exhibits a robust correlation at the group level [34], thus 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 [35].

The PHQ-9 is a short, self-administered questionnaire mainly used to screen and measure the severity of depression[33] and is sensitive to potential changes [34]. It includes the two cardinal signs of depression: anhedonia and depressed mood. We considered a risk of depression if the total score for the PHQ-9 was more than 10 (PHQ-9≥10).
The BDI is a self-administered questionnaire with 21 items, each centered around a core theme [32, 36]. Respondents are presented with 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 the BDI threshold to be positive 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 [36].

Measure of anxiety

The Generalized Anxiety Disorder questionnaire (GAD-7) is to measure or assess the severity of generalised anxiety disorder (GAD) [37].. This is a self-administered questionnaire that takes less than 5 minutes to complete, and it was especially

developed to be deployed efficiently in primary care. The optimal cutoff for the GAD-7 was found to be a cut-off for the total score of GAD-7≥10 [37].

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 [38]. 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 insomnia troubles, for the AIS scale, is 6 [39, 40].

Measures of fatigue

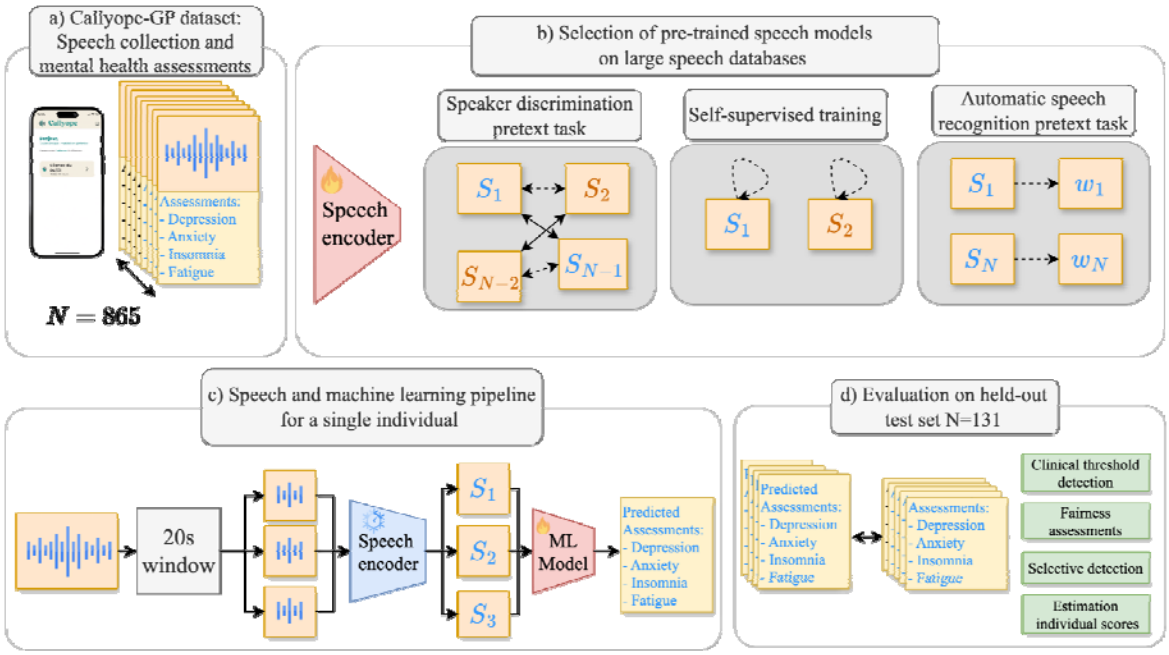
We used the Multidimensional Fatigue Inventory (MFI) to assess the different dimensions of fatigue [41–43]. It is a short self-report questionnaire (5-10 minutes) based on 20 questions to determine five dimensions of fatigue: general fatigue, physical fatigue, reduced motivation, reduced activity and mental fatigue. We also reported the total fatigue score as the sum of all sub-components.
We used the normative data from [42, 44] 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 [42], 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 [44], 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 finetuning of 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 865 included participants (b) Flowchart of the pre-training phases of our speech encoders 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=131). Symbols represent speech turn vector embeddings obtained from the speech encoder model.

Colours represent speaker identities for each speech turn and embedding, represent the words spoken in the given audio.



