Why Voice Biomarkers of Psychiatric Disorders are not used in Clinical Practice? Deconstructing the Myth of the Need for Objective Diagnoses

Vincent P. Martin[1,2,3], **Jean-Luc Rouas**[2]
[1] DDP Research Unit, Department of Precision Health, LIH, Strassen, Luxembourg
[2] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
[3] Univ. Bordeaux, CNRS, SANPSY, UMR 6033, CHU Pellegrin, F-33076 Bordeaux, France
vincentp.martin@lih.lu, jean-luc.rouas@labri.fr

Abstract

Voice biomarkers hold the promise of improving access to care and therapeutic follow-up for people with psychiatric disorders, tackling the issues raised by their high prevalence and the significant diagnostic delays and difficulties in patients follow-up. Yet, despite many years of successful research in the field, none of these voice biomarkers are implemented in clinical practice. Beyond the reductive explanation of the lack of explainability of the involved machine learning systems, we look for arguments in the epistemology and sociology of psychiatry. We show that the estimation of diagnoses, the major task in the literature, is of little interest to both clinicians and patients. After tackling the common misbeliefs about diagnosis in psychiatry in a didactic way, we propose a paradigm shift towards the estimation of clinical symptoms and signs, which not only address the limitations raised against diagnosis estimation but also enable the formulation of new machine learning tasks. We hope that this paradigm shift will empower the use of vocal biomarkers in clinical practice. It is however conditional on a change in database labeling practices, but also on a profound change in the speech processing community's practices towards psychiatry.

Keywords: Voice biomarkers, Mental health, Corpus labeling

Introduction

Context

One out of every eight individuals experiences mental health issues. Anxiety and depression are the most prevalent, affecting, respectively, 3.9% and 3.6% of the world's population[1]. These disorders exert a substantial influence on personal well-being, notably in terms of suicide risk and diminished quality of life. Additionally, they have significant repercussions on public health, resulting in considerable economic implications (The Lancet Global Health, 2020).

Given the scale of this epidemic of mental disorders in the general population, affordable and accessible tools are needed for the early diagnosis of these disorders. These tools promise improvments in therapeutic results and minimizing missed opportunities associated with the disease (Kurachi et al., 2018; Kobeissy et al., 2013; Tafalla et al., 2009).

Speech analysis has emerged as a promising avenue for detecting mental disorders on a global scale. It is cost-effective and widely accessible, as it is integrated into all smartphones, enabling passive voice recording within natural living environments of patients. Furthermore, due to the intricate interplay of numerous neuromotor (Denes and Pinson, 1963) and neurolinguistic (Kröger et al., 2020) processes involved in speech production, it is sensitive to a wide spectrum of pathologies (Fagherazzi et al., 2021).

These advantages have garnered significant attention from the voice and speech processing community, especially in the context of using voice recordings to detect psychiatric disorders. Notably, the Interspeech 2023 conference featured a dozen papers dedicated to the topic, surpassing previous records (seven in 2022, six in 2021, seven in 2020).

However, despite these promising developments and a consistent track record of high classification performance for more than a decade (Scherer et al., 2013), vocal biomarkers are not yet implemented in clinical practice. Why?

Just a problem of explainability?

We dismiss from the outset of this article a common belief prevailing in computer science laboratories that clinicians do not use AI — particularly vocal biomarkers — due to a lack of understanding of the proposed models: the opacity of these models would hinder them from comprehending the internal workings of the decision, thereby preventing their use of these tools.

However, model explainability is not the central factor in clinicians' use of a tool; they do not need

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders, consulted 4th October, 2023

LREC-COLING 2024, pages 17603–17613
20-25 May, 2024.

to understand the inner workings of their tools, but rather need to trust them (Kastner et al., 2021). How many clinicians can explain the internal functioning of an electronic thermometer? Nevertheless, through the correlations between the displayed value and the patient's condition as influenced by decisions made based on this information, clinicians establish trust in the value displayed by this electronic device. Explainability can contribute to building trust (Ferrario and Loi, 2022), but it is neither a necessary nor a sufficient condition.

In this scenario, how do we build such trust with clinicians? Our strategy has involved going into the field and observing their working methods, with particular focus on the role of diagnosis in their professional activities.

Objectives and outline

In this article, we depart from conventional arguments about database size and model performances to center our focus on the alignment between the tasks explored within the realm of vocal biomarkers, on one hand, and the clinical practice of psychiatry, on the other. To do so, we rely on insights from epistemology and sociology of psychiatry, clinical practice, and an ethnographic observation internship conducted at Saint Antoine Hospital in Paris in 2022. We develop the idea that the research community working on vocal biomarkers, by primarily emphasizing diagnostic estimation, perpetuates the fallacy of clinicians' demand for automated, objective tools for diagnosis. We explicit this idea through a literature review in Section 1.

Since the concept learnt by machine learning systems depends on the labels of the database on which they are trained, we discuss the limitations of current practices in terms of annotating databases with self-questionnaires (in Section 2) or diagnosis made by a psychiatrists (in Sections 3 and 4).

Based on these observations, we propose in Section 5 a new research paradigm for designing vocal biomarkers in psychiatry, centered around the automatic estimation of symptoms. Such an approach would not only be valuable to clinicians but would also seamlessly integrate with their professional practice, facilitating the incorporation of this tool into clinical settings.

We finally provide recommendations on how to label corpora for the estimation of symptoms in Section 6 and we conclude in Section 7.

