



Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.journals.elsevier.com/intelligence-based-medicine

Machine learning for ambulatory applications of neuropsychological testing

Chelsea Chandler^{a,*}, Peter W. Foltz^a, Alex S. Cohen^b, Terje B. Holmlund^c, Jian Cheng^d,
Jared C. Bernstein^d, Elizabeth P. Rosenfeld^d, Brita Elvevåg^{c,e}

^a University of Colorado Boulder, United States^b Louisiana State University, United States^c University of Tromsø, Norway^d Analytic Measures Inc, United States^e Norwegian Centre for eHealth Research, Norway

ARTICLE INFO

Keywords:

Machine learning

Mental illness

Neuropsychological testing

Remote assessment

ABSTRACT

Psychiatric patients, such as those suffering from depression or schizophrenia, often need to be monitored with frequent clinical interviews by trained professionals to avoid costly emergency care and preventable events. However, there simply are not enough clinicians to monitor these patients on a regular basis. Furthermore, infrequent clinical evaluations may result in clinicians missing subtle changes in patient state that occur over time. These limitations can affect both the quality, timeliness, and monetary expense of treatment. Therefore, we leveraged smart devices to implement traditional neuropsychological assessments such that they could be collected frequently, remotely, and - when viable - self-administered by the participants themselves. This approach enables the generation of an enormous quantity of data across time and different assessments. Machine learning-based methods hold the potential to automatically analyze streams of behavioral and cognitive data, such as speech and movement, and convert them to actionable events. We examined the viability of the automation of a comprehensive assessment pipeline, from administration of neuropsychological tests, to transcription of spoken responses, to an analysis of data to predict clinical states. In the present research, we examined this pipeline in 353 participants (of whom 134 were patients with a range of diagnoses of psychosis spectrum disorders, substance abuse disorders, and affective disorders, and 219 were non-patient volunteers who were presumed to be healthy). We found that machine learning-based methods can be applied to this data in order to reliably and accurately assess the neuropsychological function of individuals. Among other applications, we were able to automatically score completion of a verbal recall task and predict emotional state via spoken language, thereby opening the potential for regular, frequent analyses of cognitive and mental states.

1. Introduction

Mental illness is a public health crisis that causes a significant burden on not only patients, but also their family members, communities, and healthcare systems alike. The assessment of clinical states in mental illness is critical, but it is a complex and expensive process that could become more efficient and accurate by leveraging modern technology and analytics. According to the National Institute of Mental Health, one in five adults in the United States lives with some form of mental illness [1]. With such high prevalence, further confounded by the requirement to physically visit a doctor's office for assessments, there exist accessibility issues which contribute to an unequal access to services. Furthermore, human cognitive and mental states are dynamic over time and

context but most traditional assessment methods primarily afford cross-sectional snap-shots in time. Therefore, they potentially fail to capture a sufficiently thorough neuropsychological profile of individuals who may - by virtue of their specific clinical condition - be changing in a clinically significant manner. The dynamics in cognition and fluctuations in state warrant a new framework for assessment as such changes necessitate frequent and longitudinal monitoring.

This study researched and developed a prototype mobile automated telemental health monitoring tool for deriving frequent measurements of psychiatric patients' current neuropsychological function and then predicted patient cognitive and mental states with machine learning models. The study further compared these predictions to expert rated labels to establish how well automated analyses align with traditional clinical

* Corresponding author. Department of Computer Science, 430 UCB, 1111 Engineering Dr., Boulder, CO, 80309, United States.

E-mail address: chelsea.chandler@colorado.edu (C. Chandler).

<https://doi.org/10.1016/j.ibmed.2020.100006>

Received 17 April 2020; Received in revised form 18 September 2020; Accepted 24 September 2020

2666-5212/© 2020 Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

judgment in order to determine the feasibility of generating actionable alerts. Recent innovations, including the wide availability of mobile devices to collect continuous streams of data, combined with the advancement of machine learning methods for analyzing these data streams, promises to reshape the current manual assessment pipeline.

In this research, an iOS platform (Apple's mobile operating system) application was built to consistently and reliably collect data from a set neuropsychological tasks for the purposes of assessing the cognitive and mental state of its users. Three hundred and fifty three participants (of whom 219 were non-patient volunteers who were presumed to be healthy and 134 were adults from an inpatient unit in Louisiana serving low socioeconomic individuals with comorbid psychiatric concerns and substance use disorders¹) used this application for multiple sessions and data was collected and analyzed. For this paper, we focused our analyses on modeling tasks with speech output. As such, we produced a variety of models that were able to accurately predict certain features of a participant's task performance and mental state. First, we created support vector regression models for predicting emotion (both expert rated and self-reported) from acoustic speech features. In order to automatically transcribe speech, custom neural network automatic speech recognition systems were created for improved transcription of task-specific speech. Then, we improved our initial acoustic speech-based models with the addition of language features. Finally, we created a ridge regression model that predicted quality of story recall and a logistic regression ensemble classifier based on the same task that predicted class membership (patient vs. non-patient) of participants.

2. Application of automated methods for neuropsychological testing

Within psychiatry, clinicians are largely unaware and skeptical of the ways in which technology has the potential to make their work more efficient and accurate [2]. Nevertheless, there are at least three main areas in which the assessment of clinical states via traditional methods could be improved. First, given the vast number of patients, there are simply not enough health professionals to monitor patients as frequently as necessary. Hence, the use of technology and the automation of assessment could greatly increase the number of evaluations that could be conducted, as well as nurture equity by expanding the potential patient population that is able to participate in such assessments. Second, patients with mental illness require long-term monitoring on the scale of years which is logistically challenging for clinicians since patients all have different clinical baselines against which they need to be compared. Furthermore, current methods simply cannot accommodate the requirement of multiple assessments on this scale, and nor do the norms exist that would be required to interpret more frequently obtained measurements. An automated machine learning approach would be able to store frequently changing clinical data on a large time scale and adjust individual thresholds to account for changing state over time. Third, it is often the case that patients have met with clinicians just a few days prior to a relapse or attempt at suicide [3], thus highlighting how quickly clinical states in mental illness can change. Emerging technologies provide the technical viability for longitudinal monitoring which can track and transmit data that is determined to be key for specific patients [4], including information about affect, activation level, and suicidal ideation. Combined with suitable clinical research, machine learning could be used to generate actionable alerts to initiate human intervention [5].

¹ Although this type of sampling resulted in a design that is not optimal for clinical comparison purposes, the sampling was sufficient for the primary goal of this study which was to establish the viability of this type of data collection in terms of participants using and tolerating the system, that good quality data was possible to obtain outside of the traditional controlled laboratory, and that it was possible to apply machine learning to such data to generate predictions of mental state.

Data capture can be unobtrusive (e.g., wristband monitors and smart devices) and thus collect continuous data streams. Body movement and non-verbal aspects of speech can be modeled so as to form robust indicators of psychomotor activation and affect, while automatic analysis of content and manner of speech can be leveraged to detect morbid ideation and changes in symptoms of depression and schizophrenia [6,7].

The pattern and content of communication provides large amounts of information that can be traced back to an individual's overall cognitive and mental state. Indeed, the information conveyed in speech is core in mental health diagnosis, treatment, and in monitoring treatment success. Additionally, speech is the modality through which a large amount of neuropsychological assessment is conducted, yet traditional assessment has yet to leverage speech directly and automatically. Thus, we utilized the latest in speech technologies so as to automatically analyze audio properties and recognize speech output (i.e., via an automated speech recognition system) and then perform statistical semantic and syntactic analyses of language. We combined these audio and language analyses to predict clinically important variables and thus develop a prototype automated analysis system. We captured this data in real-time and applied analytical methods to generate scores that estimate measures of cognitive and mental states. Our conceptualization of such a 'telemental health' monitoring tool is illustrated in Fig. 1. First, a user is presented with a task item via graphics and audio output. Then, depending on the item presented, a user will either respond with speech or a screen touch. If they respond with speech, the audio is input to a speech recognition system where the words, timing, and audio features are extracted. These features are then input to machine learning models to predict content and quality of speech scores. However, if they respond with a screen touch, patterns of touch and positional features are input to machine learning models to predict content and psychomotor scores. These speech and touch scores from various task items are then combined to predict a final state estimate.