Model pretraining and tuning

Audio intensity is normalized per sample, and we compared the three main approaches to obtain representations for large scale speech models: (1) Speaker recognition ia 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 same architecture ResNet as in [45], (2) we also considered a transformer [46] architecture adapted for speech, Hubert[47], trained in a self-supervision fashion, i.e. to predict masked neighbor embeddings; finally (3) we evaluated a transformer architecture Whisper pretained to tackle automatic speech recognition.

Speaker recognition as a pretraining task has proven great results in mental health and neurology [12, 48]. The ThinResNetis pre-trained on the VoxCeleb2 dataset [49], 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. 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 [50].
We also compared these models to a self-supervised model, HubertXL which exhibit great generalization for paralinguistic tasks such as emotion recognition[51] .

Hubert was trained on 960h of Librispeech [52]. Whisper is a recent robust Automatic Speech Recognition system based on a tranformer architecture trained on 680k hours of transcribed speech. Whisper training data is much bigger and more diverse, with noisy labels than other speech models. We considered three version of the model small, medium, and large (v3).
In this work, we did not fine-tune any of the speech encoder models on the Callyope-GP dataset, which we represented in Figure 1c. by a frozen speech encoder model. For each speaker , we obtained a vector representation, a speech vector embedding denoted . We extracted segments of 20 seconds with 10 seconds overlap, and we compared different pooling of predictions, with mean and max pooling. Besides, as our data is imbalanced, we also compared classic sampling of examples to train models with undersampling of majority class, oversampling of minority classes. We used the default values of undersampling and oversampling: We found no differences with and without voice activity detection, so we used windowing for simplicity. Extraction was performed using Python (version 3.9) and the following packages were used to extract the acoustic features: pytorch 2.0.1, imbalanced learn, torchaudio 2.0.2, and voxceleb_trainer project [53].

For the finetuning for each task and each clinical score, we compared different machine learning algorithms on the validation set. For each task, once a model was selected, we retrained 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 [54].

For the speech collected in the Callyope-GP dataset, we applied the frozen speech encoder 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 pool 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.