1. Voice biomarkers as objective diagnostic tools in psychiatry

1.1. The need for objective tools

It is a recurring theme in the introductions of research papers on the use of vocal biomarkers for the assessment of psychiatric disorders that '[...] the diagnosis of depression is still largely based on the subjective judgment of a psychiatrist. A more objective measure for diagnosing depression is desirable.' (Mijnders et al., 2023). Similarly, it is common knowledge that 'Gold-standard diagnostic and assessment tools for depression and suicidality remain rooted, almost exclusively, on the opinion of individual clinicians risking a range of subjective biases.' (Cummins et al., 2015) or that 'There is an urgency to objectively diagnose, monitor over time, and provide evidence-based interventions for individuals with mental illnesses' (Low et al., 2020). However, this contrast between the subjectivity of clinicians and the objectivity of algorithms – presented as more reliable – is not new. In fact, as far back as 1967, Paul Meehl was already raising questions about the role of clinicians in light of emerging statistical techniques (Meehl, 1967). This was followed in the 1980s by Spitzer's seminal article, which not only challenged but provocatively questioned the very need for clinicians in the diagnostic process (Spitzer, 1983).

This concern seems to be well-founded due to the limited reliability of diagnoses, a matter that clinicians themselves have acknowledged. According to a 2007 survey, a significant majority of participating clinicians (87%) viewed their own diagnoses as lacking reliability (Aboraya, 2007). The predominant explanation for this lack of reliability was attributed to factors associated with the clinicians themselves, such as their education, biases, and interview techniques (63.5%). This explanation far outweighed patient characteristics (21.6%) or issues related to the definition of pathologies (14.9%). In essence, what primarily guides the diagnostic process is not just the patient and their symptoms, but also the clinician responsible for making the diagnosis.

The lack of reliability in diagnoses made by psychiatrists reinforces the perception among vocal biomarker developers of a need for objective diagnostic tools in psychiatry.

1.2. State of the art

The growing need for such tools has stimulated an entire research community investigating which vocal biomarkers can be associated with these mental conditions. In a systematic review carried out in 2020, Low et al. (2020) identified 127 studies published between 2009 and 2019 identifying psychiatric disorders through speech analysis.

Eight different pathologies have thus been identified, and among the 127 included studies, 63 (49.8%) were focused on the estimation of Major Depressive Disorder (MDD), 23 (18.1%) were about schizophrenia and 21 (16.5%) about Bipolar Disorder (BD), which are among the most prevalent and harmful psychiatric disorders (Newson et al., 2021). In the same vein, recent works have also been focusing on depression (Fara et al., 2023; Campbell et al., 2023), schizophrenia (Tang et al., 2021), bipolar disorders (Farrús et al., 2021) but also borderline trouble (Gosztolya et al., 2020), anxiety (Baird et al., 2020), or ADHD (Etter et al., 2021).

Diagnosis estimation is the most widely studied task in the literature: on 85 studies identified by Low et al. that used *machine-learning* approaches (in contrast to studies that only perform statistical tests), 61 (71.8%) of them were focused on pathology detection (binary classification). The other studies focused on the estimation of disorders severity (regression with clinical questionnaires or scores, n=19, 15.0%) and intra-speaker correlation (based on longitudinal data, n=5, 3.9%). Among them, only 5 (3.9%) proposed both classification and disorder severity estimation.

From a corpus point of view, these databases are annotated in two different ways. On the one hand, the data of 45 of the 127 studies (35.4%) rely on the diagnosis made during clinical interviews, established on the basis of international classification criteria such as those of the Diagnostic and Statistical Manual of mental disorders (DSM); or heteroquestionnaires (i.e. filled by the clinician, such as the MADRS (Montgomery and Åsberg, 1979) for depression). On the other hand, 78 (61.4%) of the studies[2] are annotated by the score to validated self-reported questionnaires (e.g. the Patient Health Questionnaire – PHQ (Kroenke et al., 2001) – for depression).

These two ways of annotating corpora have however significant epistemological limitations, which we propose to investigate in the following sections.

2. Self-reported questionnaires

Self-reported questionnaires cost nothing, are non-invasive, measurable anywhere and anytime, which makes them very suitable for data collection (Qian et al., 2020). In addition, they do not require clinical supervision during filling, so they are used to collect large databases under ecological conditions (Rutowski et al., 2022). The most studied psychiatrirc disorder studied through self-rated is depression (Low et al., 2020), usually measured using the Patient Health Questionnaire (PHQ8 or PHQ9, see Table 1). Such recent corpora encompass the corpus collected by Tymia and analysed by Fara et al. (2023), or the one collected through the RADAR-MDD study, for which voice has been analyzed by Campbell et al. (2023). The most studied corpus to date is still the DAIC-WOZ, a subset of the Distress Analysis Interview Corpus (DAIC) introduced in 2014 (Gratch et al., 2014) and labelled using the PHQ-9.

1. Little interest or pleasure in doing things
2. Feeling down, depressed, or hopeless
3. Trouble falling or staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down
7. Trouble concentrating on things, such as reading the newspaper or watching television
8. Moving or speaking so slowly that other people could have noticed? Or so fidgety or restless that you have been moving a lot more than usual?
9. Thoughts that you would be better off dead or of hurting

Table 1: Items of the PHQ-9 (Kroenke et al., 2001) corresponding to the nine symptoms of the Major Depressive Disorder in the DSM. Each item is rated from 0 'Not at all' to 4 'Nearly everyday'.

However, we identified three major limitations to annotating corpora using autoquestionnaire scores.

