BMJ Open

Protocol for Rhapsody: a longitudinal observational study examining the feasibility of speech phenotyping for remote assessment of neurodegenerative and psychiatric disorders

Elliot Hampsey [iD],[1] Marton Meszaros,[2] Caroline Skirrow,[2] Rebecca Strawbridge,[1] Rosie H Taylor,[1] Lazarus Chok,[2] Dag Aarsland,[1] Ammar Al-Chalabi [iD],[1] Ray Chaudhuri,[1,3] Jack Weston,[2] Emil Fristed,[2] Aleksandra Podlewska,[1,3] Olabisi Awogbemila,[3] Allan H Young[1]

[1]Institute of Psychiatry, Psychology, & Neuroscience, King's College London, London, UK
[2]Novoic Limited, London, UK
[3]Parkinson's Foundation Centre of Excellence, King's College Hospital NHS Foundation Trust, London, UK

Correspondence to
Elliot Hampsey;
elliot.hampsey@kcl.ac.uk

ABSTRACT

Introduction Neurodegenerative and psychiatric disorders (NPDs) confer a huge health burden, which is set to increase as populations age. New, remotely delivered diagnostic assessments that can detect early stage NPDs by profiling speech could enable earlier intervention and fewer missed diagnoses. The feasibility of collecting speech data remotely in those with NPDs should be established.

Methods and analysis The present study will assess the feasibility of obtaining speech data, collected remotely using a smartphone app, from individuals across three NPD cohorts: neurodegenerative cognitive diseases (n=50), other neurodegenerative diseases (n=50) and affective disorders (n=50), in addition to matched controls (n=75). Participants will complete audio-recorded speech tasks and both general and cohort-specific symptom scales. The battery of speech tasks will serve several purposes, such as measuring various elements of executive control (eg, attention and short-term memory), as well as measures of voice quality. Participants will then remotely self-administer speech tasks and follow-up symptom scales over a 4-week period. The primary objective is to assess the feasibility of remote collection of continuous narrative speech across a wide range of NPDs using self-administered speech tasks. Additionally, the study evaluates if acoustic and linguistic patterns can predict diagnostic group, as measured by the sensitivity, specificity, Cohen's kappa and area under the receiver operating characteristic curve of the binary classifiers distinguishing each diagnostic group from each other. Acoustic features analysed include mel-frequency cepstrum coefficients, formant frequencies, intensity and loudness, whereas text-based features such as number of words, noun and pronoun rate and idea density will also be used.

Ethics and dissemination The study received ethical approval from the Health Research Authority and Health and Care Research Wales (REC reference: 21/PR/0070). Results will be disseminated through open access publication in academic journals, relevant conferences and other publicly accessible channels. Results will be made available to participants on request.

Trial registration number NCT04939818.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Remote, partially self-administered speech-battery enabling use by those with difficulty accessing in-person healthcare.
⇒ Speech battery and study design uniquely allows for investigation of the same tasks across a wide number of indications.
⇒ Repeated assessment allows for examination of test reliability, and intrasubject and intersubject variability across key clinical groups.
⇒ Limited sample size.
⇒ Short follow-up period limits scope for longitudinal disease monitoring.

INTRODUCTION

Healthcare costs of neurological and psychiatric disorders

Neurological and psychiatric disorders (NPDs) affect 20% of older people,[1] at an estimated cost to the UK of >£68 billion annually.[2–5] With over 65s estimated to make up over a quarter of the UK population by 2050,[6] the burden of NPDs on society will increase dramatically. Alzheimer's disease accounts for 50%–70% of dementia cases, with an estimated £34.7 billion annual healthcare cost in the UK.[2] Dementia with Lewy bodies (DLB) accounts for 15%–20% of all dementia cases,[7–9] affecting >100 000 people in the UK. DLB is a much more rapidly progressing disease, with a median survival of 3.72 years postdiagnosis.[10]

While less prevalent than dementias, Parkinson's disease (PD) and motor neuron disease (MND) confer extensive costs due to high morbidity and mortality. PD affects >160 000 people in the UK,[11] costing the National Health Service (NHS) >£2.2 billion annually,[12] and over 80% may progress to dementia in the long term while the rate is higher in those with the cholinergic subtype.[13] MND affects about 5000 people in the UK,[14] with patients suffering severe impairments in addition to considerable lifespan reduction.[15]

Although affective disorders typically onset during one's late 20s,[16] they are also marked by recurrence throughout the lifespan.[17] WHO reports major depressive disorder (MDD) to have the highest disability burden of all conditions internationally,[18] incurring a care cost in the UK of >£23.8 billion annually.[4] Although less prevalent (1%–4.5% lifetime prevalence),[19 20] the burden of bipolar disorder remains considerable, costing the NHS £1.6 billion annually.[5] People with bipolar disorder have an average of 10–20 years reduced lifespan compared with the general population, due in part to the 20% of people with bipolar disorder who die by suicide.[21]

Diagnostic challenges

NPDs are complicated by challenges to accurate and early diagnosis. Estimates suggest that approximately 50% of depressive episodes go undetected,[22] while >60% of MDD and >90% of patients with bipolar disorder may be misdiagnosed.[23] Similarly, ≥50% of those who show evidence of DLB post mortem were not diagnosed with the disease during life,[24] with one study finding that 39.5% of those diagnosed clinically as free from Alzheimer's disease (AD) met the minimum histopathological threshold.[25] High rates of diagnostic errors in bipolar disorder lead to the delaying of a correct diagnosis by 5.7 years,[26] with many waiting for more than a decade.[27] Ostensibly, mis/delayed diagnosis can impair treatment provision, exacerbate course of illness, limit treatment options and reduce quality of life.[28]

Overlap between the clinical presentation of NPDs makes diagnosis challenging. For example, cognitive impairment, the hallmark symptom of prodromal AD, is also seen in all patients with DLB[29] as well as in >80% of patients with PD across all motor stages of the condition[30] and is common in people with depression. The cognitive assessments used to screen for dementia are inadequate, leaving 32% of patients with early stage AD[31] and 50% of patients with DLB[32] undiagnosed. Diagnosing other neurodegenerative disorders such as PD is particularly challenging, as there are no definitive tests.[33] Physicians instead rely on often error-prone clinical judgement,[34] with an estimated 20% of those with PD who have come to medical attention going undiagnosed.[35]

Even when diagnoses are made correctly, this can be years after symptoms have begun. Failure to intervene early is associated with a more severe impact on quality of life, such as memory, motor and psychiatric disturbances.[36–39]

Digital and remote assessment strategies

Uptake of digital and remote assessments methods has accelerated during the SARS-CoV-2 global pandemic, both in research and clinical practice.[40 41] Digital health technologies hold promise for reducing burden and improving access for those who travel to medical centres would be laborious, stressful or financially challenging.[42]