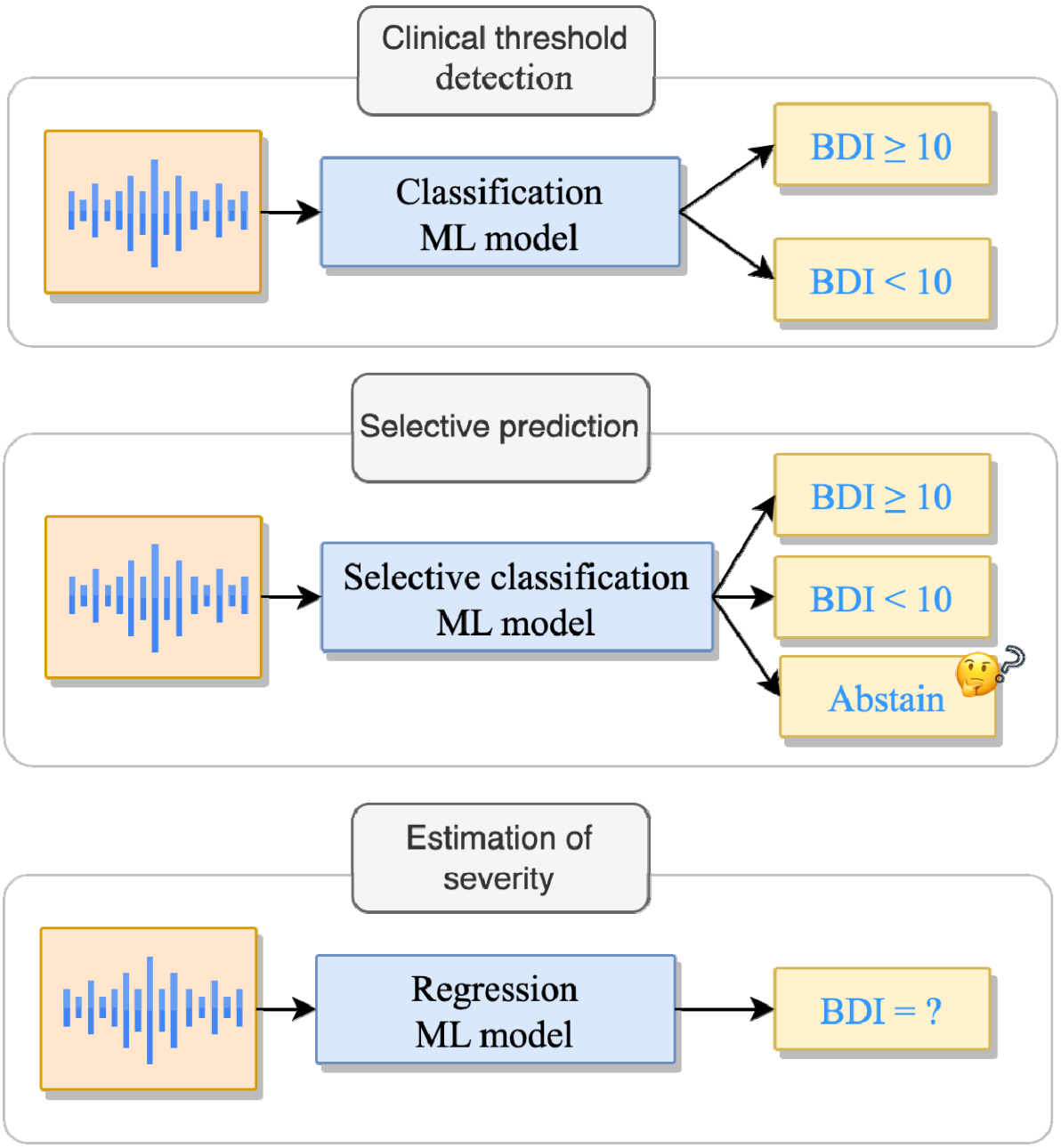


Figure 2. Schematic representation of the clinical tasks that are being assessed in this study. Each task is different in terms of the set and types of outputs. We illustrated the different tasks with the BDI clinical scale and its given threshold of 10.

Clinical threshold detection (Classification)

We first compared the predictive power of the speech encoder to discriminate between 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 macro F1 scores, along the Area under the Curve (AUC) of the Receiver Operating Characteristic on the test set,.
We compared linear-based models (for classification Logistic Regression for classification with L2 regularization, Elasticnet linear model 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.

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, and Pearson correlations (Table 4). 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.

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[28]. 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 not 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 F1 scores [57], 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 F1 score across sub-groups divided by the average F1 score on the

full test set: $= \frac{(1\ \ )}{(1)}$. We did the same for pearson correlation for regression.
The higher the DR, the better it is as it means that the model performs equally well across groups, with the perfect DR being 1, i.e. each group has the same level of performance. We used the fairlearn toolkit to perform our fairness evaluations [58].

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 appropriately the uncertainty associated with the system predictions is critical [59]. Yet, the communication of probabilities to human users is hard [60], 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 refrain from making a prediction when too uncertain (i.e., models saying "I don't know") [61, 62]. In this work, we followed the method from [63], we used the maximum output probabilities of the ML classification system as a way to measure uncertainty.
Based on a moving threshold we can obtain a specific ML system to choose to abstain when its 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.

Results

Data overview and demographics of participants

A total of 1150 participants were eligible and agreed to participate in our study. 865 completed the study, giving a recruitment yield of 75.2%. There was an equal split between Android (55%) and IOS (45%) devices used by participants.
The reasons for which participants were not included in this analysis were the following: one missing speech task, missing demographic information or one missing answer in the self-report questionnaires.

For our analyses, 605 participants were in the training set, 129 in the validation set, and 131 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>500) to evaluate error bars and predictive algorithms to avoid over-optimistic results [64]. Among the 865 participants, 275 (31.8%) were above the BDI screening threshold,

and 146 (16.9%) were above the PHQ9 threshold, 133 (15.3%) above the GAD7 threshold, 371 (42.5%) above the AIS threshold, 489 (56.5%) above the MFI general fatigue threshold, 325 (37.5%) above the MFI physical fatigue threshold, 283 (32.7%) above the MFI reduced activity threshold, 379 (43.8%) above the MFI mental fatigue threshold, 209 (24.1%) above the MFI reduced motivation threshold, and 557 (64.4%) above the threshold of the total score of the MFI. We reported co-occurrences of people at risk for each dimension in the 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	605	129	131		Statistic	P value
Demographics						
Sex (F/M/other)	386/214/5	85/44	69/63		χ^2=9.0	0.06
Education level[b]	19/107/71/19	7/21/20/81	8/24/16/83		χ^2=5.2	0.52
Age in years mean (sd) [min-max]	38.8 (18.2) [18-92]	40.2 (18.7) [18-86]	40.1 (20.0) [18-89.]		F=0.78	0.45
Clinical evaluation mean (sd) [min-max] (#negative/#positive[c])						
BDI	7.7 (7.3) [0.0 - 32.0] (413/19	7.9 (7.2) [0.0 - 29.0] (88/41)	7.9 (7.7) [0.0 - 36.0]		F=0.07	0.92