2.1. Sensibility to the disorder

Although they may be sufficiently accurate for integration into clinical settings, the combination of errors originating from autoquestionnaire measurements and classification models could potentially lead to disease detection performance levels that fall short of expectations. For instance, setting a threshold of 10 on the PHQ-9 (Kroenke et al., 2001) to identify depression yields a label with only 77% sensitivity compared to the DSM-IV definition of Major Depressive Disorder (Gilbody et al., 2007). While this level of accuracy might suffice for clinical use, the inaccuracy of this approach combined with the performance limitations of models (which typically hover around 80% at best for binary classification), place an inherent ceiling on the ability to detect depression solely through these autoquestionnaires' score. In other words, even a perfect machine learning system achieving 100% accuracy in replicating the PHQ-9 score would still exhibit a sensitivity of no more than 77% in depression detection.

[2]The annotation is unknown for 4 studies (3.1%).

2.2. Use of questionnaires by clinicians

A second limitation of annotating corpora using questionnaire scores is their absence in the standard clinical practice: clinicians seldom use questionnaires, because of lack of time but also because of a lack of training in the use of these tools (Zimmerman and McGlinchey, 2008). Indeed, information required to establish the diagnosis is usually collected during the clinical interview. Moreover, when they eventually use questionnaires, they do not apply a threshold in the same strict manner as in binary classification (e.g. PHQ > 10). On the contrary they analyze qualitatively the distribution of the responses of the patients to the different items to assess the different dimensions of the patient's complaint (Busner et al., 2011). Therefore, the usual automatic classification approaches do not suit the clinician's needs.

2.3. Validation on disorder criteria

Finally, these questionnaires are medically validated on the basis of their ability to discriminate patients belonging to different diagnostic criteria, and thus have the same limitations, which are detailed on the next section.

3. Tackling common misbeliefs about diagnosis in psychiatry

At the Interspeech 2023 conference, a new openly available dataset focusing on depression has been unveiled (Tao et al., 2023). In this context, the authors emphasize the label's trustworthiness: 64 out of 118 speakers were diagnosed with depression by professional psychiatrists, implying that 'the dataset facilitates the exploration of depression detection rather than the prediction of scores derived from questionnaires'. While this return to diagnostic criteria seems relevant for depression, which is mainly annotated with self-questionnaires, the majority of vocal biomarkers for other psychiatric disorders are identified in corpora annotated either with the diagnosis made by a psychiatrist according to a classification (DSM or ICD), or by the score of hetero-questionnaires, i.e. diagnostic interview grids, completed by psychiatrists.

For instance, previous works on estimating schizophrenia (Tang et al., 2021), Attention Deficit Disorder (Etter et al., 2021) or Autism spectrum (Briend et al., 2023) have relied on international classifications to label their data. Other works used diagnostic interview grids,such as the Hamilton Depression Rating Scale (HDRS) and the Young Mania Rating Scale (YMRS) for the estimation of bipolar disorder (Farrús et al., 2021), or the use of the PTSD checklist for DSM IV by Schultebraucks et al. (2020), among others.

Nonetheless, the annotation of corpora with diagnoses made by clinicians also suffers from significant limitations. In this section, we deconstruct widely held 'common knowledge' concerning psychiatric diagnoses.

3.1. Underlying mechanisms

The bedrock of these commonly held beliefs is the conviction regarding the existence (or in our case, the non-existence) of underlying pathophysiological mechanisms driving the pathology (e.g., psychiatric disorders being diseases of the brain). In essence, it's the faith in the existence of a 'schizococcus,'[3] a natural entity held responsible for the disorders. Overwhelming efforts have been dedicated in recent years to track down this 'schizococcus' within the realms of neuroscience and genetics, yielding ultimately mixed outcomes. As Thomas Insel, former director of the American National Institute for Mental Health (NIMH), puts it,*"I spent 13 years at NIMH really pushing on the neuroscience and genetics of mental disorders, and when I look back on that I realize that while I think I succeeded at getting lots of really cool papers published by cool scientists at fairly large costs – I think $20 billion – I don't think we moved the needle in reducing suicide, reducing hospitalizations, improving recovery for the tens of millions of people who have mental illness."* (Troisi, 2022). Whether they exist or not, these underlying pathophysiological mechanisms of mental disorders have still not been identified (Troisi, 2022; Fagerberg, 2022).

3.2. Centrality in clinical practice

Treatment plan. Contrary to other medical fields, in which the clinical outcomes (prognosis, survival rate, treatment plan) rely on the diagnosis and these mecanisms, these factors fewly rely on diagnosis in psychiatry. Particularly, the treatment plan is developed in a transdiagnostic manner based on symptoms and clinical signs (Waszczuk et al., 2017).

Attributed diagnosis. The lack of identified mechanisms underlying psychopathology has implications on classifications of mental disorders. Indeed, the assigned diagnosis is the one that most effectively accounts for the symptoms, rather than pinpointing a particular mechanism. In fact, most diagnostic criteria in the DSM contain stipulate that 'the symptoms are not better accounted for by disorder A or disorder B. For instance, in the case of MDD, criterion D specifies that 'The occurrence of the major depressive episode is not better explained by schizoaffective disorder, schizophre-

[3] I borrow this term from Dr. Philip Nuss, a psychiatrist at Saint Antoine Hospital in Paris.

nia, schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other'. The diagnosis is thus assigned as a default, lacking a more precise explanation. We are quite distant from diagnoses that reflect precise pathophysiological mechanisms!

Example: the DAIC-WOZ. This ambiguity is exemplified in the case of vocal biomarkers in the DAIC-WOZ dataset, the most commonly used corpus for depression detection from voice recordings. An information seldom mentioned in studies using the corpus, yet crucial for interpreting the task performed by machine learning algorithms, is the fact that the recorded population is a mix between general population and war vetarans (Gratch et al., 2014). Two dimensions – PTSD and depression – were originally measured using questionnaires. The PTSD score has then been truncated, and the PHQ-9 was set as binary using a threshold of 10. And the DAIC-WOZ became a reference corpus for depression. However, contrary to what is claimed by the numerous studies using DAIC-WOZ, their machine learning models do not generalize 'depression' but rather the presence of a depressive syndrome in war veterans, eventually suffering from PTSD. This could explain the difficulties in subsequently using these models in real-life scenarios.

