Journal of Alzheimer's Disease 78 (2020) 1547–1574
DOI 10.3233/JAD-200888
IOS Press

Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review

Sofia de la Fuente Garcia[a,*], Craig W. Ritchie[b] and Saturnino Luz[a]
[a] *Usher Institute, Edinburgh Medical School, The University of Edinburgh, Scotland, UK*
[b] *Centre for Clinical Brain Sciences, The University of Edinburgh, Scotland, UK*

Abstract.
Background: Language is a valuable source of clinical information in Alzheimer's disease, as it declines concurrently with neurodegeneration. Consequently, speech and language data have been extensively studied in connection with its diagnosis.
Objective: Firstly, to summarize the existing findings on the use of artificial intelligence, speech, and language processing to predict cognitive decline in the context of Alzheimer's disease. Secondly, to detail current research procedures, highlight their limitations, and suggest strategies to address them.
Methods: Systematic review of original research between 2000 and 2019, registered in PROSPERO (reference CRD42018116606). An interdisciplinary search covered six databases on engineering (ACM and IEEE), psychology (PsycINFO), medicine (PubMed and Embase), and Web of Science. Bibliographies of relevant papers were screened until December 2019.
Results: From 3,654 search results, 51 articles were selected against the eligibility criteria. Four tables summarize their findings: *study details* (aim, population, interventions, comparisons, methods, and outcomes), *data details* (size, type, modalities, annotation, balance, availability, and language of study), *methodology* (pre-processing, feature generation, machine learning, evaluation, and results), and *clinical applicability* (research implications, clinical potential, risk of bias, and strengths/limitations).
Conclusion: Promising results are reported across nearly all 51 studies, but very few have been implemented in clinical research or practice. The main limitations of the field are poor standardization, limited comparability of results, and a degree of disconnect between study aims and clinical applications. Active attempts to close these gaps will support translation of future research into clinical practice.

Keywords: Alzheimer's disease, artificial intelligence, cognitive decline, computational linguistics, dementia, machine learning, screening, speech processing

INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disease that involves decline of cognitive and functional abilities as the illness progresses [1]. It is the most common etiology of dementia. Given its prevalence, it has effects beyond just patients and

*Correspondence to: Sofia de la Fuente Garcia, Usher Institute, Edinburgh Medical School, The University of Edinburgh, Nine Edinburgh BioQuarter, 9 Little France Road, Edinburgh EH16 4UX, Scotland, UK. E-mail: sofia.delafuente@ed.ac.uk.

ISSN 1387-2877/20/$35.00 © 2020 – IOS Press and the authors. All rights reserved

carers as it also has a severe societal and economic impact worldwide [2]. Although memory loss is often considered the signature symptom of AD, language impairment may also appear in its early stages [3]. Consequently, and due to the ubiquitous nature of speech and language, multiple studies rely on these modalities as sources of clinical information for AD, from foundational qualitative research (e.g., [4, 5]) to more recent work on computational speech technology (e.g., [6–8]). The potential for using speech as a biomarker for AD is based on several prospective values, including: 1) the ease with which speech can be recorded and tracked over time, 2) its non-invasiveness, 3) the fact that technologies for speech analysis have improved markedly in the past decade, boosted by advances in artificial intelligence (AI) and machine learning, and 4) the fact that speech problems may be manifest at different stages of the disease, making it a life-course assessment that has value unlimited by disease stage.

Recent studies on the use of AI in AD research entail using language and speech data collected in different ways and applying computational speech processing for diagnosis, prognosis, or progression modelling. This technology encompasses methods for recognizing, analyzing, and understanding spoken discourse. It implies that at least part of the AD detection process could be automated (passive). Machine learning methods have been central to this research program. Machine learning is a field of AI that concerns itself with the induction of predictive models "learned" directly from data, where the learner improves its own performance through "experience" (i.e., exposure to greater amounts of data). Research on automatic processing of speech and language with AI and machine learning methods have yielded encouraging results and attracted increasing interest. Different approaches have been studied, including computational linguistics (e.g., [9]), computational paralinguistics (e.g., [10]), signal processing (e.g., [11]), and human-robot interaction (e.g., [12]).

However, investigations of the use of language and speech technology in AD research are heterogeneous, which makes consensus, conclusions, and translation into larger studies or clinical practice problematic. The range of goals pursued in such studies is also broad, including automated screening for early AD, tools for early detection of disease in clinical practice, monitoring of disease progression, and signalling potential mechanistic underpinnings to speech problems at a biological level thereby improving disease models. Despite progress in research, the small, inconsistent, single-laboratory and non-standardized nature of most studies has yielded results that are not robust enough to be aggregated and thereafter implemented toward those goals. This has resulted in gaps between research contexts, clinical potential, and actual clinical applications of this new technology.

We sought to summarize the current state of the evidence regarding AI approaches in speech analysis for AD with a view to setting a foundation for future research in this area and potential development of guidelines for research and implementation. The review has three main objectives: Firstly, to present the main aims and findings of this research, secondly to outline the main methodological approaches, and finally surmise the potential for each technique to be ready for further evaluation toward clinical use. In doing so, we hope to contribute to the development of these novel, exciting, and yet under-utilized approaches, toward clinical practice.

METHODS

The procedures adopted in this review were specified in a protocol registered with the international prospective register of systematic reviews PROSPERO (reference: CRD42018116606). In the following sections we describe the elegibility criteria, information sources, search strategy, study records management, study records selection, data collection process, data items (extraction tool), risk of bias in individual studies, data synthesis, meta-bias(es), and confidence in cumulative evidence.

Elegibility criteria

We aimed to summarize all available scientific studies where an interactive AI approach was adopted for neuropsychological monitoring. Interaction-based technology entails data obtained through a form of communication, and AI entails some automation of the process. Therefore, we included articles where automatic machine learning methods were used for AD screening, detection, and prediction, by means of computational linguistics and/or speech technology.

Articles were deemed eligible if they described studies of neurodegeneration in the context of AD. That is, subjective cognitive impairment (SCI), mild cognitive impairment (MCI), AD or other dementia-related terminology if indicated as AD-related in the full text (e.g., if a paper title reads unspecified "dementia" but the research field is AD). The included studies examined behavioral patterns that may precede overt cognitive decline as well as observable cognitive impairment in these neurodegenerative diseases. Related conditions such as semantic dementia (a form of aphasia) or Parkinson's disease (a different neurodegenerative disease) formed part of the exclusion criteria (except when in comorbidity with AD). Language was not an exclusion criterion, and translation resources were used as appropriate.

Another exclusion criterion is the exclusive use of traditional statistics in the analysis. The inclusion criteria require at least one component of AI, machine learning, or big data, even if the study encompasses traditional statistical analysis. Further exclusion criteria apply to related studies relying exclusively on neuroimaging techniques such as magnetic resonance imaging (MRI), with no relation to language or speech, even if they do implement AI methods. The same applies to biomarker studies (e.g., *APOE* genotyping). This review also excluded purely epidemiological studies, that is, studies aimed at analyzing the distribution of the condition rather than assessing the potential of AI tools for monitoring its progress.

In terms of publication status, we considered peer-reviewed journal and conference articles only. Records that were not original research papers were excluded (i.e., conference abstracts and systematic reviews). In order to avoid redundancy, we assessed research by the same group and excluded overlapping publications. This was assessed by reading the text in full and selecting only the most relevant article for review (i.e., most comprehensive and up to date). Due to limited resources, we also excluded papers when full-texts were found unavailable in all our alternative sources.

Lastly, we considered papers from a twenty-year span, from the beginning of 2000 to the end of 2019, anticipating that the closer to the end of this timeframe, the larger the number of results, as shown in Fig. 1.

Information sources

Between October and December 2019, we searched the following electronic databases: ACM, Embase, IEEE, PsycINFO, PubMed, and Web of Science. We contacted study authors by email when full-text versions of relevant papers where not available through the university library, with varying degrees of success.

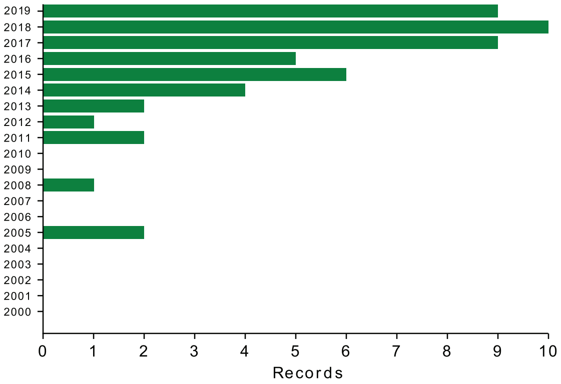

Fig. 1. Number of relevant records per year (2000–2019).

We also included relevant titles found through "forward citation tracking" with Google Scholar, screening articles references and research portal suggestions suggestions.

Search strategy

Given the heterogeneity of the field, a broad search needed to be conducted. For the health condition of interest, AD, we included terms such as dementia, cognitive decline, and Alzheimer. For the methodology, we included speech, technology, analysis and natural language processing, AI, machine learning, and big data.

The search strategy was developed collaboratively between the authors, and with the help of the University of Edinburgh's academic support librarian. After a few iterations and trials, we decided not to include the AI terms, since this seemed to constrain the search too much, yielding fewer results. Therefore, the search queries were specified as follows (example for PubMed):

- (speech AND (dementia OR "cognitive decline" OR (cognit* AND impair*) OR Alzheimer) AND (technology OR analysis)) OR ("natural language processing" AND (dementia OR "cognitive decline" OR (cognit* AND impair*) OR Alzheimer))
- Filters applied: 01/01/2000 - 31/12/2019.

Then, we applied the exclusion criteria, starting from the lack of AI, machine learning, and big data methods, usually detected in the abstract.

We used EndNote X8 [13] for study records management.

Study records selection

Screening for record selection happened in two phases, independently undertaken by two reviewers and following pre-established eligibility criteria. In the first phase, the two independent authors screened titles and abstracts against exclusion criteria using EndNote. The second phase consisted of a full-text screening for those papers that could not be absolutely included or excluded based on title and abstract information only. Any emerging titles that were deemed relevant were added to the screening process. Disagreements at any of the stages were discussed and, when necessary, a third author convened to find a resolution. Some records reported results that were redundant with a later paper of the same research group, mainly because the earlier record was a conference paper or because an extended version of the research paper had been published elsewhere at a later date. When this happened, earlier and shorter reports were excluded.

Data collection process

Our original intention was to rely on the PICO framework [14] for data collection. However, given the relative youth and heterogeneity of the research field reviewed, and the lack of existing reviews on the topic, we adapted a data extraction tool specifically for our purposes. This tool took the form of four comprehensive tables which were used to extract the relevant information from each paper. Those tables summarize general study information, data details, methodology, and clinical applicability.

The tables were initially "piloted" with a few studies, in order to ensure they were fit to purpose. Information extraction was performed independently by two reviewers and consistency was compared. When differences about extracted items was not resolved by discussion, the third author was available to mediate with the paper's full text as reference.

Data items (extraction tool)

As stated in the data collection process, data items will be extracted through the elaboration of four tables. These tables are:

- **SPICMO:** inspired in the PICO framework, it contains information on Study, Population, Interventions, Comparison groups, Methodology, and Outcomes. More details can be found just before Supplementary Table 4.
- **Data details:** dataset/subset size, data type, other data modalities, data annotation, data availability, and language. More details can be found just before Supplementary Table 5.
- **Methodology details:** pre-processing, features generated, machine learning task/method, evaluation technique, and results. More details can be found just before Supplementary Table 6.
- **Clinical applicability:** research implications, clinical potential, risk of bias, and strengths/limitations. More details can be found just before Supplementary Table 7.

Risk of bias in individual studies

Many issues, such as bias, do not apply straightforwardly to this review because it focuses on diagnostic and prognostic test accuracy, rather than interventions. Therefore, if there were to be significance tests they would be for comparisons between the results of the different methods. Besides, the scope of the review is machine learning technology, where the evaluation through significance testing is rare. Papers that rely exclusively on traditional statistics will be excluded, and therefore we expect the review to suffer from a negligible risk of bias in terms of significance testing.

The risk of bias in machine learning studies often comes from how the data is prepared in order to train your models. In a brief example, if a dataset is not split in a training and a testing subset, the model will be trained and tested on the same data. Such model is likely to achieve very good results, but chances are that its performance will drop dramatically when tested on unseen data. This risk is called "overfitting", and is assessed in Supplementary Table 7. Other risks accounted for in this table are data balance, the use of suitable metrics, the contextualization of results, and the sample size. Data balance reports whether the dataset has comparable numbers of AD and healthy participants, as well as in terms of gender or age. Suitable metrics is an assessment of whether the metric chosen to evaluate a model is appropriate, in conjunction with data balance and sample size (e.g., accuracy is not a robust metric when a dataset is imbalanced). Contextualization refers to whether their study results are compared to a suitable baseline (i.e., a measure without a target variable or comparable research results). Finally, sample size is particularly relevant because machine learning methodology was developed for large datasets, but data scarcity is a distinctive feature of this field.