	2)		(89/42)			
PHQ-9	5.6 (4.5) [0.0 - 23.0] (499/106)	5.2 (4.7) [0.0 - 25.0] (108/21)	5.1 (4.4) [0.0 - 22.0] (112/19)		F=0.85	0.42
GAD-7	5.0 (4.5) [0.0 - 21.0] (511/94)	4.8 (4.7) [0.0 - 20.0] (109/20)	5.4 (5.0) [0.0 - 21.0] (112/19)		F=0.50	0.60
AIS	5.6 (3.9) [0.0 - 24.0] (338/267)	5.0 (3.1) [0.0 - 16.0] (75/54)	5.2 (4.1) [0.0 - 19.0] (81/50)		F=1.6	0.19
MFI general fatigue	10.4 (4.2) [4.0 - 20.0] (335/270)	11.1 (4.2) [4.0 - 20.0] (80/49)	10.4 (4.1) [4.0 - 20.0] (74/57)		F=1.8	0.17
MFI physical fatigue	8.7 (3.9) [4.0 - 20.0] (376/229)	8.8 (4.0) [4.0 - 20.0] (78/51)	8.4 (3.9) [4.0 - 20.0] (86/45)		F=0.55	0.57
MFI reduced activity	8.3 (3.8) [3.0 - 20.0] (400/205)	7.9 (3.5) [4.0 - 20.0] (92/37)	8.1 (3.6) [4.0 - 18.0] (90/41)		F=0.92	0.40
MFI mental fatigue	9.2 (4.2) [4.0 -	9.5 (4.4) [4.0 - 20.0]	9.3 (4.2)		F=0.26	0.77

	20.0] (340/265)	(69/60)	[4.0 - 20.0] (77/54)			
MFI reduced motivation	7.5 (3.0) [4.0 - 20.0] (458/147)	7.2 (3.2) [4.0 - 20.0] (102/27)	7.7 (3.2) [4.0 - 18.0] (96/35)		F=0.74	0.47
MFI total score	44.2 (15.1) [19.0 - 99.0] (338/267)	44.6 (15.0) [20.0 - 90.0] (79/50)	43.9 (15.1) [20.0 - 83.0] (77/54)		F=0.09	0.91

[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 compared models and pipelines with the results on the validation set (See Table 5 in Appendix). Overall, all the Whisper models all outperformed other approaches, (WhisperM being the best with mean F1 score=0.56), while the speaker model performed the worst, even though still performing well. We found out that the performance of the pooling with the maximum prediction always outperformed the mean pooling, and undersampling and oversampling was helping. We found out that the linear-based models was outperforming random forest and gradient boosting on the validation set. Thus, we retrained linear algorithms, with max pooling and the WhisperM frozen speech encoder 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 Table 4 and Appendix Figure 2.

Clinical threshold detection (Classification)

The clinical threshold detection performed well based on the speech data and our developed system (Figure 3 and Table 2). All systems outperformed the chance levels). Based on the AUC the classification results were the highest for the BDI

score (AUC=0.78, F1=0.65) and based on the F1, it was the Total MFI score(AUC=0.68, F1=0.69) and the lowest for both metrics the MFI reduced motivation (AUC=0.61, F1=0.43).