"Pure" patients. Furthermore, studies rarely include patients who simultaneously have multiple disorders (psychiatric or somatic), which does not reflect the reality of mental disorders (Manuel et al., 2013). However, these comorbidities can also influence the voice: a patient can indeed be elderly, have Parkinson's disease, depression, and diabetes, all of which manifest in the voice.

3.3. The diagnosis is not homogeneous.

Intra-diagnosis heterogeneity. In a study published in 2021, Newson et al. have measured the symptomatic profiles of 107,349 adults with the ten most common psychiatric disorders diagnosed using DSM-5 criteria (Newson et al., 2021). They concluded that 'DSM-5 disorder criteria do not separate individuals from random when the complete mental health symptom profile of an individual is considered.' Otherwise formulated, when considering the 47 symptoms collected in this study, diagnostic criteria are as useful as a random attribution to a diagnostic group. This heterogeneity of symptomatic profiles in diagnostic criteria reaches its apogee in depression: on 3703 patients diagnosed with MDD (DSM-IV criteria), Fried and Nesse (2015) have found no less than 1030 different profiles, sometime having very few in common. This heterogeneity questions the concept generalized by models trained on small datasets (e.g. $<$2000 patients), in which inter-speaker differences could not be differentiated from intergroup variations.

Culture and time dependency. Additionally, diagnostic criteria are both unstable through cultures (e.g. the *hikikomori* diagnosis that is specific to the Japanese culture (Teo, 2010)) and unstable through time: with the advancement of scientific knowledge, updated versions of classification reference manuals are regularly published (e.g. DSM-IV in 1994, DSM-IV-TR in 2000, DSM-5 in 2013, DSM-5-TR in 2022). Since collecting this type of data requires both human and financial important investments, annotating databases with diagnostic criteria that are dedicated to some specific populations and/or that can evolve in the following years does not seem to be sustainable.

3.4. From disease to health

In the current framework of binary classification, a patient having four of the five symptoms required to be given a diagnosis of MDD, even at high intensity, would be labeled as "healthy control" in the data because he/she does not fulfill the criteria to be in the "depressive" group. This does not prevent him/her to suffer from his/her symptoms, and to have very little in common with people not having any depressive symptom.

While diseases are efficient to guide the clinical investigation (Carlat, 2017), they only allow to investigate psychiatric disorders, as opposed to mental health, which derives directly from the concept of health: "Health is a positive concept emphasizing social and personal resources, as well as physical capacities."[4] What is the meaning of a binary classification system returning a "0" for a given recording. Does this mean that the recorded person is in good health?

In regard to the promises given by the use of voice biomarkers to enhance the follow-up of mental heath disorders, we can – and we must – do better than binary classification. On the contrary, we could focus on estimating *health* dimensions, independently from any disorder, to supplement the data collected by clinicians during clinical interviews with ecological data.

4. Replacing clinicians?

"You are depressive", "You have a 75% of chance to be schizophrenic" or "The patient may be bipolar" are the majority output of the systems based on voice biomarkers. Nevertheless, this approach proves unproductive for both patients and clinicians. From the clinician point of view, the only information of the diagnosis is not sufficient to take

[4] https://www.who.int/teams/health-promotion/enhanced-wellbeing/first-global-conference, consulted 4th October 2023

any medical decision (McGorry and Nelson, 2016). Regarding patients, the delivering of the diagnosis is a delicate step and a key moment in the care of a patient suffering from a mental disorder (Cleary et al., 2009; Grajales, 2022), that should not be delivered by a smartphone application but by a trained professional.

4.1. Role of diagnosis in psychiatry

From an epistemological point of view, "the main aim of the psychiatric science is not classification as an end in itself but rather identification of causes and interventions" (Kendler, 2012). Hence, regarding diagnosis, "one of its most important goal is to facilitate communication among clinicians, researchers, administrators and patients [...] by establishing a common language."(Bolton, 2012) The diagnosis is thus an object of communication between the different parties involved in the pathology (clinicians, patient, patient's familiy, ...) but also an important element of the recognition of the patient's complaints by a health professional and by society. It is for example necessary for health insurance procedures. Playing such a critical role in the patients' life, these criteria have somehow been widen to the point that almost half of the global population is estimated to fulfill at least one criterion during their lifetime (Frances, 2013).

4.2. An objective AI model?

Even if one wanted to estimate diagnoses using "objective" tools to escape the subjectivity of psychiatrists, it would be impossible to avoid the biases of these tools. On one hand, biases coming from databases has been largely documented in the literature (e.g. (Gianfrancesco et al., 2018; Feehan et al., 2021)). On the other hand, an understudied bias comes from coders, which face the same constraints as psychiatrists (time, mood, training, interdisciplinarity) (Martin et al., 2022). Hence, the way an IA is coded influences its behavior in the same way as the training or professional culture of a psychiatrist will produce a different output than another colleague.

This so-called objectivity that would enable AI to treat all patients in a similar manner ultimately finds itself in the same position as the psychiatrist, who exhibits high internal consistency (i.e., each AI or psychiatrist consistently diagnoses in the same way) but low external consistency (similar to how two psychiatrists may provide different diagnoses for the same patient, two AIs may offer two different estimations for the same patient) (Martin et al., 2022).