The poor reporting of results and subsequent interpretation difficulties is a longstanding challenge of diagnostic test accuracy research [15]. Initially, we considered two tools for risk of bias assessment, namely the "QUADAS-2: Quality Assessment of Diagnosis Studies checklist - 2" [16] and the "PROBAST: Prediction model Risk Of Bias ASsessment Tool" [17]. However, our search covers an emerging interdisciplinary field where papers are neither diagnostic studies nor predictive ones. Additionally, the Cochrane Collaboration recently emphasized a preference for systematic reviews to focus on the performance of individual papers' on the different risk of bias criteria [18]. Consequently, we decided to assess risk of bias as part of Supplementary Table 7, according to criteria that are suitable to the heterogeneity currently inherent to the field. These criteria include the risks of bias described above, as well as an assessment of generalizability, replicability, and validity, which are standard indicators of the quality of a study. Risk of bias was independently assessed by two reviewers and disagreements were resolved by discussion.

Data synthesis

Given the discussed characteristics of the field, as well as the broad range of details covered by the tables, we anticipate a thorough discussion of all the deficiencies and inconsistencies that future research should address. Therefore, we summarize the data in narrative form, following the structure provided by the features summarized in each table. Although a meta-analysis is beyond scope at the current stage of the field, we do report outcome measures in a comparative manner when possible.

Confidence in cumulative evidence

We will assess accuracy of prognostic and diagnostic tools, rather than confidence in an intervention. Hence, we will not be drawing any conclusions related to treatment implementation.

Background on AI, cognitive tests, and databases

This section briefly defines key terminology and abbreviations referring and offers a taxonomy of features, adapted from Voleti et al. [20], to enhance the readability of the systematic review tables. This section also briefly describes the most commonly used databases and neuropsychological assessments, with the intention of making these accessible for the reader.

AI, machine learning, and speech technologies

AI can be loosely defined as a field of research that studies artificial computational systems that are capable of exhibiting human-like abilities or human level performance in complex tasks. While the field encompasses a variety of symbol manipulation systems and manual encoding of expert knowledge, the majority of methods and techniques employed by the studies reviewed here concern machine learning methods. While machine learning dates back to the 1950s, the term "machine learning" as it is used today, originated within the AI community in the late 1970s to designate a number of techniques designed to automate the process of knowledge acquisition. Theoretical developments in computational learning theory and the resurgence of connectionism in the 1980s helped consolidate the field, which incorporated elements of signal processing, information theory, statistics, and probabilistic inference, as well as inspiration from a number of disciplines.

The general architecture of a machine learning system as used in AD prediction based on speech and language can be described in terms of the learning task, data representation, learning algorithm, nature of the "training data", and performance measures. The learning task concerns the specification of the function to be learned by the system. In this review, such functions include classification (for instance, the mapping of a voice or textual sample from a patient to a target category such as "probable AD", "MCI", or "healthy control") and regression tasks (such as mapping the same kind of input to a numerical score, such as a neuropsychological test score). The data representation defines which features of the vocal or linguistic input will be used in the mapping of that input to the target category or value, and how these features will be formally encoded. Much research in machine learning applied to this and other areas focuses on data representation. A taxonomy of features used in the papers reviewed here is presented in Table 1. There is a large variety of learning algorithms available to the practitioner, and a number of them have been employed in AD research. These range from connectionist systems, of which most "deep learning" architectures are examples, to relatively simple linear classifiers such as naïve Bayes and logistic regression, to algorithms that produce interpretable outputs in the form of

Table 1
Feature taxonomy, adapted from Voleti et al. [20]

Category	Subcategory	Feature type	Feature name, abbreviation, reference
Text-based (NLP)	Lexical features	Bag of words, vocabulary analysis	*BoW*, *Vocab*.
		Linguistic Inquiry and Word Count	*LIWC* [21]
		Lexical diversity	Type-Token Ratio (*TTR*), Moving Average TTR (*MATTR*), Simpson's Diversity Index (*SDI*) Brunét's Index (*BI*), Honoré's Statistic (*HS*).
		Lexical Density	Content density (*CD*), Idea Density (*ID*), *P*-Density (*PD*).
		Part-of-Speech tagging	*PoS*.
	Syntactical features	Constituency-based parse tree scores	*Yngve* [22], *Frazier* [23].
		Dependency-based parse tree scores	
		Speech graph	Speech Graph Attributes (*SGA*).
	Semantic features	Matrix decomposition methods	Latent Semantic Analysis (*LSA*), Principal Component Analysys (PCA).
	(Word and sentence embeddings)	Neural word/sentence embeddings	*word2vec* [24]
		Topic modelling	*Latent Dirichlet Allocation* [25].
		Psycholinguistics	Reliance on familiar words (*PsyLing*).
	Pragmatics	Sentiment analysis	*Sent*.
		Use of language *UoL*	Pronouns, paraphrasing, filler words (*FW).*
		Coherence	*Coh*.
Acoustic	Prosodic features	Temporal	Pause rate (*PR*), Phonation rate (*PhR*), Speech rate (*SR*), Articulation rate (*AR*). Vocalization events.
		Fundamental Frequency	F_0 and trajectory.
		Loudness and energy	*loud*, *E*.
		Emotional content	*emo*.
	Spectral features	Formant trajectories	F_1, F_2, F_3.
		Mel Frequency Cepstral Coefficients	*MFCCs* [26].
	Vocal quality	Jitter, Shimmer, harmonic-to-noise ratio	*jitt*, *shimm*, *HNR*.
	ASR-related	Filled pauses, repetitions, dysfluencies, hesitations. fractal dimension, entropy.	*FP, rep, dys, hes, FD, entr.*
		Dialogue features (i.e., Turn-Taking)	*TT*:avg turn length, inter-turn silences.

decision trees or logical expressions, to ensembles of classifiers and boosting methods. The nature of the training data affects both its representation and the choice of algorithm. Usually, in AD research, patient data are annotated with labels for the target category (e.g., "AD", "control") or numerical scores. Machine learning algorithms that make use of such annotated data for induction of models are said to perform supervised learning, while learning that seeks to structure unannotated data is called unsupervised learning. Performance measures, and by extension the loss function with respect to which the learning algorithm attempts to optimize, usually depend on the application. Commonly used performance measures are accuracy, sensitivity (also known as recall), specificity, positive predictive value (also known as precision), and summary measures of trade-offs between these measures, such as area under the receiver operating characteristic curve and F scores. These methods and metrics are further detailed below.

Cognitive tests

This is a brief description of the traditional cognitive tests (as opposed to speech-based cognitive tasks) most commonly applied in this field, with two main purposes. On the one hand, neuropsychological assessments are one of the several factors on which clinicians rely in order to make a clinical diagnosis, which in turn results on participants being assigned to an experimental group (i.e., healthy control, SCI, MCI, or AD). On the other hand, some of these tests are recurrently used as part of the speech elicitation protocols.

Batteries used for diagnostic purposes consist of reliable and systematically validated assessment tools that evaluate a range of cognitive abilities. They are specifically designed for dementia and aimed to be time-efficient, as well as able to highlight preserved and impaired abilities. The most commonly used batteries are the Mini-Mental State Examination (MMSE) [27], the Montreal Cognitive Assessment (MoCA) [28], the Hierarchical Dementia Scale-Revised (HDS-R) [29], the Clinical Dementia Rating (CDR) [30], the Clock Drawing Test (CDT) [31], the Alzheimer's disease Assessment Scale, Cognitive part (ADAS-Cog) [32], the Protocol for an Optimal Neurpsychological Evaluation (PENO, in French) [33], or the General Practitioner Assessment of Cognition (GPCog) [34]. Most of these tests have been translated into different languages, such as the Spanish version of the MMSE (MEC) [35], which is used in a few reviewed papers.

Tools measuring general functioning, such as the General Deterioration Scale (GDS) [36] or Activities of Daily Living, such as the Katz Index [37] and the Lawton Scale [38], are also commonly used. Based on the results of these tests, clinicians usually proceed to diagnose MCI, following Petersen's criteria [39], or AD, following NINCDS-ADRDA criteria [40]. Alternative diagnoses appear in some texts, such as Functional Memory Disorder (FMD), following [41]'s criteria.

Speech elicitation protocols often include tasks extracted from examinations that were originally designed for aphasia, such as fluency tasks. Semantic verbal fluency tasks (SVF, in COWAT) [42] and are often known as "animal naming" because they require the participant generating a list of nouns from a certain category (e.g., animals) while being recorded. Another tool recycled from aphasia examinations is the Cookie Theft Picture task [43], which requires participants to describe a picture depicting a dynamic scene, and hence to also elaborate a short story. Although that is by far the most common picture used in such tests, other pictures have also been designed to elicit speech in a similar way (e.g., [44]).

Another group of tests consists, essentially, of language sub-tests (i.e., vocabulary) and immediate/delayed recall tests, extracted from batteries to measure intelligence and cognitive abilities, such as the Wechsler Adult Intelligence Scale (WAIS-III) [45] or the Wechsler Memory Scale (WMS-III) [46], respectively. Besides, the National Adult Reading Test (NART) [47], the Arizona Battery for Communication Disorders of Dementia ABCD battery (ABCD) [48], the Grandfather Passage [49] and a passage of *The Little Prince* [50] are also used to elicit speech in some articles.

Databases

Although types of data will be further discussed later, we hereby give an overview of the main datasets described. For space reasons, we only mention here those datasets which have been used in more than one study, and for which a requesting procedure might be available. For monologue data:

- *Pitt Corpus:* By far the most commonly used. It consists of picture descriptions elicited by the Cookie Theft Picture, generated by healthy participants and patients with probable AD, and linked to their neuropsychological data (i.e., MMSE). It was collected by the University of Pittsburgh [51] and distributed through DementiaBank [52].
- *BEA Hungarian Dataset:* This is a phonetic database, containing over 250 hours of multipurpose Hungarian spontaneous speech. It was collected by the Research Institute for Linguistics at the Hungarian Academy of Sciences [53] and distributed through META-SHARE.
- *Gothenburgh MCI database:* This includes comprehensive assessments of young elderly participants during their Memory Clinic appointments and senior citizens that were recruited as their healthy counterparts [54]. Speech research undertaken with this dataset uses the Cookie Theft picture description and reading tasks subsets, all recorded in Swedish.

For dialogue data, the *Carolina Conversations Collection (CCC)* is the only available database. It consists of conversations between healthcare professionals and patients suffering from a chronic disease, including AD. For dementia research, participants are assigned to an AD group or a non-AD group, if their chronic condition is unrelated to dementia (i.e., diabetes, heart disease). Conversations are prompted by questions about their health condition and experience in healthcare. It is collected and distributed by the Medical University of South Carolina [55].

In addition, some of the reviewed articles refer to the *IVA dataset*, which consists of structured interviews undertaken and recorded simultaneously by an Intelligent Virtual Agent (a computer "avatar") [56]. However, the potential availability of this dataset is unknown.

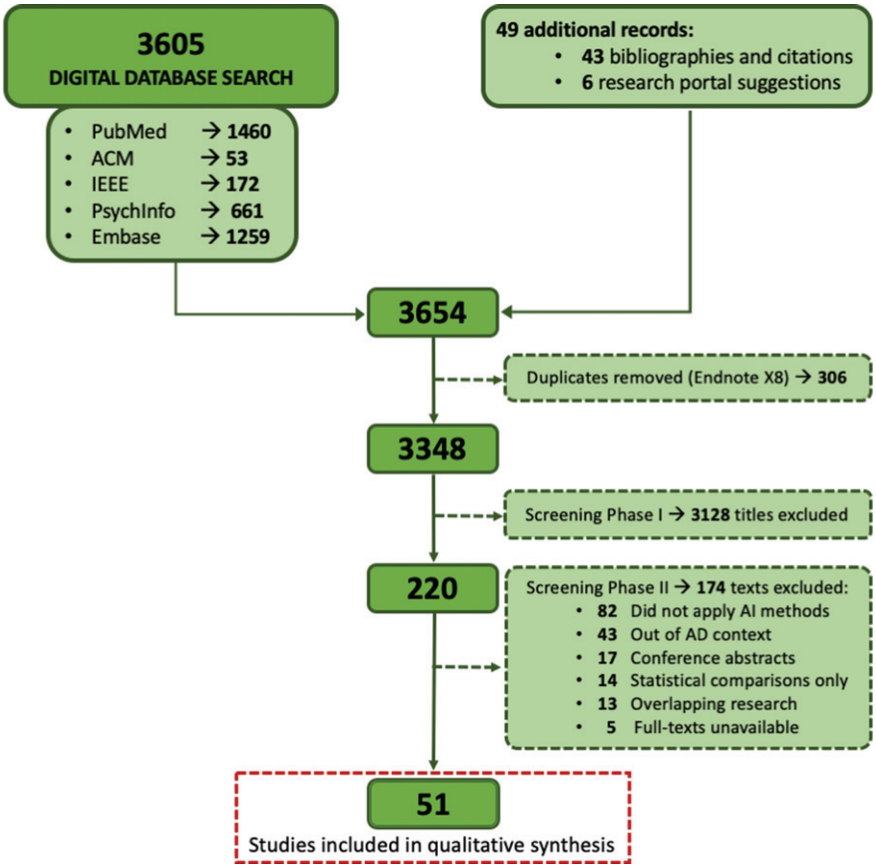
3605
DIGITAL DATABASE SEARCH
• PubMed → 1460
• ACM → 53
• IEEE → 172
• PsychInfo → 661
• Embase → 1259
49 additional records:
• 43 bibliographies and citations
• 6 research portal suggestions
3654
Duplicates removed (Endnote X8) → 306
3348
Screening Phase I → 3128 titles excluded
220
Screening Phase II → 174 texts excluded:
• 82 Did not apply AI methods
• 43 Out of AD context
• 17 Conference abstracts
• 14 Statistical comparisons only
• 13 Overlapping research
• 5 Full-texts unavailable
51
Studies included in qualitative synthesis

Fig. 2. Screening and selection procedure, following guidelines provided by PRISMA [19].