Since behavior, cognition, and clinical states are highly personal, psychometric frameworks must utilize not only traditional population thresholds but also those of individuals, which is now viable with longitudinal collection of personal data. Consider the following scenario as depicted in Fig. 2: imagine that we are capturing some number of signs of agitation from psychiatric patients' verbal content and manner of speaking. If we find in studying a large outpatient population that the best overall threshold for intervention is at 1.5 standard deviations from the average, this can guide us as a rule of thumb. Note however, that Patient A might generally be more calm and more consistently calm than the average found in the population (i.e., individual baseline statistics for Patient A suggest that a lower threshold is better as the population threshold would miss most events). That patient's intervention threshold should thus be lower than the population standard. By contrast, Patient B is an unusually energetic and emotionally labile person and thus Patient B would be best served by setting a substantially higher threshold than the threshold that is best over all the population (i.e., individual baseline statistics for Patient B suggest that a higher threshold is better as the population threshold would trigger many false alarms). We illustrate that today's measurement for Patient B (on one dimension) is well within the population range yet lower than average for their own distribution, and as such represents a measurement that may warrant further investigation. Such a conclusion may not have been possible without frequent data collection and monitoring. Since psychiatric patients are currently examined quite infrequently, changes in important mental states may fail to be detected. However, we note that detecting such clinically significant changes is just one of many critical steps in the automation pipeline. Further research is needed to identify suitable guidelines for creating thresholds for individuals with which clinicians can be alerted and action taken.

A machine learning system can implement such a tailored thresholding by learning patterns of typical individual and population performances and compare a particular participant's daily performance to both of these. We note that this is a hypothetical example and that extensive

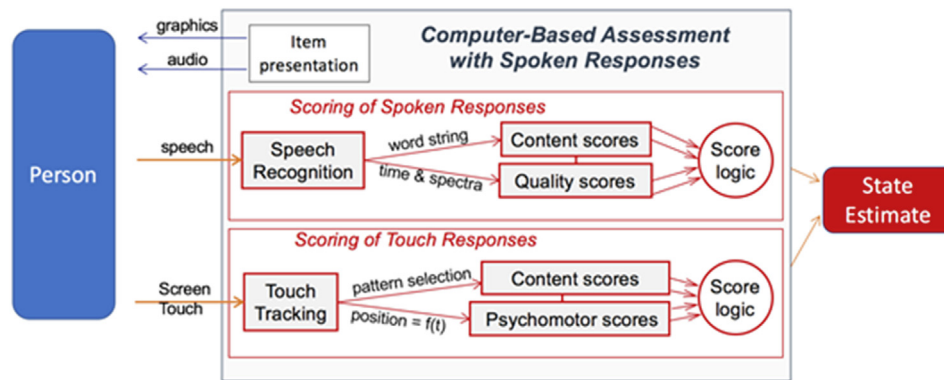


Fig. 1. Schematic view of the computer-based assessment of spoken and touch responses, illustrating the flow of information to and from a voice-interactive device.

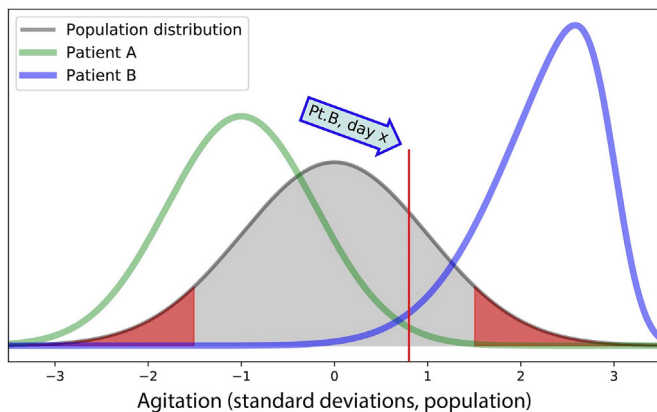


Fig. 2. Hypothetical distributions of behavioral signs of agitation from a general population (gray) and two individual psychiatric patients (green, blue). A very uncommon value for an individual patient (blue arrow, Patient B) may be well within a common range for the population and as such may be misjudged without personalized thresholding. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

clinical research is needed to produce the quantity and type of data that can confirm or disconfirm this type of conjecture and enable us to expand on the associated methodology. We suggest the above longitudinal ‘personalized medicine’ approach such that each patient becomes their own baseline, and results from other patients with ‘similar’ illnesses are used as general guides. Thus, our proposed system is to complement the efforts of busy clinicians and help them focus their attention on the most pressing situations. This is scientifically and technologically viable because an enormous amount of useful acoustic, linguistic, cognitive, and clinical assays can be derived and thus provide promise in the near future of a ‘mental blood test’.

3. Related work

Previous work on automating the assessment of neuropsychological conditions has been typically conducted within controlled laboratory settings and in modest sample sizes [8,9]. Automated assessments of narrative retelling, for example, have been conducted employing neuropsychological tests very similar to portions of our current study [10, 11]. However, as noted, the data collected in these aforementioned studies have been within controlled laboratory or clinical settings rather than our work where the focus has been on moving assessment *out* of the controlled laboratory where assessment is by an expert administrator, and *into* real world settings where tasks are administered by the participant themselves. With that said, recent research has shown that *passive* measures retrieved from general smartphone use can successfully predict

key aspects of cognitive function such as working memory, memory, executive function, language, and intelligence [12].

Natural language processing (NLP) methods provide the tools with which it is possible to model aspects of the semantics and syntax produced in speech as well as characterize the statistical properties of language use. Semantic properties of language can be computed by analyzing large corpora of text to derive estimates of the semantic relatedness of words as a function of the contexts in which they co-occur, typically through the use of probabilistic inference (e.g., Latent Dirichlet Allocation [13]), singular value decomposition (e.g., Latent Semantic Analysis [14]), or neural networks (e.g., word2vec [15]). Such estimates provide baselines that can be used to compare the generated language to measure aspects of discourse in patients. Bedi et al., for example, found that a Latent Semantic Analysis measure of semantic coherence, maximum phrase length, and the use of determiners were able to predict the subsequent development of psychosis with 100% accuracy on a small sample of psychiatric patients [8]. Similarly, Corcoran et al., found that a decrease in semantic coherence, a variance in semantic coherence, and a reduced usage of possessive pronouns predicted subsequent psychosis onset with 83% accuracy [9]. Typically these studies are interested in predicting illness onset over a matter of months or years, but our present study has a focus on immediate predictions within hours or days.

Indeed, a number of other studies have shown that in small samples of data, such approaches can predict clinical classes and clinician ratings [16–19]. However, we want to not just assess speech for the sake of the language produced, but to jointly use the speech signal as a direct modality through which to assess neuropsychological function and clinical state. The present work showcases how such techniques can move beyond a simple proof of concept and in the near future be translated into viable clinical tools.