4.3. Human relationship in healthcare

Even if AI systems were completely unbiased, it is inappropriate to reduce the diagnostic clinical interview to the psychiatrist's sole formulation of the diagnosis. Psychiatrists perform functions that are deemed 'irreplaceable' (Meehl, 1967; Spitzer, 1983), and during a medical interview, a genuine human relationship is established. This relationship forms the foundation for a therapeutic alliance, which plays a crucial role in the course and outcome of the patient's condition (Martin et al., 2000). Moreover, isolation and loneliness are significant factors in the development and persistence of mental health issues (Lim and Gleeson, 2014; Russell, 2014). The idea of breaking this isolation through smartphone applications is risky and somewhat implausible. It may even pose a risk of further isolation, thereby maintaining or exacerbating mental health problems in already isolated individuals (Nowland et al., 2018).
Thus, beyond the previously mentioned limitations of diagnosis, the continued presence of psychiatrists to initiate a break in the isolation of these patients is necessary and irreplaceable.

5. Estimating symptoms: the key to clinical utility of voice biomarkers

In light of this criticism, one might question whether vocal biomarkers will ever find utility in the field of psychiatry. To do so, we need to implement a paradigm that retains clinicians at the center of patient care, upholds their professional practices, and simultaneously fulfills their requirement for additional patient insights between appointments. Rather than emphasizing diagnostic estimation through vocal biomarkers, we propose a paradigm shift focused on estimating clinical symptoms and signs (abbreviated as symptoms in the remainder of this article). This shift in perspective not only addresses the epistemological limitations associated with diagnosis but also demonstrates a more respectful and practical approach to clinical practice.

5.1. Response to the epistemological limits of the diagnosis

Heterogeneity. Rather than assessing a combination of symptoms, our proposal entails estimating each symptom independently, thus mitigating issues associated with diagnostic heterogeneity. Moreover, it remains possible, if one so desires, to recombine the symptoms after the fact to arrive at a diagnostic estimate (e.g., (Fara et al., 2023)).

Cultural and time stability. While some symptoms or signs disappear or are still not diagnosed (e.g. hysteria (Stone et al., 2008)) and some appear with the evolution of society (e.g. misuse of

smartphones (Rho et al., 2019)), their fundamental nature makes them constant units across cultures and time (e.g. headaches during early antiquity (Magiorkinis et al., 2009) or mood disorder through history (Zarate and Manji, 2009)). As a consequence, annotating a database with symptoms is a more sustainable approach than annotating them with variable diagnostic criteria.

Transdiagnostic estimation of health. Additionally, focusing on symptoms allows for a transdiagnostic evaluation of the speaker's well-being. This means that even if an individual meets only 4 out of the 5 criteria for MDD according to the DSM, their symptoms can still be assessed, and appropriate assistance can be offered.

5.2. Respect of therapeutic relationship

A survey involving 515 psychiatrists conducted by Bourla et al. (Bourla et al., 2018) revealed a significant apprehension among psychiatrists regarding the impact of machine learning on their therapeutic relationships. The fear of being replaced by machines is heightened when focusing solely on diagnosing, whereas estimating symptoms as a complementary approach helps maintain the central role of psychiatrists in patient care.

5.3. New tasks

Shifting the focus to symptoms would redirect efforts from diagnosis toward understudied yet highly valuable tasks for clinicians.

5.3.1. Symptom severity.

While previous work has looked at the severity of the disorder (e.g. estimating the PHQ score for MDD (Rejaibi et al., 2022; Cummins et al., 2020)), no work has looked at estimating the severity of individual symptoms: identifying an increase in the severity of the disorder is of no use to clinicians without identifying the symptom(s) that cause it. Moreover, some symtpoms such as suicide ideation are of interest by themselves, even outside any disease classification.

Differential diagnosis. When two disorders are similar, clinicians sometimes have difficulty estimating the patient's diagnosis. While almost all the studies focus only on one disorder (e.g. MDD vs. Healthy Control), more recent work tends to estimate multiple disorders at the same time, e.g. (Gosztolya et al., 2020; Faurholt Jepsen et al., 2022; Pan et al., 2021; Wang et al., 2021). However, this problem has never been formulated in terms of symptoms.

Prognosis. While this task has already been proposed in an Interspeech challenge in 2021 for Alzheimer's disease (Luz et al., 2021), no article to our knowledge has addressed the estimation of the prognosis of mental disorders thanks to voice or speech descriptors (i.e. the estimation of the progression of the disease).

5.4. Specificity

Lastly, we propose that investigating the impact of symptoms on voice may offer a solution to the specificity issues identified previously in the literature (Low et al., 2020). When directly diagnosing, we are essentially assessing a combination of symptoms. However, these symptoms are not exclusive to the particular disorder under study; they are also included in the diagnostic criteria for other illnesses. For example, anhedonia or cognitive alterations manifests in both MDD and Schizophrenia. Could it be that the acoustic markers associated with these two conditions have actually captured the influence of these shared symptoms on voice, rather than the illnesses themselves? In essence, we suggest that what is expressed through voice are symptoms and that the low specificity of actual acoustic voice biomarkers regarding disorders may come from the shared symptoms between these disorders.

6. How to label corpora?

Reanalyzing existing dataset. While most of the corpora presented in the literature are concerned with the detection of mental disorders, some of them contain scores on the various items of the questionnaires used to annotate these disorders. This is the case, for example, with the DAIC-WOZ, used for the binary diagnosis of depression, but containing scores for each of the PHQ9 items used to annotate depression. Such a task was even proposed as a subchallenge during the AVEC 2017 challenge (Ringeval et al., 2017). However to our knowledge, no challenge participant has proposed a system for this subchallenge.