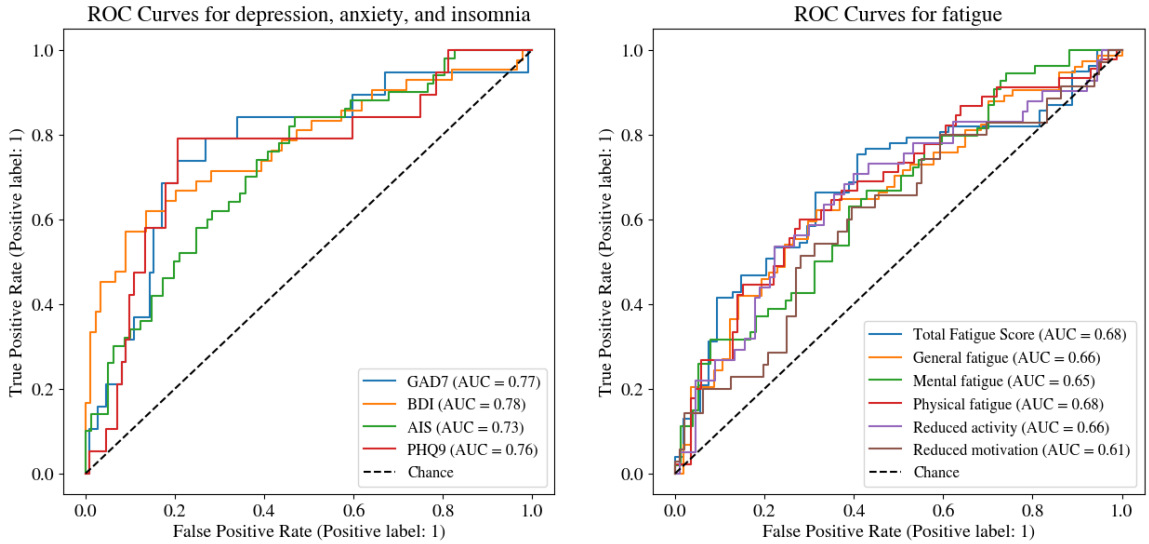


Figure 3. ROC curves for the clinical threshold detection task on the held-out test set. (Left panel) ROC curves to detect clinically relevant thresholds for depression (PHQ-9, BDI), anxiety (GAD-7), and insomnia (AIS). (Right panel): ROC curves to detect clinically relevant thresholds for fatigue components (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 2. Estimation for clinical threshold detection severity results on the held-out test set for the different considered dimensions of mental health (Classification). Higher F1 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.

	F1 score
PHQ-9	0.49
GAD-7	0.50
BDI	0.65

AIS	0.62
MFI General Fatigue	0.66
MFI Physical Fatigue	0.54
MFI Reduced Activity	0.52
MFI Mental Fatigue	0.53
MFI Reduced Motivation	0.43
MFI Total Fatigue Score	0.69

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

We computed the sensitive attributes DRs (sex, age, and education level) and assessed the differences in the quality of the c made by the speech-based system (Table 3) for classification. Overall, the classfication speech-based system had a better quality-of-service for sex (mean 0.86, SD 0.13), and the worst was for age (mean 0.33, SD 0.30). We also identified that the detection of PHQ9-7 had the worst quality-of-service disparity (Mean of DRs = 0.20, and the best quality-of-service was obtained with the score AIS (Mean of DRs = 0.71), the MFI General Fatigue (Mean of DRs = 0.75), and the MFI Physical Fatigue (Mean of DRs = 0.88). We also observed that only the MFI General Fatigue and MFI Physical Fatigue obtained a good performance for age (MFI General Fatigue Age DR=0.73; MFI Physical Fatigue Age DR=0.76), and, except for PHQ-9 and GAD-7, all DRs were satisfactory or high for sex.

Table 3. Disparity ratios (DRs) based on the F1 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	0.61	0.00	0.00		0.20
GAD-7	0.66	0.00	0.00		0.22
BDI	0.85	0.25	0.57		0.56
AIS	0.98	0.45	0.69		0.71
MFI General Fatigue	0.77	0.73	0.75		0.75
MFI Physical Fatigue	0.97	0.76	0.90		0.88
MFI Reduced Activity	0.95	0.27	0.56		0.59
MFI Mental Fatigue	0.91	0.00	0.58		0.50
MFI Reduced Motivation	0.94	0.25	0.30		0.50
MFI Total Fatigue Score	0.91	0.62	0.35		0.63
Mean (SD)	0.86 (0.13)	0.33 (0.30)	0.47 (0.30)		0.55 (0.33)

[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 observed great performances for BDI, AIS and MFI General Fatigue (Figure 4). The model that selectively predicts the risk of depression based on the BDI score achieved the best result (BDI risk-coverage AUC=0.28). The other scores could not achieve such feats of important coverage with no risk.

