# ARTIFICIAL INTELLIGENCE IN GERIATRIC MENTAL HEALTH RESEARCH AND CLINICAL CARE

EDITED BY: Ellen E. Lee, Andrea Iaboni, Helmet Karim and Ipsit Vahia
PUBLISHED IN: Frontiers in Psychiatry

frontiers Research Topics

frontiers



ISSN 1664-8714
ISBN 978-2-88974-737-5
DOI 10.3389/978-2-88974-737-5

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

# ARTIFICIAL INTELLIGENCE IN GERIATRIC MENTAL HEALTH RESEARCH AND CLINICAL CARE

Topic Editors:
**Ellen E. Lee,** University of California, San Diego, United States
**Andrea Iaboni,** University Health Network, Canada
**Helmet Karim,** University of Pittsburgh, United States
**Ipsit Vahia,** McLean Hospital, United States

**Citation:** Lee, E. E., Iaboni, A., Karim, H., Vahia, I., eds. (2022). Artificial Intelligence in Geriatric Mental Health Research and Clinical Care. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-737-5

# Table of Contents

frontiers in Psychiatry

EDITORIAL
published: 01 March 2022
doi: 10.3389/fpsyt.2022.859175

# Editorial: Artificial Intelligence in Geriatric Mental Health Research and Clinical Care

*Helmet T. Karim*[1,2], *Ipsit V. Vahia*[3,4], *Andrea Iaboni*[5,6] *and Ellen E. Lee*[7,8,9]*

[1] Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, United States, [2] Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA, United States, [3] Division of Geriatrics, McLean Hospital, Belmont, MA, United States, [4] Department of Psychiatry, Harvard Medical School, Boston, MA, United States, [5] Knowledge, Innovation, Talent, Everywhere (KITE), Toronto Rehab Institute, University Health Network, Toronto, ON, Canada, [6] Department of Psychiatry, University of Toronto, Toronto, ON, Canada, [7] Department of Psychiatry, University of California, San Diego, San Diego, CA, United States, [8] Sam and Rose Stein Institute for Research on Aging, University of California, San Diego, San Diego, CA, United States, [9] Desert-Pacific Mental Illness Research Education and Clinical Center, Veterans Affairs San Diego Healthcare System, San Diego, CA, United States

**Keywords: artificial intelligence, older adults, mental health, machine learning (ML), natural language processing (NLP)**

## Editorial on the Research Topic

## Artificial Intelligence in Geriatric Mental Health Research and Clinical Care

Though adoption of artificial intelligence (AI) has been delayed in mental health research and clinical care relative to other fields, AI could potentially enhance diagnostic, prognostic, and treatment approaches for the growing aging population. With ubiquitous usage of wearable sensors, advancements in explainable AI, and growing acceptance of AI in medicine, these approaches could support increasing clinical demands. Despite enthusiasm for AI, usage in clinical settings is tempered by validity and ethical concerns. Integrating AI in clinical settings will require collaborations between clinicians and AI experts, inclusive study samples, and rigorous evaluation (akin to clinical trials for pharmacotherapies).

This special issue is a platform to highlight new AI applications in geriatric mental health research and care and bridge clinical with AI expertise by describing conceptual and pragmatic approaches. This issue showcases varied AI approaches [machine learning, natural language processing (NLP)] applied to multiple data-streams (sensors, electronic health records, interview data, neuroimaging) from multidisciplinary international perspectives. The included papers think broadly about policy and systemic implications and respect ethical concerns and patient protections.

Two manuscripts demonstrate utility in identifying important endpoints and optimizing treatments. Grzenda et al. used two studies to predict treatment outcome in late-life depression using structural imaging. Incorporation of structural imaging improved prediction when combined with clinical markers to two different treatments (antidepressants vs. Tai Chi), indicating that structural markers improve detection of treatment resistance—which allows for more aggressive treatments earlier. Chowdhury et al. conducted a systematic review of the use of electronic health records for predicting various outcomes (primarily dementia). Despite reporting high heterogeneity in data utilized, approaches used to standardize data, and even modeling approaches, the authors acknowledged rapid growth—with 21 studies in the past 5 years. Overall, EHR-integrated AI has potential to aid clinicians—triggering cognitive screening of patients at "high-risk" for dementia or suggesting more proactive approaches for patients at "high-risk" for treatment-resistance. AI models must incorporate clinician feedback and treatment outcomes and

OPEN ACCESS

***Edited and reviewed by:***
*Su Lui,*
*Sichuan University, China*

***Correspondence:**
*Ellen E. Lee*
*eel013@health.ucsd.edu*

***Specialty section:***
*This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry*

***Received:*** *21 January 2022*
***Accepted:*** *07 February 2022*
***Published:*** *01 March 2022*

***Citation:***
*Karim HT, Vahia IV, Iaboni A and Lee EE (2022) Editorial: Artificial Intelligence in Geriatric Mental Health Research and Clinical Care. Front. Psychiatry 13:859175. doi: 10.3389/fpsyt.2022.859175*

undergo dynamic updating. These models should identify which features most strongly predict "high-risk", thus improving clinician trust and condensing and compiling complex information into a comprehensible report.

Three manuscripts focused on NLP, i.e., deciphering unstructured text to provide insights into social functioning and depression. Badal et al. identified linguistic features (first-person plural pronouns) from interviews about social relationships that predicted scores on social support and loneliness and elicited gender differences. Yamada et al. identified acoustic, prosodic, and linguistic features (inflections, pauses, second formant frequencies, filler and positive words) from interviews on daily life and functioning that were associated with higher loneliness. DeSouza et al. provide an overview of NLP in late-life depression, including tool development for real-time analysis and usage in non-clinical settings, utility as a diagnostic tool, and key ethical/legal concerns and comfort with technologies. Speech data has particular relevance as a primary evaluative technique in psychiatry. However, further work is needed to combine speech with other clinical data to refine our predictive models (e.g., longitudinal data due to individual- and language/dialect-specific issues), to navigate privacy concerns when recording speech patterns, and examine novel data sources (e.g., social media posts and videos).

Three papers analyze passive sensor data to infer how older adults live their daily lives. Recently, around 60% of seniors in North America owned smartphones (1, 2), which will continue to increase. Smart home sensor technologies that monitor living environments are increasing in acceptability and use (1). The main challenge is uncovering meaningful and actionable information within complex data. Lee et al. described using smartphone data (e.g., number of unlocks, time spent at home), that can be used to evaluate behavior and mood over time in older depressed adults receiving psychotherapy. Zulueta et al. examine relationships between keystroke dynamics and cognitive function in people at risk of bipolar disorder. Zhang et al. demonstrate use of an environmental radio sensor to monitor breathing and behaviors of older adults with COVID-19. These were associated with measures of health, cognitive function, and wellbeing. A common challenge was addressing heterogeneity of behaviors and their context where clinical interpretation was not always straightforward. For example, increased behavioral activation was associated with more time at home for some, and the opposite for others. They point to the need for larger well-characterized longitudinal studies using innovative methods for annotating sensor data with behaviors or symptoms of interest in real-time.

Two papers provide foundational information around AI and affective computing. Renn et al. provide a concise primer on various clinically applicable forms of AI. They describe potential applications for diagnostics and treatment. Smith et al. focus on affective computing, defined as "study and development of systems and devices that can recognize, interpret, process, and simulate emotion." The authors provide a detailed breakdown of clinical domains within depression and Alzheimer's disease and their quantification using markers generated by affective computing. The most consequential sections from both reviews highlight barriers and challenges to AI—namely the primarily theoretical potential of AI. While examples of the clinical impact of AI are emerging, the pace of AI tool development is tempered by concerns including absence of well-designed integration into clinical workflows and minimal cross-disciplinary training and infrastructure that is required for effective use of these tools.

One major concern shared across these reviews is the potential to build biased models based on non-representative samples. Older adults are at high-risk for exclusion from AI studies, due to decreased access and familiarity with technologies, though older adults have been shown to have capacity for learning and using tools with tailored programs (3). Datasets used to build AI algorithms must be representative of socioeconomic, regional, racial, and ethnic backgrounds to avoid building biased models with potentially negative clinical consequences. Equitable AI models will require targeted funding opportunities and an upfront focus on designing these algorithms to provide more equitable healthcare. Despite such challenges, the papers in this special issue provide insight and hope for AI tools to condense complex clinical data and incorporate novel data sources in the service of enhancing diagnostic and treatment approaches.

## AUTHOR CONTRIBUTIONS

All authors contributed to conception of the editorial, wrote sections of the manuscript, contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

HK was supported by the NIMH K01 Grant: K01MH122741. AI was supported by an Academic Scholars Award (Department of Psychiatry, University of Toronto) and the Walter and Maria Schroeder Institute for Brain Innovation and Recovery. IV was supported in parts by National Institute of Aging (NIA) Grants R01AG066670 and 3R01AG066670-02S1, The Once Upon a Time Foundation, the Eric Warren Goldman Charitable Fund of the National Philanthropic Trust and the McLean Hospital Institute for Technology in Psychiatry. EL was supported, in part, by the National Institute of Mental Health (NIMH) K23 Grant MH119375-01, the Stein Institute for Research on Aging at the University of California, San Diego, and the Veteran Affairs Healthcare System.

## ACKNOWLEDGMENTS

This editorial describes the motivation for the Research Topic, reviews the major findings, and provides a vision for future research approaches and standards.

## REFERENCES

1. Nelson Kakulla BN. *2020 Tech Trends of the 50+*. Washington DC: AARP Research (2020).
2. Statistics Canada. *Tables 22-10-0115-01 Smartphone Use and Smartphone Habits by Gender and Age Group, Inactive*. Statistics Canada (2021).
3. Fortuna KL, Torous J, Depp CA, Jimenez DE, Areán PA, Walker R, et al. A future research agenda for digital geriatric mental healthcare. *Am J Geriatr Psychiatry*. (2019) 27:1277–85. doi: 10.1016/j.jagp.2019.05.013

**Author Disclaimer:** The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychiatry

MINI REVIEW
published: 06 September 2021
doi: 10.3389/fpsyt.2021.719125

# Natural Language Processing as an Emerging Tool to Detect Late-Life Depression

*Danielle D. DeSouza*[1]*, *Jessica Robin*[1], *Melisa Gumus*[1] and *Anthony Yeung*[2]

[1] Winterlight Labs, Toronto, ON, Canada, [2] Department of Psychiatry, University of Toronto, Toronto, ON, Canada

Late-life depression (LLD) is a major public health concern. Despite the availability of effective treatments for depression, barriers to screening and diagnosis still exist. The use of current standardized depression assessments can lead to underdiagnosis or misdiagnosis due to subjective symptom reporting and the distinct cognitive, psychomotor, and somatic features of LLD. To overcome these limitations, there has been a growing interest in the development of objective measures of depression using artificial intelligence (AI) technologies such as natural language processing (NLP). NLP approaches focus on the analysis of acoustic and linguistic aspects of human language derived from text and speech and can be integrated with machine learning approaches to classify depression and its severity. In this review, we will provide rationale for the use of NLP methods to study depression using speech, summarize previous research using NLP in LLD, compare findings to younger adults with depression and older adults with other clinical conditions, and discuss future directions including the use of complementary AI strategies to fully capture the spectrum of LLD.

**Keywords: geriatric mental health, depression, speech, natural language processing, artificial intelligence, digital health, late-life depression**

OPEN ACCESS

***Edited by:***
*Ipsit Vahia,*
*McLean Hospital, United States*

***Reviewed by:***
*Sheng-Min Wang,*
*Catholic University of Korea, South Korea*
*Alexandra König,*
*University of Nice Sophia Antipolis, France*

***Correspondence:**
*Danielle D. DeSouza*
*danielle@winterlightlabs.com*

***Specialty section:***
*This article was submitted to Aging Psychiatry, a section of the journal Frontiers in Psychiatry*

***Received:*** *01 June 2021*
***Accepted:*** *11 August 2021*
***Published:*** *06 September 2021*

***Citation:***
*DeSouza DD, Robin J, Gumus M and Yeung A (2021) Natural Language Processing as an Emerging Tool to Detect Late-Life Depression. Front. Psychiatry 12:719125. doi: 10.3389/fpsyt.2021.719125*

## INTRODUCTION

Depression is one of the leading causes of disability worldwide, affecting more than 264 million people of all ages (1). Although less prevalent among older adults (2), late-life depression (LLD), also referred to as geriatric depression, remains a major public health concern due to increased risk of morbidity, suicide, physical, cognitive, and social impairments, and self-neglect (3, 4). With a progressively aging population globally, the identification and treatment of LLD is critical (5).

LLD is generally defined as depression occurring in individuals aged 60 and over, though cutoffs vary in the literature. LLD can be further divided into early onset (first depressive episode before age 60) and late onset (first depressive episode after age 60) (6). For the purposes of this review, the definition of LLD includes both early and late onset episodes (5, 7, 8). As with younger individuals, LLD can be heterogeneous ranging from subthreshold changes in mood to major depression as outlined by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). However, diagnosing LLD is more challenging due to a different symptom profile compared to younger adults, and additional medical comorbidities (5, 9, 10). Misdiagnosis can occur if classic depressive symptoms (e.g., low mood) are not verbally expressed, and instead only somatic or cognitive symptoms are reported (11). Current depression treatment guidelines also recommend the use of standardized rating scales to gauge symptom severity (5), however, these scales may over

or under emphasize the presence of somatic symptoms. One recent review of LLD scales suggested that the over-reliance on somatic items may result in a misdiagnosis of LLD due to the high prevalence of medical comorbidities in older adults (12). Additionally only a handful of assessments, including the Patient Health Questionnaire-9 (PHQ-9), Cornell Scale for Depression in Dementia, Geriatric Depression Scale, and the Hospital Anxiety and Depression Scale have specifically been validated in LLD (13–16). However, these validated scales can be susceptible to bias due to the subjective nature of scoring by the assessing clinician. These scales might also falsely identify individuals with cognitive impairment as depressed (10, 17).

To help overcome these limitations, there has been a growing interest in the development of objective measures of depression using artificial intelligence (AI) technologies such as natural language processing (NLP) (18–20). NLP approaches focus on the analysis of acoustic and linguistic aspects of human language derived from speech and text and can be integrated with machine learning approaches to classify depression and its severity (19, 21). Advantages of using these approaches to understand depression symptoms through speech include high ecological validity, low subjectivity, low cost of frequent assessments, and quicker administration of tasks compared to standard assessments. An added benefit of speech analysis using NLP is that speech data can be collected remotely, meeting a vital need for remote cognitive and behavioral assessments in the era of the coronavirus disease (COVID-19) pandemic (22). In this review we will provide rationale for the use of NLP approaches to study depression using speech, summarize previous research using NLP in LLD, compare findings to younger adults with depression and older adults with other clinical conditions, and discuss future directions including the use of complementary AI strategies to fully capture the spectrum of LLD symptoms.

To search for relevant literature related to speech analysis in individuals with depression or LLD, PubMed/MEDLINE, Web of Science, and Google Scholar databases were searched using terms including: "geriatric depression", "older adult depression", "late-life depression", "major depressive disorder", "natural language processing", "speech analysis", "speech", "acoustics", "linguistics", "voice". A sample search query used in the PubMed database is: [("geriatric depression" OR "late-life depression") AND ("speech" OR "linguistic" OR "acoustic" OR "language" or "voice")]. While this mini review was not intended to be a systematic review of all literature related to NLP and depression, we used broad search terms to capture as many studies as possible specifically related to NLP in LLD. Only English language studies were included in the search strategy and no restrictions were placed on the year of publication.

## UNDERSTANDING DEPRESSION THROUGH SPEECH ANALYSIS

Speech production is a complex process involving the communication of thoughts, ideas, and emotions by way of spoken words and phrases. Variations in physiology, cognition, and mood can produce noticeable changes in speech assessed by measures that capture *what* is being said through word selections and grammar usage (linguistic features) and *how* people sound based on acoustic waveforms (acoustic features) (**Figure 1**). For over a century, clinicians have documented subtle alterations in speech patterns in individuals with depression with early reports highlighting speech that was lower in pitch, more monotonous, slower, and more hesitant (23). These observations were most consistently seen in melancholic and psychotic depression, both of which are characterized by psychomotor retardation (24), a core feature of major depressive disorder (MDD). Other early studies investigating speech in the context of depression and psychomotor retardation reported paucity of speech, lower volume and tone, slowed responses, and monotonous speech (24, 25). Slowed speech or "speaking so slowly that other people could have noticed" is now routinely analyzed as part of self-report depression assessments such as the PHQ-9 (26).

With regard to temporal characteristics of speech in depression, speech pause time or the amount of time between utterances, has been studied since the 1940s. During this period, timing devices could be used to measure speech pause times by manually pressing switches to indicate the start and end of a pause. In one study that included a broad group of individuals with psychiatric diagnoses, it was shown that patients with depression had more silent periods compared to those with hypomania (27). Later research improved study design by more precisely grouping patients according to clinical presentation (e.g., unipolar vs. bipolar, endogenous vs. neurotic) or by using specific tasks to elicit speech. In a small pilot study by Szabadi and colleagues, speech pause time was assessed in individuals with unipolar depression and in healthy controls during a counting task (28). The results showed that participants with depression had elongated speech pause time compared to controls, whose speech pause time remained consistent over a period of 2 months. Importantly, after recovery, pause time alterations normalized in patients, suggesting a role for speech pause time as a marker of clinical improvement. In a series of follow-up studies, Greden et al. replicated the elongated speech pause time findings in larger samples of individuals with depression (29, 30). Hardy and colleagues additionally showed that changes in speech pause time between baseline and final evaluations following a course of treatment for depression were associated with clinical changes on the Retardation Rating Scale for Depression (31). Others have not replicated this finding but have instead suggested that speech pause time abnormalities may be more evident during certain speech tasks (e.g., counting task) and/or only reflect certain depression symptoms or subtypes (32).

Focusing on the linguistic aspects of speech, early studies of depression used psycholinguistic approaches to manually encode measures characterizing lexical diversity, syntactic complexity, and speech content. In one study, comparing individuals with depression to those with mania, depressed participants used more modifying adverbs, first-person pronouns, and personal pronouns. In contrast, participants with mania used more action verbs, adjectives, and concrete nouns (33). Through content analysis, it was also shown that depressed patients used more words indicating self-preoccupation, in line with the large body

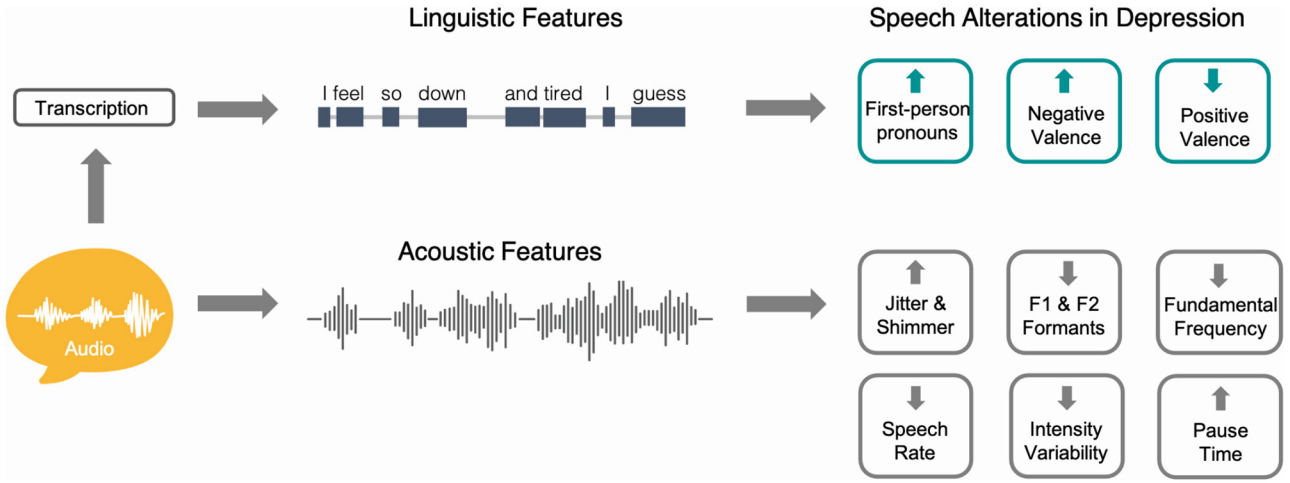


**FIGURE 1 |** Schematic representation showing how speech can be used to study linguistic and acoustic alterations in depression. To assess linguistic aspects of speech, audio samples are first transcribed using either manual or automatic speech recognition processes. Natural language processing methods can then be used to assess lexical, syntactic, and content measures of language. Audio waveforms can be directly analyzed to capture acoustic aspects of speech such as pause time, speech rate, and fundamental frequency. The right aspect of the figure shows examples of commonly reported speech alterations in the depression literature and the direction of these alterations compared to controls.

of literature indicating a role for increased self-rumination in depression [e.g., (34, 35)].

While foundational to our understanding of speech alterations in depression, many early studies relied on traditional approaches such as expert opinion or manual linguistic annotation that have known limitations including subjectivity and limited application to large-scale studies or clinical settings (36). With improved technology, supported by advances in AI, the ability to detect and objectively quantify speech alterations in depression has drastically improved. NLP is a branch of AI that specifically focuses on understanding, interpreting, and manipulating large amounts of human language and speech data. It combines computational linguistics with statistical, machine learning, and deep learning models, to take unstructured, free-form data (e.g., a voice recording or writing sample) and produce structured, quantitative acoustic and linguistic outputs. NLP has the potential to capture the speech changes in depression that reflect both physiologic changes at the basic motor level and also higher-level cognitive processing.

A substantial literature now exists examining automated assessments for depression using speech analysis, with acoustic or paralinguistic speech properties being the focus of multiple in-depth reviews (19, 21). In general, a number of acoustic measures characterized by source features from the vocal folds [e.g., jitter, shimmer, harmonics-to-noise ratio (HNR)], filter features from the vocal tract (e.g., F1 and F2 formants), spectral features [e.g., Mel Frequency Cepstral Coefficients (MFCCs)], and prosodic/melodic features [e.g., fundamental frequency (F0), speech intensity, speed, and pause duration] have been shown to be altered in individuals with depression (19, 21). While some heterogeneity exists in the direction of the reported alterations for some features (e.g., loudness), several trends corroborating earlier studies are evident. For example, greater depression severity is frequently associated with decreased F0 or pitch, intensity variability, and speech rate reflecting slower, more monotonous speech patterns (19, 37–41). Other acoustic measures such as jitter, shimmer, and HNR tend to be higher in depressed patients (19), reflecting laryngeal muscle tension, typically perceived as breathy, rough, or hoarse voices (42). In classification models, MFCCs have been shown to discriminate depressed patients from controls with high sensitivity and specificity (43–45). MFCCs have also been shown to classify the speech of stressed individuals (46). Since depressed language can also be associated with stress (47), future research to disentangle how MFCCs are altered in various cognitive and emotional states is warranted.