In the same vein, almost any corpus having been annotated with the individual items of the questionnaires can be reused to estimate symptoms from voice. However, since they are never presented in this light, we are able to give more examples here. Furhtermore, corpora that have been annotated solely with diagnostic status or overall questionnaire score cannot, unfortunately, be re-annotated with symptoms a posteriori. Since annotating a medical corpus is a costly task in human, financial and temporal terms, this article is a plea for annotating corpora with symptoms – from which it is eventually possible to re-annotate diagnoses.

Beyond questionnaires. While estimating questionnaire items is more relevant than estimating diagnoses or questionnaire scores, it is worth noting that questionnaires are "validated" based on their ability to discriminate between subjects

based on their diagnostic status. A promising approach is to move beyond questionnaires and work directly at the symptom level. Clinical psychology research, which specializes in symptom measurement and questionnaire design, can be a valuable resource for choosing labels for annotating corpora. For example, Newson et al. (2021) extracted and listed a set of 43 symptoms from 10,154 items related to the 10 most prevalent mental disorders in the DSM. This list is available online. Even better, Jover Martínez et al. (2023) not only identified a minimal set of fundamental symptoms in psychopathology through qualitative interviews and focus groups with psychiatrists but also provided a questionnaire to measure them in an Ecological Momentary Assessment context in their supplementary data[5].

7. Conclusion

To conclude, we have argued in favor of a vocal biomarker paradigm for psychiatry, moving from a focus on diagnosis to the estimation of symptoms. In particular, we have shown that diagnosis is not central in clinical practice, and that its automatic and objective estimation may not have the expected impact on the quality of patient screening and care. On the other hand, estimating symptoms using vocal biomarkers allows a transdiagnostic approach to health estimation and facilitates the integration of vocal biomarkers into current clinical practice, which has the potential to improve the management, monitoring, and quality of care provided to the many individuals suffering from mental disorders.

We hope that this article can help to convince the speech processing community of the need for dialogue with field players and understanding of their working methods, especially in a health-related context: this dialogue is the umbilical cord that connects our sometimes abstract research to the reality that we are helping to reconfigure (Callon et al., 2001).

8. Acknowledgment

This work has received financial support from the CNRS MITI PRIME 80 DSM-HEALTH and from the French Research Agency ANR through the axis “Autonom-Health” of the PEPR ”Santé Numérique”, Grant agreement n°ANR-22-PESN-000X. VM has received funding from the European Union’s Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101106577.

9. Bibliographical References

[5]https://osf.io/kcbh5/, consulted 4th October, 2023

Ahmed Aboraya. 2007. Clinicians’ opinions on the reliability of psychiatric diagnoses in clinical settings. *Psychiatry*, 4(11):31–33.

Alice Baird, Nicholas Cummins, Sebastian Schnieder, Jarek Krajewski, and Björn W. Schuller. 2020. An Evaluation of the Effect of Anxiety on Speech — Computational Prediction of Anxiety from Sustained Vowels. In *Interspeech 2020*, pages 4951–4955.

Derek Bolton. 2012. Classification and causal mechanisms: a deflationary approach to the classification problem. In *Philosophical Issues in Psychiatry II - Nosology*, oxford university press edition, pages 6–11.

Alexis Bourla, Florian Ferreri, Laetitia Ogorzelec, Charles-Siegfried Peretti, Christian Guinchard, and Stephane Mouchabac. 2018. Psychiatrists’ Attitudes Toward Disruptive New Technologies: Mixed-Methods Study. *JMIR Mental Health*, 5(4):e10240.

Frédéric Briend, Céline David, Silvia Silleresi, Joëlle Malvy, Sandrine Ferré, and Marianne Latinus. 2023. Voice acoustics allow classifying autism spectrum disorder with high accuracy. *Translational Psychiatry*, 13(1):250.

Joan Busner, Stuart L. Kaplan, Nicholas Greco, and David V. Sheehan. 2011. The use of research measures in adult clinical practice. *Innovations in Clinical Neuroscience*, 8(4):19–23.

Michel Callon, Yannick Barthe, and Pierre Lascoumes. 2001. *Agir dans un monde incertain: essai sur la démocratie technique*. Editions du Seuil, Paris. OCLC: 1354767526.

Edward L. Campbell, Judith Dineley, Pauline Conde, Faith Matcham, Katie M. White, Carolin Oetzmann, Sara Simblett, Stuart Bruce, Amos A. Folarin, Til Wykes, Srinivasan Vairavan, Richard J. B. Dobson, Laura Docio-Fernandez, Carmen Garcia-Mateo, Vaibhav A. Narayan, Matthew Hotopf, and Nicholas Cummins. 2023. Classifying depression symptom severity: Assessment of speech representations in personalized and generalized machine learning models. In *INTERSPEECH 2023*, pages 1738–1742. ISCA.

Daniel J. Carlat. 2017. *The psychiatric interview*, fourth edition edition. Wolters Kluwer, Philadelphia. OCLC: 959403655.

Michelle Cleary, Glenn E. Hunt, and Jan Horsfall. 2009. Delivering Difficult News in Psychiatric Settings. *Harvard Review of Psychiatry*, 17(5):315–321.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, James R. Williamson, Thomas F. Quatieri, and Jarek Krajewski. 2020. Generalized Two-Stage Rank Regression Framework for Depression Score Prediction from Speech. *IEEE Transactions on Affective Computing*, 11(2):272–283.

Peter B. Denes and Eliott N. Pinson. 1963. *The Speech Chain: The Physics and Biology of Spoken Language*, bell telephone laboratories, edition.