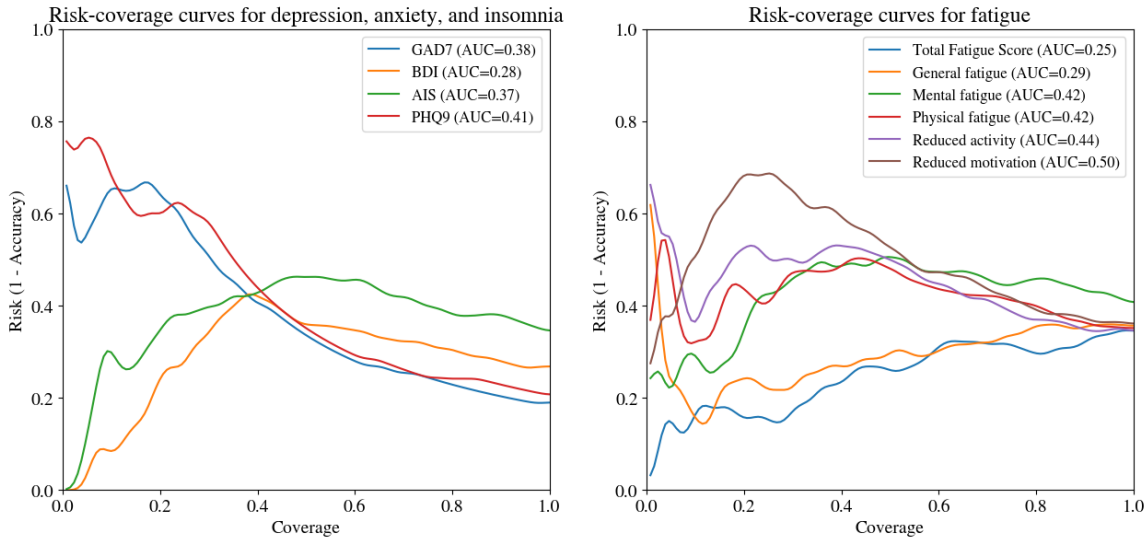


Figure 4. Risk-Coverage curves on the held-out test set for the selective clinical threshold detection task on the held-out test set illustrate the models' selective screening ability, i.e., risk detection capabilities with the ability to abstain when too uncertain. Curves are smoothed for clarity with a gaussian blur, but not used to compute AUC. 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.

Estimation of severity through predictions of individual scores (Regression)

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 $P < 1 \times 10^{-14}$). The strongest correlations

between the prediction and the actual scores on the held-out test were found for BDI score (r=0.49), GAD7(r=0.48) and the PHQ-9 (r=0.47). The lowest correlation was found for the MFI Reduced Activity (r=0.31).

Speech-based models also obtained great results in terms of absolute errors. We observed less than 3 points of MAE for AIS and reduced activity. All other scores were predicted on average with less than 5 points, except for the MFI Total Fatigue Score, since its range is [13-88].

Table 4. 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	0.47	3.1
GAD-7	0.48	3.2
BDI	0.49	4.9
AIS	0.43	2.9
MFI General Fatigue	0.38	3.3
MFI Physical Fatigue	0.32	3.0
MFI Reduced Activity	0.31	2.9
MFI Mental Fatigue	0.34	3.1
MFI Reduced Motivation	0.32	2.5
MFI Total Fatigue Score	0.44	11.3

Discussions

We aimed to explore the full capabilities and limitations of using speech data extracted from 865 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 one simple speech task performed on our smartphone application. The models were trained and calibrated on training and validation sets of participants, and we demonstrated the system's generalization on the held-out test set of 131 participants.
The results indicated that ML-based systems using speech only as input could identify participants above clinical thresholds for depression, insomnia, total fatigue components, but to a lesser extent anxiety and fatigue sub components. 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 education levels. Depression, insomnia and different dimensions of fatigue clinical threshold detection results were particularly consistent for sex, slightly less for age and to 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. Risk-coverage areas under the curves remained low, for insomnia, total fatigue, and depression detection through BDI. Finally, we showed that speech-based models could also predict the severity, with prediction of 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 0.31 to 0.49.

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 [26, 27, 65, 66]. Our recruitment and involvement of participants was in person. For medium-sized datasets with face-to-face recruitment (below 1000 participants, such as ours), it can be observed that crowd-sourced recruitment of participants via online platforms needs more training data to yield the same level of model accuracy [65, 67]. 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 [66]. This discrepancy could be explained by the fact that data quality is variable on web crowdsourcing platforming, especially for participants' psychiatric evaluations [68], 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 [35]. 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, which could be detected simultaneously through speech analysis. While prior research, like the work by [20], 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 [69]. The prevention of recurrent sleep problems can prevent other mental health troubles or relapses. The observed co-occurrence of symptoms, particularly insomnia with other clinical dimensions, highlights the interconnected nature of these symptoms and syndromes.