Computerized analysis of linguistic speech measures in depression is becoming more common with methods such as Linguistic Inquiry and Word Count (LIWC) (48), which has advantages such as high inter-rater reliability, objectivity, and cost-effectiveness compared to earlier manual approaches. Using LIWC, a recent study showed that participants with depression used more verbal utterances related to sadness compared to individuals with anxiety or comorbid depression and anxiety; however, the groups did not differ in the use of first-person singular pronouns (49). Others have used LIWC to develop composite measures tapping into first-person singular pronoun use, negative affect words, and positive affect words to capture linguistic patterns of depressed affect in nonclinical samples (47). To best capture depression symptom heterogeneity, automated speech analysis methods combining acoustic and linguistic measures may prove to be the most informative (50).

## SPEECH PATTERNS IN LLD

Given previous work establishing the relationship between speech alterations in depression, researchers have started

investigating how speech may be altered in LLD, specifically. **Table 1** summarizes recent literature on the topic and highlights different approaches used to collect and analyze speech data. While it is difficult to draw conclusions based on heterogeneous samples and methods, automated speech analysis is proving to be a promising means to tap into cognitive and depressive symptoms in LLD and can be readily adapted for naturalistic settings (55). Encouraging findings indicate that vocal measures can predict high and low depression scores in LLD between 86 and 92% of the time (54). Others have shown that LLD can be classified with 77–86% accuracy compared to age-matched controls (56). In the latter study, acoustic features contributing to accuracy values differed between sexes, highlighting the importance of taking demographic factors such as age, sex, and education into account. Speech has been shown to differ based on these variables even in the absence of clinical conditions. For example, morphological differences in vocal fold length between males and females contribute variation in acoustic features such as F0 (58).

The use of NLP in LLD presents unique challenges with regard to understanding the specificity of identified speech alterations. Older adults are more likely to have additional medical comorbidities that may also cause speech changes (59, 60). For example, depression is both an independent risk factor and a prodrome for dementia (61), which makes an underlying neurocognitive disorder a potential confounding factor in detecting speech pattern changes (62). Second, older adults are more likely to be on multiple prescribed medications, which introduces the confounding factor of medication effects on the acoustic properties of speech (63). Compared to younger adults, older adults have been found to have a lower HNR, a marker of turbulent airflow generated at the glottis during phonation (64, 65), which may be partly attributable to the effects of medications (e.g., vocal tract dryness and thickened mucosal secretions). Finally, normal age-related hormonal and structural changes (e.g., cartilage ossification, muscle degeneration) (64, 65), may also contribute to lower HNR during phonation. These factors highlight the importance of including tightly matched control groups and large sample sizes to broadly examine how different factors impact speech measures in the older adult population.

Despite these challenges, the use of NLP approaches in older adults provides opportunities to improve our understanding of depression in the context of biological sex, medical comorbidities, and other clinical factors. For example, speech changes in mild cognitive impairment (MCI) and Alzheimer's disease (AD) have been well documented in recent years (66–68), and share commonalities with speech changes related to depression. Increased pause duration, increased pronoun use and reduced lexical and syntactic complexity have all been reported as occurring in MCI and AD (66, 67, 69, 70). Due to these similar changes in speech patterns, care must be taken to differentiate dementia from depression when using speech analysis tools, to avoid misdiagnosis. One study compared LLD to those with early AD on a picture description task, and found that those with AD had reduced informativeness of their descriptions, suggesting that measures of content may be useful in differentiating depression from AD (51). Two recent studies suggested that certain acoustic speech features may help differentiate depression and dementia, or dementia with and without comorbid depression (66, 71), but this topic requires further research.

Studies comparing the speech rate of LLD vs. Parkinson's disease (PD) consistently show that rate of speech is significantly reduced in LLD, whereas the finding is not consistent in PD. Individuals with PD may exhibit decreased, increased, or typical speech rates (41, 72). These findings reflect underlying pathophysiological changes in PD such as compensation for hypokinetic muscle tone, which is not seen in LLD. These speech differences may serve as useful markers to detect or monitor for depression in PD given the high degree of comorbidity between these two diagnoses.

Finally, applying NLP using a disease-agnostic or transdiagnostic approach may also play an important role in addressing the comorbidity seen in LLD. For example, apathy is a transdiagnostic symptom seen in MDD, schizophrenia, traumatic brain injuries, AD, PD, and other neuropsychiatric disorders. A study by Konig and colleagues found that the presence of apathy was associated with shorter speech, slower speech, and lower variance of prosody (lower F0 range) (73). Thus, transdiagnostic markers like apathy may be a helpful method of discriminating which speech features are unique to a disorder vs. those that may be shared across diseases and disorders.

## COMPLEMENTARY AND NOVEL STRATEGIES TO MEASURE DEPRESSION

While current NLP approaches to capture symptoms of LLD are promising, there remain new opportunities to improve our understanding and implementation of this important research area. Over the past decade, advances in computer and mobile phone technologies have improved the quality and quantity of audiovisual input and output, allowing internet-based video clinical assessments to become more commonplace (74). The COVID-19 pandemic in particular has further led to a dramatic shift to online virtual care (75). With these shifts, NLP can potentially be applied to speech signals in real-time or asynchronously in clinical contexts. Recent studies have used NLP to generate COVID-19 phenotypes (76), track emotional distress in online cancer support groups (77), diagnose PD (78, 79), predict driving risk in older adults (80), and predict binge-eating behaviors (81).

These advances offer the potential of solving many of the challenges described in previous sections. For example, speech could be unobtrusively measured during routine visits on virtual care platforms between a healthcare provider and a patient. Data from these visits could be captured longitudinally and monitored for signs of depression, cognitive changes, or comorbid conditions. Individuals with depression in remission could conversely be monitored for signs of relapse. Preliminary studies have shown that integration of real-time audiovisual analysis into telemedicine platforms may be a feasible method

**TABLE 1 |** Summary of recent research examining speech in older adults with depression.

| References | Participants | Study objective | Speech tasks and measures | Main findings |
|---|---|---|---|---|
| Alpert et al. (39). | 22 participants (60+ years, 12 M, 10 F) meeting DSM-III-R depression criteria and 19 age-matched controls (8 M, 11 F). | To measure speech fluency and prosody in elderly depressed patients participating in a treatment trial. Participants were grouped as "agitated" or "retarded" based on clinical ratings. | Counting, reading, and free speech tasks. Acoustic measures tapping into fluency (speech productivity and pausing) and prosody (emphasis and inflection) were analyzed. | Older depressed participants had briefer utterances and less prosodic speech compared to controls. After treatment, improvement in the "retarded" group was associated with briefer pauses. |
| Murray et al. (51). | 18 participants with depression (60–90 years), 17 with Alzheimer's dementia, and 14 age-matched controls. | To determine if depression is associated with changes in discourse patterns and if this discriminates depression from early-stage Alzheimer's disease. | Picture description, sentence reading, and validated memory and language tasks. Quantitative, syntactic, and informativeness aspects of speech were analyzed. | Alzheimer's participants produced more uninformative utterances than depression and controls. No differences in informativeness between depression and controls. |
| Rabbi et al. (52). | Eight older individuals (4 M, 4 F) from a continuing care retirement community. Depression assessed using the CES-D (53). The SF-36 assessed overall well-being including mental health. | To demonstrate the feasibility of a multi-modal mobile sensing system to simultaneously assess mental and physical health in older individuals. | Authors measured the ratio of time speech was detected relative to the duration of the audio recording. | Amount of detected speech was positively associated with overall well-being. |
| Smith et al. (54). | 46 older adults (66–93 years, 10 M, 36 F) recruited from senior living communities. Depression symptoms assessed using the PHQ-9. | To determine if vocal alterations associated with clinical depression in younger adults are also indicative of depression in older adults. | Reading out loud and free speech and were collected two weeks apart. Speech measures included F0, jitter, shimmer, loudness, MFCCs, and LPCCs. | Speech features predicted high and low depression scores between 86 and 92% of the time. Changes in raw PHQ-9 scores were predicted within 1.17 points. |
| Little et al. (55). | 29 individuals with LLD meeting DSM-IV criteria (60+ years, 8 M, 21 F) and 29 matched controls with no history of depression (7 M, 22 F). MADRS, activities of daily living, and cognition scales were completed. | To test the utility of a novel wrist-worn device combined with deep learning algorithms to detect speech as an objective indicator of social interaction in LLD and in controls. | Algorithms were developed to classify: 1. speech and non-speech, and 2. wearer speech from other speech using audio recordings captured by the wearable device. | Participants with LLD produced less speech and reported poorer social and general functioning. Total speech activity and proportion of speech produced were correlated with attention and psychomotor speed but not depression severity or social functioning. |
| Lee et al. (56)[#]. | 61 individuals (60+ years, 18 M, 43 F) with major depressive disorder according to DSM-IV-TR criteria and 143 age-matched healthy controls (50 M, 93 F). | To develop a voice-based screening test for depression measuring vocal acoustic features of elderly Korean participants. | Participants read mood-inducing sentences. Variations in 2,330 acoustic speech features derived from AVEC 2013 (e.g., loudness, MFCCs, zero crossing rate) and eGeMAPS (e.g., F0, jitter, shimmer, and HNR) were assessed. | Spectral and energy-related features could discriminate men with depression with 86% accuracy. Prosody-related features could discriminate women with depression with 77% accuracy. |
| Albuquerque et al. (57). | 112 individuals (35–97 years, 56 M, 56 F). Anxiety and depression symptoms were assessed using the Hospital Anxiety Depression Scale. | To determine if variations in acoustic measures of voice are associated with non-severe anxiety or depression symptoms in adults across lifetime. | Reading vowels in disyllabic words and the "Cookie Theft" picture description task. 18 acoustic features extracted (e.g., F0, HNR, speech and pause duration measures). | Increased depression symptoms were associated with longer total pause duration and shorter total speech duration. Older participants tended to have more depressive symptoms. |

*M, male; F, female; DSM, Diagnostic and Statistical Manual of Mental Disorders; PHQ-9, Patient Health Questionnaire-9; MADRS, Montgomery-Asberg Depression Rating Scale; LLD, late-life depression; CES-D, Center for Epidemiological Study Depression Scale; MFCCs, Mel Frequency Cepstral Coefficients; LPCCs, linear predictive coding coefficients; F0, fundamental frequency, HNR, harmonics-to-noise ratio.*

[#]*This study specifically assessed speech differences between males and females (56).*

of detecting an individual's emotional state (82). Additionally, the use of smartphone and wearables technology to record these features have also demonstrated feasibility and acceptability in initial pilot studies (55, 83).

Beyond the clinician-patient interaction, NLP could also be implemented in non-clinical settings. Recent advances in smart speaker technology (e.g., Amazon Alexa, Google Assistant, Microsoft Cortana, Apple Siri) has resulted in significant consumer adoption, and these devices are currently being investigated as tools to support independent living and wellness in older adults (84). These technologies may provide an opportunity for naturalistic speech to be collected longitudinally

over time and may be a more granular and accurate method of detecting speech changes seen in depression (37). Advantages of smart speaker technologies include increased ease of use compared to computers and smartphones. Overall, these devices may hold promise to help older adults maintain independence through the use of passive monitoring for both depression and mild cognitive impairment and could serve as an "early warning system" to alert caregivers or professionals to negative symptom changes (84–86). However, the translation of AI technologies for home use in elderly populations may be limited without explicitly addressing ethical and legal considerations for patients, caregivers, and healthcare providers (87).

Regardless of the technology implemented, it is important to recognize that most of these technologies have not been specifically designed with older adults in mind. Older adults have reported hesitation about using novel technologies due to limited experience, frustration with technology, and physical health limitations (e.g., visual impairment) (88). From a privacy perspective, concerns have been raised by participants who may be unsure about how their electronic health data may be used, processed, or stored (89). Preliminary studies suggest these concerns can be mitigated with detailed informed consent from participants and by outlining privacy protocols in place (55). Ensuring that these technologies are culturally-adapted is another important consideration that can affect use and adoption (90). Finally, older adults have been shown to respond better to digital assistants with a socially-oriented interaction style (e.g., embedding informal conversation, using small talk, and encouragement) rather than assistants with a task-oriented style (e.g., structured formal responses) (91). As a result, there has been greater focus on embedding these interactions into automated social chatbots and companion robots for older adults (92). Tailoring these technologies to older adults has the potential to reduce technology hesitancy, improve adoption, and potentially increase the reliability of data that is collected as well (93).

## CONCLUSION

With an aging population globally, the identification and treatment of depression in older adults is critical. NLP approaches are proving to be a promising means to help assess, monitor, and detect depression and other comorbidities in older individuals based on speech. However, additional research is needed to fully characterize the spectrum of depression symptoms experienced by older individuals. Complementary speech collection and analysis strategies using AI, wearables, and other novel technologies may help further advance this important field.

## AUTHOR CONTRIBUTIONS

DD and AY contributed to the conception, design of the study, and wrote the first draft of the manuscript. JR and MG wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

## REFERENCES


1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet.* (2018) 392:1789–858. doi: 10.1016/S0140-6736(18)32279-7
2. Kessler RC, Birnbaum H, Bromet E, Hwang I, Sampson N, Shahly V. Age differences in major depression: results from the National Comorbidity Survey Replication (NCS-R). *Psychol Med.* (2010) 40:225–37. doi: 10.1017/S0033291709990213
3. Hodgetts S, Gallagher P, Stow D, Ferrier IN, O'Brien JT. The impact and measurement of social dysfunction in late-life depression: an evaluation of current methods with a focus on wearable technology. *Int J Geriatr Psychiatry.* (2017) 32:247–55. doi: 10.1002/gps.4632
4. Fiske A, Wetherell JL, Gatz M. Depression in older adults. *Annu Rev Clin Psychol.* (2009) 5:363–89. doi: 10.1146/annurev.clinpsy.032408.153621
5. Rodda J, Walker Z, Carter J. Depression in older adults. *BMJ.* (2011) 343:d5219–d5219. doi: 10.1136/bmj.d5219
6. Dols A, Bouckaert F, Sienaert P, Rhebergen D, Vansteelandt K. ten Kate M, et al. Early- and Late-Onset Depression in Late Life: A Prospective Study on Clinical and Structural Brain Characteristics and Response to Electroconvulsive Therapy. *Am J Geriatric Psychiat.* (2017) 25:178–89. doi: 10.1016/j.jagp.2016.09.005
7. Tedeschini E, Levkovitz Y, Iovieno N, Ameral VE, Nelson JC, Papakostas GI. Efficacy of antidepressants for late-life depression: a meta-analysis and meta-regression of placebo-controlled randomized trials. *J Clin Psychiatry.* (2011) 72:1660–8. doi: 10.4088/JCP.10r06531
8. Aziz R, Steffens DC. What are the causes of late-life depression? *Psychiatric Clinics of North America.* (2013) 36:497–516. doi: 10.1016/j.psc.2013.08.001
9. Kok RM, Reynolds CF. Management of depression in older adults: a review. *JAMA.* (2017) 317:2114. doi: 10.1001/jama.2017.5706
10. Andrews JA, Harrison RF, Brown LJE, MacLean LM, Hwang F, Smith T, et al. Using the NANA toolkit at home to predict older adults' future depression. *J Affect Disord.* (2017) 213:187–90. doi: 10.1016/j.jad.2017.02.019
11. Préville M, Mechakra Tahiri SD, Vasiliadis H-M, Quesnel L, Gontijo-Guerra S, Lamoureux-Lamarche C, et al. Association between perceived social stigma against mental disorders and use of health services for psychological distress symptoms in the older adult population: validity of the STIG scale. *Aging Ment Health.* (2015) 19:464–74. doi: 10.1080/13607863.2014.944092
12. Balsamo M, Cataldi F, Carlucci L, Padulo C, Fairfield B. Assessment of late-life depression *via* self-report measures: a review. *CIA.* (2018) 13:2021–44. doi: 10.2147/CIA.S178943
13. Alexopoulos GS, Abrams RC, Young RC, Shamoian CA. Cornell Scale for Depression in Dementia. *BIOL PSYCHIATRY*14. (1988) 23:271–84. doi: 10.1016/0006-3223(88)90038-8
14. Kenn C, Wood H, Kucyj M, Wattis J, Cunane J. Validation of the hospital anxiety and depression rating scale (HADS) in an elderly psychiatric population. *Int J Geriatr Psychiatry.* (1987) 2:189–93. doi: 10.1002/gps.930020309
15. Phelan E, Williams B, Meeker K, Bonn K, Frederick J, LoGerfo J, et al. study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam Pract.* (2010) 11:63. doi: 10.1186/1471-2296-11-63
16. Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: A preliminary report. *J Psychiatr Res.* (1982) 17:37–49. doi: 10.1016/0022-3956(82)90033-4
17. Brown LJE, Astell AJ. Assessing mood in older adults: a conceptual review of methods and approaches. *Int Psychogeriatr.* (2012) 24:1197–206. doi: 10.1017/S1041610212000075

18. Dai H-J, Su C-H, Lee Y-Q, Zhang Y-C, Wang C-K, Kuo C-J, et al. Deep learning-based natural language processing for screening psychiatric patients. *Front Psychiatry.* (2021) 11:533949. doi: 10.3389/fpsyt.2020.533949
19. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology.* (2020) 5:96–116. doi: 10.1002/lio2.354
20. Huang K-L, Duan S-F, Lyu X. Affective Voice Interaction and Artificial Intelligence: A research study on the acoustic features of gender and the emotional states of the PAD model. *Front Psychol.* (2021) 12:664925. doi: 10.3389/fpsyg.2021.664925
21. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF, et al. review of depression and suicide risk assessment using speech analysis. *Speech Commun.* (2015) 71:10–49. doi: 10.1016/j.specom.2015.03.004
22. Geddes MR, O'Connell ME, Fisk JD, Gauthier S, Camicioli R, Ismail Z. For the alzheimer society of Canada task force on dementia care best practices for COVID-19. Remote cognitive and behavioral assessment: Report of the Alzheimer Society of Canada Task Force on dementia care best practices for COVID-19. *Alzheimer's Dement.* (2020) 12: e12111. doi: 10.1002/dad2.12111
23. Kraepelin E. Manic Depressive Insanity and Paranoia. *The Journal of Nervous and Mental Disease.* (1921) 53:350. Available online at: https://journals.lww.com/jonmd/Fulltext/1921/04000/Manic_Depressive_Insanity_and_Paranoia.57.aspx (accessed May 30, 2021).
24. Buyukdura JS, McClintock SM, Croarkin PE. Psychomotor retardation in depression: Biological underpinnings, measurement, and treatment. *Progress in Neuro-Psychopharmacology and Biological Psychiatry.* (2011) 35:395–409. doi: 10.1016/j.pnpbp.2010.10.019
25. Sobin C, Sackeim HA. Psychomotor Symptoms of Depression. *Am J Psychiatry.* (1997)154:4–17. doi: 10.1176/ajp.154.1.4
26. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med.* (2001) 16:606–13. doi: 10.1046/j.1525-1497.2001.016009606.x
27. Chapple ED, Lindemann E. Clinical Implications of Measurements of Interaction Rates in Psychiatric Interviews. *Hum Organ.* (1942) 1:1–11. doi: 10.17730/humo.1.2.3r61twt687148g1j
28. Szabadi E, Bradshaw CM, Besson JA. Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression. *Br J Psychiatry.* (1976) 129:592–7. doi: 10.1192/bjp.129.6.592
29. Greden JF, Albala AA, Smokler IA, Gardner R, Carroll BJ. Speech pause time: a marker of psychomotor retardation among endogenous depressives. *Biol Psychiatry.* (1981) 16:851–9.
30. Greden JF. Biological markers of melancholia and reclassification of depressive disorders. *Encephale.* (1982) 8:193–202.
31. Hardy P, Jouvent R, Widl D. Speech Pause Time and the Retardation Rating Scale for Depression (ERD) Towards a Reciprocal Validation.5. J Affect Disord. (1984) 6:123–7. doi: 10.1016/0165-0327(84)90014-4
32. Nilsonne Å. Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatr Scand.* (1987) 76:235–45. doi: 10.1111/j.1600-0447.1987.tb02891.x
33. Andreasen NJC. Linguistic analysis of speech in affective disorders. *Arch Gen Psychiatry.* (1976) 33:1361. doi: 10.1001/archpsyc.1976.01770110089009
34. Takano K, Tanno Y. Self-rumination, self-reflection, and depression: Self-rumination counteracts the adaptive effect of self-reflection. *Behav Res Ther.* (2009) 47:260–4. doi: 10.1016/j.brat.2008.12.008
35. Nolen-Hoeksema S, Wisco BE, Lyubomirsky S. Rethinking rumination. *Perspect Psychol Sci.* (2008) 3:400–24. doi: 10.1111/j.1745-6924.2008.00088.x
36. Corcoran CM, Mittal VA, Bearden CE, Gur ER, Hitczenko K, Bilgrami Z, et al. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Res.* (2020) S0920996420302474. doi: 10.1016/j.schres.2020.04.032
37. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J Neurolinguistics.* (2007) 20:50–64. doi: 10.1016/j.jneuroling.2006.04.001
38. Horwitz R, Quatieri TF, Helfer BS, Yu B, Williamson JR, Mundt J. On the relative importance of vocal source, system, and prosody in human depression. In: *2013 IEEE International Conference on Body Sensor Networks.* Cambridge, MA, USA: IEEE. (2013) p. 1–6.
39. Alpert M, Pouget ER, Silva RR. Reflections of depression in acoustic measures of the patient's speech. *J Affect Disord.* (2001) 66:59–69. doi: 10.1016/S0165-0327(00)00335-9
40. Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. *Biol Psychiatry.* (2012) 72:580–7. doi: 10.1016/j.biopsych.2012.03.015
41. Cannizzaro M, Harel B, Reilly N, Chappell P, Snyder PJ. Voice acoustical measurement of the severity of major depression. *Brain Cogn.* (2004) 56:30–5. doi: 10.1016/j.bandc.2004.05.003
42. Farrús M, Hernando J, Ejarque P. Jitter and Shimmer Measurements for Speaker Recognition. *Eighth Annual Conference of the International Speech Communication Association.* (2007) p. 778–781.
43. Taguchi T, Tachikawa H, Nemoto K, Suzuki M, Nagano T, Tachibana R, et al. Major depressive disorder discrimination using vocal acoustic features. *J Affect Disord.* (2018) 225:214–20. doi: 10.1016/j.jad.2017.08.038
44. France DJ, Shiavi RG, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions On Biomedical Engineering.* (2000) 47:9. doi: 10.1109/10.846676
45. Cummins N, Epps J, Breakspear M, Goecke R. An Investigation of Depressed Speech Detection: Features and Normalization. *12th Annual Conference of the International Speech Communication Association.* (2011) p. 2997–3000.
46. Patro H, Senthil Raja G, Dandapat S. Statistical feature evaluation for classification of stressed speech. *Int J Speech Technol.* (2007) 10:143–52. doi: 10.1007/s10772-009-9021-0
47. Newell EE, McCoy SK, Newman ML, Wellman JD, Gardner SK. You sound so down: capturing depressed affect through depressed language. *J Lang Soc Psychol.* (2018) 37:451–74. doi: 10.1177/0261927X17731123
48. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Soc Psychol.* (2010) 29:24–54. doi: 10.1177/0261927X09351676
49. Sonnenschein AR, Hofmann SG, Ziegelmayer T, Lutz W. Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cogn Behav Ther.* (2018) 47:315–27. doi: 10.1080/16506073.2017.1419505
50. Arevian AC, Bone D, Malandrakis N, Martinez VR, Wells KB, Miklowitz DJ, et al. Clinical state tracking in serious mental illness through computational analysis of speech. *PLoS ONE.* (2020) 15:e0225695. doi: 10.1371/journal.pone.0225695
51. Murray LL. Distinguishing clinical depression from early Alzheimer's disease in elderly people: Can narrative analysis help? *Aphasiology.* (2010) 24:928–39. doi: 10.1080/02687030903422460
52. Rabbi M, Ali S, Choudhury T, Berke E. Passive and In-Situ assessment of mental and physical well-being using mobile sensors. In :*Proceedings of the 13th international conference on Ubiquitous computing - UbiComp'11.* Beijing, China: ACM Press. (2011) p. 385. doi: 10.1145/2030112.2030164
53. Andresen EM, Malmgren JA, Carter WB, Patrick DL. Screening for depression in well older adults: evaluation of a short form of the CES-D (Center for Epidemiologic Studies Depression Scale). *Am J Prev Med.* (1994) 10:77–84.
54. Smith M, Dietrich BJ, Bai E, Bockholt HJ. Vocal pattern detection of depression among older adults. (2019) 29:440–9. doi: 10.1111/inm.12678
55. Little B, Alshabrawy O, Stow D, Ferrier IN, McNaney R, Jackson DG, et al. Deep learning-based automated speech detection as a marker of social functioning in late-life depression. *Psychol Med.* (2020) 1–10. doi: 10.1017/S0033291719003994
56. Lee S, Suh SW, Kim T, Kim K, Lee KH, Lee JR, et al. Screening major depressive disorder using vocal acoustic features in the elderly by sex. *J Affect Disord.* (2021) 291:15–23. doi: 10.1016/j.jad.2021.04.098
57. Albuquerque L, Valente ARS, Teixeira A, Figueiredo D, Sa-Couto P, Oliveira C. Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. *PLoS ONE.* (2021) 16:e0248842. doi: 10.1371/journal.pone.0248842
58. Titze IR. Physiologic and acoustic differences between male and female voices. *J Acoust Soc Am.* (1989) 85:1699–707. doi: 10.1121/1.397959
59. Murton OM, Hillman RE, Mehta DD, Semigran M, Daher M, Cunningham T, et al. Acoustic speech analysis of patients with decompensated heart failure: A pilot study. *J Acoust Soc Am.* (2017) 142:EL401. doi: 10.1121/1.5007092
60. Amir O, Anker SD, Gork I, Abraham WT, Pinney SP, Burkhoff D, et al. Feasibility of remote speech analysis in evaluation of dynamic fluid overload in