4. *dMSE* application

A total of 12 unique behavioral assessment item types that were designed to assess cognition, motor skill, and language were integrated into an application called the *delta* Mental Status Exam (*delta* to indicate our interest in ‘change’; *dMSE*). The items were similar in form and structure to standardly employed neuropsychological tests [20], but were designed so that 1) the items could be easily used for frequent (e.g., daily) and remote self-administration with a smart device, 2) the items provided short engaging tasks that required the users to listen, watch, speak, and touch, and 3) the interactions could be automatically analyzed to extract a rich set of measures from each item. Nine of the 12 tasks were adaptations of popular neuropsychological tests (for an overview, see Refs. [20]), notably trail making, Stroop selective attention, immediate and delayed Logical Memory story recall (of the Wechsler Memory test, [21]), semantic verbal fluency, finger tapping, and digit, letter, and visual-spatial span tasks, and three of the tasks required participants to provide speech in response to image stills or verbal prompts. Fig. 3 shows



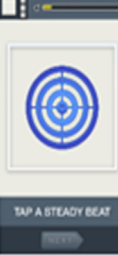
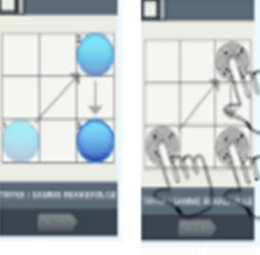




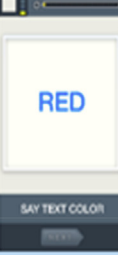

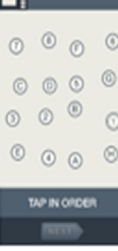

<p>1 Say what you see</p> <p>Describe what is happening in a picture. We'd like you to describe as much as possible. There will be two pictures in a row.</p>		<p>7 Name animals</p> <p>Say the name of as many animals as you can think of, as quickly as possible. You have one minute.</p>	
<p>2 Tap to the beat</p> <p>You will hear a sound with a steady rhythm. Press on the beat and continue tapping with the same rhythm when the sound goes away. You should do the task two times in a row.</p>		<p>8 Remember the order</p> <p>Follow a dot moving from space to space. Then repeat the order by pressing the screen.</p>	
<p>3 Retell the story</p> <p>Retell the story you heard the last time you used the app. Retell as many details as you can remember.</p>		<p>9 Listen and retell the story</p> <p>Listen to a story that will be read out loud. Then retell it with as many details as you can remember. Remember the story for the next time you do dMSE.</p>	
<p>4 Remember numbers</p> <p>You will see and hear a series of numbers, then press the numbers in the same order on the screen.</p>		<p>10 Write the letters</p> <p>You will hear six letters spoken to you. Write them down in the boxes in order. For example, RMBHJ. There will be four rounds.</p>	
<p>5 Say the color</p> <p>There will be differently colored words appearing on the screen. Say the color the word is written in. Say it loud, as fast as possible.</p>		<p>11 Say how you feel</p> <p>You will be asked about your current mood and mental health. Move the slider in the direction of the best answer.</p>	
<p>6 Touch the dots in order</p> <p>Touch the dots in order. There are three variations.</p>		<p>12 Suggestions</p> <p>Do you have any suggestions to improve the dMSE app?</p>	

Fig. 3. A sample sequence of tasks and brief overview of the *delta* Mental Status Examination (dMSE).

an overview of a sequence of tasks presented to a participant during each test session.

Our version of the Stroop task involved a reduction in the number of trials so as to be acceptable for ambulatory purposes yet enough trials so as to be statistically robust, and we also increased the pace of the task to make it more engaging. Our adaptation of the immediate and delayed Logical Memory story recall tasks were structurally similar to the original version, but included many more story versions and allowed for recording of the speech responses so as to apply automated speech recognition and score the responses with NLP and machine learning. These adaptations were based on solicited user feedback which allowed task redesign. As such, our versions of the tasks likely capture the similar neurocognitive processes that traditional assessment tools do, but enable much more fine grained detail to additionally be collected, analyzed, and acted upon. Although not conducted in our study, future clinical trials that additionally employ traditional neuropsychology tests and methods would be good to compare with our novel versions.

The tasks were developed in both English and Norwegian to be implemented as mobile applications on the iOS platform (for more details of the application, see Ref. [22]). For the purpose of this paper we focus on the data collected in the English language. The iOS platform provided several advantages including the ability to easily download the applications to smart devices such as iPhones and internet-connected iPods as well as update the application when needed. This allowed two versions of the application to be created; the second version incorporated changes made to better align with user feedback on the first. Additionally, the application framework permitted fast development of highly useable interface components including video and image displays, speech recording, and capture of user interaction with the touch-screen.

The usefulness and feasibility of the application was addressed by considering user receptivity to it and overall impression, their willingness to comply with its requirements, including issues such as potential fatigue or boredom as well as tolerating it over extended periods of time, and whether the tool successfully and efficiently collected useful data. Overall, our results showed that the system was easily useable by psychiatric patients and non-patients (tested on a subset of our data: 100 psychiatric patients and 125 non-patients). Seventy-two out of 100 psychiatric patients and 88 out of 125 non-patients completed all of a planned series of 5 consecutive daily sessions. Comments from the surveys were generally favorable with over 99% compliance rates (i.e., completing initiated sessions).

We solicited feedback on usability from a sub-group of psychiatric patients and clinicians ($N = 24$) about the tolerability of the test duration and efficiency of data collection (for details, see Refs. [22]). Users tolerated it well with some patients happily using it for 30–40 days/sessions (median = 5, range = 1–40). Participants were asked to indicate using a slider the magnitude (0–100) to which they believed this tool would be useful or not to monitor mental health (average = 76.5, $SD = 15.1$), and whether they liked using this tool (average = 77.0, $SD = 16.3$). The main complaint was that there were too many tasks and/or it took too long. When asked how often they believed such a tool should be used, the majority suggested several times a week if not daily and the average suggestion for length of the task was 8.75 min. Interestingly, this is only slightly shorter duration than the current version of this task (approximately 10 min). Even if the main complaint was related to duration, we see this as likely that users will comply with spending a short duration of time of testing daily in confined time periods.

It is well documented that most mental health apps are used only briefly by participants before their use is discontinued. One study of mental health apps showed that in the first ten days of app usage there is a decline of more than 80% in app open rates and then there is an additional 20% decline between days 15 and 30 [23]. We are optimistic that this will not be the case in the use of our system, although a full scale clinical trial will be necessary to establish this empirically. Our optimism stems from several sources: first, we followed the key engineering design principle of clearly establishing *a priori* what users used their current

tools to achieve and designed accordingly [24]. Second, the different types of users (patients, non-patients, clinicians) were all involved at all stages of the tool development, from the initial design to the later stages when feedback was solicited on numerous issues (e.g., the actual interface). Crucially, the improvement from our first to our second version of the tool reflected these comments. Third, the tasks are all brief and engaging (as several presentation and response modalities are employed), and this feel of novelty is further facilitated by the employment of different versions of tasks and just a subset of the entire task list appearing in each session.

5. Methods and analyses

We first investigated how well machine learning approaches could analyze unobtrusive interactions such as speech-based tasks from users and be converted to measures of cognitive function and mental state. Although there are a variety of task types in the *dMSE* application, in this paper we focus only on a select number of speech-based tasks, namely tasks 3 and 9 (verbally retelling a story immediately after hearing it and after a delay), 5 (saying the color of a word shown on the screen: the Stroop task), and additionally 11 (sliders for self-report of mood and mental health; used as prediction targets) from Fig. 3, as well as open ended response prompts. These tasks were chosen to analyze as this research specifically focused on the analysis of speech. Each task builds up speech in an incremental way; the Stroop task is simple and constrained yet with the use of ASR and speech characteristic packages we can extract important behavioral characterizations while the verbal recall task is less constrained and more challenging, both for the user and technologically. The mood and mental health sliders give key information on self-reported mental state that can be used alongside task performance.

Developing the automatic language scoring system was the critical part of the analytical research in this project and comprised four key components, namely the research and development of (i) an automated analysis of acoustic properties of speech; (ii) an automatic speech recognition system that was specific to the assessment tasks; (iii) NLP-based methods to extract and analyze semantic and syntactic features of the speech; and (iv) machine learning methods for combining features to predict clinically important variables.

In each application of machine learning methods, specific model types were chosen as they yielded the highest accuracies, were able to learn patterns in relatively small datasets, and were interpretable such that in the future clinicians will be able to ascertain the influence of individual feature inputs [25]. It is because of these constraints that more complex modeling techniques such as neural networks were not used as our final chosen models. We discuss the four components of our research below, with results of each application included in Table 1.