Nicole M. Etter, Farlah A. Cadely, Madison G. Peters, Crystal R. Dahm, and Kristina A. Neely. 2021. Speech motor control and orofacial point pressure sensation in adults with ADHD. *Neuroscience Letters*, 744:135592.

Harriet Fagerberg. 2022. Why Mental Disorders are not Like Software Bugs. *Philosophy of Science*, pages 1–22.

Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, and Vladimir Despotovic. 2021. Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. *Digital Biomarkers*, pages 78–88.

Salvatore Fara, Orlaith Hickey, Alexandra Georgescu, Stefano Goria, Emilia Molimpakis, and Nicholas Cummins. 2023. Bayesian Networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data. In *INTERSPEECH 2023*, pages 1728–1732. ISCA.

Mireia Farrús, Joan Codina-Filbà, and Joan Escudero. 2021. Acoustic and prosodic information for home monitoring of bipolar disorder. *Health Informatics Journal*, 27(1):1460458220972755. Publisher: SAGE Publications Ltd.

Maria Faurholt Jepsen, Darius Adam Rohani, Jonas Busk, Morten Lindberg Tønning, Maj Vinberg, Jakob Eyvind Bardram, and Lars Vedel Kessing. 2022. Discriminating between patients with unipolar disorder, bipolar disorder, and healthy control individuals based on voice features collected from naturalistic smartphone calls. *Acta Psychiatrica Scandinavica*, 145(3):255–267.

Michael Feehan, Leah A. Owen, Ian M. McKinnon, and Margaret M. DeAngelis. 2021. Artificial Intelligence, Heuristic Biases, and the Optimization of Health Outcomes: Cautionary Optimism. *Journal of Clinical Medicine*, 10(22):5284.

Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. *SSRN Electronic Journal*.

Allen Frances. 2013. Saving normal: An insider's look at what caused the epidemic of mental illness and how to cure it. *New York, NY: William Morrow*.

Eiko I. Fried and Randolph M. Nesse. 2015. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Medicine*, 13(1):72.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11):1544.

Simon Gilbody, David Richards, Stephen Brealey, and Catherine Hewitt. 2007. Screening for Depression in Medical Settings with the Patient Health Questionnaire (PHQ): A Diagnostic Meta-Analysis. *Journal of General Internal Medicine*, 22(11):1596–1602.

Gábor Gosztolya, Anita Bagi, Szilvia Szalóki, István Szendi, and Ildikó Hoffmann. 2020. Making a Distinction Between Schizophrenia and Bipolar Disorder Based on Temporal Parameters in Spontaneous Speech. In *Interspeech 2020*, pages 4566–4570.

Hélène Grajales. 2022. L'annonce d'un diagnostic en psychiatrie. *La Revue de l'Infirmière*, 71(281):49–50.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *LREC 2014*, pages 3123–3128, Reykjavik, Iceland.

Alberto Jover Martínez, Lotte Lemmens, Eiko I Fried, and Anne Roefs. 2023. Developing a Transdiagnostic Ecological Momentary Assessment Protocol for Psychopathology. preprint, PsyArXiv.

Lena Kastner, Markus Langer, Veronika Lazar, Astrid Schomacker, Timo Speith, and Sarah Sterz. 2021. On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 169–175, Notre Dame, IN, USA. IEEE.

Keneth S. Kendler. 2012. Classification and causal mechanisms: a deflationary approach to the classification problem - Introduction. In *Philosophical Issues in Psychiatry II - Nosology*, oxford university press edition, pages 3–5.

Firas Kobeissy, Ali Alawieh, Stefania Mondello, Rose-Mary Boustany, and Mark S. Gold. 2013. Biomarkers in psychiatry: how close are we? *Frontiers in Psychiatry*, 3.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. 2001. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Bernd J. Kröger, Catharina Marie Stille, Peter Blouw, Trevor Bekolay, and Terrence C. Stewart. 2020. Hierarchical Sequencing and Feedforward and Feedback Control Mechanisms in Speech Production: A Preliminary Approach for Modeling Normal and Disordered Speech. *Frontiers in Computational Neuroscience*, 14(573554).

Masayoshi Kurachi, Tsutomu Takahashi, Tomiki Sumiyoshi, Takashi Uehara, and Michio Suzuki. 2018. Early Intervention and a Direction of Novel Therapeutics for the Improvement of Functional Outcomes in Schizophrenia: A Selective Review. *Frontiers in Psychiatry*, 9:39.

Michelle H. Lim and John F. Gleeson. 2014. Social Connectedness Across the Psychosis Spectrum: Current Issues and Future Directions for Interventions in Loneliness. *Frontiers in Psychiatry*, 5.

Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The ADReSSO Challenge. Technical report. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Type: article.

Emmanouil Magiorkinis, Aristidis Diamantis, Dimos-Dimitrios Mitsikostas, and George Androutsos. 2009. Headaches in antiquity and during the early scientific era. *Journal of Neurology*, 256(8):1215–1220.

ChristyMaria Manuel, PavithraP Rao, Preethi Rebello, At Safeekh, and PJohn Mathai. 2013. Medical comorbidity in in-patients with psychiatric disorder. *Muller Journal of Medical Sciences and Research*, 4(1):12.

Daniel J. Martin, John P. Garske, and M. Katherine Davis. 2000. Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, 68(3):438–450.

Vincent P. Martin, Jean-Luc Rouas, Pierre Philip, Pierre Fourneret, Jean-Arthur Micoulaud-Franchi, and Christophe Gauld. 2022. How Does Comparison With Artificial Intelligence Shed Light on the Way Clinicians Reason? A Cross-Talk Perspective. *Frontiers in Psychiatry*, 13.