heart failure patients undergoing haemodialysis treatment. *ESC Heart Failure.* (2021) 8:2467–72. doi: 10.1002/ehf2.13367
61. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet.* (2020) 396:413–46. doi: 10.1016/S0140-6736(20)30367-6
62. Leyhe T, Reynolds CF. 3rd, Melcher T, Linnemann C, Klöppel S, Blennow K, et al. A common challenge in older adults: Classification, overlap, and therapy of depression and dementia. *Alzheimer's Dement.* (2017) 13:59–71. doi: 10.1016/j.jalz.2016.08.007
63. Charlesworth CJ, Smit E, Lee DSH, Alramadhan F, Odden MC. Polypharmacy among adults aged 65 years and older in the United States: 1988-2010. *J Gerontol A Biol Sci Med Sci.* (2015) 70:989–95. doi: 10.1093/gerona/glv013
64. Ferrand CT. Harmonics-to-noise ratio: an index of vocal aging. *J Voice.* (2002) 16:8. doi: 10.1016/s0892-1997(02)00123-6
65. Gorham-Rowan MM, Laures-Gore J. Acoustic-perceptual correlates of voice quality in elderly men and women. *J Commun Disord.* (2006) 39:171–84. doi: 10.1016/j.jcomdis.2005.11.005
66. Fraser KC, Rudzicz F, Hirst G. Detecting late-life depression in Alzheimer's disease through analysis of speech and language. In: *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology.* San Diego, CA, USA: Association for Computational Linguistics. (2016) p. 1–11. doi: 10.18653/v1/W16-0301
67. Ahmed S, de Jager CA, Haigh A-M, Garrard P. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology.* (2013) 27:79–85. doi: 10.1037/a0031288
68. Szatloczki G, Hoffmann I, Vincze V, Kalman J, Pakaski M. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's Disease. *Front Aging Neurosci.* (2015) 7:195. doi: 10.3389/fnagi.2015.00195
69. Mueller KD, Koscik RL, Hermann BP, Johnson SC, Turkstra LS. Declines in connected language are associated with very early mild cognitive impairment: results from the Wisconsin registry for Alzheimer's prevention. *Front Aging Neurosci.* (2018) 9:437. doi: 10.3389/fnagi.2017.00437
70. Hoffmann I, Nemeth D, Dye CD, Pákáski M, Irinyi T, Kálmán J. Temporal parameters of spontaneous speech in Alzheimer's disease. *Int J Speech Lang Pathol.* (2010) 12:29–34. doi: 10.3109/17549500903137256
71. Sumali B, Mitsukura Y, Liang K, Yoshimura M, Kitazawa M, Takamiya A, et al. Speech quality feature analysis for classification of depression and dementia patients. *Sensors.* (2020) 20:3599. doi: 10.3390/s20123599
72. Flint AJ, Black SE, Campbell-Taylor I, Gailey GF, Levinton C. Acoustic analysis in the differentiation of Parkinson's disease and major depression. *J Psycholinguist Res.* (1992) 21:383–99. doi: 10.1007/BF01067922
73. König A, Linz N, Zeghari R, Klinge X, Tröger J, Alexandersson J, et al. Detecting apathy in older adults with cognitive disorders using automatic speech analysis. *J Alzheimer's Disease.* (2019) 69:1183–93. doi: 10.3233/JAD-181033
74. Kichloo A, Albosta M, Dettloff K, Wani F, El-Amir Z, Singh J, et al. Telemedicine, the current COVID-19 pandemic and the future: a narrative review and perspectives moving forward in the USA. *Fam Med Com Health.* (2020) 8:e000530. doi: 10.1136/fmch-2020-000530
75. Gosse PJ, Kassardjian CD, Masellis M, Mitchell SB. Virtual care for patients with Alzheimer disease and related dementias during the COVID-19 era and beyond. *CMAJ.* (2021) 193:E371–7. doi: 10.1503/cmaj.201938
76. Barr PJ, Ryan J, Jacobson NC. Precision assessment of COVID-19 phenotypes using large-scale clinic visit audio recordings: harnessing the power of patient voice. *J Med Internet Res.* (2021) 23:e20545. doi: 10.2196/20545
77. Leung YW, Wouterloot E, Adikari A, Hirst G, de Silva D, Wong J, et al. Natural language processing–based virtual cofacilitator for online cancer support groups: protocol for an algorithm development and validation study. *JMIR Res Protoc.* (2021) 10:e21453. doi: 10.2196/21453
78. Frid EJ, Safra H, Hazan LL, Lokey D, Hilu L, Manevitz LO, et al. Computational diagnosis of Parkinson's disease directly from natural speech using machine learning techniques. In: *2014 IEEE International Conference on Software Science, Technology and Engineering.* (2014) p. 50–53. doi: 10.1109/SWSTE.2014.17
79. Singh S, Xu W. Robust detection of Parkinson's disease using harvested smartphone voice data: a telemedicine approach. *Telemedicine and e-Health.* (2019) 26:327–34. doi: 10.1089/tmj.2018.0271
80. Yamada Y, Shinkawa K, Kosugi A, Kobayashi M, Takagi H, Nemoto M, et al. Predicting future accident risks of older drivers by speech data from a voice-based dialogue system: a preliminary result. In: Spohrer J, Leitner C, editors. *Advances in the Human Side of Service Engineering.* Cham: Springer International Publishing. (2021) p. 131–137.
81. Funk B, Sadeh-Sharvit S, Fitzsimmons-Craft EE, Trockel MT, Monterubio GE, Goel NJ, et al. A Framework for applying natural language processing in digital health interventions. *J Med Internet Res.* (2020) 22:e13855. doi: 10.2196/13855
82. Kallipolitis A, Galliakis M, Menychtas A, Maglogiannis I. Affective analysis of patients in homecare video-assisted telemedicine using computational intelligence. *Neural Comput & Applic.* (2020) 32:17125–36. doi: 10.1007/s00521-020-05203-z
83. Abbas A, Schultebraucks K, Galatzer-Levy IR. Digital measurement of mental health: challenges, promises, and future directions. *Psychiatr Ann.* (2021) 51:14–20. doi: 10.3928/00485713-20201207-01
84. Pradhan A, Lazar A. Voice technologies to support aging in place: opportunities and challenges. *Innovation in Aging.* (2020) 4:317–8. doi: 10.1093/geroni/igaa057.1016
85. Choi YK, Thompson HJ, Demiris G. Use of an internet-of-things smart home system for healthy aging in older adults in residential settings: pilot feasibility study. *JMIR Aging.* (2020) 3:e21964. doi: 10.2196/21964
86. Rantz MJ, Scott SD, Miller SJ, Skubic M, Phillips L, Alexander G, et al. Evaluation of health alerts from an early illness warning system in independent living. *Comput Inform Nurs.* (2013) 31:274–80. doi: 10.1097/NXN.0b013e318296298f
87. Ho A. Are we ready for artificial intelligence health monitoring in elder care? *BMC Geriatr.* (2020) 20:358. doi: 10.1186/s12877-020-01764-9
88. Chung J, Bleich M, Wheeler DC, Winship JM, McDowell B, Baker D, et al. Attitudes and perceptions toward voice-operated smart speakers among low-income senior housing residents: comparison of pre- and post-installation surveys. *Gerontol Geriatr Med.* (2021) 7:233372142110058. doi: 10.1177/23337214211005869
89. Stanberry B. Telemedicine: barriers and opportunities in the 21st century. *J Intern Med.* (2000) 247:615–28. doi: 10.1046/j.1365-2796.2000.00699.x
90. Chung J, Thompson HJ, Joe J, Hall A, Demiris G. Examining Korean and Korean American older adults' perceived acceptability of home-based monitoring technologies in the context of culture. *Inform Health Soc Care.* (2017) 42:61–76. doi: 10.3109/17538157.2016.1160244
91. Chattaraman V, Kwon W-S, Gilbert JE, Ross K. Should AI-Based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior.* (2019) 90:315–30. doi: 10.1016/j.chb.2018.08.048
92. Pou-Prom C, Raimondo S, Rudzicz F. A conversational robot for older adults withAlzheimer's disease. *J Hum-Robot Interact.* (2020) 9:1–25. doi: 10.1145/3380785
93. Shishehgar M, Kerr D, Blake J. A systematic review of research into how robotic technology can help older people. *Smart Health.* (2018) 7–8:1–18. doi: 10.1016/j.smhl.2018.03.002

**Conflict of Interest:** DD, JR, and MG are employees of Winterlight Labs. AY is a medical consultant for Winterlight Labs.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychiatry

SYSTEMATIC REVIEW
published: 20 September 2021
doi: 10.3389/fpsyt.2021.738466

# Use of Machine Learning and Artificial Intelligence Methods in Geriatric Mental Health Research Involving Electronic Health Record or Administrative Claims Data: A Systematic Review

*Mohammad Chowdhury*[1], *Eddie Gasca Cervantes*[2], *Wai-Yip Chan*[2] *and Dallas P. Seitz*[1]*

[1] Department of Psychiatry, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada, [2] Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada

OPEN ACCESS

**Edited by:**
*Ipsit Vahia,*
*McLean Hospital, United States*

**Reviewed by:**
*Ellen E. Lee,*
*University of California, San Diego, United States*
*Xi Zhu,*
*Columbia University, United States*

***Correspondence:**
*Dallas P. Seitz*
*dallas.seitz@ucalgary.ca*

**Specialty section:**
*This article was submitted to Aging Psychiatry, a section of the journal Frontiers in Psychiatry*

**Received:** *08 July 2021*
**Accepted:** *26 August 2021*
**Published:** *20 September 2021*

**Citation:**
*Chowdhury M, Cervantes EG, Chan W-Y and Seitz DP (2021) Use of Machine Learning and Artificial Intelligence Methods in Geriatric Mental Health Research Involving Electronic Health Record or Administrative Claims Data: A Systematic Review. Front. Psychiatry 12:738466. doi: 10.3389/fpsyt.2021.738466*

**Introduction:** Electronic health records (EHR) and administrative healthcare data (AHD) are frequently used in geriatric mental health research to answer various health research questions. However, there is an increasing amount and complexity of data available that may lend itself to alternative analytic approaches using machine learning (ML) or artificial intelligence (AI) methods. We performed a systematic review of the current application of ML or AI approaches to the analysis of EHR and AHD in geriatric mental health.

**Methods:** We searched MEDLINE, Embase, and PsycINFO to identify potential studies. We included all articles that used ML or AI methods on topics related to geriatric mental health utilizing EHR or AHD data. We assessed study quality either by Prediction model Risk OF Bias ASsessment Tool (PROBAST) or Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) checklist.

**Results:** We initially identified 391 articles through an electronic database and reference search, and 21 articles met inclusion criteria. Among the selected studies, EHR was the most used data type, and the datasets were mainly structured. A variety of ML and AI methods were used, with prediction or classification being the main application of ML or AI with the random forest as the most common ML technique. Dementia was the most common mental health condition observed. The relative advantages of ML or AI techniques compared to biostatistical methods were generally not assessed. Only in three studies, low risk of bias (ROB) was observed according to all the PROBAST domains but in none according to QUADAS-2 domains. The quality of study reporting could be further improved.

**Conclusion:** There are currently relatively few studies using ML and AI in geriatric mental health research using EHR and AHD methods, although this field is expanding. Aside from dementia, there are few studies of other geriatric mental health conditions. The lack of consistent information in the selected studies precludes precise comparisons between

them. Improving the quality of reporting of ML and AI work in the future would help improve research in the field. Other courses of improvement include using common data models to collect/organize data, and common datasets for ML model validation.

**Keywords: geriatric, mental health, artificial intelligence, machine learning, administrative health data, electronic health records**

## INTRODUCTION

Geriatric mental health conditions such as depression or dementia are common, and it can be challenging to identify individuals with these conditions and predict outcomes associated with these conditions. Real-world data sources such as electronic health records (EHRs) and administrative health data (AHD) are increasingly used in geriatric mental health research. These data sources are available in many countries and health regions, and the details and information contained in these databases vary across the jurisdiction. The data contained in EHRs and administrative datasets are typically collected for the provision of medical care and purposing, such as financial reimbursement of providers. While this data is not collected primarily for health research purposes, EHRs and administrative data are frequently used for observational and epidemiological studies (sometimes referred to as outcomes studies). Given the non-randomized nature of studies utilizing EHR and AHD methods are used to minimize the risk of confounding and bias during the design and analysis of studies.

The typical analytic strategies employed with EHR and AHD studies involve multivariate regression models such as linear regression, logistic regression, or time-to-event models such as Cox-proportional hazards models. There is increasing interest in the potential applications of machine learning (ML) and artificial intelligence (AI) methods in mental health research (1) and in analyzing EHR and AHD data (2). ML and AI methods may provide benefits over standard biostatistical regression analysis when there is a high complexity to the underlying data (high dimensionality), which is becoming more common with EHR and AHD data sources as a greater range of information is included in these data sources (e.g., free-text clinical notes from EHR) and greater numbers of AHD sources are available for data linkage and analysis (e.g., laboratory, imaging, genetics) (3).

The application of ML and AI to the analysis of EHR and AHD in fields outside of geriatric mental health is increasing, including the development of recommendations for using ML and AI methods with these datasets (4), as well as studies of ML and AI in biomedical research (5). A recent review of EHR studies using ML or AI approaches for diagnosis or classification identified 99 unique publications across all clinical conditions (6). A review focused on the application of ML and AI approaches to dementia research using EHR identified five studies, although the review included data sources that were not routinely available in EHR and AHD, including neuroimaging or biomarker data (7). A review of the application of ML and AI approaches to research in mental health disorders, including all age groups, identified 28 studies, 6 of which utilized EHR data sources (1). To date, there are no reviews examining the application of ML and AI methods for studies using EHR and AHD in geriatric mental health research.

Our systematic review identifies the current application of ML and AI to EHR and AHD research in geriatric mental health. We identified the number of studies currently available in this field, the characteristics of study populations, data sources, and types of ML and AI methods used, along with potential strengths and limitations of studies, including the quality of study reporting. This review will highlight the current applications of ML and AI in geriatric mental health research and identify opportunities for future application of these methods to understanding geriatric mental health problems using these common data sources.

## MATERIALS AND METHODS

### Research Question

To avoid the likelihood of missing relevant articles, the inquiry is recommended to be broad (8). For this review, our research question was: what research has been undertaken to apply machine learning and artificial intelligence methods to geriatric mental health conditions using EHR or AHD? We further sought to understand the types of geriatric mental conditions included in studies, the purpose of ML or AI approaches, information on sources of data used in studies, and assessments of study quality as part of our review.

### Data Sources and Searches

We conducted this review following a pre-specified protocol and in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting guidelines [(9); **Supplementary Table 1**]. We performed an extensive search using appropriate keywords and medical subject headings to find relevant studies. A predetermined search strategy was employed in relevant databases after consultation with a librarian. We searched MEDLINE, Embase, and PsycINFO (each from inception to April 2021) to identify studies with the assistance of a research librarian. Additionally, we explored the reference lists of all relevant studies for potentially relevant citations. The search strategy was based on five key domains: artificial intelligence, geriatric, mental health, electronic health records, and administrative health data. We used free-text words and Medical Subject Headings (MeSH) terms to identify all relevant studies for each key domain. Certain text words were truncated, or wildcards were used when required. The Boolean operators "AND" and "OR" were used to combine the words and MeSH terms so that the search yielded specific yet comprehensive results. The line-by-line search strategy employed in MEDLINE (Ovid) is provided in **Supplementary Table 2**.

## Eligibility Criteria

We set specific inclusion and exclusion criteria to eliminate irrelevant studies. Only original studies that focused on ML or AI in geriatric mental health using EHR or AHD data were included in this review. We excluded reviews, editorials, commentaries, protocols, conference abstracts, and letters to the editor. We considered all types of study designs, anticipating that ML or AI techniques may use different types of study design. There were also no restrictions on the languages of the studies. However, studies conducted in populations other than older adults were excluded, and we defined older adults as study populations where the average age of participants was 65 years and older.

Studies that did not use EHR or AMD data were also excluded. We considered EHR as studies involving health records from outpatient or inpatient settings where the data was directly extracted from clinical records. Data in EHR studies could include structured data such as diagnoses, medications, or procedures, as well as unstructured data such as free-text clinical notes. Information from studies reporting imaging results within EHR was included provided that the ML or AI methods were applied to information related to the ordering of imaging tests or interpretation of results (e.g., whether an imaging test was performed, or text contained in radiology reports) as this information is commonly available in EHR. We excluded studies that utilized ML or AI approaches to raw imaging data or genetic information, which is not routinely available in abstracted EHR data. Administrative health data included information related to diagnoses, clinical services, prescribed medication, hospitalizations, and emergency department visits. AHD does not typically include free-text clinical notes. Typically, multiple sources of AHD are linked across separate databases for studies using AHD in contrast to EHR data, where all data sources are available from a single EHR record. Studies including both EHR and AHD together were also included. Further, studies that did not consider a mental health issue were also excluded. We included major neurocognitive disorders and dementia in our definition of geriatric mental health conditions in addition to other mental health conditions such as depression, schizophrenia, and suicide. The complete list of terms used in our search strategy is provided in **Supplementary Table 3**.

## Study Selection

Four reviewers (MC, DS, EG, and GC) participated in the study selection and data extraction process. Eligible articles were identified independently by the reviewers using a two-step process. First, all articles identified from the search of electronic databases were exported to Covidence (10) to remove duplicate publications. Next, the title and abstracts of non-duplicated records were independently screened by two reviewers (MC and DS). All studies identified by one of the reviewers as potentially relevant were retained (based on eligibility criteria) and included in the full-text screening. Full-text articles were further screened for eligibility by the same two reviewers (MC and DS) independently. Lastly, the selected articles in the full-text review underwent data extraction, with each article reviewed by two of the four members of the review team. Any disagreement between reviewers was resolved through consensus.