5.1. Analysis of speech

Voice recordings from structured, yet open-ended questions provide sufficient data to assess the acoustic properties of speech and characterize changes in mood and emotional valence. We chose to analyze spontaneous speech in reaction to greeting questions such as *how are you* and *how did you sleep last night*, describing still pictures and silent videos, retelling a story immediately and after a delay, free speech questions as to *how one would boil an egg*, *whether television has changed family life for the better or not*, and *why some people might prefer electric cars to the more traditional gasoline ones*, and suggestions for improving the app.

5.1.1. Detecting mood fluctuations

The first step in the evaluation of acoustic features for measuring mood fluctuations was to determine suitability of the audio recordings for acoustic analysis. The preliminary analysis showed that nearly all recordings were deemed adequate (minimal background noise and intelligible to humans), suggesting that the hardware and software technologies

Table 1

Results of each application of machine learning in the present study, ordered by the section in which it is detailed.

Assessment	Machine Learning Model	Features Used	Assessment Metric	Results
Arousal prediction	Support vector regression model with RBF kernel	Acoustic speech features	Correlation to expert rating	r = 0.70 (non-patient) r = 0.72 (patient)
Emotional valence prediction	Support vector regression model with RBF kernel	Acoustic speech features	Correlation to expert rating	r = 0.67 (non-patient) r = 0.43 (patient)
Positive affect prediction	Support vector regression model with RBF kernel	Acoustic speech and language features from story recalls	Correlation to self-report	r = 0.38 (non-patient) r = 0.44 (patient)
Negative affect prediction	Support vector regression model with RBF kernel	Acoustic speech and language features from story recalls/ Acoustic speech features from Stroop task	Correlation to self-report	r = 0.40 (non-patient) r = 0.48 (patient)/ r = 0.47 (overall)
Story recall (a measure of participant memory) rating model	Ridge regression model	(1) # unique words spoken in the recall, (2) # common words between original story and the recall, and (3) word mover's distance	Correlation to expert rating	r = 0.88
Patient vs non-patient classification model	Ensemble of logistic regression classifiers	Same features as above computed on (1) the immediate recall, (2) the delayed recall, (3) the change between the immediate and delayed recalls	Classification accuracy of whether response was from a patient or non-patient	76% (human transcribed data) 74% (ASR transcribed data)

used in the *dMSE* system provided acceptable performance. Then, we evaluated an empirically-derived limited acoustic feature set in their convergence with expert ratings of emotional arousal and valence, with many correlations in the range of 0.30–0.45 (for details, see Ref. [7]).

Using ambulatory-based acoustic analysis of natural voice recorded from relatively structured speaking tasks in participants' home environment, we were able to evaluate the consistency of acoustic signals over time, its relationship to clinically-rated symptoms and state affect, and symptom-by-state interactions. Our results suggest that acoustic signals were fairly stable over time within psychiatric patients. We conducted multi-level modelling to evaluate the degree to which demographics (gender, age), clinically-rated psychiatric symptoms (affective and mania-agitation symptoms), ambulatory self-report state (negative affect) and self-reported symptoms were related to ambulatory-based acoustic variables (dependent variables). We found that acoustic variables alone were not highly related to psychiatric symptoms, as has been shown in prior research. However, irregularities in acoustics were associated with state-by-symptom interactions. For example, high levels of stress were associated with abnormally small changes in vocal expression for a wide variety of psychiatric symptoms (i.e., affective symptoms and manic-agitation symptoms). For psychiatric patients with no active psychiatric symptoms, stress levels had normal effects on voice modulation. They spoke less when stressed and spoke louder with more emphasis and intonation. The same modulation was not seen in patients with psychiatric symptoms. Therefore, the conjecture that vocal recordings are useful for understanding serious mental illness was supported, but only if voice data is collected alongside emotional state variables [7].

5.1.2. Predicting arousal and emotional valence

Speech processing and machine learning models built on speech data obtained from greeting questions, still picture and video descriptions, immediate and delayed story recall, and suggestions for improving the app further showed that we could accurately characterize important emotional states (for details see Refs. [6]). Interestingly, the models could predict participants' *arousal* and *emotional valence* in a manner that was comparable to trained human raters (including clinicians).

A stratified sample of the data was employed so as to provide a spread across a range of response styles (according to the number of words spoken and self-reported slider values). The resulting non-patient group comprised N = 28 sessions and the patient group comprised N = 116 sessions. Two rating rubrics were developed and used to rate the level of *arousal* (degree of excitement) and *emotional valence* (positive and

negative) in each response. *Arousal* and *emotional valence* are key components of emotion and various types of emotions can be modeled within these characteristics [26]. Nine independent raters used a 1–6 scale to rate the spoken responses and each response was rated by at least two raters in order to assess inter-rater agreement. For *arousal*, the average correlation for individual responses (across all raters) was 0.62, and for *emotional valence* this was 0.59. There was less variance in the *emotional valence* ratings as a result of many “neutral” ratings, indicating that it was more difficult for humans to judge, and indeed subjectively the experience is that it is easier to detect *arousal* in speech than *emotional valence*. The average rating of each rater per response averaged over all responses per session was used as the prediction target.

Speech features were modeled to predict human ratings by extracting speech signal processing features from the responses using a state-of-the-art open source package (openSMILE [27]; which includes 6373 distinct audio features such as energy, loudness, MFCC, PLP, F0, probability of voicing, voice-quality: jitter and shimmer, formant frequencies and bandwidth (F1, F2, F3, F4, etc.), harmonics-to-noise ratio, and so on as well as statistics of features (e.g., means, extremes, moments, segments, peaks, linear and quadratic regression, percentiles, durations, onsets, etc.). A support vector regression model (with a radial basis function (RBF) kernel, degree = 3, cost = 10, eps = 0.2, loss = 0.1, and normalize = true) was then built which combined and weighted features to best predict human ratings. Support vector regression models perform well with small datasets as they work by finding hyperplanes in a derived feature space. Since our feature space is not linearly separable, the RBF kernel is used to project the samples into a higher-dimensional feature space that can then be separated with a hyperplane. Ten-fold cross-validation was used to tune model parameters on 9 out of 10 equal splits of the data. The held out fold was used for testing performance and this process was repeated for all ten folds to compute final performance measures. The model predicted *arousal* with an average correlation to the average of the human raters of r = 0.70 (non-patient) and r = 0.72 (patient). For *emotional valence*, the best model only correlated at r = 0.67 (non-patient) and r = 0.43 (patient) on average to the average human ratings (see Ref. [6] for details). Results show that *arousal* can be assessed via speech more reliably than *emotional valence*, and with more data and spread of ratings, machine scoring has the potential to match expert raters even more closely.

5.1.3. Predicting self-reported affective states

In the next experiments, we predicted self-reported emotion from the acoustics and language of the seemingly affectless task of verbally

recalling a short story. Self-reports of positive and negative affect were solicited after participants were given the story recall task in the *dMSE* application where they were asked to recite with as many details as possible a short story that was read to them verbally (more details of the verbal recall task in section 5.3). The *dMSE* application contains 7 positive affect sliders that ask the user to report on personal levels of hopefulness, calmness, appreciation, strength, ability to concentrate, happiness, and levels of energy. Similarly, 8 negative affect sliders ask the user to self-report on personal levels of anxiety, frustration, fear, sadness, stress, anger, pain, and helplessness. The final self-reported positive and negative affect values per session is the average of the slider responses from each group.

For this study, a population of 21 psychiatric patients and 79 non-patients generated 137 and 430 total sessions respectively. Similar to the above study, we used the openSMILE audio feature extractor to generate speech features from each story recall response. The language feature set included token count, type count, type token ratio, content density, mean coherence, standard deviation of coherence, and counts of particular parts of speech such as verbs, nouns, pronouns. Type token ratio is defined as the ratio of word types to word tokens. Content density is operationalized as the ratio of content words (verbs, nouns, adjectives, and adverbs) to total words. Coherence is computed by comparing adjacent windows (of size $n = 4$ words) in the text for semantic similarity using the cosine distance between word vector embeddings of the words in each window. The average and standard deviation of similarities of all adjacent windows in a recall were computed.