Digital technology can help to enhance certain aspects of assessment practices. Higher frequency assessment, with automated administration and completed remotely, can allow for more detailed assessment of behaviour and cognition over time. Higher frequency remote assessments have been successfully deployed in mood disorders and other psychiatric conditions,[43 44] and in neurological conditions, including mild cognitive impairment (MCI), mild AD[45] and PD.[46 47] These studies also typically report moderate to high levels of adherence to remote assessment and good acceptability of this method of assessment. Furthermore, digital speech capture can help to enrich analyses with more advanced text similarity analyses[48] and automated extraction of language features commonly evaluated during connected speech,[49] and furthermore incorporate vocal and acoustic features, which are sensitive to clinical status.[50–53]

While holding promise for improving convenience and access, there is concern regarding whether digital assessment is particularly challenging for certain populations, for example, in those with dementia or cognitive impairment.[54] Data integrity may be compromised by participants misunderstanding instructions (which are not or cannot be prompted for correction), or due to environmental/contextual limitations, such as distractions or ambient noise.[47] Additional care is required in the analysis of high-frequency assessments, which needs to take into account the autocorrelation of repeated observations within individuals, and potential variation within individuals over time.[47]

Novel strategies to improve illness recognition

Artificial intelligence-based techniques have shown efficacy in analysing medical data in diagnosis and could potentially detect and translate subtle, early changes in speech into predictive diagnostic models.[55] This would assist care providers and their patients, who are likely to benefit from objective profiling biomarkers that are capable of disambiguating clinically similar NPDs. Speech/Language alterations in NPDs are promising universal biomarkers as they reflect subtle cognitive, motor and mood changes.[56–59] As speech is among the first modalities to be affected in NPDs,[56 57 60–62] developing speech/language-based profiling biomarkers will also aid in early diagnosis of NPDs. Herewith, we describe the research protocol of the 'Rhapsody' study, funded by the National Institute for Health Research, which aims to assess the feasibility of recording and detecting changes in patterns of speech across a range of high burden NPDs.

Study objectives

The primary objective of the study is to evaluate the feasibility of eliciting continuous narrative speech to collect speech data remotely in three groups of NPDs: neurodegenerative cognitive disorders, other neurodegenerative disorders (PD and MND) and affective disorders. We hypothesise that due to the simple speech-based interface, participants will be able to engage with and provide speech data during virtual study visits and remote assessments. Basic feasibility will be assessed via the average length of speech elicitation for each speech task (in seconds) during the first week of self-assessment. Additionally, we will examine the proportion of participants in whom narrative speech lasting at least 20s is elicited within each group.

The study also has secondary objectives concerning the feasibility of using speech tasks to collect speech data in the remote setting, such as evaluating reliability of repeated assessments across related (comparing virtual visits vs fully remote assessments) or repeated tasks (parallel variants administered across days), examining intra-individual and inter-individual variance. We will also evaluate adherence to daily remote assessments, and participant-rated usability of the remotely administered application, measured via a brief usability questionnaire completed on the app.

Furthermore, the study will examine whether acoustic and linguistic patterns can be used to distinguish from group-specific control participants, and from other clinical indications, and whether this is impacted by relevant contextual or disease information covariates.

METHODS AND ANALYSIS

Design

The RHAPSODY study is an observational longitudinal study during which participants remotely complete a set of tasks designed to elicit speech at baseline and during the 4-week follow-up period. Participant-rated questionnaires of clinical symptoms are completed at baseline and repeated at follow-up weeks 2 and 4.

Participants

Potential participants will be identified through local clinical services, cohort-specific internal databases, advertisements on a variety of social media platforms and dedicated websites for research participation. Participant recruitment began on 15 July 2021 and will continue to 30 June 2022.

Participants will meet group-specific eligibility criteria for neurodegenerative cognitive disorders (n=50), or other neurodegenerative motor disorders (n=50), or affective disorders (n=50), in addition to matched controls (n=75). The sample size was based on a review of feasibility study group sizes.[63] Individuals must be native English speakers, have access to a smartphone and personal computer/laptop which can connect to the internet and is capable of audio recording as well as

running an operating system of macOS X with macOS 10.9 or later; or Windows 7 or above. Exclusion criteria include: current substance use disorder, stroke within the last 2 years, transient ischaemic attack or unexplained loss of consciousness within the last 12 months or current risk of suicide. The additional group-specific eligibility criteria are as follows:

Group 1: neurodegenerative cognitive disorders (n=50): participants must be aged 50–85 years, with a score of 20–30 on The Telephone Interview for Cognitive Status-modified (TICS-M)[64] and one of the following clinical diagnoses (within the last 5 years):

A. MCI due to AD or mild Alzheimer's dementia as per the National Institute of Aging—Alzheimer's Association core clinical criteria (2011).[65]

B. Probable or possible DLB as per The Dementia with Lewy Bodies Consortium definition.[66]

C. Cognitive impairment (not due to AD/DLB), either diagnosed with behavioural variant frontotemporal dementia,[67] semantic variant primary progressive aphasia or non-fluent variant primary progressive aphasia,[68] vascular dementia[69] or unspecified MCI.

Group 2: other neurodegenerative disorders (n=50):

A. Participants with MND must be aged 18–85 years and score stage 3 or less on the King's amyotrophic lateral sclerosis staging system.[70]

B. Participants with PD must be within 5 years of a diagnosis of idiopathic PD according to the UK Brain Bank Criteria, aged between 30 and 85 years, and have a Hoehn and Yahr stage[71] of ≤2.

Group 3: affective disorders (n=50): participants with affective disorders must be aged 18–85 years, diagnosed with either major depression or bipolar disorder, in a current depressive episode according to the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) criteria as assessed by the Mini-International Neuropsychiatric Interview (M.I.N.I.) V.5.0,[72] of at least moderate severity as assessed on the Clinical Global Impressions Scale.[73]

Group 4: unaffected controls (n=75): approximately 25 unaffected control participants will be recruited for each cohort, matched for gender, age and education levels. Participants will be in otherwise good health; they may experience mild disorders (of metabolic, respiratory, immunological, cardiologic and metabolic origin) that do not impair daily functioning.

Data collection procedures

Study procedures are summarised in figure 1. Potential participants will be provided with a patient information sheet by the research team, to be followed by the study visit ≥24 hours later. Participants providing consent will complete the study visit via video conferencing software. Participants will first complete the screening and baseline assessment phases, which are completed in a single sitting.

During the screening phase, relevant demographic information, medical history and information on current
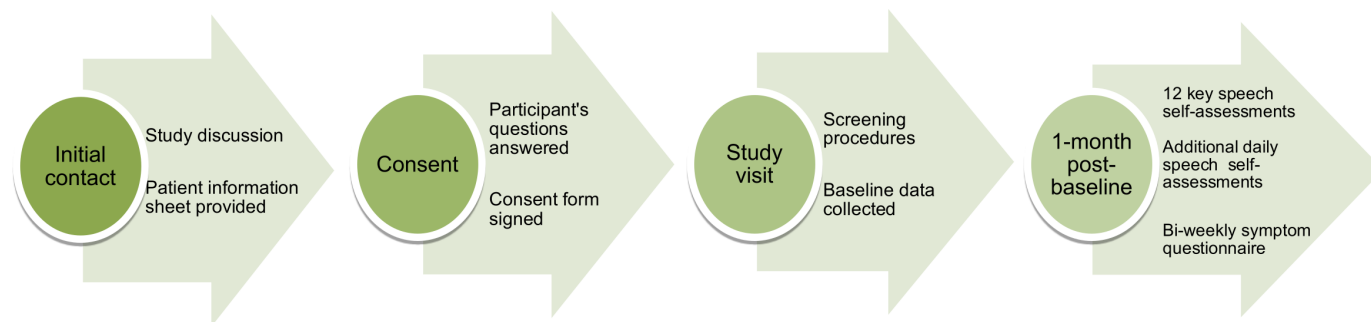
Figure 1 Flow diagram of the participant journey in the RHAPSODY study.

medications are obtained, as well as ruling out visual or hearing difficulties that would serve as exclusion criteria. Some cohort-specific assessments (table 1) will also be used to determine eligibility. Participant eligibility is then evaluated against inclusion/exclusion criteria, with ineligible individuals exiting the study at this stage. Those who are eligible to continue will then complete the baseline data collection phase during the same teleconference.