Patrick McGorry and Barnaby Nelson. 2016. Why We Need a Transdiagnostic Staging Approach to Emerging Psychopathology, Early Diagnosis, and Treatment. *JAMA Psychiatry*, 73(3):191.

Paul Meehl. 1967. What Can the Clinician Do Well? *Problems in human assessment*, pages 594–599.

Carmen Mijnders, Esther Janse, Paul Naarding, and Khiet P. Truong. 2023. Acoustic characteristics of depression in older adults' speech: the role of covariates. In *INTERSPEECH 2023*, pages 4159–4163. ISCA.

Stuart A. Montgomery and Marie Åsberg. 1979. A New Depression Scale Designed to be Sensitive to Change. *British Journal of Psychiatry*, 134(4):382–389.

Jennifer Jane Newson, Vladyslav Pastukh, and Tara C. Thiagarajan. 2021. Poor Separation of Clinical Symptom Profiles by DSM-5 Disorder Criteria. *Frontiers in Psychiatry*, 12:775762.

Rebecca Nowland, Elizabeth A. Necka, and John T. Cacioppo. 2018. Loneliness and Social Internet Use: Pathways to Reconnection in a Digital World? *Perspectives on Psychological Science*, 13(1):70–87.

Wei Pan, Liat Shenhav, Amber Afshan, Abeer Alwan, Jonathan Flint, Tianli Liu, Bin Hu, and Tingshao Zhu. 2021. The Discriminatory Power of Vocal Features in Detecting Mental Illnesses Under Complex Context. preprint, In Review.

Kun Qian, Xiao Li, Haifeng Li, Shengchen Li, Wei Li, Zuoliang Ning, Shuai Yu, Limin Hou, Gang Tang, Jing Lu, Feng Li, Shufei Duan, Chengcheng Du, Yao Cheng, Yujun Wang, Lin Gan, Yoshiharu Yamamoto, and Björn W. Schuller. 2020. Computer Audition for Healthcare: Opportunities and Challenges. *Frontiers in Digital Health*, 2:5.

Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. 2022. MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107.

Mi Jung Rho, Jihwan Park, Euihyeon Na, Jo-Eun Jeong, Jae Kwon Kim, Dai-Jin Kim, and In Young Choi. 2019. Types of problematic smartphone use based on psychiatric symptoms. *Psychiatry Research*, 275:46–52.

Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9, Mountain View California USA. ACM.

Daniel W. Russell. 2014. Loneliness and social neuroscience. *World Psychiatry*, 13(2):150–151.

Tomasz Rutowski, Amir Harati, Elizabeth Shriberg, Yang Lu, Piotr Chlebek, and Ricardo Oliveira. 2022. Toward Corpus Size Requirements for Training and Evaluating Depression Risk Models Using Spoken Language. In *Interspeech 2022*, pages 3343–3347. ISCA.

Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. 2013. Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD. In *Interspeech 2013*.

Katharina Schultebraucks, Vijay Yadav, Arieh Y. Shalev, George A. Bonanno, and Isaac R. Galatzer-Levy. 2020. Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. *Psychological Medicine*, pages 1–11.

Robert L. Spitzer. 1983. Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry*, 24(5):399–411. Place: Netherlands Publisher: Elsevier Science.

Jon Stone, Russell Hewett, Alan Carson, Charles Warlow, and Michael Sharpe. 2008. The 'disappearance' of hysteria: Historical mystery or illusion? *Journal of the Royal Society of Medicine*, 101(1):12–18. Publisher: SAGE Publications.

M. Tafalla, J. Sanchez-Moreno, T. Diez, and E. Vieta. 2009. Screening for bipolar disorder in a Spanish sample of outpatients with current major depressive episode. *Journal of Affective Disorders*, 114(1-3):299–304.

Sunny X. Tang, Reno Kriz, Sunghye Cho, Suh Jung Park, Jenna Harowitz, Raquel E. Gur, Mahendra T. Bhati, Daniel H. Wolf, João Sedoc, and Mark Y. Liberman. 2021. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *npj Schizophrenia*, 7(1):25.

Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. 2023. The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection. In *INTERSPEECH 2023*, pages 4149–4153. ISCA.

Alan R. Teo. 2010. A New Form of Social Withdrawal in Japan: a Review of Hikikomori. *International Journal of Social Psychiatry*, 56(2):178–185.

The Lancet Global Health. 2020. Mental health matters. *The Lancet Global Health*, 8(11):e1352.

Alfonso Troisi. 2022. Biological psychiatry is dead, long live biological psychiatry! *Clinical Neuropsychiatry*, 19(6):351–354.

Jinfang Wang, Ke Lv, Chang Liu, Xinli Nie, Dhananjaya Gowda, and Shuxin Luan. 2021. Automatic Assessment for Severe Self-Reported Depressive Symptoms Using Speech Cues. *IEEE Transactions on Cognitive and Developmental Systems*, 13(4):875–884.

Monika A. Waszczuk, Mark Zimmerman, Camilo Ruggero, Kaiqiao Li, Annmarie MacNamara, Anna Weinberg, Greg Hajcak, David Watson, and Roman Kotov. 2017. What do clinicians treat: Diagnoses or symptoms? The incremental validity of a symptom-based, dimensional characterization of emotional disorders in predicting medication prescription patterns. *Comprehensive Psychiatry*, 79:80–88.

Carlos A Zarate and Husseini K Manji. 2009. *Bipolar depression: molecular neurobiology, clinical diagnosis, and pharmacotherapy*. Springer.

Mark Zimmerman and Joseph B. McGlinchey. 2008. Why don't psychiatrists use scales to measure outcome when treating depressed patients? *The Journal of Clinical Psychiatry*, 69(12):1916–1919.