In each experiment variation, a support vector regression model with the same parameters and cross validation technique as reported in section 5.1.2 was employed. First we showed that by analyzing just one modality of data, there were moderate correlations with self-reported affect ($0.33 < r < 0.40$ for speech and $0.07 < r < 0.28$ for language). Second we improved on these unimodal analyses by combining the acoustic and language features from story recalls to predict a person's self-reported affect. This combination of modalities resulted in an improved model with correlations of $0.38 < r < 0.48$ to self-reports [28].

Interestingly, predictions based on variables derived from speech collected from the seemingly innocuous Stroop attentional control task were also remarkably direct assays of self-reported negative affective states (predictions correlating $r = 0.47$ with reports) as compared to predictions based on variables from speech collected during verbal self-reports on subjective state (i.e., "How do you feel today?"; $r = 0.46$). In this task, words including color words are presented in various ink colors and the participant is tasked to either read the actual word and ignore the ink color or to name the ink color and ignore the actual word; see Fig. 3, item 5 for illustration of the task in which the participant is asked to name the text color the word **RED** is written in).

In sum, we were able to measure the audible emotion in the spontaneous speech collected using *dMSE* as determined by (i) consistency of speech signals over time, (ii) clinical expert ratings of *arousal* and *emotional valence*, and (iii) self-reported positive and negative affect measurements. Thus, overall we have found that acoustic measures can model arousal in speech - especially with negative affect - and that improvement can be gained when adding language features to speech-based models.

5.2. Automatic speech recognition

In order to process lexical and semantic information in the speech of participants, it is necessary to be able to convert the speech sounds to transcribed language efficiently and accurately. Once converted to text, NLP-based approaches can be applied. In the present study, a spoken response to a task was captured by a microphone, converted to a digital signal, and then transcribed by an automatic speech recognition (ASR) system. The ASR produces a sequence of words along with ancillary information including the exact timing of the phrases, words and phonemes, and the location of pauses and disfluencies. From this time-

aligned recognition, the *dMSE* system can gather both the content of the spoken response and several aspects of quality, including the speech rate, intonation and phrasing, and gross and fine spectral characteristics of the speech. Both the linguistic content and non-lexical quality of speech are potentially useful in estimating the mental state of individuals.

An early step was to evaluate the performance of the Google speech recognition services on the speech that the *dMSE* system was capturing from the participants. Importantly, there was no risk of Personally Identifiable Information being sent to Google's API as all recordings were transcribed by humans and carefully screened prior to the automated transcription. Google's cloud speech API (<https://cloud.google.com/speech-to-text/>) can be used off-the-shelf without any particular modification. We generated Google ASR transcriptions of 193 spoken responses, comprising a total of 10,286 words. The Word Error Rate (WER) for this sample of speech was 35.5%. While this may lead one to assume that Google's language model was not sufficiently accurate for the analysis of the content or style of the responses, prior work has shown that lower WERs can still result in accurate NLP-based prediction models [29,30].

For most of the speech response tasks in *dMSE* the project, there was not enough speech data collected to train a task-specific language model. However, for our version of the Stroop attentional control task, an accurate language model could be deployed, which allowed a comparison of a task-specific ASR system (using a custom Stroop-specific language model; details provided in Refs. [31]) with Google's cloud speech ASR. The custom ASR system used the Kaldi speech recognition toolkit [32] to train a Deep Neural Network - Hidden Markov Model (DNN-HMM) acoustic model. The ASR comparison was performed using a set of 175 responses from non-patients in our study for which the project had human transcriptions. We measured recognition accuracy as WER. The WER was 6.26% for the custom ASR model and 17.9% for the Google cloud speech API version.

Although Google's cloud speech recognition API can be used off-the-shelf without any additional modifications, the WER is rather high on the Stroop task (while it has a lower WER than that of general speech, given the small number of words spoken and the importance of exact word matching in this task, the WER is much more significant here). Since we assume that Google's acoustic model is very well trained for many kinds of speech, the apparent reason for the accuracy deficit is Google's language model, which is designed for recognizing general English discourse on any topic and this language context does not play any important role in our version of the Stroop task. We have further examined this potential on the story recall task and also compared NLP analyses conducted on transcriptions created by humans versus those generated automatically with ASR (more details in next section and Ref. [29]). The findings were positive and we conclude that ASR is now at a point where it can be used to transcribe language for a range of behavioral and cognitive tasks and that the noise is sufficiently low that the added value makes ambulatory monitoring viable.

5.3. Semantic and structural features of story recall

The semantic content and syntactical structure of language have been shown to be effective indicators of severity of cortical disorders. For example, in assessing psychiatric patients' recall of previously heard stories, the story recall task (called the Logical Memory subtest) of the globally used Wechsler Memory Scale [21] puts emphasis on human raters counting the number of "story units" to measure the accuracy of recall. This assessment of human memory is one of the most important ways in which neurocognitive function is established. However, such methods rely on the human raters to find close to exact matches of words in the recalls, which can be time consuming and may miss subtle changes in wording or in recall organization. Automatically extracting features from transcribed speech holds the promise of overcoming this bottleneck (see Ref. [29] for further details).

There are a number of classes of variables that can encode characteristics of text that can allow automated measurement of the quality of a recall. One class of measures are considered *surface features* of language. Such measures include counts of words, noun phrases, and words related to cognitive and affective processes. A second class of measures looks at *structural features* of language, which analyzes features of how the language is put together (e.g., the manner of the language) and can include parses of the syntactic structure and the probabilities that word pairs would likely occur together in the English language (e.g., n-grams). Finally, *semantic features* are classes of variables that assess the meaning expressed in texts. These measures can include the choice of words as they relate to a specific topic as well as encoding the underlying meaning of words, sentences, or whole passages. Such measures often incorporate more general knowledge of the world or a domain and are able to measure meaning at a conceptual level rather than through counting direct overlap (e.g., detecting the similarity in meaning between words like “grocery store” vs. “market”). Small transformations in wording of concepts (e.g., “a man”, “a gentleman”, “a guy”) can be counted as equivalent concepts. However, changes in the amount of detail (e.g., “a brown clock”, “a clock”) would be considered as having a different number of concepts.

We derived a range of features from spoken story recall responses (described below) that are automated, fair and objective. We have applied NLP techniques to analyze the spoken responses to verbal recall tasks to evaluate to what extent they can provide new ways of testing and scoring this task, what additional information they provide regarding how individuals perform the task, and whether such techniques offer promise of being useful in ambulatory settings where frequent, remote and self-administration of tasks is necessary [29,33].

5.4. Predicting clinically important variables from story recall features

Ten stories for memory recall were developed that were structurally similar to the widely used traditional prose recall task in the Logical Memory subtest of the Wechsler Memory Scale-III [21]. The stories were between 61 and 82 words in length (average length = 72 words). Among the ten, five were informational (e.g., how to bathe a guinea pig) and five were narrative (e.g., a story about a lost balloon at a party). Each narrative had two characters, a setting, an action that happened in the setting causing a problem, and then a resolution. Each informational passage presented a purposeful sequence of actions or an explanation of a process to accomplish a goal. The participants were asked to verbally recall the story immediately after hearing it as well as after a one day delay of time, resulting in 750 recalls generated by the psychiatric patient class (575 immediate and 175 delayed) and 427 recalls generated by the non-patient class (216 immediate recalls and 211 delayed).

A set of computational methods were validated on a subset of the data. We used a set of 193 recall responses that each had been independently rated by seven judges on the quality of the content and the theme of the recall on a six-point scale. Overall, the raters correlated with each other at an average of $r = 0.83$ (ranging between 0.75 and 0.89). This indicated that human raters were able to employ the scale reliably. All responses were human-transcribed as well as automatically transcribed using ASR. The ASR engine used was Google's Speech API (as before, the human transcribers initially screened the recordings for Personally Identifiable Information before sending them to Google's Speech API). The ASR system generated a total WER of 20.90%. Two transcribers were used for the human transcriptions, resulting in a human transcriber WER of 7.2%.