A. (All groups): the Patient Health Questionnaire-9 item[74] is a short, self-report tool which uses a 4-point Likert scale to assess nine items relating to depressive symptomology.

B. (Group 3): M.I.N.I. is a structured interview for mental health diagnoses according to the DSM-IV criteria. The suicidality scale will be completed by any participant scoring ≥10 on the Patient Health Questionnaire-9 item.

C. (Group 3): the Clinical Global Impressions Scale is a 3-item, researcher-administered tool that provides an overall 'global' judgement concerning the participant's severity of depressive illness.

D. (Group 1): TICS-M is a short, researcher-administered assessment of cognitive functioning validated for use over the phone.

E. (Group 3): the Inventory of Depressive Symptomatology (30-item Clinician Rated)[75] is a researcher-rated instrument that provides detailed assessment of depression severity over the previous 7 days.

Verbal cognitive and speech tasks

All participants will complete the same sequence of verbal cognitive and speech assessments. An overview of speech tasks is provided in table 2. Some of the tasks are optimised to elicit continuous narrative speech, such as reading text and recalling a presented story. Others elicit specific categories of words, or repetition of sentences or sounds.

These tasks combined allow for evaluation of a range of cognitive, linguistic and acoustic features elicited during spoken responses. Differential impairments are expected across groups and tasks, for example, while neurodegenerative cognitive disorders such as AD are

Study procedure	Screening/Baseline	Week 1	Week 2	Week 3	Week 4
Medical history, concomitant medication overview and demographics	X				
Telephone Interview for Cognitive Status-modified (group 1 only)	X				
The Mini-International Neuropsychiatric Interview* (psychiatric diagnostic evaluation) (group 3 only)	X				
Clinical Global Impression (group 3 only)	X				
Speech/Language assessments (see table 2)	X	X	X	X	X
Inventory of Depressive Symptomatology—Clinician Rated (group 3 only)	X				
Patient Health Questionnaire-9 item†	X		X		X
Generalised Anxiety Disorder Assessment-7 item†	X		X		X
Maudsley 3-Item Depression Visual Analogue Scale†	X		X		X
Altman Self-Rating Mania Scale† (groups 2 and 3 only)	X		X		X

Table 1 Screening tools and questionnaires completed at baseline and follow-up

(above the table) Self-assessments header spanning Week 1–Week 4.

*Suicide questionnaire administered if participant scores >10 on the Patient Health Questionnaire-9 item.
†Patient-rated.

Table 2 Overview of speech tasks used in RHAPSODY (main and/or remote assessments)

Speech and cognitive tasks	Task details	Task characteristics and selected evidence from prior research
Count 1–10 or 10–1	Participants are recorded counting from 1 to 10 or from 10 to 1.	The counting task was used as a brief familiarisation task for participants.
The Automatic Story Recall Task (ASRT) with immediate and delayed recall	The ASRT has 36 parallel story variants (18 'short' and 18 'long' stories), presented at a steady reading rate and matched for linguistic and discourse measures, including number of words, number of sentences, number of dependent clauses, mean sentence length and ratio of dependent clauses to t-units.[45] Prerecorded stories are played to the participants, who are be asked to recall the stories in as much detail as possible immediately and again after a delay.	Story recall is often used to evaluate verbal episodic memory. Prior meta-analytic evidence suggests that lesser impairment would be expected in story recall performance for individuals with DLB than those with AD-MCI.[91] A smaller head-to-head research study indicates that lesser impairment may be expected for individuals with PD than AD-MCI.[92] Meta-analytic evidence suggests that the task may be sensitive to mood disorders, with individuals with a history of depression having task performance deficits in story recall, with large effect sizes reported.[93]
Category fluency task	Participants will be given a category and required to give as many examples of that category as they can in 60s. Participants will complete parallel variants of this task, commonly implemented in the research literature, including animals, vegetables, fruits.	Category fluency evaluates semantic memory function and taps into attentional and executive functions. Impairments in category fluency have been noted in a rage of the clinical indications, including MCI/AD and DLB,[94] PD,[93] MND,[95] MDD[93] and BD,[96] but with different reported effect sizes.
Letter fluency task	Participants will be given a letter of the alphabet and required to give as many examples of words beginning with that letter as they can in 60s. Letters prompted include F, A and S.	Letter fluency tasks evaluate lexical access through the phonological route and are believed to more strongly depend on executive functions than category fluency measures. Prior research shows that task performance is impaired in adults with MCI,[97] PD,[94] MND,[95] following MDD[93] and in BD,[96] but with different reported effect sizes.
Action fluency task	Participants are tasked with naming as many of examples of things that 'people do', that is, generate verbs such as 'run' or 'work', in 60s.	Impairments in verb fluency tasks are thought to more greatly reflect frontostriatal neuropathology and neurochemical deterioration known to occur with progression of PD.[98] A systematic review[99] reports more prominent impairment for action fluency than for category fluency for PD, differential task performance for AD and DLB participants in action fluency and no difference in action fluency for comparisons of MCI and healthy control participants.
Digit span forwards and backwards	In this abbreviated digit span task, in forwards digit span, a series of 5 digits is presented to participants (eg, 8-3-1-9-6), which they are asked to repeat back in the same order. In backwards digit span, a series of 5 digits is presented to participants (eg, 8-3-1-9-6), which they are asked to repeat in backwards order (eg, 6-9-1-3-8). Forwards and backwards digit span were completed three times each with three different 5-digit sequences.	Digit span tests are associated with auditory attention and short-term memory function, with greater reliance on working memory for the backwards span variant. Meta-analyses show greater impairments in backwards span than in forwards span for MCI/AD[91]; PD.[100] In affective disorders, meta-analyses have emphasised deficits in backwards span in the absence of deficits in forwards digit span.[77 93]
Stroop	In this abbreviated Stroop task, participants are first presented with a grid of 50 colours written out in text (eg, 'BLUE') and asked to read these back as quickly as they can. They are then presented with a panel of colours presented in blocks and are asked to name these as quickly as possible. Finally, they are presented with a panel of 50 words typed with the typeface in a colour incongruent with the written word (eg, 'GREEN' written in the colour red). They are asked to state the colour of the letters as quickly as possible.	The Stroop test is an extensively used test to evaluate inhibition to verbal interference. Prior meta-analyses have shown impairments in Stroop interference performance in individuals with amnestic MCI[101]; PD[94]; MND[95]; MDD[77] and BD.[78]
Procedural discourse questions	Participants are tasked with describing, in as much detail as possible, how they would go about doing the dishes by hand (in main assessment), how they would go about posting a letter and how they would make a cup of tea (each assessed on one occasion during remote assessments).	The procedural discourse task elicits naturalistic speech, used to express temporal and hierarchical steps in a behavioural sequence. The task has been found to be sensitive to speech differences in individuals with MCI and mild AD in comparison with healthy control participants.[102]
Picture description task	Participants verbally describe a picture in as much detail as they can. Four different pictures will be used, the first being the 'cookie theft' picture taken from the Boston Diagnostic Aphasia Examination battery[103] administered during main assessment, and the other two being the 'rescue' picture[104] and two simple emotion eliciting pictures administered during remote assessments.[105]	The picture description task produces a structured output that can be scored according to the completeness of response, and is useful for measuring and detecting differences in language content, syntax, pragmatics and acoustic features.[49] Research shows changes in this task for individuals with MCI and AD dementia, lesser impairment in individuals with depression, and differences in speech error corrections between AD and PD groups.[49]