## Data Extraction and Synthesis

For each selected article, two out of the four reviewers independently extracted data using Covidence (10). The following information from each study was extracted: study ID, country, the purpose of the ML/AI, study design, type of dataset used and data format, sample size, ML or AI methods used, predictors (features) used by ML/AI, comparison with other methods, the performance of ML/AI, and main finding(s). As we anticipated substantial heterogeneity in study designs, study populations, and methods, we did not plan to conduct a quantitative meta-analysis of results.

## Study Quality Assessment

ML or AI techniques are generally used for either prediction or diagnostic/classification purposes. Considering this, we assessed each study either by Prediction model Risk OF Bias ASsessment Tool [PROBAST; (11, 12)] or Quality Assessment of Diagnostic Accuracy Studies [QUADAS-2; (13)] checklist depending on the purpose of the selected study. Each reviewer independently assessed study quality.

PROBAST is designed to evaluate the risk of bias and concerns regarding the applicability of diagnostic and prognostic prediction model studies. PROBAST contains 20 questions under four domains: participants, predictors, outcome, and analysis, facilitating judgment of risk of bias and applicability. The overall risk of bias (ROB) of the prediction models was judged as "low", "high", or "unclear", and overall applicability of the prediction models was considered as "low concern", "high concern", and "unclear" according to the PROBAST checklist (11, 12). When a prediction model evaluation is judged as low on all domains, then the model is treated as having "low ROB" or "low concern regarding applicability". On the other hand, when model evaluation is judged as high for at least one domain, then the model is treated as having "high ROB" or "high concern regarding applicability". Finally, when a prediction model evaluation is judged as unclear in one or more domains and is judged as low in the rest, then the model is treated as having "unclear ROB" or "unclear concern regarding applicability".

QUADAS-2 (13) is the modified version of QUADAS, a tool used in systematic reviews to evaluate the risk of bias and applicability of diagnostic accuracy studies. QUADAS-2 consists of 11 signaling questions in four key domains: patient selection, index test(s), reference standard, and flow and timing. Signaling questions helps to judge the ROB as "low", "high", or "unclear". Similar to PROBAST, when a study is judged as "low" on all domains, then the overall ROB and applicability of that study is judged as "low risk of bias" and "low concern regarding applicability". However, if a study is rated "high" or "unclear" in one or more domains, the study may be classified as "at risk of bias" or "concerns regarding applicability". Both PROBAST and QUADAS-2 are used to assess the risk of bias and concerns regarding applicability. In our review, we have used the tools for the assessment of the risk of bias.

## Data Analysis

Descriptive synthesis was undertaken to describe the existing literature on artificial intelligence techniques in geriatric mental

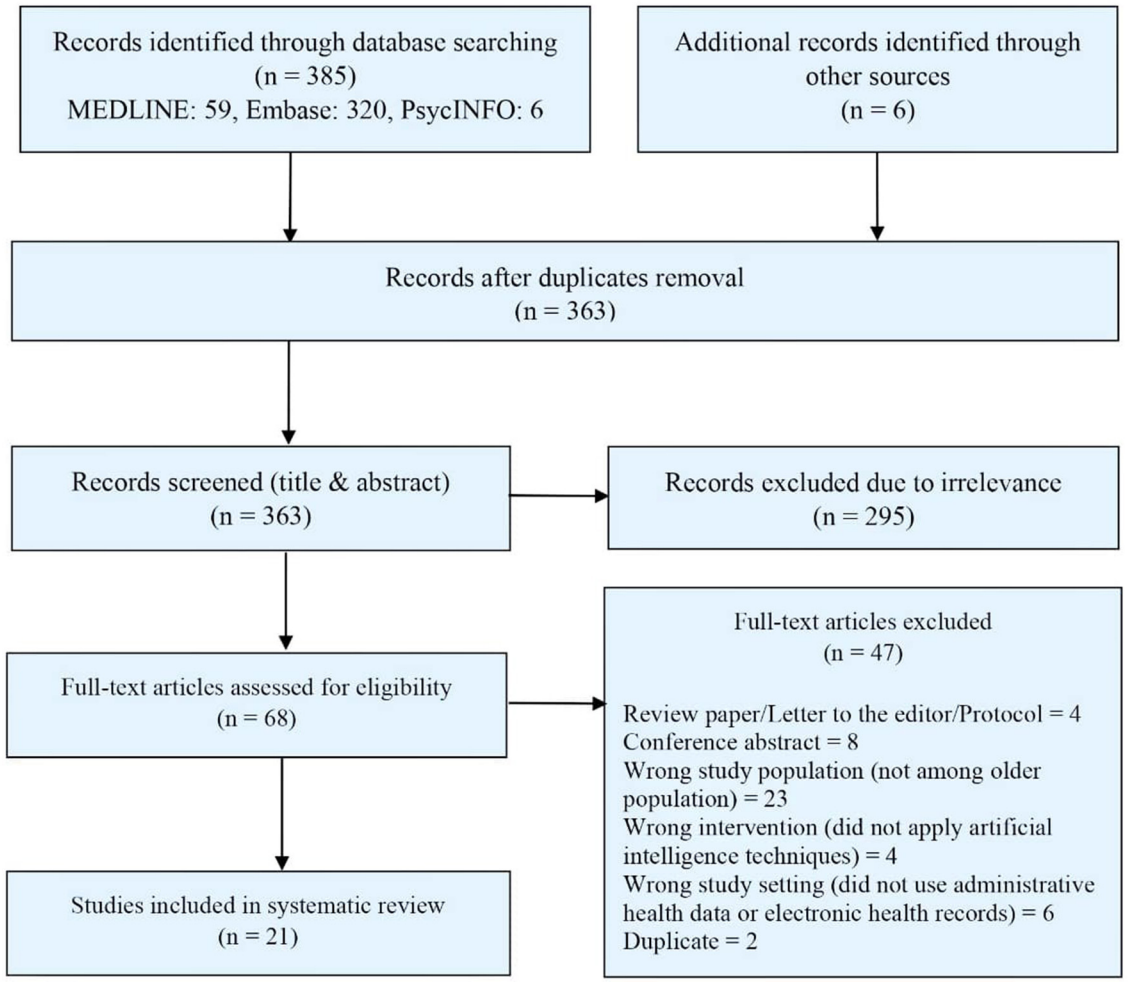


**FIGURE 1 |** PRISMA diagram for the systematic review of studies on artificial intelligence in geriatric mental health using EHR or AHD.

health using AHD or EHR. Using the PRISMA flow diagram (14), we summarized the number of studies identified and those excluded (with the reason for exclusion) and included in the systematic review. The results of included studies were summarized in tables and synthesized in a narrative format.

# RESULTS

## Study Identification and Selection

We identified 385 articles through our electronic database search and an additional six articles through reference search. After removing duplicates, 363 titles and abstracts were screened for eligibility, and from there, 68 articles were selected for full-text screening. After assessing full-texts, 21 articles were finally selected for the systematic review (7, 15–34). The detailed study selection process is summarized in **Figure 1**.

## Study Characteristics

The detailed characteristics of the studies included in this review are presented in **Table 1**. Among the studies we identified, most of the studies were conducted in the USA ($N = 12$). The remaining studies were conducted in Spain ($N = 3$), UK ($N = 2$), South Korea ($N = 2$), France ($N = 1$), and Austria ($N = 1$). Study designs mainly were cohort ($N = 14$) followed by case-control studies ($N = 5$) and cross-sectional studies ($N = 1$). The study design was not reported in one study. The sample size of the individual studies varied between 1,909 and 17,227,820.

EHR was the most used data type and was used as the sole data source by 14 different studies. Five studies used only AHD, while two studies used both EHR and AHD. The dataset pattern was structured in nature in most studies ($N = 11$), unstructured free-text in only three studies, and seven studies used both structured and unstructured free-text data.

## ML and AI Methods Used in Studies

There were considerable variations among the specific ML or AI methods used by different studies. Random forest was the most common ML technique and was used by seven studies (7, 20, 28, 30–32, 34). Both natural language processing (NLP) (18, 21, 25, 29) and logistic regression modeling techniques

**TABLE 1 |** Study characteristics of the included studies.

| Study ID | Country | Purpose of the ML/AI | Study design | Type of datasets used and data format | Sample size | ML or AI methods used | Predictors (features) used by ML/AI | Comparison with other methods | Performance of ML/AI | Main finding |
|---|---|---|---|---|---|---|---|---|---|---|
| Kim et al. (15) | South Korea | Prediction of dementia | Cohort | AHD, structured | 11,443 | SVM, WEKA | Demographics, diagnoses, clinical measurements, laboratory values | Two sets of features compared | Longitudinal Model 1: Best accuracy = 72.0% | SVM approaches can be used to predict who will develop dementia |
| Nori et al. (16) | USA | Predict dementia incidence | Nested case-control | AHD, structured | Cases = 44,945, Controls = 760,646 | Lasso, logistic regression | 50 predictors | No | AUC = 0.693 | On large datasets, ML methods can automatically recruit many predictors |
| Fisher et al. (27) | USA | Model progression of Alzheimer's disease | Cohort study | EHR, structured | 1,909 | Conditional Restricted Boltzmann Machine | 44 variables, cognition laboratory, clinical information | RF | Accuracy, correlation coefficient, R square, AUC, and the root mean square errors | Predict disease progression of Alzheimer's disease |
| Wang et al. (29) | USA | Predict death among people with dementia | Cohort | EHR, structured and unstructured | 26,921 | DNN, LSTM, and NLP | 500 features, patient demographic variables | No | AUC = 0.956 [0.955–0.956] (1-year model) | DNN can be accurate in predicting mortality and could be used as a proxy for selecting patients |
| Mar et al. (30) | Spain | Predict dementia-related neuropsychiatric symptoms | Cohort | EHR, unstructured | 4,003 | RF | 62 variables including demographics, chronic disease, prescriptions | No | AUC = 0.80 for the psychotic cluster model, AUC = 0.74 for the depressive cluster model | Predict the presence of psychotic and/or depressive symptoms in dementia-diagnosed patients |
| Park et al. (31) | South Korea | Predict incidence of Alzheimer's disease | Cohort | AHD, structured | 40,736 | RF, SVM logistic regression | Demographics, lab tests, medication prescriptions, diagnoses | Logistic regression model | Best AUC of 0.898, 0.775, and 0.725 in predicting baseline, 1- and 4-year incident AD | RF outperformed SVM and logistic regression for every prediction year in predicting AD |
| Jauk et al. (32) | Austria | Identify patients at high risk for delirium | Cohort | EHR, structured | 4,663 | RF | Demographics, diagnoses, laboratory, nursing notes | Clinical expert risk scores | Sensitivity = 74.1%, Specificity = 82.2%, AUC = 0.86, Calibration = Poor | Demonstrated ML's predictive power for delirium |
| Miled et al. (7) | USA | Predict onset of dementia | Case-control | EHR, structured and unstructured | Training: 7,644, Testing: 17,760 | RF | 235 features, prescriptions, diagnosis, medical notes, demographics | No | The best model (highest accuracy) accuracy of 77.43% | Model which is generalizable and can predict dementia within 1 year |
| Nori et al. (33) | USA | Predict the risk for incident dementia and mild cognitive impairment | Nested case-control | AHD and EHR, structured | 561,093 cases 16,666,727 controls | Gradient boosted trees, feed-forward network, recurrent neural network | Medications, diagnoses, procedures, demographics, and health service utilization | Comparison of deep learning and ML model | AUC on test data at year 0: 92.4% (BT), 93.1% (FFN), 94.4% (RNN); AUC on validation data at year 8: 79.9% (BT), 80.7% (FFN), 77.0% (RNN) | FFN performed best among the three deep network models, but marginally better than boosted trees |
| Mar et al. (34) | Spain | Identify clusters of depressive and psychotic symptoms of dementia | Cohort | AHD, structured | 631,949 | RF | Demographics, medical conditions, and medications | No | AUC = 0.8 | Estimations of psychotic symptoms, depressive symptoms, incidence, prevalence of dementia |
| Tsang et al. (17) | UK | Predict hospitalization among dementia patients | Cohort | EHR, structured | 59,298 | Entropy regularization with ensemble DNN | 54,649 unique features or event codes | RF | Accuracy = 0.759 (reduced features), Accuracy = 0.755 (full feature) | Method for predicting hospitalization of patients suffering from dementia |
| Anzaldi et al. (18) | USA | Determine the presence of geriatric syndromes and frailty | Cohort | AHD and EHR, structured and unstructured | 18,341 | NLP | Seed phrases for each syndrome | No | Correlations between frailty and NLP definitions of syndromes were between 0.07 and 0.27 | Patients identified as "frail" had higher healthcare utilization and geriatric syndromes |

*(Continued)*

**TABLE 1** | Continued

| Study ID | Country | Purpose of the ML/AI | Study design | Type of datasets used and data format | Sample size | ML or AI methods used | Predictors (features) used by ML/AI | Comparison with other methods | Performance of ML/AI | Main finding |
|---|---|---|---|---|---|---|---|---|---|---|
| Wang et al. (19) | USA | Identify important themes in provider's notes | Cohort | EHR, unstructured | 7,875 | Latent Dirichlet allocation topic model | 250 topics | No | The trends of the topics from the clinical notes were compared to the structured data using Pearson's correlation | Analyze the disease course during the last 2 years of life |
| Halladay et al. (20) | USA | Develop a prediction rule for delirium | Cohort | EHR, structured | 39,377 | RF | 16 features, demographics, cognition, infection, lab tests, diagnosis | Electronic delirium prediction rules | AUC compared to clinical rules = 0.81–0.91 | A prevalent delirium prediction rule was developed |
| Kharrazi et al. (21) | USA | Assess the value of EHR to identify geriatric syndromes | Cohort | EHR, structured and unstructured | 18,341 | NLP | Unstructured electronic health record data (free-text clinical notes) | No | Sensitivity: 87.5–100% across the geriatric syndromes, Specificity: 95.4–100% across the geriatric syndromes | NLP method was incorporated in identification of individuals with geriatric syndromes |
| Chen et al. (22) | USA | Identify older adults automatically with geriatric syndromes from free text EHRs | Cohort | EHR, unstructured | 18,341 | End-to-end neural architecture, DNN | Diagnoses, target sentence, surrounding sentences, document | No | At sentence level: best model achieved a micro-F1 of 0.605. At patient level: best model achieved a micro-F1 of 0.843 | EHR free text can be used to identify older adults with geriatric syndromes using proposed deep learning system |
| Violán et al. (23) | Spain | Identify multimorbidity pattern | Cross-section | EHR, structured | 916,619 | Fuzzy c-means cluster analysis | 62–49, for a "no" to 2% prevalence threshold, respectively | No | O/E ratio and exclusivity was reported for different diseases within eight different patterns | Multimorbidity patterns were obtained in an elderly population |
| Shao et al. (24) | USA | Detection of dementia | Case-control | EHR, structured and unstructured | 11,166 | Topic modeling and logistic regression model | Total 853 features: non-dementia diagnoses, medications, and clinical notes | No | Optimal sensitivity = 0.825 and specificity = 0.832. AUC "Unclear = Dementia" = 0.912 | Demonstrated automated methods feasibility to identify topics from EHR that can be used to assign a dementia risk score |
| Topaz et al. (25) | USA | Identify common neuropsychiatric symptoms of dementia | Cohort | EHR, structured and unstructured | 89,459 | NLP | Home healthcare free-text clinical notes | No | Average performance: F-score = 0.88, Precision = 0.87, Recall = 0.91 | Identification of neuropsychiatric symptoms of dementia |
| Cabeli et al. (26) | France | Develop an approach to uncover relationships between mixed-type data from medical records | Not reported | EHR, structured | 1,628 | MIIC network learning algorithm | 107 variables of different types (19 continuous and 88 categorical variables) | No | Not reported | Provides a user-friendly global visualization tool which could help other practitioners visualize and analyze effects from patient medical records |
| Ford et al. (28) | UK | Automatically detect undiagnosed dementia | Case-control | EHR, structured and unstructured | 93,426 | Logistic regression, naïve Bayes, RF | Eight diagnostic codes and nine keywords | Logistic regression | AUC Random Forest = 0.94, AUC Naive Bayes = 0.90, AUC Logistic Regression = 0.94 using both codes and keywords | Identified undiagnosed patients with dementia |

*AHD, administrative health data; DNN, deep neural network; HER, electronic health record; ML, machine learning; NLP, natural language processing; RF, random forest; SVM, support vector machine; AUC, area under the curve; AD, Alzheimer's disease; LSTM, the long short-term memory; MIIC, multivariate information-based inductive causation.*

(16, 24, 28, 31) were used by four studies. Three studies used deep learning approaches, including deep neural networks (17, 22, 29). Support vector machine (15, 31) and topic modeling (19, 24) were used by two studies each. Finally, lasso (33), naïve Bayes (28), multivariate information-based inductive causation (MIIC) network learning algorithm (26), fuzzy c-means cluster analysis (23), conditional restricted Boltzmann machine (27), WEKA (15), and gradient boosted trees (33) were applied by one study each.

Various sets of features, or sets of predictors, were considered in the identified studies. They include patient demographics, body measurements, history of family illness, personal disease history, administrative, diagnoses, laboratory, prescriptions, medications, medical notes, procedures, health service utilization, clinical and background information, topics in clinical notes, and ICD codes.

ML and AI methods were used for a variety of reasons. However, most of the studies applied the techniques either for prediction purposes ($N = 9$) or for classification or diagnosis purposes ($N = 9$). Among other objectives, one study used the methods to estimate the population incidence and prevalence of dementia and neuropsychiatric symptoms, one study to compare patients who are described in clinical notes as "frail" to other older adults concerning geriatric syndrome burden and healthcare utilization, and one study to compute and assess the significance of multivariate information between any combination of mixed-type variables and their relation to medical records.

Different outcomes were considered by different studies while incorporating ML or AI techniques. Major neurocognitive disorder, dementia, and Alzheimer's disease were the most common conditions among the articles included in our study and were reported in 11 studies (7, 15, 16, 24, 25, 27, 28, 30, 31, 33, 34). Among the other outcomes, a geriatric syndrome (falls, malnutrition, dementia, severe urinary control issues, absence of fecal control, visual impairment, walking difficulty, pressure ulcers, lack of social support, and weight loss) in 3 studies (18, 21, 22), delirium in 2 studies (20, 32), mild cognitive impairment (33), cognitive disorder (26), multimorbidity pattern (23), mortality (29), hospital admission (17), and themes/topics mentioned in care providers' notes (19) were considered once. Outcomes were primarily defined using ICD diagnosis codes in AHD and EHR.

The majority of the studies did not compare the ML and AI algorithms with any other biostatistical methods except a few where a comparison with logistic regression was made. ML/AI techniques outperformed logistic regression in one study while performed similarly in another study. The area under the receiver operating characteristics curve (AUC) was the most commonly reported performance measure of ML and AI algorithms, with values ranging from 0.69 to 0.98. None of the studies performed model validation in an external population where the performance of a model's prediction or classification can be generalized to unseen new data. As such, we could not evaluate any of the model's generalizability in this study.

**TABLE 2 |** Study quality assessment using PROBAST.

| Study ID | Domain 1. Participants | Domain 2. Predictors | Domain 3. Outcome | Domain 4. Analysis |
|---|---|---|---|---|
| Kim et al. (15) | + | + | ? | – |
| Nori et al. (16) | + | + | + | + |
| Fisher et al. (27) | + | + | ? | ? |
| Wang et al. (29) | + | + | + | ? |
| Mar et al. (30) | + | + | + | ? |
| Park et al. (31) | + | + | + | + |
| Jauk et al. (32) | + | + | ? | ? |
| Ben Miled et al. (7) | + | + | – | ? |
| Nori et al. (33) | + | ? | + | + |
| Mar et al. (34) | + | + | + | + |
| Tsang et al. (17) | + | + | + | ? |

*+ indicates low ROB; – indicates high ROB; ? indicates unclear ROB.*

## Study Quality Assessment

Study quality was assessed either by PROBAST or QUADAS-2, depending on the nature of the outcome. For example, in studies where the purpose was a prediction, we assessed quality with PROBAST. Nevertheless, when the study purpose was identification or classification, we assessed quality using QUADAS-2. Thus, PROBAST was applied in 11 studies, while QUADAS-2 was involved in 10 studies.

When PROBAST was applied, ROB was assessed as low in 100% of the studies according to the signaling questions of the "participants" domain, 91% of the studies according to the "predictors" domain, and 64% of the studies according to the "outcome" domain (**Table 2**). However, only in 36% of the studies ROB was assessed as low according to the "analysis" domain's signaling questions and was unclear in 55% of the studies. ROB was estimated as high in terms of both outcome and analysis in one study when PROBAST was applied. Only in three studies (16, 31, 34) low ROB was observed in all of the PROBAST domains.

When QUADAS-2 was applied, ROB was low in 80% of the studies according to the signaling questions of the "patient selection" domain (**Table 3**). However, in most of the studies, ROB was unclear according to the signaling questions of the other three domains of QUADAS-2. For example, ROB was unclear in 50% of the studies according to the "index test", 60% of the studies according to the "reference standard", and 80% of the studies according to the "flow and timing" domain's signaling questions. In each of the domains of QUADAS-2 except "flow and timing", there was one study where ROB was assessed as high. In none of the studies, low ROB was observed according to all the QUADAS-2 domains.

# DISCUSSION

Our review identified the application of ML and AI techniques in geriatric mental health using EHR or AHD data. We were able to identify 21 studies with all studies published recently within the past 5 years, suggesting the increasing application

**TABLE 3 |** Study quality assessment using QUADAS-2.

| Study ID | Domain 1. Patient selection | Domain 2. Index test(s) | Domain 3. Reference standard | Domain 4. Flow and timing |
|---|---|---|---|---|
| Anzaldi et al. (18) | + | + | + | ? |
| Wang et al. (19) | + | ? | ? | ? |
| Halladay et al. (20) | + | − | + | + |
| Kharrazi et al. (21) | + | ? | + | ? |
| Chen et al. (22) | + | + | ? | ? |
| Violán et al. (23) | + | ? | ? | ? |
| Shao et al. (24) | + | + | ? | ? |
| Topaz et al. (25) | + | + | − | + |
| Cabeli et al. (26) | ? | ? | ? | ? |
| Ford et al. (28) | − | ? | ? | ? |

*+ indicates low ROB; – indicates high ROB; ? indicates unclear ROB.*

of ML and AI in this topic. As anticipated, ML or AI techniques were predominantly used either for prediction or classification purposes, and dementia was the most frequent condition considered in the studies. Both EHR and AHD data were considered; however, EHR data was the most frequent data source. There were considerable variations among ML, and AI techniques applied, ranging from more traditional ML techniques such as random forest to more advanced deep neural networks. The quality of study reporting was variable, with the majority of studies having unclear reporting of elements related to study quality. The relative advantages of ML or AI techniques compared to biostatistical methods were not assessed in most studies.