For the story recall task, a ridge regression model was built to predict human ratings on recall accuracy. Ridge regression models are especially good at avoiding overfitting on a training set as the coefficients of input features are regularized and thus result in better predictions on new, unseen data. In a similar process to that employed in sections 5.1.2 and 5.1.3, we used 10-fold cross-validation to find the most suitable model parameters for training and testing. The best regularization strength was

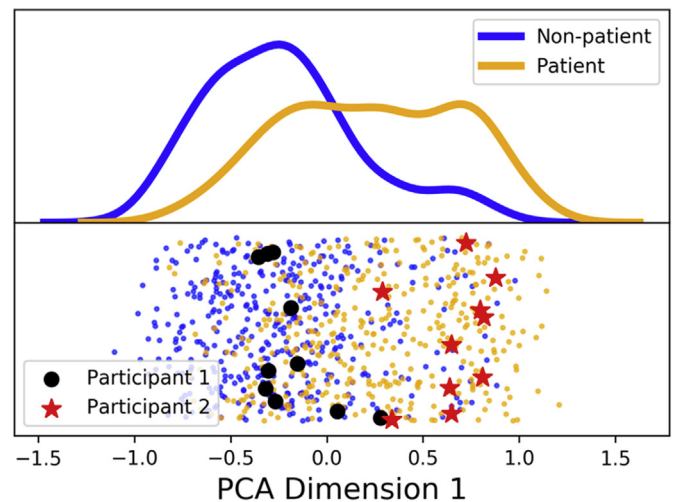


Fig. 4. Distribution plots of the one-dimensional PCA reduced classification model data with random jitter to make the point distribution visible. Colors reflect the class that each point belongs to (orange for the patient class on the right hand side and blue for the non-patient class on the left hand side). The sessions of two participants were chosen to illustrate not only how they compare to the greater distributions but also how varied individual performance on verbal recall can be. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

found to be 0.01. The ridge regression model was built using the three best language features that accounted for different aspects of speech that human raters employ. These three features were the number of unique words spoken in a recall (mean regression coefficient = 2.47), the number of common words between the original story and the recall (mean regression coefficient = 3.14), and the word mover's distance of the word2vec vectors of the original story and the recall (mean regression coefficient = -2.71 [34]). The regression model correlated with human ratings with an average over the 10 test folds of $r = 0.88$. When ASR transcriptions were used instead of human transcriptions, the average correlation was $r = 0.86$.

Additionally, an ensemble classification model which comprised three logistic regression classifiers (one for immediate recalls, one for delayed recalls, and one for the change between immediate and delayed recalls) was created in order to predict whether a participant was in the psychiatric patient class ($N = 105$) or the non-patient class ($N = 120$). Logistic regression models are commonly used to convert standard regression problems to classification problems. Thus each individual model contained in the ensemble returned a tuple of the probability that the current participant belonged to each class. The model then generated a weighted combination of each probability tuple, resulting in a final class membership probability estimate. Leave-one-out cross-validation was performed across data from individual participants, training on all participants but one, and testing on the one left out. Using the same features as the regression model, the ensemble classifier predicted with an accuracy of 76% on human transcriptions and 74% on ASR transcriptions. This is a significant result given the diversity in the patient class as well as the fact that diagnoses are typically based on an entire battery of tests and other contextual information.

Principal Component Analysis (PCA) was used to project the 8-dimensional feature set used in the classification model to 1 dimension for visualization purposes. PCA is a tool commonly used in exploratory data analysis, as it allows for the visualization of high dimensional spaces. Fig. 4 shows the distributions of the two classes' reduced data. The plots show two clear distributions with some overlap in the middle. Specifically, the non-patient class is more consistent in their verbal recall performance, with a high peak and a lower spread. On the other hand, as expected, psychiatric patients vary more in their performance, and tend

to have lower scores. Two participants from the patient class were chosen to showcase their performance in relation to the entire dataset. Participant 2 is in line with the distribution of the patient class, whereas Participant 1 falls right in the peak of the non-patient class. From these plots, we can also see how a participant's performance on a particular day compares to populations as a whole and to their own prior sessions. Unsurprisingly, the classifier correctly predicted Participant 2 as a patient and incorrectly predicted Participant 1 as a non-patient.

Overall, these results indicate that we can automatically derive a range of semantic and surface level features from spoken recalls, and that these features can be harnessed to accurately predict the ratings of expert humans as well as provide accurate classifications of psychiatric patient and non-patient participants. If we examine the results across the different measures computed in this paper (see Table 1), we see a strong pattern of machine learning-based approaches being successful at characterizing performance across a range of tasks using speech and language features.

6. Future directions

Finally, we discuss whether the carefully crafted features (e.g., arousal predictions, semantic similarity measurements, self-report scales) can be configured into a functional system with multiple parameters to monitor a psychiatric patient's clinical state with sufficient accuracy as to be useful. One central benefit of the dMSE application is the ability to consistently and reliably track patients' performances on a longitudinal

basis. We have previously demonstrated the complete pipeline of automating single components of assessment such that actionable inferences about cognitive state may be taken [29,33].

However, in the same way that expert clinicians do not make such critical decisions based upon one observation or type of data, so too is it necessary for machines to emulate the process of using multiple data channels, including the temporal aspect of data (i.e., time is a core component in most diagnostic categories in psychiatry such as the Diagnostic and Statistical Manual of Mental Disorders-5; [35,36]). Indeed, as it is not beneficial to merely compare the performance of one patient with others on a single task, the application allows for the tracking and monitoring of the fluctuations of one patient's performance on several tasks over time. This enables the necessary examination of whether the multi-dimensional data are equally relevant, namely whether various data types collected at a specific time point are similarly valid, a process that requires complex weighting. This in turn allows clinicians to have access to data streams to monitor the highs and lows of mental state as a whole. Fig. 5 displays real results from 2 participants (Participant A from the patient class and Participant B from the non-patient class) who interacted with the dMSE application over a period of 11 and 7 sessions, respectively. We see more consistency in the performance of the non-patient and more fluctuations in the patient. The fluctuations and correlations (or lack thereof) of performance on various tasks can be subjected to machine learning so as to produce critical and objective health summaries for clinicians that are displayed in an easy to interpret dashboard-like format.