Speech and cognitive tasks	Task details	Task characteristics and selected evidence from prior research
Sequence narration task	Participants must describe their narrative interpretation of a series of images presented as a storyboard, taken from the 'Argument' sequence.[104] Participants have 1 min to look at the picture sequence and are then asked to tell the story represented by the picture sequence.	The task produces a relatively structured output, where analysis outputs can include completeness of the narrative. The task is also used to evaluate lexico-semantic deficits, and syntactic complexity. The pictorially provided structure is thought to place a decreased load on working memory.[106]
Reading a script	Participants will read a short passage aloud: 'The boy who cried wolf'.[107] A passage designed for phonetic description and acoustic research on varieties of English.	This task allows the evaluation of pronunciation variations in the English language in different accents and dialects.
Sustained phonation task	Participants are required to phonate several sounds (/a:/, /i:/ and /u:/) for as long and steadily as possible, in one breath.	Sustained phonation allows for the measurement of voice quality. Changes in vowel articulation has been noted as a potential early marker for PD, and research shows changes in the audio speech measures distinguish PD participants from controls and a likely increase in voice impairment following developmental trajectory of the individuals.
Sentence reading task	During sentence repetition, a sentence (containing 11 or 15 words) is presented auditorily to the participant is asked to must repeat it back in the same way and in one breath. The sentences each contain two /u:/, two /i:/ and two /a:/ long corner vowels to measure vowel space area, following voiceless consonants. During each assessment, two different sentences are presented.	Sentence repetition allows for the measurement of voice quality within a more naturalistic speech response. Research indicates that sentence repetition may be more sensitive to audio speech changes in PD than sustained phonation.[108]
Syllable repetition task	Participants are required to repeatedly phonate a syllable as quickly as possible for one breath. They are first asked to compete this for the syllable /ba/ and then for three consecutive syllables /pa/ /ta/ /ka/.	These tasks assess difficulties with phonology, articulation and working memory. Prior research shows changes in acoustic features in the /pa/ /ta/ /ka/ task in PD (including jitter, pause rates, intensity variation, alternating motion rate and pitch variation).[108] Other research shows changes acoustic features (including alternating motion rate, jitter, frequency and an overall variable) for MND in repeated phonation of the syllable /ba/.[109]

AD, Alzheimer's disease; BD, bipolar disorder; DLB, dementia with Lewy bodies; MCI, mild cognitive impairment; MDD, major depressive disorder; MND, motor neuron disease; PD, Parkinson's disease.

commonly associated with anomia and decreased information content, other neurodegenerative conditions such as PD show greater impairment in volume and articulation.[53 56] Language-based changes have also been noted in PD including rates of pauses, phrase length and changes in sentence generation and construction.[76] Similarly, cognitive and speech changes in mood disorders are commonly described.[52 77–79]

The consistent use of tasks across groups allows for a head-to-head comparison of task performance and speech measures for different clinical indications and different tasks. The tasks are selected due to a rich research background showing their involvement in specific clinical indications (described in table 2). However, with a few notable exceptions, cross-diagnostic evaluations are not commonly reported.

Speech tasks will be administered and recorded using the application, with all participants completing the same set of speech tasks. Participants will be supported with using the Novoic application on their smartphone. Audio recordings will be transferred to a secure server and then deleted from the smartphone device.

Recordings of the combined screening and baseline visit will also be made simultaneously by video conferencing software. Zoom (https://zoom.us) will be used due to its ability to turn off audio manipulation effects. The main assessment phase begins with a series of speech tasks, self-recorded using a smartphone application,

that the researcher guides the participant to complete. An overview of the specific tasks completed in the main assessment is provided in table 3.

Following the completion of the main sequence speech tasks, all participants will then be asked to complete online self-rated questionnaires via the Qualtrics platform (www.qualtrics.com) (see table 1):

A. The Generalised Anxiety Disorder Assessment-7 item[80] is a brief tool for assessing symptoms of anxiety. Participants rate seven items relating to core symptoms of anxiety on a 4-point Likert scale concerning the frequency at which those symptoms occur.

B. The Maudsley 3-Item Depression Visual Analogue Scale[81] is a researcher-rated assessment of depressive illness which detects symptom severity and suicidality.

C. The Altman Self Rating Mania Scale[82] is a 5-item instrument designed to self-assess the presence and severity of manic and hypomanic symptoms by comparing how they feel currently to their non-affected baseline state.

Participants will then complete follow-up data collection over a period of 4 weeks using the smartphone application. Participants will be notified via the mobile application to complete 12 brief (approximately 15 min) unsupervised speech and cognition assessments over the month following the baseline visit. Participants will receive a different set of speech tasks each day from the application's built-in task-bank, including the same main

Table 3 Assessments completed during main (supervised) assessment, and remotely via the application on the participants' own smartphones

Task	Main (supervised) assessment	Remote assessment day											
		1	2	3	4	7	8	9	10	11	14	21	28
Counting 1–10 or 10–1	X	X	X	X	X	X	X	X	X	X	X	X	X
Automated Story Recall Task immediate and delayed recall	X	X	X	X	X	X	X	X	X	X	X*	X*	X*
Category fluency	X	X					X				X*	X*	X*
Letter fluency	X		X					X			X*	X*	X*
Action fluency	X										X*		
Digit span forward and backward				X					X				
Stroop							X						
Procedural discourse	X				X					X			
Picture description	X		X					X					
Sustained phonation	X										X*		
Syllable repetition	X	X*									X*		
Sentence repetition	X	X*									X*		
Sequence narration	X												
Reading a script	X												X*
Usability questionnaire	X		X							X			
Vision and hearing questionnaire	X			X									
Baseline mood, sleep, attention and effort questionnaire	X	X											
Daily mood, sleep, attention and effort questionnaire	X	X	X	X	X	X	X	X	X	X	X	X	X

Remote assessment day corresponds to the number of days after the virtual visit. Test order is shown in descending order. All Automated Story Recall Task immediate and delayed recall assessments are completed with brief distractor tasks occurring in-between.
*Repeated stimuli.

speech and cognition tasks described above, and abbreviated test of executive function (Stroop, digit span).