Finally, , 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 [70]. Age factor was the least preserved in our grouped performances, this can be attributed to change in voice changes due to hormones [71], and normal aging affects the different parts of the vocal production system: larynx, respiratory system, resonators, saliva system and the individual's global emotional status [72]. 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-sized dataset (300<N<1000) needs more diverse data to achieve better generalizability, and the non-longitudinal design misses insights on symptom evolution. Besides, the use of self-assessment scales introduces a potential bias because they rely heavily on the insight of participants. A limitation of our study is the use of a fixed train-dev-test split. This split approach, while convenient for comparing results across different tasks, can introduce bias and limit the

generalizability of our findings. Future research with larger samples, longitudinal designs, and the inclusion of pathological data is crucial for exploiting the full potential of voice biomarkers in mental health. Studying various and diverse speech tasks and prompts also holds the potential for multiple benefits related to user adherence. Indeed, tailored tasks can address specific mental health needs, cater to individual preferences and boost engagement. Varying prompts can reduce user fatigue and sustain 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 applications and ensuring fairness across the population demographics, our findings pave the way for further development and responsible integration of such tools into clinical settings, advancing personalized mental health assessment and intervention. In future work, we will extend this study by including longitudinal data, adding more diverse linguistic and geographic data, and including more severely affected patients who are 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 a former 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
AUC: Area Under the Curve

Multimedia Appendix 1

Figure Appendix 1. Co-occurrences percentage of people at risk for each dimension, normalization is done per row: first row can be read as 57% of individuals who are at risk for depression based on the PHQ-9 are at risk of GAD-7.