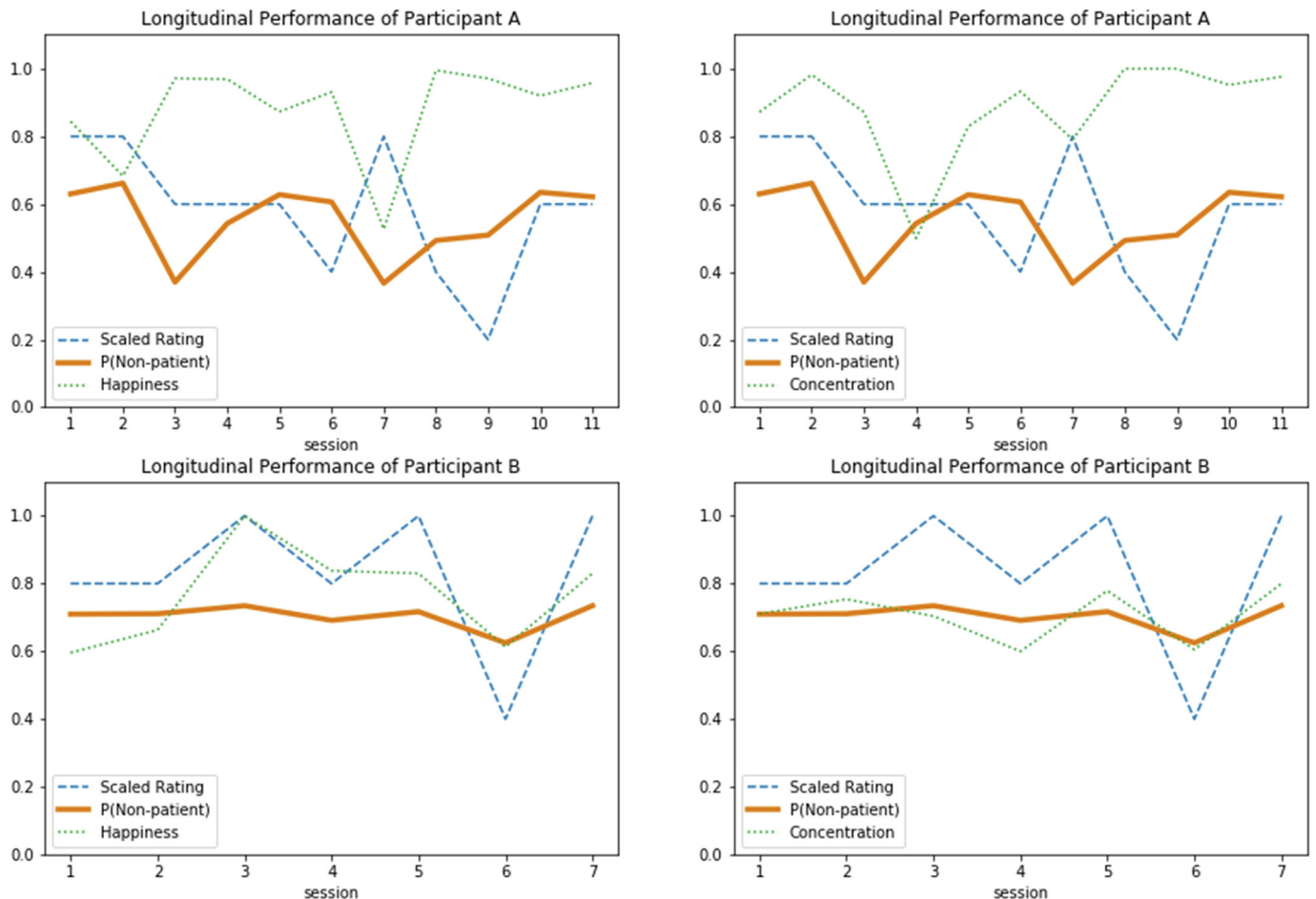


Fig. 5. Plots showing the mutual variation of performance on the immediate recall task over sessions (note: 1–6 rating scaled to 0–1), response to a slider question which was one of the following: a) how happy the participant was (plots on left hand side) and b) how concentrated the participant was (plots on right hand side), and the calculated probability that the participant was a non-patient participant on the given session, evaluated on their immediate recall content, delayed recall content, and change between the immediate and delayed recall.

To date, this automated combination of multi-modal data streams over time has yet to be addressed by the field. This proof of concept has much promise, but whether this translates into medical gain is still an open question, as discussed further in the next section.

7. Discussion

The present study has shown that automated machine learning approaches to neuropsychological assessment can be used for understanding clinical judgment. Good quality data can be collected by a smart device application that users self-administer. Our framework allows us to address the three main issues previously raised: the number of clinicians needed to monitor the vast amount of patients with mental disorders can potentially be reduced, the ability to monitor patients frequently and consistently, while taking into account an extensive amount of past data is enabled, and the vast amount of data collected paired with knowledge of life events allows for machine learning analysis of mental state both at an individual level and at a general population level. Additionally, this framework lays the groundwork for much needed increased equity in mental health services as it can provide patients the assistance they need regardless of any personal characteristics such as gender, age, ethnicity, location, or socioeconomic status (see Ref. [25]). Clearly, full-scale randomized clinical trials are needed to refine and validate these approaches. Such randomized control trials will further need to address the legal and technical agile e-health frameworks needed for successful development, deployment, and data analytic methods to be used for remote monitoring.

A number of challenges remain to be solved in furthering this approach (for issues related to tracking language in real time, see Ref. [37]). Notably, any machine learning model will need exposure to vast quantities of patient data in order to learn subtle trends and differences so as to be able to weigh the different data channels accordingly based upon clinical condition and various baseline demographics (e.g., medical history, gender, age, and so on), and thus learn the nuances that are both part of normal variance as well as those that are indicative or predictive of significant clinical change. The machine learning methods used are bounded by small datasets and the requirement that they must be explainable. While advanced approaches such as long short term memory networks or recurrent neural networks for temporal data hold great prediction power in many fields, they cannot be harnessed with the current limited datasets. Longitudinal analyses require hundreds of measures in order to see change, not just a handful of snapshots over a few weeks. Thus, more measurements (over both types and time) are needed.

To collect the necessary quantities of data for machine learning approaches to be harnessed will require consortium efforts in order to develop the clinical decision support systems. To achieve this, it remains essential to establish how this vast quantity of data can be combined into a meaningful whole, notably with other external measures. Much future work is necessary to train machines to emulate the relative weightings of the cognitive and emotional features expressed in individual patients and present the findings in an understandable way to clinicians. Additionally, these neuropsychological measurements must be contextual (i.e., incorporate information about a person's life on that day), and thus this new approach to psychometrics will also require the collection of other related context variables concurrently.

Indeed, our current work is focused on how to best combine metrics from audio and semantic analyses, as well as data from other behavioral streams, to predict clinically important variables. This is an extremely difficult, but critically important, task. It is difficult because the clinical end points themselves are not perfect (i.e., current gold standard labels as evaluated by clinicians are imperfect and participants' self-reported states are not necessarily related to the actual diagnosis and prognosis). Nonetheless, we regard the approach illustrated above that enables the calculation of the probability of illness to be in line with current medical notions about the continuum of diagnostic entities. Furthermore, the

incorporation of temporal data, such that the dynamic nature of cognition and mental state can be captured and analyzed, provides the much needed framework for the empirical investigations necessary to deliver the precision promises of personalized medicine [38]. The future goal of a 'telemental health' monitoring tool needs to be supported by a randomized clinical trial and implementation of a core e-health system that tracks the clinical state of psychiatric outpatients and, when appropriate, alerts clinical staff to contact that patient.

Finally, although many mental health centered mobile phone applications are downloaded and used by consumers, most have been found to be quickly abandoned [23]. More studies are needed to ascertain which qualities of applications retain the most users over time. With data science we can find the useful measures and reduce the amount of testing needed. Therefore in the future, the full suite of testing can be reduced and a resulting mobile phone application can be shortened and thus more tolerable to users.

8. Conclusion

In this study, we employed smart devices such that frequent and remote monitoring was practically viable. We see this as a first step towards individualized, longitudinal, illness specific, and contextually based machine learning methods. We leveraged machine learning methods for automated analysis and scoring, which is necessary with the amount of information collected, and afforded the analysis of content and pattern in speech, thus increasing both objectivity and sensitivity. Thus far, most AI-based modeling of psychiatric cognitive biomarkers has relied on small laboratory-based datasets. This approach provides a pathway to collecting the size of data needed for truly generalizable AI methods and to be able to better measure and understand population and individual variability for better predictive models. It furthermore allows for the examination of patient data in a very nuanced manner so as to develop individualized personalized baselines to detect clinically meaningful change. However, in order to define 'meaningful' it is necessary to calibrate by developing models that weigh different behavioral streams appropriately.

Given that the gold standard external measures in psychiatry are flawed, it is difficult to establish exactly what the objective behavioral metrics should be anchored to and thus how to calibrate change, but this approach enables more nuanced design of new psychometrics and tests (e.g., Refs. [39]). The present work provides a pathway to individualized models and shows how they can be realized with a machine learning-based approach by analyzing speech and language, as well as self-reports. Our analyses show strong promise, but will require substantial and highly complex analysis and additional research to properly understand this issue. For such automated methods to be used in medicine, the methods must be explainable, transparent, and generalizable. We argue that there is an urgency for the development of a framework with which to evaluate such complex methodologies [25]. Importantly, the knowledge, results, and tools developed from the current project form the much needed foundation for the development of clinical decision support systems, which are a major part of future medical systems that will afford the user (patient, clinicians, and family members) the unprecedented opportunity to better manage and regulate mental health based upon daily performance and health metrics.

Declarations of competing interest

None.