RESULTS

Adding up all digital databases, the searches resulted in 3,605 records. Another 43 papers were identified by searching through bibliographies and citations and 6 through research portal suggestions, adding up to 3,654 papers in total. Of those, 306 duplicates were removed using EndNote X8, leaving 3,348 for the first screening phase. In this first phase, 3,128 papers were excluded based on title and abstract, and therefore 220 reached the second phased of screening. Five of these papers did not have a full-text available, and therefore 215 papers where fully screened. Finally, 51 papers were included in the review (Fig. 2).

Existing literature

The review by [20] is, to our knowledge, the only published work with a comparable aim to the the present review, although there are important scope differences. First of all, the review by Voleti et al. differs from ours in terms of methodological scopes. While their focus was to create a taxonomy for speech and language features, ours was to survey diagnosis and cognitive assessment methods that are used in this field and to assess the extent to which they are successful. In this sense, our search was intentionally broad. There are also differences in the scope of medical applications. Their review studies a much broader range of disorders, from schizophrenia to depression and cognitive decline. Our search, however, targeted cognitive decline in the context of dementia and AD. It is our belief that these reviews complement each other in providing systematic accounts of these emerging fields.

Data extraction

Tables with information extracted from the papers are available as Supplementary Material. There are four different tables: a general table concerning usual clinical features of interest (after the PICOS framework), and three more specific tables concerning data details, methodology details, and implications for clinicians and researchers. Certain conventions and acronyms were adopted when extracting article information, and should be considered when interpreting the information contained on those tables. These conventions are available in the Supplementary Material, prior to the tables.

DISCUSSION

In this section, the data and outcomes of the different tables are synthesized in different subsections and put into perspective. Consistent patterns and exceptions are outlined. Descriptive aspects are organized by column names, following table order and referencing their corresponding table in brackets.

Study aim and design (Supplementary Table 4: SPICMO)

Most of the reviewed articles aim to use acoustic and/or linguistic features in order to distinguish the speech produced by healthy participants from the one produced by participants with a certain degree of cognitive impairment. The majority of studies attempt binary models to detecting AD and, less often, MCI, in comparison to HC. A few studies also attempt to distinguish between MCI an AD. Even when the dataset contains three or four groups (e.g., HC, SCI, MCI, AD), most studies only report pairwise group comparisons [57–61]. Out of 51 reviewed papers, only seven did attempt three-way [50, 62–64] or four-way [12, 65, 66] classification. Their results are inconclusive and present potential biases related to the quality of the datasets (i.e., low accuracy on balanced datasets, or high accuracy on imbalanced datasets).

Slightly different objectives are described by [67], the only study predicting conversion from MCI to AD, and by [68], the only study predicting progression from HC to any form of cognitive impairment. While these studies also learned classifiers to detect differences between groups, they differ from other studies in that they use longitudinal data. There is only one article with a different aim than classification. This is the study by Duong et al. [69], who attempt to describe AD and HC discourse patterns through cluster analysis.

Despite many titles mentioning cognitive monitoring, most research addresses only the presence or absence of cognitive impairments (41, out of 51 papers). Outside of those, seven papers are concerned with three or four disease stages [12, 50, 62–66], two explore longitudinal cognitive changes (although still through binary classification) [67, 68], and one describes discourse patterns [69]. We note that future research could take further advantage of this longitudinal aspect to build models able to generate a score reflecting risk of developing an impairment.

Population (Supplementary Table 4: SPICMO)

The target population are elderly people who are healthy or exhibit certain signs of cognitive decline related to AD (i.e., SCI, MCI, AD). Demographic information is frequently reported, most commonly age, followed by gender and years of education.

Cognitive scores such as MMSE are often part of the descriptive information provided for study participants as well. This serves group assignment purposes and allows quantitative comparisons of participants' degree of cognitive decline. In certain studies, MMSE is used to calculate the baseline against which classifier performance will be measured [44, 70, 71]. However, despite being widely used in clinical and epidemiological investigations, MMSE has been criticized for having ceiling effects, especially when used to assess pre-clinical AD [72].

Some studies report no demographics [6, 66, 68, 73, 74], only age [10, 75], only age and gender [61, 76, 77], or only age and education [70, 78]. An exception is the dataset AZTIAHORE [79, 80], which contains the youngest healthy group (20–90 years old) and a typical AD group (68–98 years old), introducing potential biases due to this imbalance. Demographic variables are established risk factors for AD [81], therefore demographics reporting is essential for this type of study.

Interventions (Supplementary Table 4: SPICMO)

Study interventions almost invariably consist of a speech generation task preceded by a health assessment. This varies between general clinical assessments, including medical and neurological examinations, and specific cognitive testing. The comparison groups are based on diagnosis groups, which in turn are established with the results of such assessments. Therefore, papers lacking that information do not specify their criteria for group assignment [60, 61, 75, 79, 80, 82–85]. This could be problematic, since the field currently revolves around diagnostic categories, trying to identify such categories through speech data. Consequently, one should ensure that standard criteria have been used and that models are accurately tuned to these categories.

Speech tasks are sometimes part of the health assessment. For instance, speech data are often recorded during the language sub-test of a neuropsychological battery (e.g., verbal fluency, story recall, or picture description tasks). Another example of speech generated within clinical assessment is the recording of patient-doctor consultations [8, 85, 86] of cognitive examinations (e.g., MMSE [83]). There are also studies where participants are required to perform language tests outwith the health assessment, for speech elicitation purposes only. Exceptionally, two of these studies work with written rather than spoken language [87, 88]. Alternative tasks for this purpose are reading text passages aloud (e.g., [89]), recalling short films (e.g., [63]), retelling a story (e.g., [90]), retelling a day or a dream (e.g., [91]), or taking part in a semi-standardized (e.g., [68]) or conversational (e.g., [10]) interview.

Most of these are examples of constrained, laboratory-based interventions, which seldom include spontaneously generated language. There are advantages to collecting speech under these conditions, such as ease of standardization, better control over potential confunding factors, and focus on high cognitive load tasks that may be more likely to elicit cognitive deficits. However, analysis of spontaneous speech production and natural conversations also has advantages. Spontaneous and conversational data can be captured in natural settings over time, thus mitigating problems that might affect performance in controlled, cross-sectional data, such as a participant having an “off day” or having slept poorly the night before the test.

Comparison groups (Supplementary Table 4: SPICMO)

This review targets cognitive decline in the context of AD. For its purpose, nomenclature heterogeneity has been homogenized into four consistent groups: HC, SCI, MCI, and AD; with an additional group, CI, to account for unspecified impairment (see Supplementary Table 1). As an exception to this nomenclature are Mirheidari et al. [8, 12, 86], who compare participants with an impairment caused by neurodegenerative disease (ND group, including AD) to an impairment caused by functional memory disoders (FMD); and Weiner and Schultz [68] and Weiner et al. [74], who introduce a category called age-associated cognitive decline (AACD).

Furthermore, some studies add subdivisions to these categories. For instance, there are two studies that classify different stages within the AD group [79, 80]. Another study divides the MCI group between amnesic single domain (aMCI) and amnesic multiple domain (a+mdMCI), although classification results for two groups are not very promising [57]. Within-subject comparisons have also been attempted, comparing participants who remained in a certain cognitive status to those who changed [67, 74].

Most studies target populations where a cohort has already been diagnosed with AD or a related condition, looking for speech differences between those and healthy cohorts. Therefore, little insight is offered into pre-clinical stages of the disease.

Outcomes of interest (Supplementary Table 4: SPICMO)

Given the variety of diagnostic categories and types of data and features used, it is not easy to establish state-of-the-art performance. For binary classification, the most commonly attempted task, the reported performance ranges widely depending in the data use, the recording conditions, and the variables used in modelling. For instance Lopez-de Ipiña et al. [80] reported an accuracy that varied between 60% and 93.79% using only acoustic features that were generated *ad hoc*. Although the second figure is very promising, their dataset is small, 40 participants, and remarkably imbalanced in terms of both diagnostic class and age. In terms of class, even though they initially report 20 AD and 20 HC, the AD group is divided in three different severity stages, with 4, 10, and 6 participants, respectively, whereas the control group remains unchanged (20). In terms of age, 25% percent of their healthy controls fall within a 20–60 years old age range, while 100% of the AD group are over 60 years old. In contrast, Haider et al. [11] reported 78.7% accuracy, using also acoustic features only, but generated from standard feature sets that had been developed for computational paralinguistics. Besides, this figure appears as more robust because the dataset is much larger (164 participants) and it is balanced for class, age and gender, as well as audio enhanced. Guo et al. [92] obtained 85.4% accuracy on the same dataset as [11], but using text-based features only and without establishing class, age, or gender balance. All the figures quoted so far refer to monologue studies. The state-of-the-art accuracy for dialogue data is 86.6%, obtained by Luz et al. [10] using acoustic features only.

Regarding other classification experiments, we see that Mirzaei et al. [50] reports 62% for a 3-way classification, discriminating HC, MCI, and AD. They are also among the few to appropriately report accuracy, since they work with a class-balanced dataset, while many other studies report overall accuracy in class-imbalanced datasets. Accuracy figures can be very misleading in the presence of class imbalance. A trivial rejector (i.e., a classifier that trivially classifiers all instances as negative with respect to a class of interest), would achieve very high accuracy on a dataset that contained, say, 90% negative instances. For example, Nasrolahzadeh et al. [65] report really high accuracy with a 4-way classifier, 97.71%, but in a highly imbalanced dataset. However, Mirheidari et al. [12] reported 62% accuracy and 0.815 AUC for a 4-way classifier in a slightly more balanced dataset and Thomas et al. [66], also 4-way, only 50%, on four groups of MMSE scores. Other studies attempting 3-way classification experiments in balanced datasets are Egas López et al. [62], 56% and Gosztolya et al. [63] with 66.7%. Kato et al. [64], however, reports 85.4% 3-way accuracy in an imbalanced dataset.

These results are diverse, and it stands clear that some will lead to more robust conclusions than others. Notwithstanding, numerical outcomes are always subject to the science behind them, the quality of the datasets and the rigor of the method. This disparity of results therefore highlights the need for improved standards of reporting in this kind of study. Reported results should include metrics that allow the reader to assess the trade-off between false positives and false negatives in classification, such as specificity, sensitivity, fallout, and F scores, as well measures that are less sensitive to class imbalance, widely used in other applications of computational paralinguistics, such as unweighted average recall. Contingency tables and ROC curves should also be provided whenever possible. Given the difficulties in reporting, comparing and differentiating the results for the 51 reviewed studies on an equal footing, we refer the reader to Supplementary Tables 4 and 6.

Size of dataset or subset (Supplementary Table 5: Data Details)

Within a machine learning context, all the reviewed studies use relatively small datasets. About 31% train their models with less than 50 participants [8, 10, 50, 64, 68, 71, 79, 80, 82, 83, 85, 86, 89, 91, 93, 94], while only 27% have 100 or more participants [9, 11, 57, 60, 67, 70, 73, 75–77, 92, 95–97]. In fact, 5 report samples with less than 30 participants [68, 79, 83, 89, 94].

It is worth noting that those figures represent the dataset size in full, which is then divided in two, three or four groups, most of the times unevenly. There are only 6 studies where not only the dataset, but also each experimental group contains 100 or more participants/speech samples [6, 9, 11, 85, 92, 95]. All of these studies used the *Pitt Corpus*.

The *Pitt Corpus* is the largest dataset available. It is used in full by Ben Ammar and Ben Ayed [95], and contains 484 speech samples, although it is not clear to how many unique participants these samples belong. With the same dataset, Luz [6] reports 398 speech samples, but again, no number of unique participants. However, another study working with the *Pitt Corpus* does report 473 speech samples from 264 participants [9]. It is important for studies to report numbers of unique participants in order to allow the reader to assess the risk that the machine learning models might actually be simply learning to recognize participants rather than their underlying cognitive status. This risk can be mitigated, for example, by ensuring that all samples from each participant are in either the training set or the test set, but not both.

Data type (Supplementary Table 5: Data Details)

This column refers to the data used in each reviewed study, indicating if these data consist of monologues or dialogues, purposefully elicited narratives or speech obtained through a cognitive test. It also includes whether data was recorded or recorded and transcribed, and how this transcription was done (i.e., manual or automatic).

Of the reviewed studies, 82% used monologue data, and most of them (36) obtained speech through a picture description task (e.g., *Pitt Corpus*). These are considered relatively spontaneous speech samples, since participants may describe the picture in whichever way they want, although the speech content is always constrained. Among other monologue studies, eight work with speech obtained through cognitive tests, frequently verbal fluency tasks. Only two papers rely on truly spontaneous and natural monologues, prompted with an open question instead of a picture description [60, 65].

Dialogue data are present less frequently, in 27% of the studies, and elicited more heterogeneously.