Additional, brief self-report measures, incorporated into remote assessments, included:

► A 4-item questionnaire on the usability of the smartphone application.
► A brief, 3 min long questionnaire on whether the participant has any visual or hearing impairments, and whether they had related difficulties with the administered tasks.
► A baseline 4-item questionnaire: examining typical mood, sleep, mind wandering and effort.
► A brief, daily 4-item questionnaire on current state: mood, sleep, mind wandering and effort.

The remote assessment schedule is provided in table 3. Participants will also complete follow-up questionnaires online via Qualtrics during weeks 2 and 4, which are shown in table 1.

Statistical analysis plan

The main analysis objectives of the study include examining: (1) feasibility of eliciting and collecting speech data within different clinical groups, (2) reliability, and intrasubject and intersubject variance of speech task performance, (3) adherence to remote unsupervised assessments, (4) app usability and (5) to evaluate whether acoustic and linguistic patterns can be used to distinguish between clinical indications and controls.

Feasibility will be assessed as the length of speech generated during speech elicitation tasks. The primary end point is calculated as the number of participants who successfully complete Automated Story Recall Task (ASRT) assessments as a fraction of the disease cohort, which may indicate likely suitability of this procedure within each disease group. Participants are considered successful where at least one of the immediate story recalls produces a spoken response spanning ≥20 s from the first to the last word. The effect of demographic confounders (age, sex and education) on task feasibility will be evaluated.

Reliability will be measured by examining the stability of equivalent tasks completed over the assessment period. This will include examining reliability across parallel test variants that vary by assessment day and by setting (baseline vs remote follow-up).

Practice and learning effects of parallel test variants over repeated assessment will be characterised with linear mixed effects models. Correlational analysis will be completed between parallel test variants, and intraclass correlations will be carried out to examine test–retest reliability of the same test variants administered across days and between settings (virtual study visit vs remote assessment). Coefficients of individual agreement will be used as a measure of interparticipant and intraparticipant variability.

Usability will be examined via the scores on the usability questionnaire for each disease cohort and their matched controls to examine disease cohort-specific problems or difficulties in completing remote assessments. Adherence will be defined as the proportion of participants engaging in daily remote assessments in each indication. Data will be analysed using t-tests, analysis of variance or non-parametric equivalents, as appropriate. Age and education differences between groups will be controlled for as appropriate.

For the end points which require training classifiers or regressors to make predictions, the entire data set will be used for both training machine learning models and validating their statistical properties. Machine learning methods used may include supervised, unsupervised, semi-supervised and reinforcement learning methods for forming intermediate representations of the data and performing the final classification/regression analyses. Speech measures will be derived using several methods from the fields of signal/speech processing and natural language processing to investigate which speech markers are most predictive of clinical status. Some of the planned methods for analyses are described here, but in this broad rapidly evolving field, others may be used where suitable and as they become available.

The sample rate, bit rate and codec compression of smartphones along with speech intelligibility measures will be used to probe dependence of speech features on audio quality measures. Data from the mPower study suggest that phonation data from a bring-your-own-device setup is sufficiently consistent to predict motor disorders to some degree.[83]

Acoustic measures will be extracted via Surfboard, an automated audio feature extraction library,[83] which extracts a range of acoustic speech features (such as mel-frequency cepstrum coefficients, F0 contour, formant frequencies, intensity, loudness); and Vector-Quantized Prosody (VQP), a self-supervised contrastive model for non-timbral prosody.[84] Linguistic and pause analyses will be completed after automated transcription for onward textual analysis. Text features (including number of words, noun and pronoun rate, idea density) will be extracted using Stanza.[85] Speaking time, articulation rate and pause information will be extracted directly from the number of words and timestamps in the transcriptions. Group prediction from text and audio features will be evaluated within the python machine learning package scikit learn[86] with k-fold cross-validation. For ASRTs,

textual analyses will be completed with ParaBLEU, a paraphrase evaluation model,[87] and G-match, evaluating the cosine similarity of textual embeddings between two texts.[45]

Deep learning methods will be used, where the speech measures are derived as internal representations of the models. Methods used here will include unsupervised context-aware word embeddings, a way of distinguishing between identical words with different contextual meanings,[87] and audio representation learning.[84]

Training and validation of the ML models will be performed using cross-validation techniques. Performance of classification analysis will be characterised by the area under the receiver operating characteristic curve (AUC), sensitivity, specificity and Cohen's kappa. Demonstration that a binary classifier performs better than the random baseline will be done via mapping AUC estimates to a p value via the Mann-Whitney U statistic.[88] Results will be contrasted with demographic comparisons (comprising sex, education and age) as in prior analyses,[48] to evaluate the contribution of demographic imbalances.

Methodological issues and limitations
The primary potential limitation of the study is poor compliance with the self-administered speech tasks over the follow-up period. Efforts to address this potential issue include reminder notifications from the smartphone application and allowing participants to complete these self-assessments at flexible timepoints. It should also be noted that adherence to the self-administered tasks is itself an outcome measure of the study, and so missing data in this regard are informative.

Practice effects on repeated tasks are likely to have an impact on task performance. Parallel test variants, such as those for the ASRT, can minimise practice effects on repeated exposure. However, even with parallel test variants, repeated administration is likely to see task improvements over time related to repeated exposure to the same task and greater familiarity with the test structure and method.[45] For some conditions and some tasks such as category fluency, the practice effects themselves (or lack thereof) may be of interest for identifying specific diagnostic groups, such as individuals with MCI on tasks such as category fluency.[89 90]

Other potential limitations include the fact that recruitment is limited by the requirement for participants to be able to access and use electronic devices for speech tasks and follow-up questionnaires, as not all patients own or have access to connected digital platforms,[40] even with assistance from others.

Data management and oversight
All aspects of the study will be overseen by the Study Management Group. Primary investigators will oversee and manage all aspects of their study in their respective cohorts. Their responsibility includes monitoring adverse events in their respective cohorts and managing participant discontinuation.

Audio data will be recorded by the app and by the video conference software. On completion of each assessment, the audio will be securely transferred to Novoic's cloud servers, which are fully secure and compliant under relevant security standards (including ISO27001) and privacy regulations (including General Data Protection Regulation). Other clinical data will be stored in a secure password-protected server, compliant with all applicable laws and regulations, including ICH E6 Good Clinical Practice, EU Annex 11, GDPR.

Data access

The speech data obtained during the study will not be made available to a repository due to the potential for participant identification.

Patient and public involvement

We have involved patients and members of the public in planning of this study, having sought service users' perspectives on protocol documentation which resulted in amendments to language used and patient facing documentation. We have also implemented changes based on carers' feedback from a focus group of 10–15 individuals. We also actively collect usability feedback from service users. During dissemination, we will invite service users and carers to contribute to a public perspective on the interpretation of trial findings. Results will be made available to participants on request.