A recent review by Graham et al. (1) on a broader topic (AI for mental health and mental illness) identified 28 studies. However, their review is different from our systematic review in many different ways. First, the review by Graham et al. was not a systematic review, and the search was performed in PubMed and Google Scholar only with studies published between 2015 and 2019. In contrast, we performed a systematic review utilizing three databases without any time constraints. Second, their search was also not restricted to EHR and AHD data as ours; instead, they considered studies with data from all sources, including social media platforms, novel monitoring systems such as smartphone and video and brain imaging data. Third, their review included studies with participants from all the age groups starting from 14 years as opposed to our study, which focused on geriatric mental health where study participants were older adults. Fourth, neurocognitive disorders (e.g., dementia) were the primary outcome in most of our included studies. In contrast, Graham et al. did not consider studies with neurocognitive disorders in their review, and depression was identified as the most common mental illness. Nevertheless, supervised machine learning (e.g., random forest) was the most commonly used AI technique according to their review, similar to our findings. Another recent study by Elizabeta et al. (35) reviewed AI in the healthcare of older adults. They did not mention any specific number of studies; instead, they discussed some studies where ML or AI approaches were applied in the medical care of older people and concerns associated with AI use in medicine. However, the study is fundamentally different from our study in the sense that they consider overall healthcare, whereas our focus is only on mental health. Our review provides additional information about AI and ML in healthcare focused on the mental health of older adults and applications specific to EHR and AHD data sources.

EHR and AHD are rich resources that capture information of all the medical investigators involved in patients' healthcare records and provide ample opportunity to utilize this information for research, including mental health research. However, there are challenges associated with EHR and AHD data mainly due to the large sample sizes available, the volume of longitudinal data on participants, incompleteness, and inconsistency (6). Therefore, there is a potential role for automated analytic methods for disease diagnosis and prognosis or prediction from EHR and AHD data. Data-driven ML and AI techniques can overcome the challenges related to the volume, potential complexity, or dimensionality of EHR data. Information stored in EHR and AHD can be fall under two broad categories: structured (e.g., diagnosis, prescriptions, medical tests, etc.) and unstructured free texts [e.g., medical notes; (7)]. The use of structured data (i.e., diagnostic codes, prescription medications) is more extensive in many areas of health, primarily due to its limited pre-processing requirement compared to unstructured data. On the other hand, unstructured data primarily derived from medical notes poses additional challenges due to the difficulty of transferring free text into structured features. Nevertheless, unstructured data has also been applied to build models for different disease conditions, including dementia (7). NLP-based AI methods can translate unstructured text data into structured forms more amenable to machine inference or perform inference without explicit intervening translation. Combining structured and unstructured EHR data and using them to build ML models can produce better performance than each data source independently (7). Our study also noticed seven studies used combined data in predicting mortality and dementia and the diagnosis of geriatric syndromes and dementia.

Recently, increased emphasis has been put on using ML or AI tools in clinical research, particularly related to precision medicine (36, 37). Modern ML techniques offer benefits over traditional statistical methods due to their ability to detect complex non-linear relations, high-dimensional interactions among the features, and their capability to handle gigantic amounts of data. Since machine learning tools are more recent, advanced, and have the reputation of producing more accurate predictive performance, we anticipated studies using these tools might demonstrate improved predictive performance compared with the studies using more common biostatistical analytic methods. In our review, we identified only two studies (28, 31) comparing ML approaches with statistical methods. One study, Park et al. (31), found that the predictive performance of ML techniques random forest and support vector machine were superior compared to logistic regression in predicting Alzheimer's disease. In contrast, a similar predictive performance between random forest and naïve Bayes ML techniques and logistic regression was observed in predicting dementia in the study by Ford et al. (28). Overall this is in keeping with other findings that ML algorithms tend to provide mixed evidence for improving predictive performance compared with conventional statistical models in the other domains of health (38–42).

One of the considerations related to identifying situations where ML or AI may outperform biostatistical approaches include situations where the dataset is large and there are many complex and interrelated features or predictors. In situations where these conditions are not present, the performance of ML and AI techniques may not provide more accurate results when compared to biostatistical methods, even when they require additional expertise and computing resources to realize. While AI and ML may offer benefits in some situations, the potential limitations of these methods and optimal strategies for employing them in mental health research need to be carefully considered. Moreover, the inference of some ML models, such as neural networks, is hard to explain. Behavioral and performance explainability of ML models is a critical issue pertaining to whether "black box" models can be trusted, whether they appropriately infer from their input features, and whether they generalize well to "unseen" data (43).

Our review identified a lack of standard reporting in this area of literature. Authors often reported different aspects of the ML algorithms and in varying ways, which created difficulty for data collection and standardization within this review. In addition, the results of the ML study findings are often insufficiently reported primarily due to the inherent complexity of machine learning methods and the flexibility in specifying machine learning models, which hinders reliable assessment of study validity and consistent interpretation of study findings (5). Recently, new guidelines have been introduced with a list of reporting items to be included in a research article and a set of practical sequential steps to be followed for developing predictive models using ML (5). A new initiative to establish a TRIPOD (44) statement specific to machine learning (TRIPOD-ML) is also underway to guide authors to develop, evaluate and report ML algorithms properly (45). These reporting guidelines may assist authors in improving reporting in future studies in this area, particularly for research studies published in clinically oriented publications in contrast to engineering or computer science-focused publications.

The clinical implications of our findings include considerations related to the future application of ML and AI in geriatric mental health research (3, 46). ML and AI algorithms are typically used to classify or predict, translating to clinical applications related to diagnosis and prognosis (47). Mental health diagnoses are clinical compared to some other fields of medicine where diagnoses may be based on quantitative assessments or laboratory investigations. ML and AI analytic approaches may be well-suited to improve diagnostic accuracy in complex classification problems such as mental health diagnoses. To date, much of the research on this topic is confined to diagnosing dementia, although, as our review indicated, there is some research now related to the identification of geriatric syndromes or patterns of behavioral symptoms in dementia. ML and AI approaches require further study in diagnosing addictions and mental health problems in older adults. While only a few studies directly compare ML or AI approaches to more commonly used biostatistical methods, ML and AI may provide promising advances in disease state classification, particularly with more complex data.

Similarly, ML and AI may also provide improved prediction of the onset or progression of addictions and mental health problems. To date, the main clinical conditions that these methods have been applied to have been the onset of dementia. However, ML and AI approached may also facilitate improved prognostic models for predicting the onset of other geriatric mental health disorders. Predicting treatment response for an individual and personalizing therapeutic interventions based on this information, also known as precision medicine, is another potential application of ML and AI in geriatric mental health (47). Finally, ML and AI methods may be well-suited to analyzing unstructured data such as free-text clinical notes increasingly available in EHR. Incorporating clinician-generated data from unstructured data sources is likely to improve diagnostic or predictive performance when compared analyses conducted using only structured data such as diagnoses or laboratory values. Our review has identified current clinical applications of ML and AI approaches and highlights potential future areas for research and clinical applications related to research using EHR and AHD in geriatric mental health. While research on ML and AI in geriatric mental health is in its early stages it is anticipated that these methods will be increasingly used and have the potential to transform research and clinical care in this field as in other fields of medicine (48).

To the best of our knowledge, this is the first systematic review on the application of ML or AI in geriatric mental health conditions using EHR or AHD data, and a detailed critical appraisal of the applications has been performed. One of our study's strengths is the extent of the systematic search, which includes massive use of keywords and MeSH terms while searching three different databases and extensive use of the reference lists of the identified studies. We did not place any restrictions on language, geographical location, or time periods to keep our search broad. Consequently, there was little chance

that any relevant study was missed. Nevertheless, our study also has limitations. We did not perform a search on gray literature. A search on electronic databases along with the gray literature could make the search more comprehensive. Although many of our identified studies were diagnostic or prognostic models and a meta-analysis of performance measures (e.g., AUC) of the models could provide a comprehensive summary of the performance of these models (49), we did not perform any meta-analysis from the studies due to their high heterogeneity. Failing to assess publication bias amongst the studies is another potential limitation of this study. Nevertheless, we assessed ROB associated with the studies using PROBAST and QUADAS-2 checklists.

In conclusion, we identified existing research on geriatric mental health in this study where ML or AI techniques were applied using EHR or AHD data. We were able to locate a relatively small number of studies that suggest ML or AI application in geriatric mental health is relatively uncommon at present, although this field is rapidly expanding throughout healthcare research. Outside of the clinical topic of dementia, there are few studies of other geriatric mental health conditions such as depression, anxiety, or suicide where ML and AI may be helpful. The lack of consistent information in the selected studies indicates that improvements in the quality of reporting of ML and AI in the future may also help improve research in this field. Additional information on how ML or AI approaches may be best utilized in EHR and AHD studies is required, including information about when these approaches are more or less likely to produce more accurate results compared to typical biostatistical analyses. Overall, ML and AI tools can play a vital role in redefining the diagnosis of mental illness using a secondary data source, thus facilitating early disease detection, a better understanding of disease progression, optimizing medication/treatment dosages, and uncovering novel treatments for geriatric mental health conditions.

## AUTHOR CONTRIBUTIONS

All authors contributed to this work. DS contributed to the conception and design of the review. MC and DS read and screen abstracts and titles of potentially relevant studies, and screened the full-text papers. MC, DS, W-YC, and EC extracted data and rating the quality independently. MC performed the analysis and drafted the manuscript. DS and W-YC critically reviewed it and suggested amendments before submission. All authors approved the final version of the manuscript and take responsibility for the integrity of the reported findings.

## FUNDING

This research received no grant from any funding agency in public, commercial, or not-for-profit sectors. Partial support for this project was provided by the Canadian Institutes of Health Research—Canadian Consortium on Neurodegeneration in Aging.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.738466/full#supplementary-material

## REFERENCES

1. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial intelligence for mental health and mental illnesses: an overview. *Curr Psychiatry Rep.* (2019) 21:116. doi: 10.1007/s11920-019-1094-0
2. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol.* (2019) 188:2222–39. doi: 10.1093/aje/kwz189
3. Thesmar D, Sraer D, Pinheiro L, Dadson N, Veliche R, Greenberg P. Combining the power of Artificial intelligence with the richness of healthcare claims data: opportunities and challenges. *Pharmacoeconomics.* (2019) 37:745–52. doi: 10.1007/s40273-019-00777-6
4. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. *Annu Rev Public Health.* (2019) 41:21–36. doi: 10.1146/annurev-publhealth-040119-094437
5. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res.* (2016) 18:1–10. doi: 10.2196/jmir.5870
6. Latif J, Xiao C, Tu S, Rehman SU, Imran A, Bilal A. Implementation and use of disease diagnosis systems for electronic medical records based on machine learning: a complete review. *IEEE Access.* (2020) 8:150489–513. doi: 10.1109/ACCESS.2020.3016782
7. Ben Miled Z, Haas K, Black CM, Khandker RK, Chandrasekaran V, Lipton R, et al. Predicting dementia with routine care EMR data. *Artif Intell Med.* (2020) 102:101771. doi: 10.1016/j.artmed.2019.101771
8. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol Theory Pract.* (2005) 8:19–32. doi: 10.1080/1364557032000119616
9. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ.* (2009) 339:b2535. doi: 10.1136/bmj.b2535
10. Babineau J. Product review: covidence (systematic review software). *J Can Heal Libr Assoc.* (2014) 35:68. doi: 10.5596/c14-016
11. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med.* (2019) 170:51–8. doi: 10.7326/M18-1376
12. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Int Med.* (2019) 170:W1–33. doi: 10.7326/M18-1377
13. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* (2011) 155:529–36. doi: 10.7326/0003-4819-155-8-201110180-00009
14. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med.* (2009) 6:e1000100. doi: 10.1371/journal.pmed.1000100
15. Kim H, Chun HW, Kim S, Coh BY, Kwon OJ, Moon YH. Longitudinal study-based dementia prediction for public health. *Int J Environ Res Public Health.* (2017) 14:983. doi: 10.3390/ijerph14090983
16. Nori V, Hane C, Martin D, Kravetz A, Sanghavi D. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. *bioRxiv.* (2018) 14:396127. doi: 10.1101/396127
17. Tsang G, Zhou SM, Xie X. Modeling large sparse data for feature selection: hospital admission predictions of the dementia patients using primary

care electronic health records. *IEEE J Transl Eng Heal Med.* (2021) 9:3000113. doi: 10.1109/JTEHM.2020.3040236
18. Anzaldi LJ, Davison A, Boyd CM, Leff B, Kharrazi H. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. *BMC Geriatr.* (2017) 17:1–7. doi: 10.1186/s12877-017-0645-7
19. Wang L, Lakin J, Riley C, Korach Z, Frain LN, Zhou L. Disease trajectories and end-of-life care for dementias: latent topic modeling and trend analysis using clinical notes. *AMIA Annu Symp Proc.* (2018) 2018:1056–65.
20. Halladay CW, Sillner AY, Rudolph JL. Performance of electronic prediction rules for prevalent delirium at hospital admission. *JAMA Netw Open.* (2018) 1:e181405. doi: 10.1001/jamanetworkopen.2018.1405
21. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. *J Am Geriatr Soc.* (2018) 66:1499–507. doi: 10.1111/jgs.15411
22. Chen T, Dredze M, Weiner JP, Kharrazi H. Identifying vulnerable older adult populations by contextualizing geriatric syndrome information in clinical notes of electronic health records. *J Am Med Informatics Assoc.* (2019) 26:787–95. doi: 10.1093/jamia/ocz093
23. Violán C, Foguet-Boreu Q, Fernández-Bertolín S, Guisado-Clavero M, Cabrera-Bean M, Formiga F, et al. Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a Mediterranean population. *BMJ Open.* (2019) 9:e029594. doi: 10.1136/bmjopen-2019-029594
24. Shao Y, Zeng QT, Chen KK, Shutes-David A, Thielke SM, Tsuang DW. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Med Inform Decis Mak.* (2019) 19:1–11. doi: 10.1186/s12911-019-0846-4
25. Topaz M, Adams V, Wilson P, Woo K, Ryvicker M. Free-text documentation of dementia symptoms in home healthcare: a natural language processing study. *Gerontol Geriatr Med.* (2020) 6:233372142095986. doi: 10.1177/2333721420959861
26. Cabeli V, Verny L, Sella N, Uguzzoni G, Verny M, Isambert H. Learning clinical networks from medical records based on information estimates in mixed-type data. *PLoS Comput Biol.* (2020) 16:e1007866. doi: 10.1371/journal.pcbi.1007866
27. Fisher CK, Smith AM, Walsh JR, Simon AJ, Edgar C, Jack CR, et al. Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Sci Rep.* (2019) 9:1–14. doi: 10.1038/s41598-019-49656-2
28. Ford E, Sheppard J, Oliver S, Rooney P, Banerjee S, Cassell JA. Automated detection of patients with dementia whose symptoms have been identified in primary care but have no formal diagnosis: a retrospective case-control study using electronic primary care records. *BMJ Open.* (2021) 11:e039248. doi: 10.1136/bmjopen-2020-039248
29. Wang L, Sha L, Lakin JR, Bynum J, Bates DW, Hong P, et al. Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions. *JAMA Netw Open.* (2019) 2:e196972. doi: 10.1001/jamanetworkopen.2019.6972
30. Mar J, Gorostiza A, Ibarrondo O, Cernuda C, Arrospide A, Iruin A, et al. Validation of random forest machine learning models to predict dementia-related neuropsychiatric symptoms in real-world data. *J Alzheimer's Dis.* (2020) 77:855–64. doi: 10.3233/JAD-200345
31. Park JH, Cho HE, Kim JH, Wall MM, Stern Y, Lim H, et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *NPJ Digit Med.* (2020) 3:46. doi: 10.1038/s41746-020-0256-0
32. Jauk S, Kramer D, Großauer B, Rienmüller S, Avian A, Berghold A, et al. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. *J Am Med Informatics Assoc.* (2020) 27:1383–92. doi: 10.1093/jamia/ocaa113
33. Nori VS, Hane CA, Sun Y, Crown WH, Bleicher PA. Deep neural network models for identifying incident dementia using claims and EHR datasets. *PLoS ONE.* (2020) 15:e0236400. doi: 10.1371/journal.pone.0236400
34. Mar J, Gorostiza A, Arrospide A, Larrañaga I, Alberdi A, Cernuda C, et al. Estimation of the epidemiology of dementia and associated neuropsychiatric symptoms by applying machine learning to real-world data. *Rev Psiquiatr Salud Ment.* (2021). doi: 10.1016/j.rpsm.2021.03.001. [Epub ahead of print].
35. Elizabeta B M-L, Tracy H, John M. Artificial intelligence in the healthcare of older people. *Arch Psychiatry Ment Heal.* (2020) 4: 7–13. doi: 10.29328/journal.apmh.1001011
36. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. *BMC Med Inform Decis Mak.* (2018) 18:1–15. doi: 10.1186/s12911-018-0719-2
37. Chowdhury MZI, Turin TC. Precision health through prediction modelling: factors to consider before implementing a prediction model in clinical practice. *J Prim Health Care.* (2020) 12:3–9. doi: 10.1071/HC19087
38. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open.* (2020) 3:e1918962. doi: 10.1001/jamanetworkopen.2019.18962
39. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol.* (2013) 66:398–407. doi: 10.1016/j.jclinepi.2012.11.008
40. Tollenaar N, van der Heijden PGM. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *J R Stat Soc Ser A Stat Soc.* (2013) 176(Pt 2):565–84. doi: 10.1111/j.1467-985X.2012.01056.x
41. Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. *Stud Health Technol Inform.* (2004) 107(Pt 1):736–40. doi: 10.3233/978-1-60750-949-3-736
42. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol.* (2017) 2:204–9. doi: 10.1001/jamacardio.2016.3956
43. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x
44. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* (2015) 13:1–10. doi: 10.1186/s12916-014-0241-z
45. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet.* (2019) 393:1577–9. doi: 10.1016/S0140-6736(19)30037-6
46. Lindsell CJ, Stead WW, Johnson KB. Action-Informed Artificial intelligence - matching the algorithm to the problem. *JAMA J Am Med Assoc.* (2020) 323:2141–2. doi: 10.1001/jama.2020.5035
47. Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. *Am J Med.* (2019) 132:795–801. doi: 10.1016/j.amjmed.2019.01.017
48. Matheny M, Israni ST, Ahmed M. Artificial intelligence in health care. *Natl Acad Med.* (2018) 1–269.
49. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ.* (2017) 356:i6460. doi: 10.1136/bmj.i6460

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 20 October 2021
doi: 10.3389/fpsyt.2021.738494

# Machine Learning Prediction of Treatment Outcome in Late-Life Depression

*Adrienne Grzenda[1]*, William Speier[2], Prabha Siddarth[1,3], Anurag Pant[3], Beatrix Krause-Sorio[1,3], Katherine Narr[3,4] and Helen Lavretsky[1,3]*

*[1] Department of Psychiatry and Biobehavioral Science, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States, [2] Medical Imaging and Informatics Group, Department of Radiological Sciences, University of California, Los Angeles, Los Angeles, CA, United States, [3] Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, CA, United States, [4] Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States*

OPEN ACCESS

***Edited by:***
*Helmet Karim,*
*University of Pittsburgh, United States*

***Reviewed by:***
*Joseph Kazan,*
*University of Pittsburgh, United States*
*Joanne C. Beer,*
*University of Pennsylvania, United States*
*Amin Zandvakili,*
*Warren Alpert Medical School of Brown University, United States*

****Correspondence:***
*Adrienne Grzenda*
*agrzenda@mednet.ucla.edu*

***Specialty section:***
*This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry*

***Received:*** *08 July 2021*
***Accepted:*** *20 September 2021*
***Published:*** *20 October 2021*

***Citation:***
*Grzenda A, Speier W, Siddarth P, Pant A, Krause-Sorio B, Narr K and Lavretsky H (2021) Machine Learning Prediction of Treatment Outcome in Late-Life Depression. Front. Psychiatry 12:738494. doi: 10.3389/fpsyt.2021.738494*

**Background:** Recent evidence suggests that integration of multi-modal data improves performance in machine learning prediction of depression treatment outcomes. Here, we compared the predictive performance of three machine learning classifiers using differing combinations of sociodemographic characteristics, baseline clinical self-reports, cognitive tests, and structural magnetic resonance imaging (MRI) features to predict treatment outcomes in late-life depression (LLD).

**Methods:** Data were combined from two clinical trials conducted with depressed adults aged 60 and older, including response to escitalopram ($N = 32$, NCT01902004) and Tai Chi ($N = 35$, NCT02460666). Remission was defined as a score of 6 or less on the 24-item Hamilton Rating Scale for Depression (HAMD) at the end of 24 weeks of treatment. Features subsets were constructed from baseline sociodemographic and clinical features, gray matter volumes (GMVs), or both. Three classification algorithms were compared: (1) Support Vector Machine-Radial Bias Function (SVMRBF), (2) Random Forest (RF), and (3) Logistic Regression (LR). A repeated 5-fold cross-validation approach with a wrapper-based feature selection method was used for model fitting. Model performance metrics included Area under the ROC Curve (AUC) and Matthews correlation coefficient (MCC). Cross-validated performance significance was tested by permutation analysis. Classifiers were compared by Cochran's Q and *post-hoc* pairwise comparisons using McNemar's Chi-Square test with Bonferroni correction.

**Results:** For the RF and SVMRBF algorithms, the combined feature set outperformed the clinical and GMV feature sets with a final cross-validated AUC of 0.83 ± 0.11 and 0.80 ± 0.11, respectively. Both classifiers passed permutation analysis. The LR algorithm performed best using GMV features alone (AUC 0.79 ± 0.14) but failed to pass permutation analysis using any feature set. Performance of the three classifiers differed significantly for all three features sets. Important predictive features of treatment response included anterior and posterior cingulate volumes, depression characteristics, and self-reported health-related quality scores.