For instance, in structured dialogues (4 studies), both speakers (i.e., patient and professional) are often recorded while taking a cognitive test [8, 12, 83, 94]. Semi-structured dialogues (5 studies) are interview-type conversations where questions are roughly even across participants. From our point of view, the most desirable data type are conversational dialogues (5 studies), where interactive speech is prompted with the least possible constraints [10, 66, 79, 80, 98]. A few studies have collected dialogue data through an intelligent virtual agent (IVA) [8, 12, 94] showing the potential for data to be collected remotely, led by an automated computer system.

In terms of data modalities (e.g., audio, text, or both), two studies are the exception where data was directly collected as written text [87, 88]. A few studies (6) work with audio files and associated ASR transcriptions [12, 44, 62, 63, 77, 99]. Another group of studies (14) use solely voice recordings [50, 57, 60, 61, 64, 65, 71, 79, 80, 82, 84, 89, 97, 100]. More than half of the studies (55%) rely, at least partially, on manually transcribed data. This is positive for data sharing purposes, since manual transcriptions are usually considered golden standard quality data. However, methods that rely on transcribed speech may have limited practical applicability, as they are costly and time-consuming, and often (as when ASR is used) require error prone (see section on pre-processing below) intermediate steps compared to working directly with the audio recordings.

Other modalities (Supplementary Table 5: Data Details)

The most frequently encountered data modality, apart from speech and language, is structured data related to cognitive examinations, largely dominated by MMSE and verbal fluency scores. Another modality is video, which is available in some datasets such as CCC [10, 98], AZTITXIKI [79], AZTIAHORE [60, 80], IVA [12, 85], or the one in Tanaka et al. [94], although it is not included in their analysis. Other analyzed modalities include neuroimaging data, such as MRI [67] and fNIRS [64], eye-tracking [7, 94], or gait information [71].

In order to develop successful prediction models for pre-clinical populations, it is likely that future interactive AI studies will begin to include demographic information, biomarker data, and lifestyle risk factors [101].

Data annotation (Supplementary Table 5: Data Details)

Group labels and sizes are presented in this section of the Data Details table, the aim of which is to give information about the available speech datasets. Accordingly, labels remain as they are reported in each study, as opposed to the way in which we homogenized them to describe Comparison Groups in Supplementary Table 4. In other words, even though the majority of studies annotate their groups as HC, SCI, MCI, and AD, some do not. For example, the HC group is labelled as CON (control) [91], NC (normal cognition) [57, 64, 88, 99], CH (cognitively healthy) [82], and CN (cognitively normal) [67]. SCI can also be named SMC [96], and there is a similar but different category (AACD) reported in two other studies [68, 74]. MCI and AD are more homogeneous due to being diagnostic categories that need to meet certain clinical criteria to be assigned, although some studies do refer to AD as *dementia* [62, 95]. Another heterogeneous category is CI (i.e., unspecified cognitive impairment), which is annotated as *low* or *high* MMSE scores [93], or as *mild dementia* [89]. *Mild dementia* may sound similar to MCI, however the study did not report diagnostic criteria for MCI to be considered.

This section offers insight into another aspect in which lack of consensus and uniformity is obvious. Using accurate terminology (i.e., abiding by diagnosis categories) when referring to each of these groups could help establish the relevance of this kind of research to clinical audiences.

Data balance (Supplementary Table 5: Data Details)

Only 39% (20) of the reviewed studies present class balance, that is, the number of participants is evenly distributed across the two, three, or four diagnostic categories [7, 8, 11, 50, 60, 62–64, 75, 78, 82, 84, 86, 88–91, 94, 95, 98]. Among these 20 studies, one reports only between-class age and gender balance [94]; another one reports class balance, within-class gender balance, and between-class age and gender balance [11]. A few report balance for all features except for within-class gender balance, which is not specified [62, 63, 88]. Lastly, there is only one study that, apart from class balance, also reports gender balance within and between classes, as well as age and education balance between classes [87]. Surprisingly, nine other studies fail to report one or more demographic aspects.

Sometimes gender is reported per dataset, but not per class (e.g., [93]), and therefore not accounted for in the analysis, even though is one of the main risk factors [81]. Often, *p*-values are appropriately presented to indicate that demographics are balanced between groups (e.g., [62]). Unfortunately, almost as often, no statistical values are reported to argue for balance between groups (e.g., [83]). There are also cases where where the text reports demographic balance but neither group distributions nor statistical tests are presented (e.g., [91]). Another aspect to take into account is the differences between raw and pre-processed data. For instance, Lopez-de Ipiña et al. [79, 80] describe a dataset where 20% of the HC speech data, but 80% of the AD speech data, is removed during pre-processing. Hence, even if these datasets had been balanced before (they were not) they will definitely not be balanced after pre-processing has taken place.

It is also worth discussing the reasons behind participant class imbalance when the same groups are class balanced in terms of samples. Fraser et al. [9], for example, work with a subset of the *Pitt Corpus* of 97 HC participants and 176 AD participants; however, the number of samples is 233 and 240, respectively. Similar patterns apply to other studies where the number of participants and samples are reported [92, 98]. Did HC come for more visits, or did perhaps AD participants fail to come to later visits or drop out of the study? These incongruities could be hiding systematic group biases.

Conclusions drawn from imbalanced data are subject to a greater probability of bias, especially in small datasets. For example, certain performance metrics to evaluate classifiers are more robust (e.g., *F1*) than others (e.g., *acc*) against this imbalance. Accordingly, in this table, the smaller the dataset, the more strict we have been when evaluating the balance of its features. Moving forward, it is desirable that more emphasis is placed on data balance, not only in terms of group distribution, but also in terms of those demographic features established risk factors (i.e., age, gender, and years of education).

Data availability (Supplementary Table 5: Data Details)

Strikingly, very few studies make their data available, or even report on its (un)availability, even when using available data hosted by a different institution (e.g., studies using the *Pitt Corpus*). The majority (77%, 39 studies) fail to report on data availability. From the remaining 12 studies, nine use data from DementiaBank (*Pitt Corpus* or *Mandarin_Lu*) and do report data origin and availability. However, only [75, 90] share the exact specification of the subset of *Pitt Corpus* used for their analysis, in order for other researchers to be able to replicate their findings, taking advantage of the availability of the corpus. The same applies to Luz et al. [10], who made available their identifiers for the CCC dataset. One other study, Fraser et al. [7], mentions that data are available upon request to authors.

Haider et al. [11], one of the studies working on the *Pitt Corpus*, has released their subset as part of a challenge for INTERSPEECH 2020, providing the research community with a dataset matched for age and gender and with enhanced audio. In such an emerging and heterogeneous field, shared tasks and data availability are important progression avenues.

Language (Supplementary Table 5: Data Details)

As expected, a number of studies (41%) were conducted through English. However, there is a fair amount of papers using data in a variety of languages, including: Italian [91], Portuguese [57, 90], Chinese and Taiwanese [82], French [50, 69, 77, 96, 102], Hungarian [62, 63, 99], Spanish [83, 89, 100], Swedish [7, 59, 87], Japanese [64, 71, 94], Turkish [85], Persian [65], Greek [61, 88], German [68, 74], or reported as multilingual [60, 79, 80].

This is essential if screening methodologies for AD are to be implemented worldwide [103]. The main caveat, however, is not the number of studies conducted in a particular language, but the fact that most of the studies conducted in languages other than English do not report on data availability. As mentioned, only Dos Santos et al. [90] and Fraser et al. [7] report their data being accessible upon request, and Chien et al. [82] works with data available from DementiaBank. For speech-based methodology aimed at AD detection, it would be a helpful practice to make these data available, so that other groups are able to increase the amount of research done in any given language.

Pre-processing (Supplementary Table 6: Methodology)

Pre-processing includes the steps for data preparation prior to data analysis. It is essential to determine in which shape any given data is introduced in the analysis pipeline, and therefore, the outcome of it. However, surprisingly little detail is reported in the reviewed studies.

Regarding text data, the main pre-processing procedure is transcription. Transcription may happen manually or through ASR. The Kaldi speech recognition toolkit [104], for instance, was used in several recent papers (e.g., [12, 62]). Where not specified, manual transcription is assumed. Although many ASR approaches do extract information on word content (e.g., [8, 44, 71, 85, 86, 96]), some focus on temporal features, which are content-independent (e.g., [63, 82]). Some studies report their transcription unit, that is, word-level transcription (e.g., [9]), phone-level transcription (e.g., [63]) or utterance-level transcription (e.g., [91]). Further text pre-processing involves tokenization [73, 82, 90, 94], lemmatization [87], and removal of stopwords and punctuation [87, 90]. Depending on the research question, dysfluencies are also removed (e.g., [87, 90]), or annotated as relevant for subsequent analysis (e.g., [59]).

Currently, commercial ASRs are optimized to minimize errors at word level, and therefore not ideal for generating non-verbal acoustic features. Besides, it seems that AD patients are more likely to generate ungrammatical sentences, incorrect inflections and other subtleties that are not well handled by such ASR systems. In spite of this, only a few papers, by the same research group, rely on ASR and report WER (word error rate), DER (diarisation eror rate), or WDER (word diarisation error rate) [8, 12, 85]. It is becoming increasingly obvious that off-the-shelf ASR tools are not readily prepared for dementia research, and therefore some reviewed studies developed their own custom ASR systems [44, 63].

Regarding acoustic data, pre-processing is rarely reported outside the audio files being put through an ASR. When reported, it mainly involves speech-silence segmentation with voice activity deteciton algorithms (VAD), including segment length and the acoustic criterion chosen for segmentation thresholds (i.e., intensity) [6, 11, 44, 50, 60, 61, 64, 65, 68, 76, 79, 80, 96, 102]. It should also include any audio enhancement procedures, such as volume normalization or removal of background noise, only reported in Haider et al. [11] and Sadeghian et al. [44].

We concluded from the reviewed papers that it is not common practice for authors in this field to give a complete account of the data pre-processing procedures they followed. As these procedures are crucial to reliability and replicability of results, we recommend that further research specify these procedures more thoroughly.

Feature generation (Supplementary Table 6: Methodology)

Generated speech features are divided into two main groups, text-based and acoustic features, and follow the taxonomy presented in Table 1. Some studies work with multimodal feature sets, including images [94] and gait [71] measurements.

Text-based features comprise a range of NLP elements, commonly a subset consisting of lexical and syntactical indices such as type-token ratio (*TTR*), *idea density* or *Yngve* and *Frazier* indices. *TTR* is a measure of lexical complexity, calculated by taking the total number of unique words, also called lexical items (i.e., types) and dividing by the total number of words (i.e., tokens) in a given language instance [105]. *Idea density* is the number of ideas expressed in a given language instance, with 'ideas' understood as new information and adequate use of complex propositions. High early *idea density* seems to be a lower risk predictor for developing AD later in life, whereas lower idea density appears associated with brain atrophy [106]. *Yngve* [22] and *Frazier* [23] scores indicate syntactical complexity by calculating the depth of the parse tree that results from the grammatical analysis of a given language instance. Both indices have been associated with working memory [107] and showed a declining pattern in the longitudinal analysis of the written work by Iris Murdoch, a novelist who was diagnosed with AD [108].

In some studies, the research question targets a specific aspect of language, such as syntactical complexity [59], or a particular way of representing it, such as speech graph attributes [57]. Fraser et al. [9] present a more comprehensive feature set, including some acoustic features. Similar to Fraser et al. [9], although less comprehensive, a few other studies combine text-based and acoustic features [8, 44, 71, 78, 86, 89, 91, 92, 94, 96]. However, most published research is specific to one type of data or another.

The most commonly studied acoustic features are prosodic temporal features, which are almost invariably reported, followed by ASR-related features, specifically pause patterns. There is also focus on spectral features (features of the frequency domain representation of the speech signal obtained through application of the Fourier transform), which include *MFCCs* [62]. The most comprehensive studies include spectral, ASR-related, prosodic temporal, voice quality features [8, 50, 60, 79, 84, 92, 100], as well as features derived from the Higuchi Fractal Dimension [80] or from higher order spectral analysis [65]. It is worth noting here that Tanaka et al. [94] extract F_0's coefficient of variation per utterance. The decision to not extract F_0's mean and SD was due to their association with individual differences and sex. Similarly, Gonzalez-Moreira et al. [89] report F_0 and functionals in semitones, because research argues that using semitones to express F_0 reduces gender differences [109], which is corroborated by the choice of semitones in the standardized eGeMAPS [11].

Studies using spoken dialogue recordings extract turn-taking patterns, vocalization instances, and speech rate [10, 94]. Those focusing on transcribed dialogues also extract turn-taking patterns, as well as dysfluencies [8, 12, 86]. Guinn et al. [98] work with longitudinal dialogue data but do not extract specific dialogue or longitudinal features.

With regards to feature selection, 30% of the studies do not report feature selection procedures. Among those that do, the majority (another 30%) report using a filter approach based on a statistical index of feature differences between classes, such as *p*-values, Cohen's *d*, *AUC*, or *Pearson's* correlation. Others rely on wrapper methods [50], RFE [8, 86], filter methods based on information gain [65, 95], PCA [64], best first greedy algorithm [44], and cross-validation, seeking through the iterations for which feature type contributes more to the classification model [80].