Acknowledgements

This project was funded by grant 231395 (2014–2017) from the Research Council of Norway awarded to Brita Elvevåg.

References

- [1] NIMH - mental illness. <https://www.nimh.nih.gov/health/statistics/mental-illness.shtml>. [Accessed 13 July 2020].
- [2] Cohen AS, Fedechko T, Schwartz E, Le TP, Foltz PW, Bernstein J, Cheng J, Rosenfeld E, Elvevåg B. Psychiatric risk assessment from the clinician's perspective: lessons for the future. *Community Ment Health J* 2019;55:1165–72. <https://doi.org/10.1007/s10597-019-00411-x>.
- [3] Gonda X, Fountoulakis KN, Kaprinis G, Rihmer Z. Prediction and prevention of suicide in patients with unipolar depression and anxiety. *Ann Gen Psychiatr* 2007;6: 23.
- [4] Lai X, Liu Q, Wei X, Wang W, Zhou G, Han G. A survey of body sensor networks. *Sensors* 2013;13:5406–47.
- [5] Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. *Psychiatr Rehabil J* 2015;38(3):218–26. <https://doi.org/10.1037/prj0000130>.
- [6] Cheng J, Bernstein J, Rosenfeld E, Foltz PW, Cohen AS, Holmlund TB, Elvevåg B. Modeling self-reported and observed affect from speech. In: *Proceedings of interspeech, Hyderabad, India*; 2018. p. 3653–7.
- [7] Cohen AS, Fedechko TL, Schwartz EK, Le TP, Foltz PW, Bernstein J, Cheng J, Holmlund TB, Elvevåg B. Ambulatory vocal acoustics, temporal dynamics and serious mental illness. *J Abnorm Psychol* 2019;128:97–105. <https://doi.org/10.1037/abn0000397>.
- [8] Bedi G, Carrillo F, Cecchi GA, Fernández-Slezak D, Sigman M, Mota NB, Ribeiro S, Javitt DC, Copelli M, Corcoran CM. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophr* 2015;1:15030.
- [9] Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, Bearden CE, Cecchi G. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatr* 2018;17(1):67–75. <https://doi.org/10.1002/wps.20491>.
- [10] Lehr M, Prud'hommeaux E, Shafran I, Roark B. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In: *Proceedings of the 13th Annual Conference of the International Speech Communication Association*; 2012. p. 2.
- [11] Lehr M, Shafran I, Prud'hommeaux E, Roark B. Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment. In: *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics. Human Language Technologies*; 2013. p. 211–20.
- [12] Dagum P. Digital biomarkers of cognitive function. *NPJ Digital Med.* 2018;1(1):1–3.
- [13] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003; 3(Jan):993–1022.
- [14] Landauer TK, Foltz PW, Laham D. Introduction to latent semantic analysis. *Discourse Process* 1998;25:259–84.
- [15] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: *arXiv preprint arXiv:1301.3781, in Workshop Proc. ICLR*; 2013.
- [16] Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr Res* 2007;93:304–16.
- [17] Elvevåg B, Foltz PW, Rosenstein M, DeLisi L. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *J Neurolinguistics* 2010;23:270–84.
- [18] Rosenstein M, Foltz PW, DeLisi LE, Elvevåg B. Language as a biomarker in those at high-risk for psychosis. *Schizophr Res* 2015;165:249–50.
- [19] Iter D, Yoon J, Jurafsky D. Automatic detection of incoherent speech for diagnosing schizophrenia. In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*; 2018. p. 136–46.
- [20] Lezak MD, Howieson DB, Bigler ED, Tranel D. *Neuropsychological assessment*. fifth ed. Oxford University Press; 2012.
- [21] Wechsler D. *Wechsler memory scale*. In: *WMS-III: administration and scoring manual*. third ed. San Antonio, TX: The Psychological Corporation; 1997.
- [22] Holmlund TB, Foltz PW, Cohen AS, Johansen HD, Sigurdson R, Fugelli P, Bergsager D, Cheng J, Bernstein J, Rosenfeld E, Elvevåg B. Moving psychological assessment out of the controlled laboratory setting and into the hands of the individual: practical challenges. *Psychol Assess* 2019;31(3):292–303. <https://doi.org/10.1037/pas0000647>.
- [23] Baumeister A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. *J Med Internet Res* 2019;21(9):e14567. <https://doi.org/10.2196/14567>.
- [24] Nielsen J. *Usability engineering*. Boston: Academic Press; 1993. <https://doi.org/10.1016/C2009-0-21512-1>.
- [25] Chandler C, Foltz PW, Elvevåg B. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophr Bull* 2020;46:11–4. <https://doi.org/10.1093/schbul/sbz105>.
- [26] Russell J. A circumplex model of affect. *J Pers Soc Psychol* 1980;39(6):1161–78.
- [27] Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, Chetouani M, Wengler F, Eyben F, Marchi E, Mortillaro M, Salamin H, Polychroniou A, Valente F, Kim S. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: *Interspeech*; 2013. p. 148–52.
- [28] Chandler C, Foltz PW, Cheng J, Cohen AS, Holmlund TB, Elvevåg B. Predicting self-reported affect from speech acoustics and language. In: *Proceedings of the LREC 2020 workshop on: resources and processing of linguistic, para-linguistic and extra-linguistic data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID-3)*; 2020. p. 9–14.
- [29] Chandler C, Foltz PW, Cheng J, Bernstein JC, Rosenfeld EP, Cohen AS, Holmlund TB, Elvevåg B. Overcoming the bottleneck in traditional assessments of verbal memory: modeling human ratings and classifying clinical group membership. In: *Proceedings of the sixth workshop on computational linguistics and clinical psychology*. Minneapolis, Minnesota, USA; 2019. June (pp. 137–147). Available at: <https://aclweb.org/anthology/volumes/proceedings-of-the-sixth-workshop-on-computational-linguistics-and-clinical-psychology/>.
- [30] Foltz PW, Laham RD, Derr M. Automated speech recognition for modeling team performance. *Proceedings of the 47th annual human factors and ergonomic society meeting*. Santa Monica, CA: Human Factors and Ergonomic Society; 2003.
- [31] Holmlund TB, Cheng J, Foltz PW, Cohen AS, Bernstein J, Rosenfeld E, Laeng B, Elvevåg B. Using automated speech processing for repeated measurements of attentional bias and control. Manuscript submitted for publication; 2020.
- [32] Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, Silovsky J, Stemmer G, Vesely K. The Kaldi speech recognition toolkit. *IEEE 2011 workshop on automatic speech recognition and understanding*. In: *IEEE catalog no.: CFP11SRQ-USB*; 2011.
- [33] Holmlund TB, Chandler C, Foltz PW, Cohen AS, Cheng J, Bernstein JC, Rosenfeld EP, Elvevåg B. Applying speech technologies to assess verbal memory. *NPJ Digital Med.* 2020;3:33. <https://doi.org/10.1038/s41746-020-0241-7>.
- [34] Kusner M, Sun Y, Kolkin NI, Weinberger KQ. From word embeddings to document distances. In: *Proceedings of the 32nd international conference on machine learning*; 2015. p. 957–66.
- [35] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. fifth ed. Arlington, VA: American Psychiatric Publishing; 2013.
- [36] Weinberger DR, Goldberg TE. RDOCS redux. *World Psychiatr* 2014;13:1.
- [37] Holmlund TB, Fedechko TL, Elvevåg B, Cohen AS. Chapter 28: tracking language in real time in psychosis. In: *Badcock JC, Paulik-White G, editors. A clinical introduction to psychosis: foundations for clinical and neuropsychologists*. Elsevier; 2020b.
- [38] Insel TR. Digital phenotyping: technology for a new science of behavior. *J Am Med Assoc* 2017;318(13):1215–6. <https://doi.org/10.1001/jama.2017.11295>.
- [39] Chandler C, Holmlund TB, Foltz PW, Cohen A, Elvevåg B. Deciding on the best verbal memory test for research purposes: lessons from machine learning. Manuscript submitted for publication; 2020.