**Conclusion:** This preliminary exploration into the use of ML and multi-modal data to identify predictors of general treatment response in LLD indicates that integration of clinical and structural MRI features significantly increases predictive capability. Identified features are among those previously implicated in geriatric depression, encouraging future work in this arena.

**Keywords: machine learning, pharmacology, prediction model, computational modeling, late-life depression (LLD)**

# INTRODUCTION

Late-life depression (LLD) is a common disorder among community elderly associated with poor quality of life, increased risk for cognitive decline, and increased mortality, including suicide (1–3). Medical comorbidities and polypharmacy increase the complexity of treatment selection due to drug-drug interactions and heightened risk of adverse events (4). Decreased efficacy of antidepressants is observed with increasing age, likely attributable to increased somatic illness burden, ischemic or neurodegenerative brain changes, and/or suboptimal dosing by prescribers (5).

LLD treatment selection is currently guided by patient preference and trial and error. The search for treatment-response biomarkers has generated a wealth of genomic and neuroimaging data, however no candidate markers have transcended into routine clinical practice. Structural magnetic resonance imaging (MRI) features are appealing due the non-invasiveness of acquisition and relatively low cost. In LLD compared to healthy controls, gray matter volume (GMV) reductions are frequently observed in the fronto–striatal–limbic regions (6–9). Differences in GMV often associate to differences in antidepressant treatment response (10–13).

Early and aggressive intervention in LLD is critical to mitigating its devastating consequences. Machine learning algorithms have significantly advanced diagnostic and prognostic modeling of structural MRI data in numerous psychiatric disorders (14). Predictions from unimodal data, however, have produced often mixed results when applied to new data with high accuracy sometimes limited to the most severe forms of illness (15). Models that integrate multiple data modalities (e.g., clinical, imaging, biological), have shown superiority in diagnostic classification tasks (16–20). Such models, however, require a higher degree of expertise than unimodal models, both in design and in interpretation of results, especially when using "small" data (<100 observations (19). In the current study, we hypothesized that a multi-modal feature set would better predict depressive remission in patients with LLD compared feature sets containing only clinical or GMV variables.

# METHODS

## Data Sources

Data were derived from two completed clinical trials of treatment of LLD (NCT01902004; NCT02460666, **Supplementary Table S1**) (21, 22). NCT01902004 spanned from January 2013 to January 2019, while NCT02460666 spanned January 2016 to November 2020. Informed consent was obtained from all participants prior to engaging in any research procedures and all procedures were approved by the Institutional Review Board at UCLA. Both studies employed a similar study protocol. Exclusion criteria were: (1) history of any psychiatric disorder (except for stable comorbid anxiety or stable comorbid insomnia); (2) acute suicidal ideation or suicide attempt within the past year; (3) severe or acute unstable medical illness or neurological disorder; or (4) dementia. Both studies required a diagnosis of major depressive disorder as defined by Diagnostic and Statistical Manual (DSM)-IV-TR or DSM-5. For the current analysis, inclusion criteria were set at: (1) age $\geq$ 60 years; (2) normal cognitive functioning as defined by a Mine Mental Status Exam (MMSE) score of 24 or greater; and (3) at least mild-moderate depression at treatment initiation.

## Treatments and Clinical Assessments

For NCT01902004, participants were required to be free of antidepressant medication prior to enrollment, then randomized to receive either escitalopram/placebo or escitalopram/memantine (12, 22). For NCT02460666, participants continued their current but ineffective antidepressant or psychotherapy treatment and were randomized to receive either Tai chi or health education (23). Treatment duration was 24 weeks for both trials. Participants completed a battery of self-reported and cognitive measures (see **Supplementary Table S2**) pre- and post-treatment. The primary measure of depression remission in both studies was a HAMD score of 6 or less by end of treatment. The distribution of sociodemographic and illness characteristics did not differ significantly between the two studies (**Supplementary Table S1**). Most patients in NCT02460666 were maintained on a selective serotonin reuptake inhibitor (SSRI, 20/35, 57.1%), while the remainder received a serotonin norepinephrine reuptake inhibitor (SNRI, 7/35, 20%), norepinephrine and dopamine reuptake inhibitor (NDRI, 2/35, 5.7%), or other treatment (8/35, 22.9%). A total of 28/67 (42%) participants in the combined sample achieved remission of depression by the end of treatment (NCT01902004: 56%; NCT02460666: 29%).

## Image Acquisition

A high-resolution T1-weighted structural brain scan was collected at baseline for each participant using the MPRAGE sequence (3D multi-echo magnetization-prepared rapid gradient-echo sequence). Scans were acquired using Siemens 3T Trio or Prisma systems (Siemens, Erlangen, Germany) with a 32-channel head coil (HEA, HEP) at the Ahmanson and

Lovelace Brain Mapping Center at UCLA. Prisma settings: 0.8 $mm^3$ isotropic voxel size, TR = 2,500 ms, TE = 1.81:1.79:7.18 ms; FoV = 256 mm; 256 × 256 matrix; TI 1,000 ms; flip angle = 8°. Trio settings: 1 $mm^3$ isotropic voxel size, TR = 2,150 ms, TE = 1.74 ms, 3.6, 5.46, and 7.32 ms; FoV = 256 mm; 256 × 256 matrix; TI 1,260 ms; flip angle = 7°. Acquisition time was 8.22 min for Prisma and 5.18 min for Trio scans.

## Image Preprocessing

Freesurfer (version 6.0) (http://surfer.nmr.mgh.harvard.edu) was used for reconstruction of gray matter volumetric measurements at both sites (24). The data cleaning pipeline included the correction of magnetic field in homogeneities, removal of non-brain tissues, segmentation of gray matter from white matter and cerebrospinal fluid, and parcellation of cortical regions using the Desikan–Killany atlas. The reconstructed scans were then carefully inspected for tissue misclassifications and manually corrected as needed. A simple least-square linear regression between raw volumes and the estimated total intracranial volume (eTIV) generated adjusted volumes, a method shown to greatly reduce sex-based volume differences (25).

## Feature Sets

In total, there were seven socio-demographic features, nine medical and mental health illness features, 18 baseline self-reported measures, six cognitive test, and 68 GMV features available in the training and external validation datasets (see **Supplementary Table S2**). Three feature sets were created: (1) socio-demographic, medical and mental health illness features, and baseline self-reported measures and cognitive tests (designated the "clinical" feature set), (2) GMV features, and (3) combination of all available features.

## Classification Analysis

All analyses were performed in Python (v. 3.8) using the *scikit-learn* (v. 0.23.2) and *mlxtend* packages at default settings (26, 27). Three popular classifiers were selected for comparison with the three feature sets: (1) Support Vector Machine Classifier—Radial Bias Function Kernel (SVMRBF), (2) Random Forest (RF), and (3) L2-regularized Logistic Regression (LR). These algorithms have demonstrated high performance on small datasets in the literature (17, 28). A repeated 5-fold (i.e., 5-folds, 5-repeats) cross-validation approach was used to train and evaluate the classifiers. During splitting, folds were stratified to preserve the proportion of subjects in each target class (e.g., remitter, non-remitter). Data pre-processing steps occurred on the training and test folds independently to avoid against data leakage. Features were filtered to remove those with an absolute intercorrelation of 0.9 (with the features with lesser correlation with predicted target retained) or low variance. Given the excess of features to observations, a wrapper feature selection method was employed. The Boruta algorithm determines relevant features by comparing their predictive performance in a random forest classifier to copies permutated with noise (shadow) (29). Features are ranked and those falling below the maximum importance score of the shadow features or a designated threshold are removed. For the current study, the top 20 features as ranked by the Boruta algorithm were retained for each feature set. Categorical variables were one-hot encoded with 24 missing values imputed by the median value of all other observations. Continuous features were scaled according to the individual feature's quantile range (enables robustness to outliers) and non-normally distributed features were transformed by quantile transformation.

Model performances were estimated by the Area under the ROC Curve (AUC) and Matthews correlation coefficient (MCC) (30). MCC is a more reliable metric than accuracy in binary classification problems as the MCC score is high only if the prediction yields good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportional to the size of positive elements and the size of negative elements in the dataset (30). Scores were averaged across all folds to determine training and testing performance. The classifiers were refit on the entire training data to calculate final AUC scores and visualized by receiver operator curve.

## Classifier Comparison, Significance Testing, and Feature Information

The significance of the cross-validated performance scores was assessed by permutation analysis. Briefly, predicted targets were permutated 1,000 times to generate a randomized dataset. The percentage of permutations for which the AUC obtained on the randomized data is greater than that obtained using the true data yields the *p*-value. A low *p*-value signifies low likelihood that the model predictions are obtained by chance. Cochran's *Q*-test was performed to determine if the three classifiers differed significantly from each other in performance, followed by *post-hoc* McNemar's Chi-Square test with Bonferroni correction. For all tests, $p < 0.01$ determined significance. The impact of features to model output was explored by calculating Shapley values *via* the *SHAP* package (v. 0.39.0) and visualized by beeswarm plot (31).

# RESULTS

The receiver operator curves and final cross-validated AUC scores for each classifier and feature set combination are shown in **Figure 1** and **Supplementary Table S3**. On the clinical feature set, the classifiers performed as follows: LR (Train: AUC 0.84 ± 0.04; Test: AUC 0.65 ± 0.16, MCC 0.19 ± 0.30; Overall: AUC 0.64 ± 0.16); RF (Train: AUC 0.99 ± 0.01; Test: AUC 0.79 ± 0.14, MCC 0.41 ± 0.22; Overall: 0.79 ± 0.14) and SVMRBF (Train: AUC 0.99 ± 0.01; Test: 0.64 ± 0.16, MCC 0.13 ± 0.22; Overall: 0.58 ± 0.18). On the GMV feature set, the classifiers performed as follows: LR (Train: AUC 0.81 ± 0.03; Test: AUC 0.68 ± 0.12, MCC 0.32 ± 0.22; Overall: 0.68 ± 0.12); RF (Train: AUC 0.99 ± 0.01; Test: AUC 0.79 ± 0.10, MCC 0.38 ± 0.24; Overall: 0.79 ± 0.10); and SVMRBF (Train: AUC 0.98 ± 0.01; Test: 0.81 ± 0.10, MCC 0.45 ± 0.20; Overall: 0.81 ± 0.10). On the combined feature set, the classifiers performed as follows: LR (Train: AUC 0.92 ± 0.03; Test: AUC 0.66 ± 0.15, MCC 0.27 ± 0.33; Overall: 0.66 ± 0.15); RF (Train: AUC 0.99 ± 0.00; Test: AUC 0.84 ± 0.11, MCC 0.47 ± 0.29; Overall: 0.83 ± 0.11); and SVMRBF (Train:

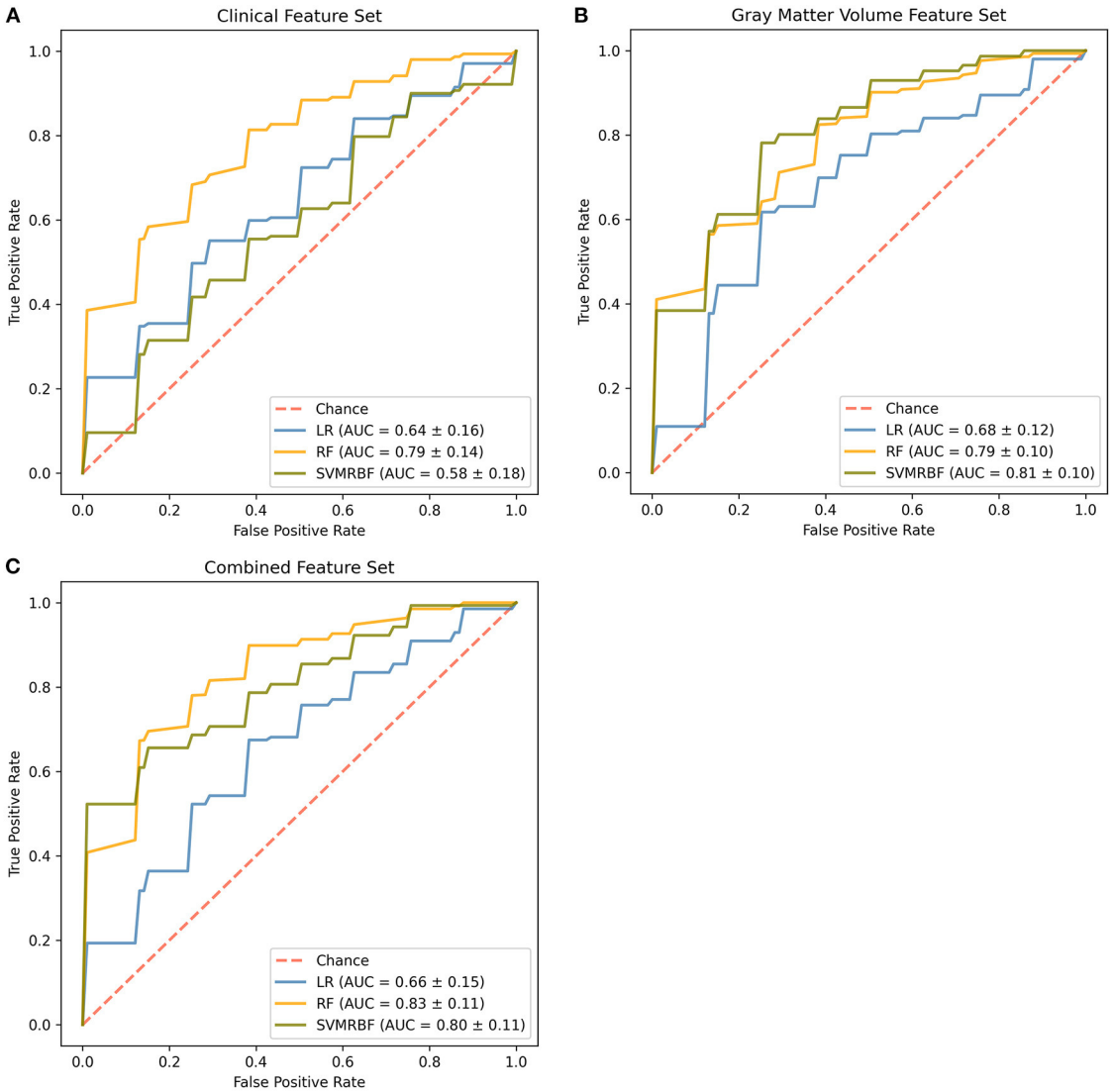


**FIGURE 1 |** Comparison of classifiers by algorithm and feature set. Evaluated feature sets included **(A)** sociodemographic and clinical features only, **(B)** gray matter volumes only, or **(C)** a combination of all available features. Features were ranked by feature importance and the top 20 from each feature set used for classifier training. The final mean cross-validated area under the cover (AUC) scores are shown for the Logistic Regression (LR, blue), Random Forest (RF, gold), and Support Vector Machine-Radial Bias Function (SVMRBF, olive) classifiers.

AUC 0.99 ± 0.00; Test: 0.81 ± 0.11, MCC 0.52 ± 0.22; Overall: 0.80 ± 0.11).

At a $p < 0.01$ for significance, permutation analysis (**Figure 2**) indicates that the LR classifier did not achieve performance above chance for any feature set (Clinical: $p = 0.042$; GMV: $p = 0.019$; Combined: $p = 0.028$), the RF classifier achieved significance for all feature subsets (Clinical: $p = 0.002$; GMV: $p = 0.001$; Combined: $p = 0.001$), and the SVMRBF classifier was significant for the GMV and combined feature sets (Clinical: $p = 0.050$; GMV: $p = 0.001$; Combined: $p = 0.001$). Comparison across classifiers using Cochran's test found significance differences for the clinical ($Q$: 18.9, $p < 0.01$), GMV ($Q$: 13.1, $p < 0.01$), and combined feature sets ($Q$: 16.1, $p < 0.01$). For the clinical feature set, *post-hoc* McNemar's Chi-Squared testing found that LR vs. SVMRBF and RF vs. SVMRBF did not differ significantly (Chi2: 3.4, $p = 0.07$; Chi2: 6.1, $p = 0.01$, respectively), but LR vs. RF differed (Chi2: 13.5, $p < 0.01$). For the GMV feature set, LR vs. RF differed significantly (Chi2: 7.6, $p < 0.01$), but not LR vs. SVM RBF (Chi2: 5.8, $p = 0.02$) or RF vs. SVMRBF (Chi2: 0.12, $p = 0.72$). Finally, for the combined feature set, LR vs. SVMRBF and LR vs. RF differed significantly (Chi2: 7.7, $p < 0.01$; Chi2: 7.7, $p < 0.01$), but not RF vs. SVM (Chi2: 0.25; $p = 0.62$).

SHAP (SHapley Additive exPlanation) values were calculated for the RF classifier with the combined feature set (**Figure 3**). SHAP values reflect the magnitude of a feature's influence on model predictions, not a decrease in model performance as with

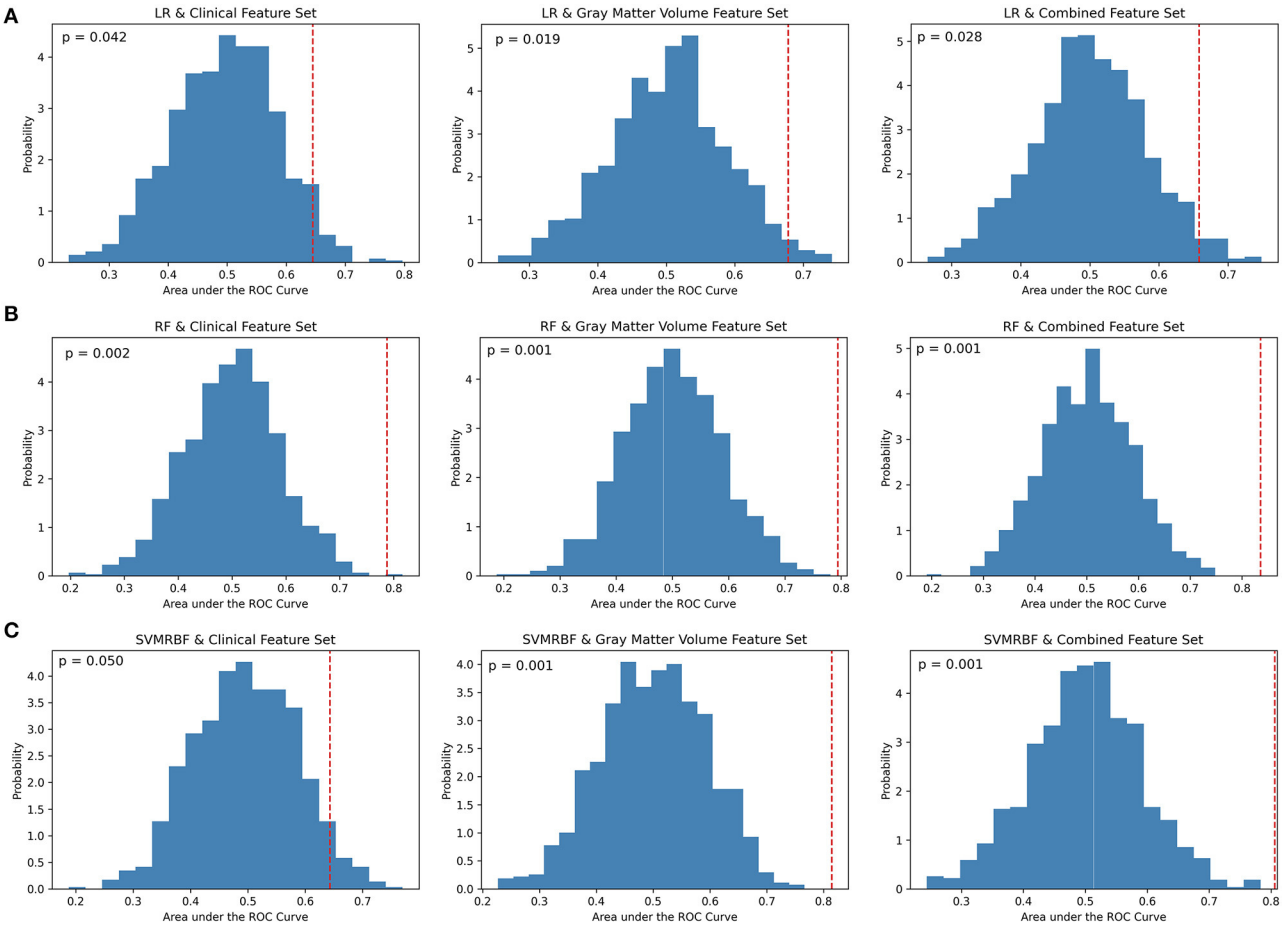


**FIGURE 2 |** Permutation analysis of cross-validation scores by classifier and feature set. **(A)** Logistic Regression (LR), **(B)** Random Forest (RF), and **(C)** Support Vector Machine-Radial Bias Function (SVMRBF) performance on 1,000 permuted datasets vs. true data was used to calculate the percentage of AUC scores occurring by chance (*p*-value). A $p < 0.01$ determined significance. The dashed red line denotes the mean of AUC scores on the true data compared to the probability distribution of AUC scores calculated on the permuted data.

permutation-based feature performance measures. The most influential feature on prediction of depressive remission was the left-hand caudal anterior cingulate volume, which changes the predicted absolute depression remission probability, on average, by 7%. Other high-ranking features included current age, age of depression onset, baseline HAMD score, current episode duration, and cardiovascular risk factor score, all of which altered remission probability by 2–4%. SHAP values do not permit inference of causality, only correlation with the predicted target.

## DISCUSSION

LLD, like other mood disorders, involves a complicated, multi-directional interplay between biology, psychological, environmental, and social mediators. Considerable heterogeneity exists in clinical phenotypes among patients with LLD, reflective of differing psychobiological pathways to illness. Here, we have demonstrated prediction of treatment response in LLD is improved using a combination of feature types. Our results mirror that of Patel et al. (17), where the authors integrated clinical, cognitive, and MRI data toward improved prediction of diagnosis and treatment response to a 12-week open trial of several different antidepressants in LLD.