Despite certain similarities and a few features being common to most acoustic works (i.e., prosodic temporal), there is striking heterogeneity among studies. Since they usually obtain features using *ad hoc* procedures, these studies are seldom comparable, making it difficult to ascertain the state-of-the-art in terms of performance, as pointed out before, and assess further research avenues. However, this state of affairs may be starting to change as the field matures. Haider et al. [11], for instance, chose to employ standardized feature sets (i.e., ComPare, eGeMAPS, emobase) obtained through formalized procedures [110] which are extensively documented and can be easily replicated. Furthermore, one of these feature sets, eGeMAPS, was developed specifically to target affective speech and underlying physiological processes. Utilizing theoretically informed, standardized feature sets increases the reliability of a study, since the same features have been previously applied (and can continue to be applied) to other engineering tasks, always extracted in the exact same way. Likewise, we argue that creating and utilizing standardized feature sets will improve this field by allowing cross-study comparisons. Additionally, we recommend that the approach to feature generation should be more consistently reported to enhance study replicability and generalizability.

Machine learning task/method (Supplementary Table 6: Methodology)

Most reviewed papers employ supervised learning, except for a study that uses cluster analysis to investigate distinctive discourse patterns among participants [69].

As regards choice of machine learning methods, very few papers report the use of artificial neural networks [91, 95], recurrent neural networks [82], multi-layer perceptron [44, 67, 79, 80, 86], or convolutional neural networks [60, 85]. This is probably due to the fact that most datasets are relatively small, and these methods require large amounts of data. Rather, most studies use several conventional machine learning classifiers, most commonly SVM, NB, RF, and *k*-NN and then compare their performance. Although these comparisons must be assessed cautiously, a clear pattern seems to emerge with SVM consistently outperforming other classifiers.

Cognitive scores, particularly MMSE, are available with many datasets, including the most commonly used, *Pitt Corpus*. However, these scores mostly remain unused except for diagnostic group assignments, or more rarely, as baseline performance [44, 70, 71], in studies that conclude that MMSE is not more informative than speech based features. All supervised learning approaches work toward classification and no regression over cognitive scores is attempted. We regard this as a gap that could be explored in future research.

It is worth noting, however, that some attempts at prediction of MMSE score have been presented in workshops and computer science conferences that are not indexed in the larger biobliography databases. These approaches achieved some degree of success. Linz et al. [111], for instance, trained a regression model that used the SVF to predict MMSE scores and obtained a mean absolute error of 2.2. A few other works used the *Pitt Corpus* for similar purposes, such as Al-Hameed et al. [112], who extracted 811 acoustic features to build a regression model able to predict MMSE scores with an average mean absolute error of 3.1; or Pou-Prom and Rudzicz [113], who used a multiview embedding to capture different levels of cognitive impairment and achieved a mean absolute error of 3.42 in the regression task. Another publication with the *Pitt Corpus* is authored by Yancheva et al. [114], who extracted a more comprehensive feature set, including lexicosyntactic, acoustic, and semantic measures, and used them to predict MMSE scores. They trained a dynamic Bayes network that modeled the longitudinal progression observed on these features and MMSE over time, reporting a mean absolute error of 3.83. This is, actually, one of the very few works attempting a progression analysis over longitudinal data.

Evaluation techniques (Supplementary Table 6: Methodology)

A substantial proportion of studies (43%) do not present a baseline against which study results can be compared. Among the remaining papers, a few set specific results from a comparable work as their baseline [6, 65] or from their own previous work [75]. Others calculate their baseline by training a classifier with all the generated features, that is, before attempting to reduce the feature set with either selection or extraction methods [83, 95, 99], with cognitive scores only [7, 44, 70, 71] or by training a classifier with demographic scores only [63]. Some baseline classifiers are also trained with a set of speech-based features that excludes the feature targeted by the research question. Some examples are studies investigating the potential of topic model features [87], emotional features [79], fractal dimension features [80], higher order spectral features [65], or feature extracted automatically, as opposed to manually [73, 85, 96]. Some studies choose random guess or naive estimations (ZeroR) [10, 11, 66, 74, 88] as their baseline performance.

While several performance metrics are often reported, *accuracy* is the most common one. While it seems straightforward to understand a classifier's performance by knowing its *accuracy*, it is not always appropriately informed. Since *accuracy* is not robust against dataset imbalances, it is only appropriate when diagnostic groups are balanced, such as when reported in Roark et al. [78] and Khodabakhsh and Demiroğlu [84]. This is especially problematic for works on imbalanced datasets where accuracy is the only metric reported [9, 12, 44, 60, 66, 77, 83, 92, 93, 100]. Clinically relevant metrics such as *AUC* and *EER* (e.g., [61, 102]), which summarize the rates of false alarms and false negatives, are reported in less than half of the reviewed studies.

Cross-validation (CV) is probably the most established practice for classifier evaluation. It is reported in all papers but five, of which two are not very recent [66, 93], another two do not report CV but report using a hold-out set [82, 91], and only one reports using neither CV nor a hold-out set procedure [95]. There is a fair amount of variation within the CV procedures reported, since datasets are limited and heterogeneous. For example, leave-one-out CV is used in one third of the reviewed papers, as an attempt to mitigate the potential bias caused by using a small dataset. Several other studies choose leave-pair-out CV instead [7, 70, 73, 75, 78, 97], since it produces unbiased estimates for *AUC* and also reduces potential size bias. There is also another research group who attempted to reduce the effects of their imbalance dataset by using stratified CV [68, 74]. Lastly, no studies report hold-out set procedures, except for the two mentioned above, with training/test sets divided at 80/20% and 85/15%, respectively, and another study where the partition percentages are not detailed [97].

There is a potential reporting problem in that many studies do not clearly indicate whether their models' hyper-parameters were optimized on the test set within or outside each fold of the CV. However, CV is generally considered the best method of evaluation when working with small datasets, where held-out set procedures would be even less reliable, since they would involve testing the system on only a few samples. CV is therefore an appropriate choice for the articles reviewed. The lack of systematic model validation on entirely separate datasets, and the poor practice of using accuracy as the single metric in imbalanced datasets, could compromise the generalizability of results in this field. While it is worth noting that the former issue is due to data scarcity, and therefore more difficult to address, a more appropriate selection of performance metrics could be implemented straight away to enhance the robustness of current findings.

Results overview (Supplementary Table 6: Methodology)

Performance varies depending on the metric chosen, the type of data and the classification algorithm used. Hence, it is very difficult to summarize these results. The evaluated classifiers range between 50%

or even lower in some cases, up to over 90% accuracy. However, as we have pointed out, performance figures must be interpreted with caution due to the potential biases introduced by dataset size, dataset imbalances and non standardized *ad hoc* feature generation. Since these biases cannot be fully accounted for and models are hardly comparable to one another, we do not think it is meaningful to further highlight the best performing models. Such comparisons will become more meaningful when all conditions for evaluation can be aligned, such as in the ADReSS challenge [115], which provides a benchmark dataset (balanced and enhanced) and commits to a reliable study comparison.

Further research on the methodology and how different algorithms behave with certain types of data will shed light on why some classifiers perform even worse than random while others are close to perfect. This could simply be because the high performing algorithms were coincidentally tested on 'easy' data (e.g., better quality, simpler structures, very clear diagnoses), but the problem could also be classifier specific and therefore differences would be associated with the choice of algorithm. Understanding this would influence the future viability of this sort of technology.

Research implications (Supplementary Table 7: Clinical applicability)

This section reviews the papers in terms of novelty, replicability, and generalizability, three aspects key to future research.

As regards **novelty**, the newest aspect of each research paper is succinctly presented in the tables. This is often conveyed by the title of an article, although caution must be exercised with regards to how this information is presented. For example, Tröger et al.'s title (2018) reads "Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment", but only when you read the full text does it become clear that such telephone screening has been simulated.

There is often novelty in pilot studies, especially those presenting preliminary results for a new project, hence involving brand new data [80, 91] or tests for a newly developed system or device [60]. Outside of those, assessing novelty in a systematic review over a 20-year span can be complicated what was novel 10 years ago might not be novel since today. For example, 3-way classification entailed novelty in Bertola et al. (2014) [57], as well as 4-way classification did in Thomas et al. (2005) [66] with text data and little success, and later in Nasrolahzadeh et al. (2018) [65] with acoustic data and an improved performance. Given its low frequency and its naturalness, we have chosen to present the use of dialogue data [10, 68, 84, 85, 94, 98] as a novelty relevant for future research. Other examples of novelty consist of automated neuropsychological scoring, either by automating traditional scoring [57] or by generating a new type of score [67, 70].

Methodological novelty is also present. Even though most studies apply standard machine learning classifiers to distinguish between experimental groups, two approaches do stand out: Duong et al.'s (2005) unique use of cluster analysis (a form of unsupervised learning) with some success, and the use of ensemble [67, 90] and cascaded [7] classifiers, with much better results. Some studies present relevant novelty for pre-processing, generating their own custom ASR systems [44, 61, 63, 99], which offers relevant insight about off-the-shelf ASR. While this is based on word accuracy, some of the customized ASR systems are phone-based [63, 99] and seem to work better with speech generated by participants with AD. Another pre-processing novelty is the use of dynamic threshold for pause behavior [76], which could be essential for personalized screening. With regards to feature generation, "active data representation" is a novel method utilized in conjunction with standardized feature sets by Haider et al. [11], who confirmed the feasibility of a useful tool that is open software and readily available (i.e., ComParE, emobase and eGeMAPS). A particularity of certain papers is their focus on emotional response, analyzed from the speech signal [79, 80]. This could be an avenue for future research, since there are other works presenting interesting findings on emotional prosody and AD [116, 117]. Last, but not least, despite the mentioned importance of early detection, most papers do not target early diagnosis, or do it in conjunction with severe AD only (i.e., if the dataset contains participants at different stages). Consequently, Lundholm Fors et al. (2018) [59] introduced a crucial novelty by not only assessing, but actively recruiting and focusing on participants at the pre-clinical stage of the disease (SCI).

Another essential novelty is related to longitudinal aspects of data [68, 77, 97]. The vast majority of studies work on monologue cross-sectional data, although some datasets do include longitudinal information (i.e., each participant has produced several speech samples). This is sometimes discarded, either by treating each sample as a different participant, which generates subject dependence across samples [74]; or by cross-observation averaging, which misses longitudinal information but does not generate this dependence [75, 97]. Other studies successfully used this information to predict change of cognitive status within-subject [68, 118]. Guinn et al. [98] work with longitudinal dialogue data that becomes cross-sectional after pre-processing (i.e., they conglomerate samples by the same participant) and they do not extract specific dialogue features.

The novelty with most clinical potential is, in our view, the inclusion of different types of data, since something as complex as AD is likely to require a comprehensive model for successful screening. However, only a few studies combine different sources of data, such as MRI data [67], eye-tracking [7], and gait [71]. Similarly, papers where human-robot interaction [8, 12, 85, 94] or telephone-based systems [96, 97] are implemented also offer novel insight and avenues for future research. These approaches offer a picture of what automatic, cost-effective screening could look like in a perhaps not so distant future.

On a different front, **replicability** is assessed based on whether the authors report complete and accurate procedures of their research. Replicability has research implications because, before translating any method into clinical practice, its performance needs to be confirmed by other researchers being able to reproduce similar results. In this review, replicabilility is labelled as *low*, *partial*, and *full*. When we labelled an article as *full* with regards to replicability, we meant that their methods section was considered to be thorough enough to be reproduced by an independent researcher, from the specification of participants demographics and group size to the description of pre-processing, feature generation, classification, and evaluation procedures. Only three articles were labeled as *low* replicability [66, 74, 85], as they lacked detail in at least two of those sections (frequently data information and feature generation procedures). Twenty-two and twenty-five studies were labelled as *partial* and *full*, respectively. The elements most commonly missing in the *partial* papers are pre-processing (e.g., [83]) and feature generation procedures (e.g., [78]), which are essential steps in shaping the input to the machine learning classifiers. It must be highlighted that all *low* replicability papers are conference proceedings, where text space is particularly restricted. Hence, it does not stand out as one of the key problems of the field, even though it is clear that the description of pre-processing and feature generation must be improved.

The last research implication is **generalizability**, which is the degree to which a research approach may be attempted with different data, different settings, or real practice. Since generalizability is essentially about how translatable research is, most aspects in this last table are actually related to it:

- Whether *external validation* has been attempted is directly linked to generalizability;
- *feature balance:* results obtained in imbalanced datasets are less reliable and therefore less generalizable to other datasets;
- *contextualization of results:* for something to be generalizable is essential to know where it comes from and how does it compare to similar research;
- *spontaneous speech*: speech spontaneity is one aspect of naturalness, and the more natural the speech data, the more representative of “real” speech and the more generalizable;
- *conversational speech*: we propose that conversational speech is more representative of “real” speech;
- *content-independence:* if the classifier input includes features that are tied up with task content (e.g., lexical, syntactic, semantic, pragmatics), some degree of generalizability is lost;
- *Transcription-free:* a model that needs no transcription is free from ASR or manual transcription constraints, relying only on acoustic features. We suggest this to increase generalizability, for example, by being language-independent, therefore facilitating method uzability with non-English language for which corpus training is less feasible due to even more severe data scarcity. Transcription-free methods also facilitate the protection of users’ privacy, as they do not focus on speech content, which could encourage ethics committees to reduce restrictions on data collection, thereby addressing data scarcity.