Ethics and dissemination

The study received REC approval from the London—Queen Square Research Ethics Committee on 15 April 2021, and HRA approval from HRA and Health and Care Research Wales on 23 April 2021 (REC reference: 21/PR/0070). Findings will be published in peer-reviewed journals and presented at conferences focused on neuroscience and psychiatry, and machine learning and natural language processing. Novoic will create a project webpage containing key findings, published papers and open-source software packages. Results will be made available to participants on request.

Contributors ERH (ORCID ID: 0000-0001-6985-5646): contributed to the writing of the study protocol, in addition to the writing and editing of the manuscript for journal submission. MM (ORCID ID: 0000-0003-4937-3062): main contributor on designing the study and protocol. Contributed to writing the full study protocol and editing the manuscript for journal submission. CS (ORCID ID: 0000-0001-8692-7787): contributed to the study design, writing and editing of the manuscript for journal submission. RS (ORCID ID: 0000-0002-2984-1124): contributed to the editing of the study protocol and manuscript for journal submission, in addition to study set up activities. RHT (ORCID ID: 0000-0001-6742-4842): contributed to editing of the manuscript for journal submission. LC (ORCID ID: 0000-0002-9514-364X): contributed to the writing of the study protocol, in addition to the editing of the manuscript for journal submission. DA (ORCID ID: 0000-0001-6314-216X): contributed to study and protocol design. Principal investigator for neurodegenerative disease cohort. AA-C (ORCID ID: 0000-0002-4924-7712): contributed to study and protocol design. Principal investigator for motor disorders cohort. RC (ORCID ID: 0000-0003-2815-0505): contributed to study and protocol design. Principal investigator for motor disorders cohort. JW (ORCID ID: 0000-0001-5344-7840): contributed to the writing of the study protocol, in addition to editing of the manuscript for journal submission. EF (ORCID ID: 0000-0002-9590-7275): main contributor on designing the study and protocol. Contributed to writing the full study protocol and editing the manuscript for journal submission. AP (ORCID ID: 0000-0002-0805-3356): contributed to editing of the manuscript for journal submission. OA (ORCID ID: 0000-0002-6814-3097): contributed to editing of the manuscript for journal submission. AY (ORCID ID: 0000-0003-2291-6952): chief investigator. Contributed to study and protocol design and edited the manuscript for journal submission.

ORCID iDs
Elliot Hampsey http://orcid.org/0000-0001-6985-5646
Ammar Al-Chalabi http://orcid.org/0000-0002-4924-7712

REFERENCES

1 Who.int. Mental health of older adults [Internet], 2021. Available: https://www.who.int/news-room/fact-sheets/detail/mental-health-of-older-adults [Accessed 30 Sep 2021].
2 Wittenberg R, Hu B, Barraza-Araiza L, *et al*. Projections of older people with dementia and costs of dementia care in the United Kingdom, 2019–2040. *Care Policy and Evaluation Centre Working Paper 5* 2019.
3 Weir S, Samnaliev M, Kuo T-C, *et al*. Short- and long-term cost and utilization of health care resources in Parkinson's disease in the UK. *Mov Disord* 2018;33:974–81.
4 No health without mental health: a cross-government mental health outcomes strategy for people of all ages 2011.
5 McCrone P, Dhanasiri S, Patel A, *et al*. *Paying the price: the cost of mental health care in England to 2026*, 2008.
6 Ons.gov.uk. Overview of the UK population - Office for National Statistics [Internet], 2021. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/august2019 [Accessed 30 Sep 2021].
7 McKeith I, Mintzer J, Aarsland D, *et al*. Dementia with Lewy bodies. *Lancet Neurol* 2004;3:19–28.
8 The Lewy Body Society. *A guide to Lewy body dementia*, 2019.
9 O'Brien J, Taylor J, Thomas A. Improving the diagnosis and management of neurodegenerative dementia of Lewy body type in the NHS (DIAMOND-Lewy programme). *NIHR Journals Library Publications*.
10 McKeith IG, Rowan E, Askew K, *et al*. More severe functional impairment in dementia with Lewy bodies than Alzheimer disease is related to extrapyramidal motor dysfunction. *Am J Geriatr Psychiatry* 2006;14:582–8.

11 Gumber A, Ramaswamy B, Ibbotson R, *et al*. Economic, Social and Financial Cost of Parkinson's on Individuals, Carers and their Families in the UK. *Project Report. Centre for Health and Social Care Research, Sheffield Hallam University* 2017.

12 McCrone P, Allcock LM, Burn DJ. Predicting the cost of Parkinson's disease. *Mov Disord* 2007;22:804–12.

13 Marras C, Chaudhuri KR. Nonmotor features of Parkinson's disease subtypes. *Mov Disord* 2016;31:1095–102.

14 Gowland A, Opie-Martin S, Scott KM, *et al*. Predicting the future of ALS: the impact of demographic change and potential new treatments on the prevalence of ALS in the United Kingdom, 2020-2116. *Amyotroph Lateral Scler Frontotemporal Degener* 2019;20:264–74.

15 Westeneng H-J, Debray TPA, Visser AE, *et al*. Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *Lancet Neurol* 2018;17:423–33.

16 Solmi M, Radua J, Olivola M, *et al*. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Mol Psychiatry* 2022;27:281-295.

17 Kessing LV. Recurrence in affective disorder. *British Journal of Psychiatry* 1998;172:29–34.

18 James SL, Abate D, Abate KH, *et al*. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet* 2018;392:1789–858.

19 Fajutrao L, Locklear J, Priaulx J, *et al*. A systematic review of the evidence of the burden of bipolar disorder in Europe. *Clin Pract Epidemiol Ment Health* 2009;5:3..

20 Hirschfeld RM. Differential diagnosis of bipolar disorder and major depressive disorder. *J Affect Disord* 2014;169 Suppl 1:S12–16.

21 Dome P, Rihmer Z, Gonda X. Suicide risk in bipolar disorder: a brief review. *Medicina* 2019;55:403.

22 Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet* 2009;374:609–19.

23 Vermani M, Marcus M, Katzman MA. Rates of detection of mood and anxiety disorders in primary care. *Prim Care Companion CNS Disord* 2011.

24 Hohl U, Tiraboschi P, Hansen LA, *et al*. Diagnostic accuracy of dementia with Lewy bodies. *Arch Neurol* 2000;57:347.

25 Beach TG, Monsell SE, Phillips LE, *et al*. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on aging Alzheimer disease centers, 2005-2010. *J Neuropathol Exp Neurol* 2012;71:266–73.

26 Morselli PL, Elgie R, GAMIAN-Europe. GAMIAN-Europe/BEAM survey I--global analysis of a patient questionnaire circulated to 3450 members of 12 European advocacy groups operating in the field of mood disorders. *Bipolar Disord* 2003;5:265–78.

27 Dagani J, Signorini G, Nielssen O, *et al*. Meta-Analysis of the interval between the onset and management of bipolar disorder. *Can J Psychiatry* 2017;62:247–58.