The features identified as influencing classifier prediction in the current study corroborate several prior findings in the literature. Age of depression onset and cardiovascular health are among the most notable. LLD encompasses both individuals with early-onset depression (EOD), who develop depressive symptoms before the age of 25 and experience recurrent episodes across lifetime, and individuals with first presentation after age 50–65, or late-onset depression (LOD). The LOD phenotype displays less heritability and a stronger association with underlying cerebrovascular disease with a clinical profile of fronto-subcortical dysfunction, apathy, higher likelihood of progression to dementia, and increased antidepressant resistance (32–34).

Self-reported health-related quality of life (HRQOL) measures (SF36—energy, SF36—emotional well-being) as well as baseline

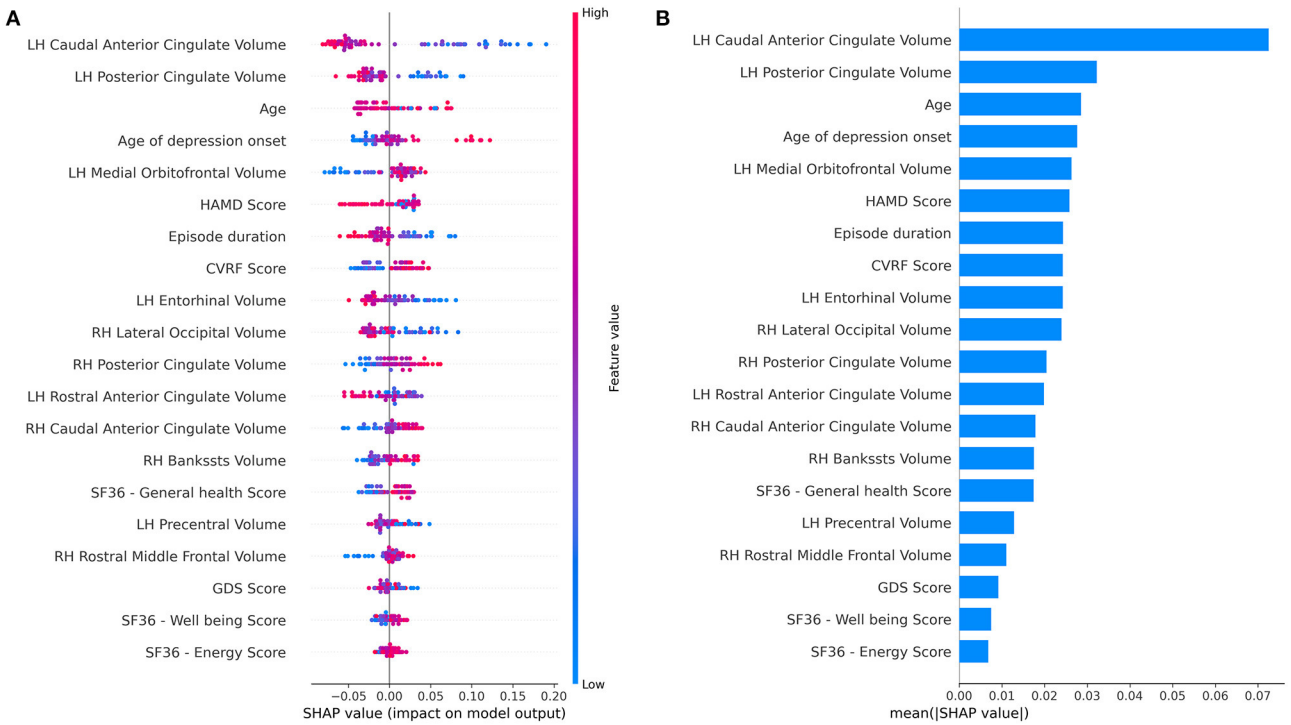


**FIGURE 3 |** Feature importance summary of random forest classifier using the combined feature set. **(A)** SHAP (SHapley Additive exPlanation) values are ordered by value of a feature to the predictions made by the classifier. The position on the x-axis on shows whether the effect of that value is associated with a higher or lower prediction for a given observation. Red color indicates the feature is high for that observation or low (blue). **(B)** Summary of mean SHAP values or overall magnitude of a feature's impact on prediction of depressive remission.

depression severity and chronicity also emerged as informative to prediction, consistent with prior investigations (35–38). Chronic physical disability associates to poor prognosis (39–41). Among the GMVs identified, dysfunction and differences in the anterior cingulate in LLD is well-established (42–44). Entorhinal volume also associates to multiple aspects of LLD, including somatic symptoms and cognitive impairment/conversion to dementia (45–47). Volume of the entorhinal cortex is inversely associated with the number of years since the first episode of depression and associates with treatment-resistant depression in females (6, 45).

The type of response predicted in the current study is general rather than treatment-specific. While the character of the two clinical trial cohorts did not differ substantially in demographics or illness features, the treatment modalities and conditions varied with one group initiating a new SSRI while the other continued their existing antidepressant or therapy and received a new add-on health intervention. Differential treatment response prediction is the goal of the precision medicine approach. However, just as there are converging and diverging pathways to depression, converging and diverging pathways in treatment response (and resistance) are anticipated. Certain data types may offer differing levels of discriminatory predictive power. For example, in a recent study in a sample of 81,630 adults, treatment-specific predictive models from electronic health record data did not perform better than general treatment response models (48). A classifier capable of predicting treatment response to a focused range of options (e.g., SSRIs) could arguably hold higher clinical utility in practice than one that predicts response to a single agent (49, 50).

The current work has several strengths, including the rigor of the analysis. Machine learning algorithms possess known variability in their tolerance for number of features, multi-collinearity, and noise. The RF classifier, for example, performed well-across all feature sets and demonstrated the least degradation in performance (generalization error) on the testing data. The primary limitations of the study are the small sample size and lack of a dataset with similar features for external validation. Cross-validation is only an estimate of performance on unseen data. The generalizability of a model cannot be fully determined without validation in an external dataset (51). Additionally, "small" data is prone to overfitting, even with robust feature selection and cross-validation. For the current work, a static number of features were employed in each feature set to permit comparison across classifiers. In moving from exploratory analysis to development of an optimized model, features could be even more aggressively reduced, hyperparameters tuned (e.g., limiting the maximum depth of the branching of the RF classifier, the number of support vectors for SVMRBF), and models combined (ensemble modeling) to further reduce overfitting.

## CONCLUSION

The current preliminary study into the use of ML to identify predictors of treatment response in late-life depression indicates that integration of clinical and structural MRI significantly increases predictive capability. Timely treatment selection in LLD is critical to preservation of quality of life and cognitive capacity. The current results suggest machine learning coupled with multimodal data are a promising avenue for the development of a non-invasive, precision approach to illness management.

## DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Protected health information, clinical trial data. Requests to access these datasets should be directed to Helen Lavretsky (hlavretsky@mednet.ucla.edu).

## AUTHOR CONTRIBUTIONS

HL and KN conceived of the study. AG completed the analyses. AP and BK-S assisted with data processing. All authors contributed to the writing of the manuscript and interpretation of results.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.738494/full#supplementary-material

## REFERENCES


1. Riddle M, Potter GG, Mcquoid DR, Steffens DC, Beyer JL, Taylor WD. Longitudinal cognitive outcomes of clinical phenotypes of late-life depression. *Am J Geriatr Psychiatry*. (2017) 25:1123–34. doi: 10.1016/j.jagp.2017.03.016
2. Steffens DC. Late-life depression and the prodromes of dementia. *JAMA Psychiatry*. (2017) 74:673–4. doi: 10.1001/jamapsychiatry.2017.0658
3. Szanto K, Galfalvy H, Vanyukov PM, Keilp JG, Dombrovski AY. Pathways to late-life suicidal behavior: cluster analysis and predictive validation of suicidal behavior in a sample of older adults with major depression. *J Clin Psychiatry*. (2018) 79:17m11611. doi: 10.4088/JCP.17m11611
4. Holvast F, Van Hattem BA, Sinnige J, Schellevis F, Taxis K, Burger H, et al. Late-life depression and the association with multimorbidity and polypharmacy: a cross-sectional study. *Fam Pract*. (2017) 34:539–45. doi: 10.1093/fampra/cmx018
5. Kok RM, Reynolds CF. Management of depression in older adults: a review. *JAMA*. (2017) 317:2114–22. doi: 10.1001/jama.2017.5706
6. Andreescu C, Butters MA, Begley A, Rajji T, Wu M, Meltzer CC, et al. Gray matter changes in late life depression-a structural MRI analysis. *Neuropsychopharmacology*. (2008) 33:2566–72. doi: 10.1038/sj.npp.1301655
7. Chang C-C, Yu S-C, Mcquoid DR, Messer DF, Taylor WD, Singh K, et al. Reduction of dorsolateral prefrontal cortex gray matter in late-life depression. *Psychiatry Res*. (2011) 193:1–6. doi: 10.1016/j.pscychresns.2011.01.003
8. Sexton CE, Mackay CE, Ebmeier KP. A systematic review and meta-analysis of magnetic resonance imaging studies in late-life depression. *Am J Geriatr Psychiatry*. (2013) 21:184–95. doi: 10.1016/j.jagp.2012.10.019
9. Du M, Liu J, Chen Z, Huang X, Li J, Kuang W, et al. Brain grey matter volume alterations in late-life depression. *J Psychiatry Neurosci*. (2014) 39:397–406. doi: 10.1503/jpn.130275
10. Droppa K, Karim HT, Tudorascu DL, Karp JF, Reynolds CF, Aizenstein HJ, et al. Association between change in brain gray matter volume, cognition, and depression severity: pre- and post- antidepressant pharmacotherapy for late-life depression. *J Psychiatr Res*. (2017) 95:129–34. doi: 10.1016/j.jpsychires.2017.08.002
11. Laird KT, Siddarth P, Krause-Sorio B, Kilpatrick L, Milillo M, Aguilar Y, et al. Anxiety symptoms are associated with smaller insular and orbitofrontal cortex volumes in late-life depression. *J Affect Disord*. (2019) 256:282–7. doi: 10.1016/j.jad.2019.05.066
12. Krause-Sorio B, Siddarth P, Kilpatrick L, Laird KT, Milillo MM, Ercoli L, et al. Combined treatment with escitalopram and memantine increases gray matter volume and cortical thickness compared to escitalopram and placebo in a pilot study of geriatric depression. *J Affect Disord*. (2020) 274:464–70. doi: 10.1016/j.jad.2020.05.092
13. Pimontel MA, Solomonov N, Oberlin L, Kanellopoulos T, Bress JN, Hoptman MJ, et al. Cortical thickness of the salience network and change in apathy following antidepressant treatment for late-life depression. *Am J Geriatr Psychiatry*. (2020) 29:241–8. doi: 10.1016/j.jagp.2020.06.007
14. Mateos-Pérez JM, Dadar M, Lacalle-Aurioles M, Iturria-Medina Y, Zeighami Y, Evans AC. Structural neuroimaging as clinical predictor: a review of machine learning applications. *Neuroimage Clin*. (2018) 20:506–22. doi: 10.1016/j.nicl.2018.08.019
15. Ramasubbu R, Brown MR, Cortese F, Gaxiola I, Goodyear B, Greenshaw AJ, et al. Accuracy of automated classification of major depressive disorder as a function of symptom severity. *Neuroimage Clin*. (2016) 12:320–31. doi: 10.1016/j.nicl.2016.07.012
16. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Trans Biomed Eng*. (2015) 62:1132–40. doi: 10.1109/TBME.2014.2372011
17. Patel MJ, Andreescu C, Price JC, Edelman KL, Reynolds CF 3rd, Aizenstein HJ. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *Int J Geriatr Psychiatry*. (2015) 30:1056–67. doi: 10.1002/gps.4262
18. Pearlson GD, Liu J, Calhoun VD. An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. *Front Genet*. (2015) 6:276. doi: 10.3389/fgene.2015.00276
19. Calhoun VD, Sui J. Multimodal fusion of brain imaging data: a key to finding the missing link(s) in complex mental illness. *Biol Psychiatry Cogn Neurosci Neuroimaging*. (2016) 1:230–44. doi: 10.1016/j.bpsc.2015.12.005
20. Hilbert K, Lueken U, Muehlhan M, Beesdo-Baum K. Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: a multimodal machine learning study. *Brain Behav*. (2017) 7:e00633. doi: 10.1002/brb3.633
21. Siddarth D, Siddarth P, Lavretsky H. An observational study of the health benefits of yoga or tai chi compared with aerobic exercise in community-dwelling middle-aged and older adults. *Am J Geriatr Psychiatry*. (2014) 22:272–3. doi: 10.1016/j.jagp.2013.01.065
22. Lavretsky H, Laird KT, Krause-Sorio B, Heimberg BF, Yeargin J, Grzenda A, et al. A randomized double-blind placebo-controlled trial of combined escitalopram and memantine for older adults with major depression and subjective memory complaints. *Am J Geriatr Psychiatry*. (2020) 28:178–90. doi: 10.1016/j.jagp.2019.08.011
23. Lavretsky H, Milillo MM, Kilpatrick L, Grzenda A, Wu P, Nguyen SA, et al. A randomized controlled trial of Tai Chi Chih or health education for geriatric depression. *Am J Geriatr Psychiatry*. (2021). doi: 10.1016/j.jagp.2021.07.008. [Epub ahead of print].
24. Fischl B. FreeSurfer. *Neuroimage*. (2012) 62:774–81. doi: 10.1016/j.neuroimage.2012.01.021
25. Voevodskaya O, Simmons A, Nordenskjold R, Kullberg J, Ahlstrom H, Lind L, et al. The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease. *Front Aging Neurosci*. (2014) 6:264. doi: 10.3389/fnagi.2014.00264

26. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. *arXiv:1309.0238 [arXiv preprint]* (2013).
27. Raschka S. MLxtend: providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J Open Source Softw.* (2018) 3:638. doi: 10.21105/joss.00638
28. Leaver AM, Wade B, Vasavada M, Hellemann G, Joshi SH, Espinoza R, et al. Fronto-temporal connectivity predicts ECT outcome in major depression. *Front Psychiatry.* (2018) 9:92. doi: 10.3389/fpsyt.2018.00092
29. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw.* (2010) 36:1–13. doi: 10.18637/jss.v036.i11
30. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* (2020) 21:6. doi: 10.1186/s12864-019-6413-7
31. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, CA) (2017). p. 4768–77.
32. Tedeschini E, Levkovitz Y, Iovieno N, Ameral VE, Nelson JC, Papakostas GI. Efficacy of antidepressants for late-life depression: a meta-analysis and meta-regression of placebo-controlled randomized trials. *J Clin Psychiatry.* (2011) 72:1660–8. doi: 10.4088/JCP.10r06531
33. Richard E, Reitz C, Honig LH, Schupf N, Tang MX, Manly JJ, et al. Late-life depression, mild cognitive impairment, and dementia. *JAMA Neurol.* (2013) 70:374–82. doi: 10.1001/jamaneurol.2013.603
34. Mussele SVD, Fransen E, Struyfs H, Luyckx J, Marien P, Saerens J, et al. Depression in mild cognitive impairment is associated with progression to Alzheimer's disease: a longitudinal study. *J Alzheimers Dis.* (2014) 42:1239–50. doi: 10.3233/JAD-140405
35. Joel I, Begley AE, Mulsant BH, Lenze EJ, Mazumdar S, Dew MA 3rd, et al. Dynamic prediction of treatment response in late-life depression. *Am J Geriatr Psychiatry.* (2014) 22:167–76. doi: 10.1016/j.jagp.2012.07.002
36. Smagula SF, Butters MA, Anderson SJ, Lenze EJ, Dew MA, Mulsant BH, et al. Antidepressant response trajectories and associated clinical prognostic factors among older adults. *JAMA Psychiatry.* (2015) 72:1021–8. doi: 10.1001/jamapsychiatry.2015.1324
37. Yang W-C, Lin C-H, Wang F-C, Lu M-J. Factors related to the improvement in quality of life for depressed inpatients treated with fluoxetine. *BMC Psychiatry.* (2017) 17:309. doi: 10.1186/s12888-017-1471-3
38. Karp JF, Weiner D, Seligman K, Butters M, Miller M, Frank E, et al. Body pain and treatment response in late-life depression. *Am J Geriatr Psychiatry.* (2005) 13:188–94. doi: 10.1097/00019442-200503000-00003
39. Shmuely Y, Baumgarten M, Rovner B, Berlin J. Predictors of improvement in health-related quality of life among elderly patients with depression. *Int Psychogeriatr.* (2001) 13:63–73. doi: 10.1017/S1041610201007463
40. Comijs HC, Nieuwesteeg J, Kok R, Van Marwijk HW, Van Der Mast RC, Naarding P, et al. The two-year course of late-life depression; results from the Netherlands study of depression in older persons. *BMC Psychiatry.* (2015) 15:20. doi: 10.1186/s12888-015-0401-5
41. Collard RM, Arts MHL, Schene AH, Naarding P, Oude Voshaar RC, Comijs HC. The impact of frailty on depressive disorder in later life: findings from the Netherlands Study of depression in older persons. *Eur Psychiatry.* (2017) 43:66–72. doi: 10.1016/j.eurpsy.2017.01.003
42. De Asis JM, Stern E, Alexopoulos GS, Pan H, Van Gorp W, Blumberg H, et al. Hippocampal and anterior cingulate activation deficits in patients with geriatric depression. *Am J Psychiatry.* (2001) 158:1321–3. doi: 10.1176/appi.ajp.158.8.1321
43. Alexopoulos GS, Gunning-Dixon FM, Latoussakis V, Kanellopoulos D, Murphy CF. Anterior cingulate dysfunction in geriatric depression. *Int J Geriatr Psychiatry.* (2008) 23:347–55. doi: 10.1002/gps.1939
44. Katz R, De Sanctis P, Mahoney JR, Sehatpour P, Murphy CF, Gomez-Ramirez M, et al. Cognitive control in late-life depression: response inhibition deficits and dysfunction of the anterior cingulate cortex. *Am J Geriatr Psychiatry.* (2010) 18:1017–25. doi: 10.1097/JGP.0b013e3181d695f2
45. Furtado CP, Maller JJ, Fitzgerald PB. A magnetic resonance imaging study of the entorhinal cortex in treatment-resistant depression. *Psychiatry Res.* (2008) 163:133–42. doi: 10.1016/j.pscychresns.2007.11.005
46. Leal SL, Noche JA, Murray EA, Yassa MA. Disruption of amygdala-entorhinal-hippocampal network in late-life depression. *Hippocampus.* (2017) 27:464–76. doi: 10.1002/hipo.22705
47. O'Shea DM, Dotson VM, Woods AJ, Porges EC, Williamson JB, O'Shea A, et al. Depressive symptom dimensions and their association with hippocampal and entorhinal cortex volumes in community dwelling older adults. *Front Aging Neurosci.* (2018) 10:40. doi: 10.3389/fnagi.2018.00040
48. Hughes MC, Pradier MF, Ross AS, Mccoy TH Jr, Perlis RH, Doshi-Velez F. Assessment of a prediction model for antidepressant treatment stability using supervised topic models. *JAMA Netw Open.* (2020) 3:e205308. doi: 10.1001/jamanetworkopen.2020.5308
49. Rajpurkar P, Yang J, Dass N, Vale V, Keller AS, Irvin J, et al. Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: a prespecified secondary analysis of a randomized clinical trial. *JAMA Netw Open.* (2020) 3:e206653. doi: 10.1001/jamanetworkopen.2020.6653
50. Athreya AP, Brückl T, Binder EB, John Rush A, Biernacka J, Frye MA, et al. Prediction of short-term antidepressant response using probabilistic graphical models with replication across multiple drugs and treatment settings. *Neuropsychopharmacology.* (2021) 46:1272–82. doi: 10.1038/s41386-020-00943-x
51. Grzenda A, Kraguljac NV, Mcdonald WM, Nemeroff C, Torous J, Alpert JE, et al. Evaluating the machine learning literature: a primer and user's guide for psychiatrists. *Am J Psychiatry.* (2021) 78:715–29. doi: 10.1176/appi.ajp.2020.20030250


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

CASE REPORT
published: 27 October 2021
doi: 10.3389/fpsyt.2021.754169

# Contactless In-Home Monitoring of the Long-Term Respiratory and Behavioral Phenotypes in Older Adults With COVID-19: A Case Series

*Guo Zhang[1†], Ipsit V. Vahia[2,3*†], Yingcheng Liu[1†], Yuzhe Yang[1], Rose May[2], Hailey V. Cray[2], William McGrory[4] and Dina Katabi[1]*

[1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, United States, [2] Division of Geriatric Psychiatry, McLean Hospital, Belmont, MA, United States, [3] Department of Psychiatry, Harvard Medical School, Boston, MA, United States, [4] Mary Ann Morse at Heritage Affordable Senior Living, Framingham, MA, United States

OPEN ACCESS

**Edited by:**
*Tao Wang, Northwestern Polytechnical University, China*

**Reviewed by:**
*Richard B. Reilly, Trinity College Dublin, Ireland*
*Bing Zhang, Nanjing Drum Tower Hospital, China*

***Correspondence:**
*Ipsit V. Vahia*
*ivahia@mclean.harvard.edu*

*[†] These authors have contributed equally to this work and share first authorship*

**Specialty section:**
*This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry*

**Received:** *06 August 2021*
**Accepted:** *04 October 2021*
**Published:** *27 October 2021*

**Citation:**
*Zhang G, Vahia IV, Liu Y, Yang Y, May R, Cray HV, McGrory W and Katabi D (2021) Contactless In-Home Monitoring of the Long-Term Respiratory and Behavioral Phenotypes in Older Adults With COVID-19: A Case Series. Front. Psychiatry 12:754169. doi: 10.3389/fpsyt.2021.754169*

Currently, there is a limited understanding of long-term outcomes of COVID-19, and a need for in-home measurements of patients through the whole course of their disease. We study a novel approach for monitoring the long-term trajectories of respiratory and behavioral symptoms of COVID-19 patients at home. We use a sensor that analyzes the radio signals in the room to infer patients' respiration, sleep and activities in a passive and contactless manner. We report the results of continuous monitoring of three residents of an assisted living facility for 3 months, through the course of their disease and subsequent recovery. In total, we collected 4,358 measurements of gait speed, 294 nights of sleep, and 3,056 h of respiration. The data shows differences in the respiration signals between asymptomatic and symptomatic patients. Longitudinally, we note sleep and motor abnormalities that persisted for months after becoming COVID negative. Our study represents a novel phenotyping of the respiratory and behavioral trajectories of COVID recovery, and suggests that the two may be integral components of the COVID-19 syndrome. It further provides a proof-of-concept that contactless passive sensors may uniquely facilitate studying detailed longitudinal outcomes of COVID-19, particularly among older adults.