Just as replicability, it is labelled as *low*, *moderate*, and *high*, depending of how many of the aforementioned criteria each study meets. Different to what we described with regards to replicability, the majority of studies (20) are labelled with *low* generalizability, 17 as *moderate*, and 14 as *high*. The most common reasons for decreased generalizability are dependence on content, followed by dependence on ASR or other transcription methods, although the two are related. Content-dependence makes it difficult to apply to other tasks or data (e.g., [12, 91]). This is even more pronounced in those studies where the approach heavily relies on word content, such as *n*-grams (e.g., [75]). Linguistic models that target only one linguistic aspect are also *low* generalizability, particularly if this aspect is language-dependent (e.g., syntaxis [59]). Examples of *high* generalizability include models relying solely on acoustic features, therefore free of content and transcription constrains (e.g., [89]), and especially if a standardized available feature set is used [11]. Other generalizable studies present more than one dataset (e.g., [62]), different languages in the same study (e.g., [87]), conversational data (e.g., [10]), a system designed for direct real application (e.g., [94]), and/or data from real scenarios [8].

Clinical potential (Supplementary Table 7: Clinical applicability)

The clinical applicability table aims to directly assess whether reviewed research could be translatable into clinical practice. Generalizability (discussed above) is essential for this purpose, but it will not be included here to avoid redundancy. We also note that clinical applicability of a diagnostic test is a somewhat vague construct in that one might need applicability in a clinical population *or* applicability for a clinician to understand its use. From a clinician's perspective, the translational steps from research on speech and language biomarkers to clinical use are not unlike those of any other diagnostic tool. This highlights, as we point out in the conclusion, that this translational development pathway would benefit from joint development between clinicians, speech and language experts, and AI experts working in concert. The other systematic aspects chosen to evaluate clinical potential are:

- **External validation:** In the majority of studies, data are collected detached from clinical practice and later analyzed for result reporting. The majority of papers (84%) present neither external validation procedures nor a system design that involves them. Only four studies, all of them by the same group [8, 12, 56, 85], collect their data in a real life setting (doctor-patient consultations). Another four studies take into account feasibility for clinical screening within their system design, for example, collecting data directly with a computerized decision support system [60], through human-robot interaction [94], a computer-supported screening tool [77], or simulating telephone-based data [96].
- **Potential application:** 78% of the reviewed papers present a method that could be applied as a diagnosis support system for MCI (e.g., [87]) or AD (e.g., [9]). The remaining studies work on disease progression by including SCI participants [59], predicting within-subject change [68] or discriminating between HC, MCI, and AD stages (e.g., [102]).
- **Global Health:** Although this could include a broad range of aspects, for the purpose of this review we have chosen to mention the language of the study and the processing unit of choice. This is because most research is done in English (41%), and work published in other languages helps towards methods being more universally applicable. Also, because smaller the processing units (i.e., phoneme versus word), tend to be more generalizable across languages. The most common processing unit is the sentence (63%), followed by conversations (16%), words (8%), syllables (4%), and phonemes (4%).
- **Remote application:** For such a prevalent disease, remote screening could significantly reduce the load on health systems. The majority of the studies, 67%, do not mention the possibility of their method being used remotely or having being designed for remote use, and only 25% suggest this as a possibility when motivating their project or discussing the results. Only four studies (2%), actually bring this into practice by experimenting with multi-modal human-robot interaction [93], infrastructure-free [77], or telephone-based [95, 96] approaches.

A further aspect, not explicitly included on this table, is *model interpretability*. While the accepted opinion is that the clinicians' ability to be able to interpret an AI model is essential for the adoption of AI technologies in medicine, the issue is still the subject of lively debate, with influential machine learning researchers like Geoff Hinton arguing that "clinicians and regulators should not insist on explainability" [119]. In terms of biomarkers of disease, very few if any clinicians understand the fine detail of an MRI report; it is the results presented to them that clinicians contextualize rather than the statistical or AI journey these results have been on to be presented to them. It could be argued that the case of speech and language biomarkers is no different. Of the papers reviewed here, only 4 mention interpretability or model interpretation explicitly [7, 9, 57, 97]. However, inherently interpretable models are used in a number of studies. Such interpretable methods were indicated in the above section on AI methods and include: linear regression, logistic regression, generalized linear and additive models, decision trees, decision rules, RuleFit, naive Bayes, and K-Nearest neighbors [120], and in some cases linear discriminant analysis. As shown in Supplementary Table 6, 57% of the studies reviewed included at least one of these types of models in their evaluation, even though most such inclusions were made for comparison purposes.

With regards to the selected criteria, the result tables highlight that research undertaken using non-English speech data almost invariably includes acoustic features, either as part of a larger feature set, such as Beltrami et al. [91] in Italian; or exclusively relying on acoustic features, such as Nasrolahzadeh et al. [65] in Persian, Weiner and Schultz [68] in German, Lopez-de Ipiña et al. [80], Espinoza-Cuadros et al. [83], Gonzalez-Moreira et al. [89], and Meilan et al. [100] in Spanish, and [64] in Japanese. Apart from English, only Portuguese and Chinese have been researched exclusively with text features [57, 82, 90].

Some of the field's needs clearly arise here. Firstly, there is a need for actual attempts to use these models in real clinical practice. For twenty years, conclusions and future directions of these research papers have suggested this, but very few published studies do bring it into a realization. Secondly, there is a need for enhanced focus on disease progression and risk prediction. Most studies mention the need for AD to be diagnosed earlier than it is now, and yet not many do actually work in that direction. Thirdly, further investment on research performed on languages other than English is needed, and increased focus on smaller language units, which are more generalizable to other languages or other samples of the same language. Alternatively, we suggest that a shift toward acoustic features only would potentially address these difficulties. Finally, one of the most obvious advantages of using AI for cognitive assessment is the possibility of using less infrastructure and less personnel. In order for this to become a reality, the remote applicability of these methods requires more extensive research.

Risk of bias (Supplementary Table 7: Clinical applicability)

This column highlights sources of potentially systematic errors or other circumstances that may introduce bias in the inferred conclusions. These can be summarized as follows:

- **Feature balance:** Class, age, gender, and education balances are essential for experimental results to be unbiased. Only 13 studies (25%) are balanced for these main features, and another five are balanced in terms of class but not in terms of other features. In the studies that seek to address class imbalance in their datasets, the main strategies used are subsampling [11, 59], use of statistical methods such as stratified CV [74], and careful choice of evaluation methods including use of the UAR metric [62] and ROC curve analysis [97].
- **Suitable metrics:** Equally important for bias prevention is choosing the right performance metrics to evaluate machine learning classifiers. For example, with a class-imbalanced dataset, accuracy is not a robust metric and should therefore not be used, or at least, complemented with other measures. However, 18 studies (35%) working with an imbalanced dataset report accuracy only.
- **Contextualized results:** Referring mainly to whether the reported research is directly and quantitatively compared to related works, or, ideally, whether a baseline against which results can be compared is provided. Only 61% of the studies reviewed provide such context.
- **Overfitting:** Studies would apply both CV and held-out sets to ensure their models do not overfit. CV should be applied when tuning machine learning hyper parameters when training the model, and the held-out set should be used to test the model on strictly unseen data. The majority of the studies do report CV (78%), but even more studies (90%) do not report hold-out set. Hence, there is a high risk that the reviewed models are, to some degree, overfitted to the data they have been trained with. Ideally, models should also be validated on entirely separate datasets. Only one of the studies reviewed carries out this kind of validation, although their method aims to use speech alignment in order to automatically score a cognitive task, instead of investigating the potential for demen-

tia prediction of the linguistic or speech features themselves [70].

- **Sample size:** Labelled as up to 50 participants ($ds \leq 50$), up to 100 participants ($ds \leq 100$) or over 100 participants ($ds > 100$). The results show that 13 studies were carried on smaller datasets (i.e., $ds \leq 50$), 24 studies carried on medium-sized datasets (i.e., $ds \leq 100$) and 14 studies carried on modestly larger datasets (i.e., $ds > 100$). However, seven of the studies carried on a medium-sized dataset and one study carried on a larger dataset attempted 3-way or even 4-way classification. Therefore, the group sizes of these studies are further reduced by the fact that the original dataset size needs to be divided into three or four groups, instead of the two groups used for binary classification.

We decided to use these numerical labels to classify the datasets, instead of assigning categories such as small or large, because even the largest dataset of the reviewed studies is relatively small when put into a machine learning context. All in all, there is a clear need for larger available datasets that are also balanced in terms of class and main risk factors. On larger datasets, it should be more straightforward to increase methodological rigor (e.g., by using CV, hold-out sets) and to seek for more active and systematic ways to prevent overfitting.

Strengths/Limitations (Supplementary Table 7: Clinical applicability)

In our view, a few desirable qualities should be present in AI research for AD, in order for it to be finally translatable into clinical practice. These are:

- **Spontaneous speech:** We consider spontaneous speech data to be more representative of real life spoken language. Although speech data obtained through non-spontaneous, constrained cognitive tasks present methodological advantages, we argue that spontaneous speech is desirable for cognitive monitoring due to its ubiquity, naturalness, and relative ease of collection. Under this criterion, we seek not only to explore the advantages of using speech for cognitive screening, but also the suitability for continuous and longitudinal collection. 65% of the papers meet this criterion with this by using open question data (e.g., free episodic recalls, discourses prompted by a picture, conversational dialogues). The remaining papers rely on constrained data, obtained for example by recording the words produced in a fluency test.
- **Conversational speech:** Similarly, we deem conversational speech to be more representative of real life spoken language than monologue speech. Here again we find a trade-off between naturalness and standardization. While monologueas are easier to handle (by requiring fewer preprocessing steps) and may avoid potential confunding factors present in dialogues (e.g., relationships between speakers, conversational style, cultural norms surrounding doctor-patient conversations), some methods may take advantage of these very factors for cognitive screening as they enrich the cognitive mechanisms involved in the interaction [120]. Of the reviewed papers, only 18% report the use of dialogue (i.e., structured, semi-structured, or conversational).
- **Automation:** Most of the reviewed papers claim some degree of automation in their procedure, but looking closely, only 37% describe a fully (or nearly fully) automatic method, from transcription to classification. Another 37% describe a partially automatic procedure, frequently automating feature generation and/or classification steps, but with a manual transcription and/or manual feature set reduction. The rest describe methods that require manual interference at almost every stage, and were therefore deemed to not be automatic.
- **Content-independence:** This is desirable in order for successful methods to be equally successful when speech is elicited in different ways (i.e., with different tasks, which imply different content). 55% of the papers report procedures that do rely on content-related characteristics of speech, such as word content. The rest either rely solely on acoustic features or phoneme based transcribed features, unrelated to word content.
- **Transcription-free:** As mentioned above, ASR methods are an automatic alternative to manual transcription, but they are not free of constrains. Therefore, we consider transcription free approaches to offer a more relevant contribution to the clinical application of AI for AD detection. Under this criterion, 35% of the reviewed papers use a transcription-free approach, whereas the rest rely on either manual or ASR transcriptions.

Only two studies meet all five criteria with a "yes" [10, 84]. In our view, the field needs to further explore the use spontaneous speech (ideally conversational), and indeed we have observed renewed interest in its use during the time span of this review, as AI becomes increasingly involved, as shown in Fig. 1. Automation also needs to be pursued by trying to bridge the gaps where automation becomes challenging, namely, during transcription, as well as during feature generation and feature set reduction (i.e., feature selection and feature extraction). Seeking automation entails a complex trade-off, since there is clearly valuable information about a person's cognitive status reflected in the content of what they say, as well as how they say it and how they choose to structure it. In addition, not all linguistic features are content-dependent and metrics such as word frequency, vocabulary richness, repetitiveness, and syntactic complexity are not linked to semantic content or meaning. However, processing language to obtain these metrics makes automation and generalization more difficult, specially as regards non-English data. While content-related information can offer insights into the nature of the disease and its development, reviewing the potential for AI systems in terms of practical usefulness in clinical settings for cognitive health monitoring requires considerations of content-independent and transcription-free approaches due to their ease of implementation, successful performance and more straightforward generalizability.

Overall conclusions

We have conducted the first systematic review on the potential application of interactive AI methods to AD detection and progression monitoring using natural language processing and speech technology to extract "digital biomarkers" for machine learning modelling.

Given the somewhat surprising quantity and variety of studies we found, it seems reasonable to conclude that this is a very promising field, with potential to gradually introduce changes into clinical practice. Almost all studies report relatively high performance, despite the difficulties inherent to the type of data used and the heterogeneity of the methods. When compared to neuropsychological assessment methods, speech and language technology were found to be at least equally discriminative between different groups. It is worth noting that the most commonly used neuropsychological test, MMSE, has been criticized [72] due to its inherent biases and lack of sensitivity to subtle symptoms. In this context, interactive AI could offer the same or better performance as a screening method, with the additional advantages of being implemented automatically and, possibly, remotely. Notwithstanding, while most of the papers hereby reviewed highlight the potential of AI and machine learning methods, no actual translation into clinical practice has been achieved. One might speculate that this slow uptake, despite nearly 20 years of research in this field, is due to difficulties in attaining meaningful interdisciplinary cooperation among between AI/machine learning research experts and clinicians. We expect that the growing interest in and indeed adoption of AI/machine learning methods in medicine will provide the stimulus needed for effective translation to take place in this field as it has in others. Despite an unexpectedly high number of records found eligible to review (51), the field remains highly heterogeneous with respect to the available data and methodology. It is difficult to compare results on an equal footing when their conclusions are drawn from monologue, dialogue, spontaneous, and non-spontaneous speech data. Similarly, different choices of processing units (varying from phoneme and syllable to a word, sentence, or a longer instance) pose additional comparability challenges. Furthermore, while machine learning methodology is somewhat standardized through a wealth of open-source tools, the feature generation and feature set reduction procedures are not. Feature generation varies greatly, with the same feature falling into slightly different categories depending on the study. Consequently, abiding by a standard taxonomy like the one proposed by Voleti et al. [20], which we adapted in Table 1, becomes essential in order to make cross-study comparisons. Surprisingly, many studies do not report on their approach to feature set reduction, or do it very vaguely, giving less than enough detail for replication. To our knowledge, only one study [11] relies on standardized feature sets available to the research community, while all other articles extract and calculate speech and language indices in an *ad hoc*, non-consensual way.