28 Rasmussen J, Langerman H. Alzheimer's Disease - Why We Need Early Diagnosis. *Degener Neurol Neuromuscul Dis* 2019;9:123–30.

29 Goldman JG, Williams-Gray C, Barker RA, *et al*. The spectrum of cognitive impairment in Lewy body diseases. *Mov Disord* 2014;29:608–21.

30 Aarsland D, Andersen K, Larsen JP, *et al*. Prevalence and characteristics of dementia in Parkinson disease: an 8-year prospective study. *Arch Neurol* 2003;60:pp.:387–92.

31 NHS Digital. *Recorded dementia diagnoses*. London: NHS England, 2020.

32 Price A, Farooq R, Yuan J-M, *et al*. Mortality in dementia with Lewy bodies compared with Alzheimer's dementia: a retrospective naturalistic cohort study. *BMJ Open* 2017;7:e017504.

33 Jankovic J. Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry* 2008;79:368–76.

34 Rizek P, Kumar N, Jog MS. An update on the diagnosis and treatment of Parkinson disease. *CMAJ* 2016;188:1157–65.

35 Schrag A, Ben-Shlomo Y, Quinn N. How valid is the clinical diagnosis of Parkinson's disease in the community? *J Neurol Neurosurg Psychiatry* 2002;73:529–34.

36 Sperling RA, Dickerson BC, Pihlajamaki M, *et al*. Functional alterations in memory networks in early Alzheimer's disease. *Neuromolecular Med* 2010;12:27–43.

37 Kulisevsky J, Pagonabarraga J. Cognitive impairment in Parkinson's disease: tools for diagnosis and assessment. *Mov Disord* 2009;24:1103–10.

38 D'Onofrio G, Sancarlo D, Panza F, *et al*. Neuropsychiatric symptoms and functional status in Alzheimer's disease and vascular dementia patients. *Curr Alzheimer Res* 2012;9:759–71.

39 NICE. *Motor neurone disease Briefing paper*. London: NICE, 2015.

40 De Marchi F, Contaldi E, Magistrelli L, *et al*. Telehealth in neurodegenerative diseases: opportunities and challenges for patients and physicians. *Brain Sci* 2021;11:237.

41 Patel SY, Mehrotra A, Huskamp HA, *et al*. Trends in outpatient care delivery and telemedicine during the COVID-19 pandemic in the US. *JAMA Intern Med* 2021;181:388.

42 National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Sciences Policy; Forum on Drug Discovery, Development, and TranslationShore C, ed. *Virtual clinical trials: challenges and opportunities: proceedings of a workshop*. National Academies Press (US), 2019.

43 Cormack F, McCue M, Taptiklis N, *et al*. Wearable technology for high-frequency cognitive and mood assessment in major depressive disorder: longitudinal observational study. *JMIR Ment Health* 2019;6:e12814.

44 Goodday SM, Atkinson L, Goodwin G, *et al*. The true colours remote symptom monitoring system: a decade of evolution. *J Med Internet Res* 2020;22:e15188.

45 Skirrow C, Meszaros M, Meepegama U. Validation of a remote and fully automated story recall task to assess for early cognitive impairment in older adults: a longitudinal case-control observational study. *medRXiv* 2022.

46 Broen MPG, Marsman VAM, Kuijf ML, *et al*. Unraveling the relationship between motor symptoms, affective states and contextual factors in Parkinson's disease: a feasibility study of the experience sampling method. *PLoS One* 2016;11:e0151195.

47 Omberg L, Chaibub Neto E, Perumal TM, *et al*. Remote smartphone monitoring of Parkinson's disease and individual response to therapy. *Nat Biotechnol* 2022;40:480–7.

48 Fristed E, Skirrow C, Meszaros M. Evaluation of a speech-based AI system for early detection of Alzheimer's disease remotely via smartphones. *medRxiv* 2021.

49 Mueller KD, Hermann B, Mecollari J, *et al*. Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. *J Clin Exp Neuropsychol* 2018;40:917–39.

50 Codina-FilbJ, Escalera S, Escudero J, *et al*. Mobile eHealth Platform for Home Monitoring of Bipolar Disorder. In: Springer C, ed. *et al. MultiMedia Modeling. MMM 2021. Lecture Notes in Computer Science*. 12573, 2021.

51 Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, *et al*. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. *Front Psychol* 2021;12:620251.

52 Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investig Otolaryngol* 2020;5:96–116.

53 Smith KM, Caplan DN. Communication impairment in Parkinson's disease: impact of motor and cognitive symptoms on speech and language. *Brain Lang* 2018;185:38–46.

54 Sano M, Lapid MI, Ikeda M, *et al*. Psychogeriatrics in a world with COVID-19. *Int Psychogeriatr* 2020;32:1101–5.

55 Patel UK, Anwar A, Saleem S, *et al*. Artificial intelligence as an emerging technology in the current care of neurological disorders. *J Neurol* 2021;268:1623–42.

56 Boschi V, Catricalà E, Consonni M, *et al*. Connected speech in neurodegenerative language disorders: a review. *Front Psychol* 2017;8:269.

57 Engelman M, Agree EM, Meoni LA, *et al*. Propositional density and cognitive function in later life: findings from the precursors study. *J Gerontol B Psychol Sci Soc Sci* 2010;65:pp.:706–11.

58 Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis* 2016;49:407–22.

59 Cummins N, Scherer S, Krajewski J, *et al*. A review of depression and suicide risk assessment using speech analysis. *Speech Commun* 2015;71:10–49.

60 Riley KP, Snowdon DA, Desrosiers MF, *et al*. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun study. *Neurobiol Aging* 2005;26:341–7.

61 Cuetos F, Arango-Lasprilla JC, Uribe C, *et al*. Linguistic changes in verbal expression: a preclinical marker of Alzheimer's disease. *J Int Neuropsychol Soc* 2007;13:433–9.

62 Snowdon DA*et al*. Linguistic ability in early life and cognitive function from the nun and Alzheimer's Disease in late life study. *JAMA* 1996;275:528–32.

63 Hertzog MA. Considerations in determining sample size for pilot studies. *Res Nurs Health* 2008;31:180–91.

64 Welsh KA, Breitner JC, Magruder-Habib KM. Detection of dementia in the elderly using telephone screening of cognitive

status. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology* 1993;6:103–10.

65 Albert MS, DeKosky ST, Dickson D, *et al*. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;7:270–9.

66 McKeith IG, Boeve BF, Dickson DW, *et al*. Diagnosis and management of dementia with Lewy bodies. *Neurology* 2017;89:88–100.

67 Rascovsky K, Hodges JR, Knopman D, *et al*. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain* 2011;134:2456–77.

68 Mesulam MM. Primary progressive aphasia. *Ann Neurol* 2001;49:425–32.

69 Román GC, Tatemichi TK, Erkinjuntti T, *et al*. Vascular dementia: diagnostic criteria for research studies. Report of the NINDS-AIREN International workshop. *Neurology* 1993;43:250.

70 Roche JC, Rojas-Garcia R, Scott KM, *et al*. A proposed staging system for amyotrophic lateral sclerosis. *Brain* 2012;135:847–52.