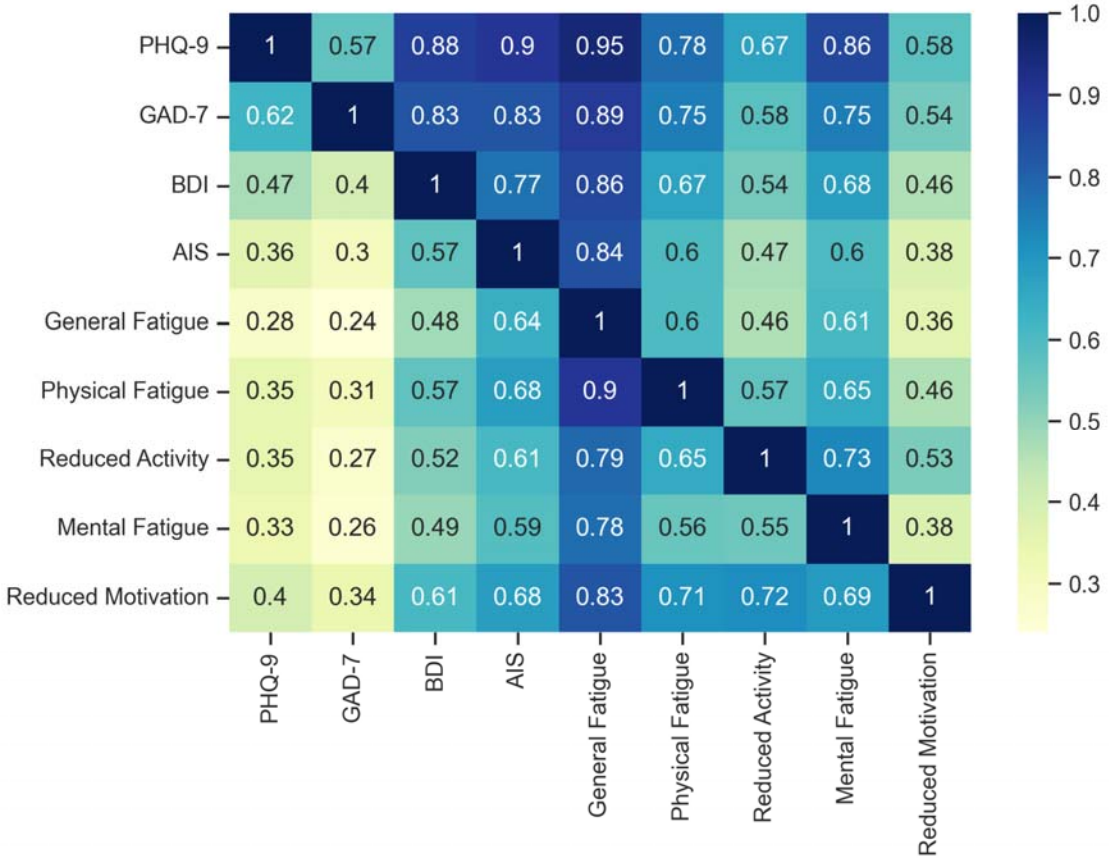


Figure Appendix 2. ROC curves for the clinical threshold detection task on the held-out test set.
(Left panel) PR curves to detect clinically relevant thresholds for depression (PHQ-9, BDI), anxiety (GAD-7), and insomnia (AIS). (Right panel): ROC curves to detect clinically relevant thresholds for fatigue 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 5. Results on the validation set for the clinical threshold detection task on the validation set for the F1 score and AUC score, averaged across clinical scales. We displayed the top 3 approach for each model. We underlined the best approach per frozen speech model, and bolded the best overall.

Frozen speech model	Pooling	Sampling	Classifier	F1 score
Random classifier (Dummy)	n/a	n/a	n/a	0.15
Speaker Recognition ResNetSE34	max	undersample	Logistic Regression	0.518
	max	undersample	Random Forest	0.514
	max	oversample	Logistic Regression	0.510
Self-supervised HubertXL	max	oversample	Logistic Regression	0.532
	max	undersample	Logistic Regression	0.528
	max	undersample	Gradient boosting	0.522
WhisperS	max	oversample	Logistic Regression	0.535

	max	undersample	Logistic Regression	0.522
	max	None	Logistic Regression	0.513
WhisperM	max	oversample	Logistic Regression	0.560
	max	oversample	Gradient boosting	0.547
	max	undersample	Logistic Regression	0.542
WhisperL	max	oversample	Logistic Regression	0.548
	max	undersample	Logistic Regression	0.545
	max	undersample	Gradient boosting	0.527

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. Bard HA, O'Driscoll C, Miller CB, et al (2023) Insomnia, depression, and anxiety symptoms interact and individually impact functioning: A network and relative importance analysis in the context of insomnia. Sleep Med 101:505–514. https://doi.org/10.1016/j.sleep.2022.12.005
11. 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
12. 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
13. 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
14. 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
15. 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
16. 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
17. 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
18. 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
19. 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
20. Thoret E (2024) Sleep deprivation detected by voice analysis. PLoS Comput Biol 20:e1011849. https://doi.org/10.1371/journal.pcbi.1011849
21. Levelt WJM (1989) Speaking: from intention to articulation. MIT Press, Cambridge, Mass.
22. Guenther FH (2016) Neural Control of Speech. MIT Press
23. 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
24. 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
25. 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
26. 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
27. 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
28. Barocas S, Hardt M, Narayanan A (2023) Fairness and machine learning: Limitations and opportunities. MIT Press
29. 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
30. 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
31. 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
32. 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
33. 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
34. 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
35. 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

36. 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
37. 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
38. 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
39. 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
40. 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
41. 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
42. Schwarz R, Krauss O, Hinz A (2003) Fatigue in the General Population. Onkologie 26:140–144. https://doi.org/10.1159/000069834
43. 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
44. 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
45. 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
46. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is All you Need. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.
47. Hsu W-N, Bolte B, Tsai Y-HH, et al (2021) HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEEACM Trans Audio Speech Lang Process 29:3451–3460. https://doi.org/10.1109/TASLP.2021.3122291
48. 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
49. Chung JS, Nagrani A, Zisserman A (2018) VoxCeleb2: Deep Speaker Recognition. In: Interspeech 2018. pp 1086–1090
50. 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

51. Morais E, Hoory R, Zhu W, et al (2022) Speech Emotion Recognition Using Self-Supervised Features. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Singapore, Singapore, pp 6922–6926
52. Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, South Brisbane, Queensland, Australia, pp 5206–5210
53. (2024) clovaai/voxceleb_trainer
54. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
55. 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
56. 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
57. 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
58. Weerts H, Dudík M, Edgar R, et al (2023) Fairlearn: Assessing and Improving Fairness of AI Systems. J Mach Learn Res 24:1–8
59. 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
60. 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
61. 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
62. El-Yaniv R, Wiener Y On the Foundations of Noise-free Selective Classification
63. Hendrycks D, Gimpel K (2016) A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks
64. Varoquaux G (2018) Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 180:68–77
65. 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
66. 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
67. 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
68. 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
69. 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
70. 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
71. 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
72. 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