Furthermore, although cross-validation is implemented in most publications as an evaluation technique, many studies proceed with feature set reduction outside a cross-validation setting. That is, both training and testing data are used to find the relevant features that will serve as input to the classifier input. Additionally, although it is standard practice to tune machine learning models using a preferred performance metric (i.e., *acc, EER, AUC, F1*), we must recognize the potential effect this might have on the reliability and generalizability of such models. If CV is done correctly (i.e., not optimizing hyper-parameter tuning within the test set of each fold), the models created in any given fold of the CV procedure are tested on unseen data, although many studies do not provide this information. Barely any of the reviewed studies reported a hold-out set procedures or experiments on an entirely separate dataset, which would be the ideal scenario for robust model validation.

One of the reasons behind this lack of rigor is the size and variable quality of the datasets, which prevents adequate subsets to be generated while the size and integrity of the experimental groups is maintained. Consequently, we are confident that establishing certain standards on data and methodology will also increase the strictness of study evaluation. With regards to data type and availability, firstly, we should mention that data collection in this field is particularly difficult due to ethic constraints, due to the personally-identifying nature of speech data. Secondly, a benchmark dataset is essential to set the long overdue baselines of the field. Such baselines should not only refer to detection performance for SCI, MCI, and AD classes, but also to regression models able to predict cognitive scores, which is repeatedly proposed but hardly ever done, and prediction of progression and risk. Thirdly, we note that conversational dialogue (i.e., natural dialogue) is an under-explored resource in this field. As noted before, although monologue data presents methodological advantages, dialogue data has the potential to offer richer results precisely due to factors that under certain methodological frameworks might be dismissed as confounds. That is, an AI system trained to evaluate speakers interaction, cultural norms, and conversational styles has potential to be more versatile in monitoring cognitive health for different people, in different settings and at different times of the day. Furthermore, dialogue data could be easier and more natural to collect in real life (i.e., we spend part of our day interacting with somebody else), as well as more representative of a broader range psycholinguistic aspects such as alignment and entrainment at different linguistic levels [121], which might be relevant to AD detection.

With regards to methodology, we recommend a wider use of standardized feature sets, such as eGeMAPS [122], purposefully developed to detect physiological changes in voice production. Needless to say, other feature sets should also be built and tested systematically, for the field to move toward finding a golden standard. Further benefits of a consensual set of features entail the possibility of tracking those features "back to the brain", in order to find their neural substrate and hence contributing to knowledge of the neuropathology of AD.

In terms of aims and objectives, research suggests that embedded devices installed in the home to monitor patient health and safety may delay institutionalization [123], and therefore more emphasis should be placed on the feasibility of remote evaluations. To this end, we propose that future research focuses on natural conversations, which are straightforward to collect passively, continuously, and longitudinally in order to monitor cognitive health through an AI system. We also argue that focusing on cohorts already diagnosed with AD is no longer a relevant task for AI. As noted earlier, the majority of studies reviewed in this paper focus on diagnosis. We argue that emphasis should shift toward earlier stages of the disease, when pre-clinical unobserved changes start. Future research should therefore attempt to include healthy populations at higher risk of developing AD in larger scale longitudinal studies, as well as compare those populations to lower risk populations. There is good potential for interactive AI technology to contribute at those stages, given its increasingly ubiquitous presence in our lives, through wearable devices, smartphones, "smart homes", and "smart cities".

In addition, novel AI/machine learning digital biomarkers [124] could be used in combination with established biomarkers to target populations at risk of later dementia onset, as has already been proposed [101]. It needs to be emphasized that recorded data are considered personal data (i.e., with potential to identify a subject), with the ethical and regulatory hurdles this entails as regards data collection and analysis. We suggest that the field would benefit from revised ethics agreements to facilitate speech data collection, as well as from data sharing across institutions until datasets reach sufficient size to support complex machine learning structures and results are robust enough to encourage clinical applications. Increased collaboration between clinicians and AI experts should favor these developments.

ACKNOWLEDGMENTS

We thank Marshall Dozier, the Academic Support Librarian for her help with the search queries and the PROSPERO protocol. This research was supported by the Medical Research Council (MRC), grant number MR/N013166/1.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/20-0888r1).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JAD-200888.

REFERENCES

[1] American Psychiatric Association (2013) *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[2] World Health Organization (2013) Mental health action plan 2013-2020. *WHO Library Cataloguing-in-Publication Data*, pp. 1-44.

[3] Ross GW, Cummings JL, Benson DF (1990) Speech and language alterations in dementia syndromes: Characteristics and treatment. *Aphasiology* **4**, 339-352.

[4] Watson CM (1999) An analysis of trouble and repair in the natural conversations of people with dementia of the Alzheimer's type. *Aphasiology* **13**, 195-218.

[5] Bucks RS, Singh S, Cuerden JM, Wilcock GK (2000) Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology* **14**, 71-91.

[6] Luz S (2017) Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data. In *Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on*, pp. 45-46.

[7] Fraser KC, Lundholm Fors K, Eckerstrom M, Ohman F, Kokkinakis D (2019) Predicting MCI Status From Multimodal Language Data Using Cascaded Classifiers. *Front Aging Neurosci* **11**, 205.

[8] Mirheidari B, Blackburn D, Walker T, Reuber M, Christensen H (2019) Dementia detection using automatic analysis of conversations. *Comput Speech Lang* **53**, 65-79.

[9] Fraser KC, Meltzer JA, Rudzicz F (2016) Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis* **49**, 407-422.

[10] Luz S, la Fuente SD, Albert P (2018) A method for analysis of patient speech in dialogue for dementia detection. In Proceedings of the LREC 2018 Workshop "Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)", Kokkinakis D, ed. ELRA, Paris.

[11] Haider F, De La Fuente Garcia S, Luz S (2019) An assessment of paralinguistic acoustic features for detection of Alzheimer's Dementia in spontaneous speech. *IEEE J Sel Top Signal Process* **14**, 272-281.

[12] Mirheidari B, Blackburn D, O'Malley R, Walker T, Venneri A, Reuber M, Christensen H (2019) Computational Cognitive Assessment: Investigating the Use of an Intelligent Virtual Agent for the Detection of Early Signs of Dementia. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2732-2736.

[13] Clarivate Analytics (2016) Endnote x8. Philadelphia, United States.

[14] Richardson WS, Wilson MC, Nishikawa J, Hayward RS (1995) The well-built clinical question: a key to evidence-based decisions. *ACP J Club* **123**, A12-13.

[15] Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM (2008) Systematic reviews of diagnostic test accuracy. *Ann Intern Med* **149**, 889-897.

[16] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM (2011) Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* **155**, 529-536.

[17] Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S (2019) Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* **170**, 51-58.

[18] Higgins JP and Altman DG (2008) Assessing risk of bias in included studies. In *Cochrane handbook for systematic reviews of interventions: Cochrane book series*, Higgins JPT, Green S, eds. The Cochrane Collaboration, pp. 187-241.

[19] Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Ann Intern Med* **151**, 264-269.

[20] Voleti RNU, Liss J, Berisha V (2019) A review of automated speech and language features for assessment of cognition and thought disorders. *IEEE J Sel Top Signal Process* **14**, 282-298.

[21] Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: Liwc and computerized text analysis methods. *J Lang Soc Psychol* **29**, 24-54.

[22] Yngve VH (1960) A model and an hypothesis for language structure. *Proc Am Philos Soc* **104**, 444-466.

[23] Frazier L (1985) *Syntactic Complexity*. Cambridge University Press, Cambridge, UK.

[24] Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[25] Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* **3**, 993-1022.

[26] Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust* **28**, 357-366.

[27] Folstein MF, Folstein SE, McHugh PR (1975) "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* **12**, 189-198.

[28] Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H (2005) The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* **53**, 695-699.

[29] Cole M, Dastoor D, Simpson T (2015) *Hierarchic Dementia Scale – Revised: Instruction Manual*. Dementia Training Study Centre.

[30] Morris JC (1993) The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology* **43**, 2412-2414.

[31] Freedman M, Leach L, Kaplan E, Shulman K, Delis DC (1994) *Clock drawing: A neuropsychological analysis*. Oxford University Press, USA.

[32] Rosen WG, Mohs RC, Davis KL (1984) A new rating scale for Alzheimer's disease. *Am J Psychiatry* **141**, 1356-1364.

[33] Joanette Y, Ska B, Poissant A, Belleville S, Bellavance A, Gauthier S (1995) Évaluation neuropsychologique et profils cognitifs des démences de type Alzheimer: dissociations transversales et longitudinales. In *Neuropsychologie Clinique des Démences*, Eustache DF, Agniel A, eds. Solal, Marseille, pp. 91–106.

[34] Brodaty H, Pond D, Kemp NM, Luscombe G, Harding L, Berman K, Huppert FA (2002) The gpcog: a new screening test for dementia designed for general practice. *J Am Geriatr Soc* **50**, 530-534.

[35] Peña MM, Carrasco PM, Luque ML, García AIR (2012) Evaluación y diagnóstico del deterioro cognitivo leve. *Rev Logopedia Foniatr Audiol* **32**, 47-56.

[36] Reisberg B, Ferris SH, De Leon M, Crook T (1988) Global deterioration scale (GDS). *Psychopharmacol Bull* **24**, 661-663.

[37] Wallace M, Shelkey M, Hartford Institute for Geriatric Nursing (2007) Katz index of independence in activities of daily living (ADL). *Urol Nurs* **27**, 93-94.

[38] Graf C (2009) The Lawton instrumental activities of daily living (IADL) scale. *Gerontologist* **9**, 179-186.

[39] Petersen RC, Smith GE, Waring SC, Ivnik RJ, Kokmen E, Tangelos EG (1997) Aging, memory, mild cognitive impairment. *Int Psychogeriatr* **9**(Suppl. 1), 65-69.

[40] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939-939.

[41] Schmidtke K, Pohlmann S, Metternich B (2008) The syndrome of functional memory disorder: definition, etiology, natural course. *Am J Geriatr Psychiatry* **16**, 981-988.

[42] Benton A, Hamsher K, Sivan A (1976) *Multilingual aphasia examination*. University of Iowa, Iowa City.

[43] Goodglass H, Kaplan E (1983) *The Boston Diagnostic Aphasia Examination*. Lea & Febinger, Philadelphia.

[44] Sadeghian R, Schaffer JD, Zahorian SA (2017) Speech processing approach for diagnosing dementia in an early stage. *Proc Interspeech*, pp. 2705-2709.

[45] Wechsler D (1997) *Wechsler adult intelligence scale– third edition manual WAIS-III*. Psychological Corporation.

[46] Wechsler D (1997) *WMS-III: Wechsler memory scale administration and scoring manual*. Psychological Corporation.

[47] Nelson HE, Willison J (1991) *National adult reading test (NART)*. Nfer-Nelson Windsor.

[48] Bayles KA, Tomoeda CK, Dharmaperwira-Prins R (1993) *ABCD: Arizona battery for communication disorders of dementia*. Canyonlands Publishing.

[49] Darley FL, Aronson AE, Brown JR (1975) *Motor speech disorders*. Saunders.

[50] Mirzaei S, El Yacoubi M, Garcia-Salicetti S, Boudy J, Kahindo C, Cristancho-Lacroix V, Kerherve H, Rigaud AS (2018) Two-stage feature selection of voice parameters for early Alzheimer's disease prediction. *IRBM* **39**, 430-435.

[51] Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL (1994) The natural history of Alzheimer's disease. *Arch Neurol* **51**, 585.

[52] MacWhinney B (2019) Understanding spoken language through TalkBank. *Behavior Research Methods* **51**(4), 1919-1927. doi:10.3758/s13428-018-1174-9

[53] Gósy M (2013) Bea–a multifunctional hungarian spoken language database. *Phonetician* **105**, 50-61.

[54] Wallin A, Nordlund A, Jonsson M, Lind K, Edman Å, Göthlin M, Stålhammar J, Eckerström M, Kern S, Börjesson-Hanson A, Carlsson M, Olsson E, Zetterberg H, Blennow K, Svensson J, Öhrfelt A, Bjerke M, Rolstad S, Eckerström C (2016) The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *J Cereb Blood Flow Metab* **36**, 114-131.

[55] Pope C, Davis BH (2011) Finding a balance: The Carolinas Conversation Collection. *Corpus Linguist Linguist Theory* **7**, 143-161.

[56] Mirheidari B, Blackburn DJ, Harkness K, Walker T, Venneri A, Reuber M, Christensen H (2017) An avatar-based system for identifying individuals likely to develop dementia. In *Interspeech 2017*, pp. 3147-3151.