71 Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. *Neurology* 1967;17:427.

72 Sheehan DV, Lecrubier Y, Sheehan KH, *et al*. The Mini-International neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 1998;59 Suppl 20:22-33:34–57.

73 Guy W. *Clinical global impressions, ECDEU assessment manual for psychopharmacology, revised (DHEW Publ. No. ADM 76-338*. Rockville: National Institute of Mental Health, 1976: 218–22.

74 Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. *J Gen Intern Med* 2001;16:606–13.

75 Rush AJ, Carmody T, Reimitz P-E. The inventory of depressive symptomatology (IDS): clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms. *Int J Methods Psychiatr Res* 2000;9:45–59.

76 Suárez-González A, Cassani A, Gopalan R. When it is not primary progressive aphasia: a scoping review of spoken language impairment in other neurodegenerative dementias. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 2021;7.

77 Bora E, Harrison BJ, Yücel M, *et al*. Cognitive impairment in euthymic major depressive disorder: a meta-analysis. *Psychol Med* 2013;43:2017–26.

78 Torres IJ, Boudreau VG, Yatham LN. Neuropsychological functioning in euthymic bipolar disorder: a meta-analysis. *Acta Psychiatr Scand Suppl* 2007;434:17–26.

79 Bourne C, Aydemir Ö, Balanzá-Martínez V, *et al*. Neuropsychological testing of cognitive impairment in euthymic bipolar disorder: an individual patient data meta-analysis. *Acta Psychiatr Scand* 2013;128:149–62.

80 Spitzer RL, Kroenke K, Williams JBW, *et al*. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006;166:1092.

81 Moulton CD, Strawbridge R, Tsapekos D, *et al*. The Maudsley 3-item visual analogue scale (M3VAS): validation of a scale measuring core symptoms of depression. *J Affect Disord* 2021;282:280–3.

82 Altman EG, Hedeker D, Peterson JL, *et al*. The Altman self-rating mania scale. *Biol Psychiatry* 1997;42:948–55.

83 Lenain R, Weston J, Shivkumar A, *et al*. Surfboard: Audio Feature Extraction for Modern Machine Learning [Internet], 2020. arxiv.org. Available: https://arxiv.org/abs/2005.08848 [Accessed 12 Apr 2022].

84 Weston J, Lenain R, Meepegama U. &amp. Learning de-identified representations of prosody from raw audio. Proceedings of the 38th International Conference on machine learning, in. *Proceedings of Machine Learning Research* 2021;139:11145 https://proceedings.mlr.press/v139/weston21a.html

85 Qi P, Zhang Y, Zhang Y, *et al*. Stanza : A Python Natural Language Processing Toolkit for Many Human Languages [Internet], 2003. Available: https://arxiv.org/pdf/2003.07082.pdf [Accessed 12 Apr 2022].

86 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine Learning in Python [Internet], 2011. Available: https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf [Accessed 12 Apr 2022].

87 Weston J, Lenain R, Meepegama U. Generative pretraining for paraphrase evaluation, 2022. Available: https://arxiv.org/pdf/2107.08251.pdf [Accessed 12 Apr 2022].

88 Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 1947;18:50–60.

89 Cooper DB, Epker M, Lacritz L, *et al*. Effects of practice on category fluency in Alzheimer's disease. *Clin Neuropsychol* 2001;15:125–8.

90 Cooper D, Lacritz L, Weiner M. Category fluency in mild cognitive impairment. *Alzheimer Disease & Associated Disorders* 2004;18:120–2.

91 Besser L, Litvan I, Monsell S. Mild cognitive impairment in Parkinson's disease versus Alzheimer's disease. *Parkinsonism & Related Disorders* 2016;27:54–60.

92 Semkovska M, Quinlivan L, O'Grady T, *et al*. Cognitive function following a major depressive episode: a systematic review and meta-analysis. *Lancet Psychiatry* 2019;6:851–61.

93 Kudlicka A, Clare L, Hindle JV. Executive functions in Parkinson's disease: systematic review and meta-analysis. *Mov Disord* 2011;26:2305–15.

94 Gurnani A, Gavett B. The Differential Effects of Alzheimer's Disease and Lewy Body Pathology on Cognitive Performance: a Meta-analysis. *Neuropsychology Review* 2016;27:1–17.

95 Beeldman E, Raaphorst J, Klein Twennaar M, *et al*. The cognitive profile of ALS: a systematic review and meta-analysis update. *Journal of Neurology, Neurosurgery & Psychiatry* 2015;87:611–9.

96 Raucher-Chéné D, Achim AM, Kaladjian A, *et al*. Verbal fluency in bipolar disorders: a systematic review and meta-analysis. *J Affect Disord* 2017;207:359–66.

97 Bauer K, Malek-Ahmadi M. Meta-Analysis of controlled oral word association test (COWAT) Fas performance in amnestic mild cognitive impairment and cognitively unimpaired older adults. *Appl Neuropsychol Adult* 2021:1–7.

98 Piatt AL, Fields JA, Paolo AM, *et al*. Action (verb naming) fluency as an executive function measure: convergent and divergent evidence of validity. *Neuropsychologia* 1999;37:1499–503.

99 Beber B, Chaves M. The basis and applications of the action fluency and action naming tasks. *Dementia & Neuropsychologia* 2014;8:47–57.

100 Ramos AA, Machado L. A comprehensive meta-analysis on short-term and working memory dysfunction in Parkinson's disease. *Neuropsychol Rev* 2021;31:288–311.

101 Rabi R, Vasquez BP, Alain C, *et al*. Inhibitory control deficits in individuals with amnestic mild cognitive impairment: a meta-analysis. *Neuropsychol Rev* 2020;30:97–125.

102 Clarke N, Barrick TR, Garrard P. A Comparison of Connected Speech Tasks for Detecting Early Alzheimer's Disease and Mild Cognitive Impairment Using Natural Language Processing and Machine Learning. *Front Comput Sci* 2021;3.

103 Goodglass H, Kaplan E. *The assessment of aphasia and related disorders*. 2nd edn. Philadelphia, PA: Lea and Febiger, 1983.

104 Nicholas LE, Brookshire RH. A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *J Speech Hear Res* 1993;36:338–50.

105 Teh EJ, Yap MJ, Liow SJR. PiSCES: pictures with social context and emotional scenes with norms for emotional valence, intensity, and social engagement. *Behav Res Methods* 2018;50:1793–805.

106 Sherratt S, Bryan K. Textual cohesion in oral narrative and procedural discourse: the effects of ageing and cognitive skills. *International Journal of Language & Communication Disorders* 2018;54:95–109.

107 Deterding D. The North wind versus a wolf: short texts for the description and measurement of English pronunciation. *J Int Phon Assoc* 2006;36:187–96.

108 Rusz J, Cmejla R, Tykalova T, *et al*. Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. *J Acoust Soc Am* 2013;134:2171–81.

109 Rong P, Yunusova Y, Richburg B, *et al*. Automatic extraction of abnormal lip movement features from the alternating motion rate task in amyotrophic lateral sclerosis. *Int J Speech Lang Pathol* 2018;20:610–23.