[57] Bertola L, Mota NB, Copelli M, Rivero T, Diniz BS, Romano-Silva MA, Ribeiro S, Malloy-Diniz LF (2014) Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Front Aging Neurosci* **6**, 185.

[58] Konig A, Satt A, Sorin A, Hoory R, Derreumaux A, David R, Robert P (2018) Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Curr Alzheimer Res* **15**, 120-129.

[59] Lundholm Fors K, Fraser KC, Kokkinakis D (2018) Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment. *Stud Health Technol Inform* **247**, 705-709.

[60] Martinez de Lizarduy U, Calvo Salomon P, Gomez Vilda P, Ecay Torres M, Lopez de Ipina K (2017) ALZUMERIC: A decision support system for diagnosis and monitoring of cognitive impairment. *Loquens* **4**, https://doi.org/10.3989/loquens.2017.037.

[61] Satt A, Sorin A, Toledo-Ronen O, Barkan O, Kompatsiaris I, Kokonozi A, Tsolaki, M (2013) Evaluation of speech-based protocol for detection of early-stage dementia. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1692-1696.

[62] Egas López JV, Tóth L, Hoffmann I, Kálmán J, Pákáski M, Gosztolya G (2019) Assessing Alzheimer's disease from speech using the i-vector approach. In *International Conference on Speech and Computer*, pp. 289-298. Springer.

[63] Gosztolya G, Vincze V, Tóth L, Pakaski M, Kalman J, Hoffmann I (2019) Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Comput Speech Lang* **53**, 181-197.

[64] Kato S, Endo H, Homma A, Sakuma T, Watanabe K (2013) Early detection of cognitive impairment in the elderly based on Bayesian mining using speech prosody and cerebral blood flow activation. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, pp. 5813-5816.

[65] Nasrolahzadeh M, Mohammadpoory Z, Haddadnia J (2018) Higher-order spectral analysis of spontaneous speech signals in Alzheimer's disease. *Cogn Neurodyn* **12**, 583-596.

[66] Thomas C, Keselj V, Cercone N, Rockwood K, Asp E (2005) Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *IEEE International Conference Mechatronics and Automation, 2005*, pp. 1569-1574.

[67] Clark DG, McLaughlin PM, Woo E, Hwang K, Hurtz S, Ramirez L, Eastman J, Dukes RM, Kapur P, DeRamus TP, Apostolova LG (2016) Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimers Dement (Amst)* **2**, 113-122.

[68] Weiner J, Schultz, T (2016) Detection of intra-personal development of cognitive impairment from conversational speech. In *Speech Communication; 12. ITG Symposium*, pp. 1-5.

[69] Duong A, Giroux F, Tardif A, Ska B (2005) The heterogeneity of picture-supported narratives in Alzheimer's disease. *Brain Lang* **93**, 173-184.

[70] Prud'hommeaux ET, Roark B (2015) Graph-based word alignment for clinical language evaluation. *Comput Linguist* **41**, 549-578.

[71] Shinkawa K, Kosugi A, Nishimura M, Nemoto M, Nemoto K, Takeuchi T, Numata Y, Watanabe R, Tsukada E, Ota M, Higashi S, Arai T, Yamada Y (2019) Multimodal behavior analysis towards detecting mild cognitive impairment: Preliminary results on gait and speech. *Stud Health Technol Inform* **264**, 343-347.

[72] Carnero-Pardo C (2014) Should the mini-mental state examination be retired? *Neurología* **29**, 473-481.

[73] Prud'Hommeaux ET, Roark B (2011) Alignment of spoken narratives for automated neuropsychological assessment. *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011*, pp. 484-489.

[74] Weiner J, Herff C, Schultz T (2016) Speech-based detection of Alzheimer's disease in conversational German. In *17th Annual Conference of the International Speech Communication Association*, pp. 1938-1942.

[75] Orimaye SO, Wong JSM, Golden KJ, Wong CP, Soyiri IN (2017) Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics* **18**, 34.

[76] Rochford I, Rapcan V, D'Arcy S, Reilly RB (2012) Dynamic minimum pause threshold estimation for speech analysis in studies of cognitive function in ageing. *Conf Proc IEEE Eng Med Biol Soc* **2012**, 3700-3703.

[77] Tröger J, Linz N, Alexandersson J, König A, Robert P (2017) Automated speech-based screening for Alzheimer's disease in a care service scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, Barcelona, Spain. ACM.

[78] Roark B, Mithcell M, Hosom J, Hollingshead K, Kaye J (2011) Spoken language derived measures for detecting mild cognitive impairment. *N Engl J Med* **19**, 2081-2090.

[79] Lopez-de Ipiña K, Alonso-Hernandez JB, Sole-Casals J, Travieso-Gonzalez CM, Ezeiza A, Faundez-Zanuy M, Calvo PM, Beitia B (2015) Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of Alzheimer's disease. *Neurocomputing* **150**, 392-401.

[80] Lopez-de Ipiña K, Alonso JB, Solé-Casals J, Barroso N, Henriquez P, Faundez-Zanuy M, Travieso CM, Ecay-Torres M, Martinez-Lage P, Egiraun H (2015) On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognit Comput* **7**, 44-55.

[81] Mielke MM, Vemuri P, Rocca WA (2014) Clinical epidemiology of Alzheimer's disease: assessing sex and gender differences. *Clin Epidemiol* **6**, 37.

[82] Chien Y, Hong S, Cheah W, Fu L, Chang Y (2018) An assessment system for Alzheimer's disease based on speech using a novel feature sequence design and recurrent neural network. In *2018 IEEE International Conference on Systems, Man, Cybernetics (SMC)*, pp. 3289-3294.

[83] Espinoza-Cuadros F, Garcia-Zamora MA, Torres-Boza D, Ferrer-Riesgo CA, Montero-Benavides A, Gonzalez-Moreira E, Hernandez-Gómez LA (2014) A spoken language database for research on moderate cognitive impairment: Design and preliminary analysis. In *Advances in Speech and Language Technologies for Iberian Languages, IberSpeech 2014*, volume 8854, Navarro Mesa, JL, Ortega, A, Teixeira, A, Hernández Pérez, E, Quintana Morales, P, Ravelo García, A, Guerra Moreno, I, and Toledano, DT, editors, Springer International Publishing, Cham, pp. 219–228.

[84] Khodabakhsh A, Demiroğlu C (2015) Analysis of speech-based measures for detecting and monitoring Alzheimer's disease. *Methods Mol Biol* **1246**, 159-173.

[85] Mirheidari B, Blackburn D, Walker T, Venneri A, Reuber M, Christensen H (2018) Detecting signs of dementia using word vector representations. In *Interspeech*, pp. 1893-1897.

[86] Mirheidari B, Blackburn D, Harkness K, Walker T, Venneri A, Reuber M, Christensen H (2017) Toward the automation of diagnostic conversation analysis in patients with memory complaints. *J Alzheimers Dis* **58**, 373-387.

[87] Fraser KC, Lundholm Fors K, Kokkinakis D (2019) Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Comput Speech Lang* **53**, 121-139.

[88] Rentoumi V, Paliouras G, Danasi E, Arfani D, Fragkopoulou K, Varlokosta S, Papadatos S (2017) Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 33-38.

[89] Gonzalez-Moreira E, Torres-Boza D, Kairuz H, Ferrer C, Garcia-Zamora M, Espinoza-Cuadros F, Hernandez-Gómez L (2015) Automatic prosodic analysis to identify mild dementia. *Biomed Res Int* **2015**, 916356.

[90] Dos Santos LB, Corrêa EA, Oliveira ON, Amancio DR, Mansur LL, Aluísio SM (2017) Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In *ACL 2017 – 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, pp. 1284-1296. Association for Computational Linguistics (ACL).

[91] Beltrami D, Calzà L, Gagliardi G, Ghidoni E, Marcello N, Favretti RR, Tamburini F (2016) Automatic Identification of Mild Cognitive Impairment through the Analysis of Italian Spontaneous Speech Productions. In *Proceedings of the Tenth International Conference*

on Language Resources and Evaluation (LREC 2016), pp. 2086-2093.

[92] Guo Z, Ling Z, Li Y (2019) Detecting Alzheimer's disease from continuous speech using language models. *J Alzheimers Dis* **70**, 1163-1174.

[93] D'Arcy S, Rapcan V, Penard N, Morris ME, Robertson IH, Reilly RB (2008) Speech as a means of monitoring cognitive function of elderly speakers. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, number January 2008, pp. 2230-2233.

[94] Tanaka H, Adachi H, Ukita N, Ikeda M, Kazui H, Kudo T, Nakamura S (2017) Detecting dementia through interactive computer avatars. *IEEE J Transl Eng Health Med* **5**, 2200111.

[95] Ben Ammar R, Ben Ayed Y (2018) Speech processing for early Alzheimer disease diagnosis: machine learning based approach. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1-8.

[96] Tröger J, Linz N, König A, Robert P, Alexandersson J (2018) Telephone-based dementia screening I: automated semantic verbal fluency assessment. *PervasiveHealth '18: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pp. 59-66.

[97] Yu B, Williamson JB, Mundt JC, Quatieri TF (2018) Speech-based automated cognitive impairment detection from remotely-collected cognitive test audio. *IEEE Access* **6**, 40494-40505.

[98] Guinn C, Singer B, Habash A (2014) A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. In *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, pp. 98-103.

[99] Tóth L, Hoffmann I, Gosztolyac G, Vinczec V, Szatlóczkid S, Bánrétib Z, Pákáskid M, Kálmán J (2018) A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr Alzheimer Res* **15**, 130-138.

[100] Meilan JJ, Martinez-Sanchez F, Carro J, Lopez DE, Millian-Morell L, Arana JM (2014) Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dement Geriatr Cogn Disord* **37**, 327-334.

[101] de la Fuente S, Ritchie C, Luz S (2019) Protocol for a conversation-based analysis study: Prevent-ED investigates dialogue features that may help predict dementia onset in later life. *BMJ Open* **9**, e026254.

[102] Konig A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, Manera V, Verhey F, Aalten P, Robert PH, David R (2015) Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement (Amst)* **1**, 112-124.

[103] Danso S, Terrera G, Luz S, Ritchie C (2019) Application of big data and artificial intelligence technologies to dementia prevention research: an opportunity for low-and-middle-income countries. *J Glob Health* **9**, 020322.

[104] Madikeri S, Dey S, Motlicek P, Ferras M (2016) Implementation of the standard i-vector system for the kaldi speech recognition toolkit. Technical report, Idiap.

[105] Biber D, Connor U, Upton T (2007) Discourse on the move. *Using corpus analysis to describe discourse structure*.

[106] Riley KP, Snowdon DA, Desrosiers MF, Markesbery WR (2005) Early life linguistic ability, late life cognitive function, neuropathology: findings from the nun study. *Neurobiol Aging* **26**, 341-347.

[107] Resnik P (1992) Left-corner parsing and psychological plausibility. In *Proceedings of the 14th conference on Computational linguistics-Volume 1*, pp. 191-197. Association for Computational Linguistics.

[108] Pakhomov S, Chacon D, Wicklund M, Gundel J (2011) Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. *Behav Res Methods* **43**, 136-144.

[109] 't Hart J (1981) Differential sensitivity to pitch distance, particularly in speech. *J Acoust Soc Am* **69**, 811-821.

[110] Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459-1462.

[111] Linz N, Tröger J, Alexandersson J, Wolters M, König A, Robert P (2017) Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 719-728.

[112] Al-Hameed S, Benaissa M, Christensen H (2017) Detecting and predicting Alzheimer's disease severity in longitudinal acoustic data. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, pp. 57-61.

[113] Pou-Prom C, Rudzicz F (2018) Learning multiview embeddings for assessing dementia. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2812-2817.

[114] Yancheva M, Fraser KC, Rudzicz F (2015) Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pp. 134-139.

[115] Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney, B (2020) Alzheimer's dementia recognition through spontaneous speech: the ADReSS challenge. *In INTERSPEECH* (pp. 2172-2176). ISCA. https://doi.org/10.21437/Interspeech. 2020-2571

[116] Haider F, de la Fuente Garcia S, Albert P, Luz S (2020) Affective speech for Alzheimer's dementia recognition. In *LREC 2020 Workshop: Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID-3 @ LREC). Marseille, France*, pp. 191-197. European Language Resources Association.

[117] Horley K, Reid A, Burnham D (2010) Emotional prosody perception and production in dementia of the Alzheimer's type. *J Speech Lang Hear Res* **53**, 1132-1146.

[118] Pakhomov SVS, Hemmy LS (2014) A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. *Cortex* **55**, 97-106.

[119] Wang F, Kaushal R, Khullar D (2019) Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med* **172**, 59-60.

[120] Molnar, C (2019) *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/

[121] Pickering MJ, Garrod S (2004) Toward a mechanistic psychology of dialogue. *Behav Brain Sci* **27**, 169-190; discussion 190-226.

[122] Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, Devillers LY, Epps J, Laukka P, Narayanan SS, Truong KP (2016) The Geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing. *IEEE Trans Affect Comput* **7**, 190-202.

[123] Fredericks EM, Bowers KM, Price KA, Hariri RH (2018) Cal: A smart home environment for monitoring cognitive decline. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pp. 1500-1506.

[124] Coravos A, Khozin S, Mandl KD (2019) Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ Digit Med* **2**, 1-5.