**Keywords: COVID-19, long-term outcomes, phenotypes, contactless monitoring, respiration, behavior, older adults, case report**

## INTRODUCTION

As the COVID-19 pandemic approaches the 2-year mark, a growing body of literature suggests that even after recovery from the acute viral illness, there may be a range of long-term neuropsychiatric sequelae (1–3). Further, the long-term outcomes of COVID-19 are affected by a broad range of factors including the presence of pre-existing medical or psychiatric conditions (4, 5), the nature and severity of the acute respiratory illness (6), the quality of care received in the short and long term (7), the patient's socioeconomic status (4, 8), and age (9). As a result, we see a very heterogeneous range of outcomes with COVID-19. The impact of this heterogeneity is especially pronounced in the context of behavior symptoms, which can manifest themselves as predisposing

factors (10, 11), a core symptom of COVID-19 (12, 13), a virus-induced long-term symptom (2, 3), or an independent secondary effect (2, 14, 15).

These findings highlight the need for tracking the trajectory of COVID-19 disease through its acute and post-acute phase. Today, however, we lack solutions for collecting objective and longitudinal measurements from COVID-19 patients at home. Existing solutions for collecting data from patients at home rely mainly on self-reporting (9). Yet, self-reporting can be highly subjective, particularly when considering behavioral symptoms. Further, monitoring COVID-19 patients through the course of their disease is complicated by distancing and isolation protocols during the acute phase (16). It is also complicated by the difficulty of sustaining patient engagement in longitudinal post-acute studies (17, 18). While wearable devices and mobile phones may help collecting longitudinal data, such devices are ill-suited to older adults, and individuals who suffer from impaired memory and/or cognition (19).

We study a new solution for passively collecting objective and continuous measurements of COVID-19 patients recovering at home, during their active disease phase and post-acute recovery. We specifically focus on monitoring the trajectory of respiratory and behavioral phenotypes in older patients. We use a novel wireless sensor that sits in the background of the home (akin to a Wi-Fi router). The sensor, called Emerald, transmits very low power wireless signals, and analyzes their reflections from nearby humans and inanimate objects using machine learning (20, 21). It infers physiological and behavioral markers, including respiratory signals, gait speed, sleep patterns, and the time spent in different locations at home (activity graph). It can collect data continuously for prolonged periods, without physical contact with patients, and operates passively without burdening patients or caregivers. This sensor has been validated for monitoring sleep, gait speed, location, and respiration (20, 22, 23), and has been piloted in clinical studies of agitation (24), dementia (19), and Parkinson's disease (25). **Figure 1** illustrates the operation of the Emerald sensor.

We present the results of 24/7 monitoring of three COVID-19 patients in an assisted living facility for 3 months. Our sensor technology offers a pragmatic solution to several of the challenges around long-term tracking in COVID-19, especially in older adults. It does not require any active engagement from the monitored person, and collects data passively and continuously, without exposing caregivers or others to the patient. We show how this approach facilitates continuous phenotyping of changes in physiological and behavioral parameters through the acute and post-acute phases of COVID-19.

Today's clinical studies are limited to episodic measurements, typically conducted in the clinic. In contrast, the approach described herein enables zooming in on individual patients in their natural living environment to obtain detailed and clinically meaningful measurements of their condition over long periods of time, and without interfering with patients' lives.

# RESULTS

In total we monitored the subjects for 327 days. We processed the data to identify missing measurements due to accidental device unplugging, device malfunctioning, or patient being away in the hospital. After accounting for missing measurements, we collected 294 nights of sleep, 3,056 h of respiration signals, and 4,358 measurements of gait speed. A detailed description of subject recruitment and methods can be found in **Supplementary Material**.

## Clinical Course

In **Table 1**, we report the demographic and clinical characteristics of the study samples. Subject 1 is an 88yo male. His initial COVID symptoms were sore throat, fever and muscle ache. He was hospitalized for these symptoms and his monitoring began on day 0, upon return to the residential facility. He did not require any additional hospitalization. Subject 2 was an 81yo female with a pre-COVID history of COPD, generalized anxiety, and mild cognitive impairment. Her COVID symptoms started with fever, fatigue, and a sore throat. During the monitoring period, she was hospitalized due to breathing distress on day 7's early morning. She recovered gradually after discharge from hospital on day 14. Subject 3 is a 73yo female who tested positive for SARS Cov-2 infection but demonstrated no clinical symptoms of COVID throughout the study period. She had a prior history of bipolar disorder.

## Respiratory Changes

**Figures 2A–C** plots the daily respiratory rate (RR) of the three subjects, while being COVID positive, and as they transition to negative testing. The three subjects have different recovery experiences. The first two subjects are symptomatic, whereas subject 3 is asymptomatic. This is reflected in their RR in the figure, which shows that subjects 1 and 2 have significantly less stable RR than subject 3. Focusing on the symptomatic patients, subject 1 had a smooth gradual recovery, whereas subject 2 had respiration distress and was hospitalized for about a week. As shown in **Figure 2A**, subject 1 started with an elevated RR with a median of 25 breaths per minute (BPM) on day 1, and 26.3 BPM on day 2. Over the next 6 days, his median RR gradually decreased by about 3 BPM, and stabilized at a baseline around 21 BPM. In contrast, and as in **Figure 2B**, subject 2's RR initially decreased slowly to a median of 19 BPM; but on day 7, the RR suddenly jumped to 22 BPM. On that day the subject reported a medical emergency, was subsequently admitted to the hospital, and received medical treatment for breathing difficulty. After subject 2 came back from the hospital, her RR began to drop from a median of 19.5 BPM on day 14 to a baseline of 18.0 BPM on day 23. As for the asymptomatic patient, i.e., subject 3, her RR was stable for the whole recovery period with a daily median of 11.5 BPM, as shown in **Figure 2C**.

Next, we check whether the differences in RR for the symptomatic subjects are statistically meaningful. We divide the RR samples into two groups for each subject, where the first group refers to RR samples captured before becoming COVID negative and the second group refers to RR samples after the subject became COVID negative. We run a single-sided Mann–Whitney U test for each subject and calculate the effect size by Cohen's *d*. As expected, the RR elevation for the symptomatic subjects, i.e., subject 1 ($U = 3.0E7$, $p = 1.8E\text{-}268$) and subject 2 ($U = 5.8E7$, $p = 2.6E\text{-}241$), is statistically significant, but is

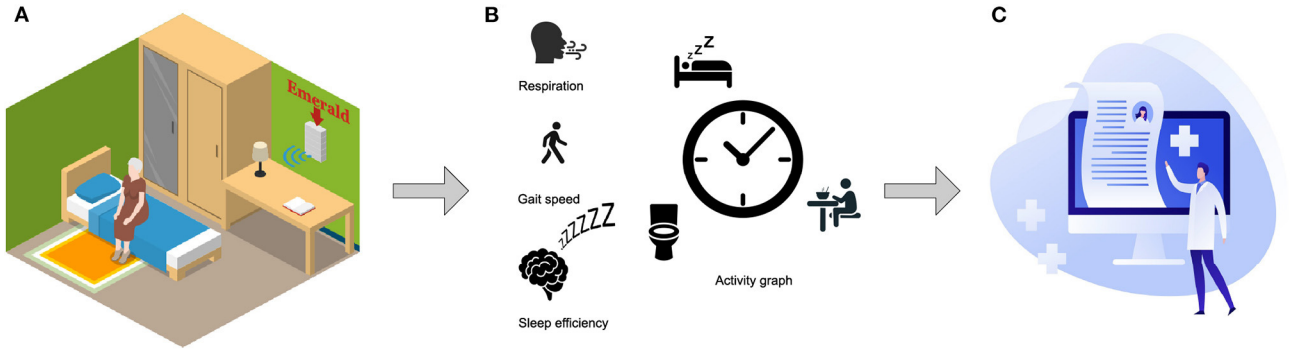


**FIGURE 1 |** System overview. **(A)** The Emerald device sits in the background of the home like a Wi-Fi router. It sends out a wireless signal that is 1,000 times weaker than the home Wi-Fi, and analyses the reflected signal 24/7 without burdening the monitored subject. **(B)** Measurements of the trajectories of respiration, gait speed, sleep efficiency, and daily activities of the subject are generated automatically. **(C)** Clinicians and researchers (with the proper credentials) can access the data remotely at home or in the office.

**TABLE 1 |** Demographic and clinical characteristics of the study samples ($N = 3$).

| Variables | Subject 1 | Subject 2 | Subject 3 |
|---|---|---|---|
| Age | 88 | 81 | 73 |
| Sex | M | F | F |
| Monitoring begins | 04/27/2020 | 04/07/2020 | 04/16/2020 |
| Day confirmed COVID positive with respect to the first day of monitoring* | Day -11 | Day -6 | Day 0 |
| Day confirmed COVID negative with respect to the first day of monitoring* | Day 8 | Day 17 | Day 8 |
| Monitoring ends* | Day 106 | Day 115 | Day 106 |
| Acute COVID-19 symptoms | • Fever<br>• Fatigue<br>• Sore throat<br>• Muscle pain | • Fever<br>• Fatigue<br>• Sore throat | None |
| Other comorbid conditions | Obsessive compulsive disorder, Major depressive disorder | Mild cognitive impairment, Generalized anxiety disorder (GAD), Benzodiazepine abuse in remission, CHF, COPD | Major neurocognitive impairment (mixed type), Bipolar disorder |

**Days are relative to the first day of monitoring.*

insignificant for the asymptomatic subject 3 (U = 2.0E7, $p = 0.9986$). The Cohen's $d$ for subject 1 ($d = 0.52$) and subject 2 ($d = 0.53$) shows medium effect size, while for subject 3 ($d = -0.04$) the effect size is minimum. Further, when comparing the hospitalization day of subject 2 (i.e., day 7) to days 1 to 6, we find the elevation is significant (single-sided Mann–Whitney $U$ test, $U = 1.2E6$, $p = 2.2E{-}90$), and the effect size is large (Cohen's $d$, $d = 2.0$).

## Sleep Efficiency

Sleep efficiency is the ratio of the total sleep time to the total time in bed (26). Sleep efficiency is highly correlated with the mental status of an individual. Usually, the more anxious an individual is, the lower the sleep efficiency will be (27). **Figures 2D–F** reports the sleep efficiency for all 3 subjects. As shown in **Figure 2D**, subject 1 has low sleep efficiency that worsens around day 37 to 64 in the middle of the study but subsequently stabilizes. One-way ANOVA on these 7 groups in the figure verifies there is significant difference ($df = 94$; $F = 15.57$; $p = 0.016$) between the groups. More specifically, sleep efficiency from day 37∼64 is significantly worse (single-sided $t$-test; $df = 94$; $t = 3.85$; $p = 2.4E\text{-}4$) than the other 10 weeks. This is consistent with the assisted living facility (ALF) staff's observations that subject 1 demonstrated some initial anxiety, but over time, his anxiety subsided, and he was able to recover. As shown in **Figure 2E**, subject 2's sleep efficiency decreased slightly over the monitoring period. Her sleep efficiency for the first 6 weeks is higher than the later 8 weeks (single-sided $t$-test; $df = 78$; $t = 1.60$; $p = 0.057$). Subject 3 had the best sleep efficiency overall (single-sided $t$-test; $df = 264$; $t = 6.01$; $p = 3.7E\text{-}9$); however, her sleep efficiency worsened from day 51 to day 106 (the end of the study) compared with day 9 to day 50 (single-sided $t$-test; $df = 90$; $t = 2.69$; $p = 0.0044$). This aligns with the ALF staff's observation that subject 3 has shown fewer health issues than the other two patients, though in the last 2 months of the study, she has experienced anxiety.

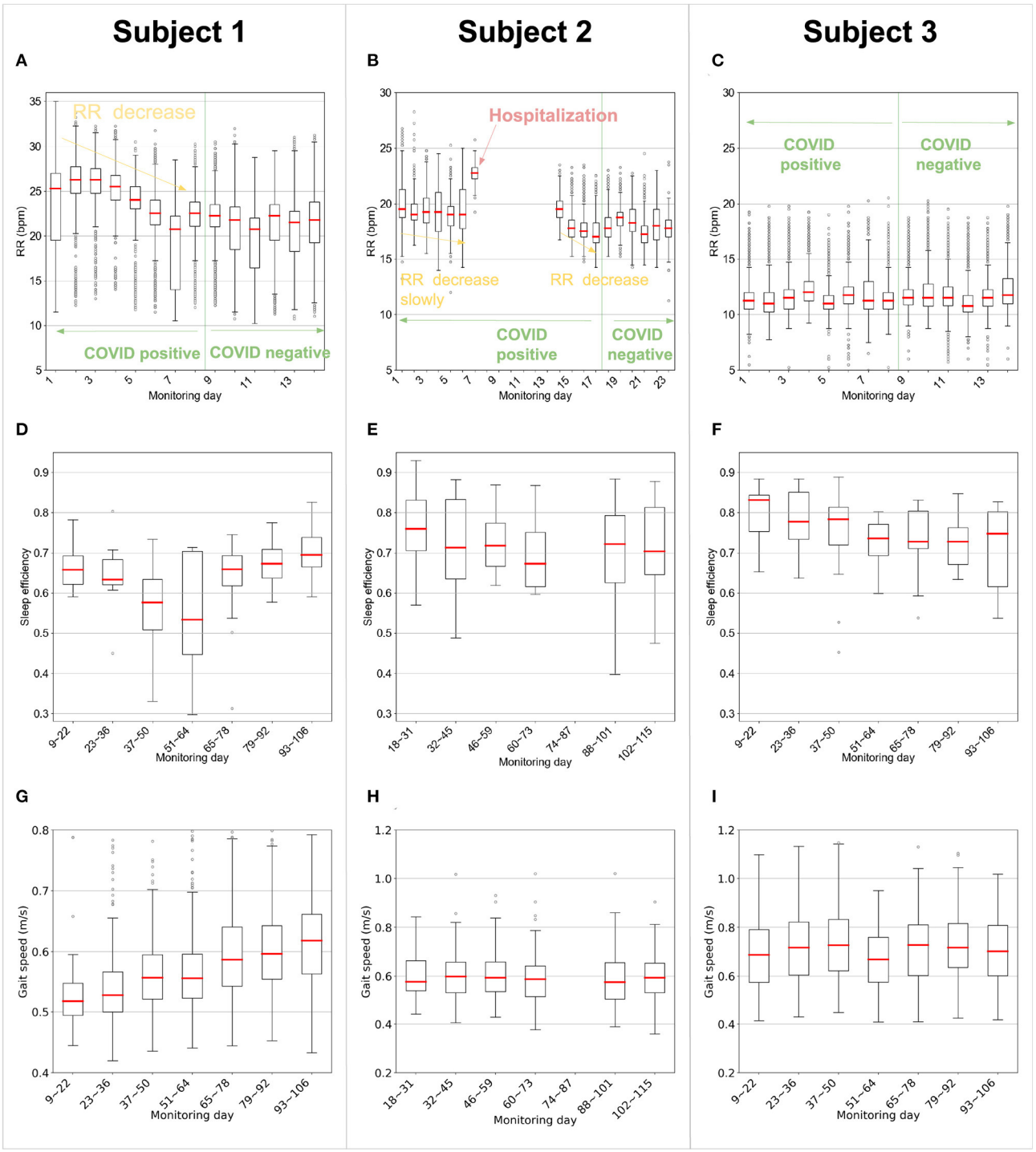


**FIGURE 2 |** Contactless measurements of respiration rate, sleep efficiency, and gait speed. **(A–C)** The daily respiration rate (RR) of the patients during sleep as they recovered (for each boxplot sequentially, $n = 962$, 703, 964, 831, 766, 762, 863, 935, 1,193, 1,032, 998, 973, 1,218, and 1,137 for subject 1; $n = 1{,}433$, 1,003, 1,332, 1,393, 1,779, 1,694, 177, 1,059, 950, 1,077, 1,234, 1,001, 932, 1,715, 1,019, 1,403, and 783 for subject 2; $n = 994$, 928, 1,045, 1,012, 864, 815, 832, 820, 1,139, 920, 929, 851, 919, and 977 for subject 3). The figure shows that while being COVID positive, the symptomatic subjects experienced an elevated RR in comparison to their baselines. In contrast, the asymptomatic subject had an RR similar to her baseline. Further, the figure shows that prior to hospitalization, Subject 2 experienced an unusually elevated RR. **(D–F)** Sleep efficiency of the subjects computed as the ratio of the sleeping time to the time in bed. Each box refers to data for two consecutive weeks (for each boxplot sequentially, $n = 14$, 14, 14, 14, 14, 14 and 11 for subject 1; $n = 14$, 14, 14, 13, 12 and 12 for subject 2; $n = 14$, 14, 14, 14, 14, 12 and 9 for subject 3.). **(G–I)** Gait speed of the subjects, where each boxplot represents two consecutive weeks (for each boxplot sequentially, $n = 40$, 226, 309,

*(Continued)*

**FIGURE 2 |** 526, 356, 403, and 335 for subject 1; $n = 58$, 107, 133, 122, 136, and 126 for subject 2; $n = 237$, 123, 133, 94, 121, 174, and 183 for subject 3.). **(D–I)** From day 74–day 87, subject 2 has 12 days of data missing due to accidental unplugging of the device by the housing staff. There are only 2 days left for these 2 weeks, so the box is not plotted for this period. **(A–I)** In each box plot, the central line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to 1.5 times the interquartile range. Points beyond the whiskers are plotted individually using the circle symbol.

## Gait Speed

**Figures 2G–I** compares the trajectories of gait speed of the 3 participants. As shown in **Figure 2G**, subject 1 has a relatively low initial median walking speed of 0.52 m/s. However, his walking improves steadily in the following months to reach 0.62 m/s by the end of the monitoring period. Regression analysis shows that the rate of improvement in his gait speed was 0.0076 m/s per 7 days ([0.0068, 0.0085] 95% CI, $p = 2.9E\text{-}66$). The ALF staff reported that subject 1's walking was impaired initially, most likely due to his very limited mobility while in quarantine. Subsequently with physical therapy, his walking speed improved. **Figure 2H** shows that subject 2 did not exhibit a significant change in her gait speed over the course of the monitoring period (−0.0015 m/s per 7 days [−0.0035, 0.00037] 95% CI, $p = 0.11$). Similar to subject 2, subject 3's gait speed, in **Figure 2I**, is relatively steady during the observation period (−0.00076 m/s per 7 days [−0.0024, 0.00089] 95% CI, $p = 0.37$).

## Activity

**Figure 3** shows general behavior patterns for all three subjects. Each circle is one day (12 a.m. at the top refers to midnight, and 12 p.m. at the bottom refers to noon) the inner most circle is the first day after becoming COVID negative, and the outermost circle is the last day of monitoring. The graphs provide a longitudinal view of subjects' basic activities after they became COVID negative.

Based on **Figure 3A**, we note that subject 1 started to leave the room on the day he was declared COVID negative. This can be inferred from the white patches, which refer to time intervals during which the subject is outside the coverage of the Emerald wireless signals, i.e., outside his room at the ALF. The subject, however, continued to take his meals in his room rather than in the dining hall as indicated by the yellow cones (i.e., being on the chair) around 8 a.m., 12 p.m., and 5 p.m., in his activity graph. This continued until 29 days before the end of the study, when we see the yellow cones are replaced by white-colored region, which indicates that he started to take all of his meals in the dining room. He spent almost 12 h per day in his bed during the entire post-acute monitoring period. Subject 1 also left his room regularly around 10 a.m. every day, which coincides with daily physical activity classes (as confirmed by the ALF staff).

In **Figure 3B**, we note that subject 2 began to leave her room right after her quarantine period ended on day 17, as indicated by the white regions in her activity graph. Subject 2 seems to be in her room around mealtime; however, her eating patterns are not as regular as those of the other patients. This can be inferred from the lack of clear yellow cones (i.e., time on chair) around meal times, like those observed in the activity graphs of the other two patients. At the end of the study, subject 2 started leaving her room around breakfast time, which indicates that she started joining other residents for breakfast. During nighttime, it is relatively common for subject 2 to wake up in the middle of the night and leave her bed and the whole room as indicated by the white lines interspersed in the blue region. This behavior is confirmed with the staff at the ALF.

As shown in **Figure 3C**, subject 3 remained in her room for the first 2 weeks after lifting the quarantine. Later, she demonstrated a routine of leaving her room in the morning for breakfast then for a physical activity class. However, she remained in her room sitting on a chair for large parts of the remainder of the day and appeared to leave her room in the afternoon and evening (also for meals) only in the last few days that she was monitored. She has a regular diurnal cycle but it is relatively common for her to wake up at night and move from her bed to her chair (as indicated by the yellow lines in the blue region).

# DISCUSSION

Today, clinical studies are limited to a few episodic measurements of each patient, which hampers their ability to track detailed longitudinal disease progression. Our primary aim was to conduct digital phenotyping of the acute and post-acute phases of COVID-19 using a novel radio sensor capable of simultaneously and continuously monitoring multiple physiological and behavioral parameters in a passive and contactless manner. Our results demonstrate that such an approach can facilitate remote tracking of changes in behavior and respiration, which in turn may be markers of the recovery process. The inherent properties of the Emerald device—i.e., eliminating the need for interaction with the device or active data entry on the part of patient, and no requirement for ongoing maintenance such as regular charging—have facilitated the collection of continuous longitudinal data.

Among our subjects, we noted that subject 2, who had the most significant symptoms from COVID-19, appeared to demonstrate the most severe longitudinal disruption of respiration, sleep and daily routine. While we cannot comment on which aspects may be related to COVID-19 vs. exacerbations of premorbid neuropsychiatric symptoms, subject 2's data point to the possibility that especially in older adults, there may be a relationship between behavior symptoms and the COVID-19 disorder itself. Subject 1 had a less severe initial manifestation of COVID-19. However, longitudinal phenotyping indicates a physiological consequence to prolonged quarantine (i.e., worsening mobility and changes to sleep). It is unclear whether this was a direct consequence of COVID-19 infection or born out of the strains of isolation. The data did, however, enable the staff to identify and promptly attend to these issues. In the case of subject 3 who was asymptomatic to COVID-19, phenotyping data shed light on how a combination of the

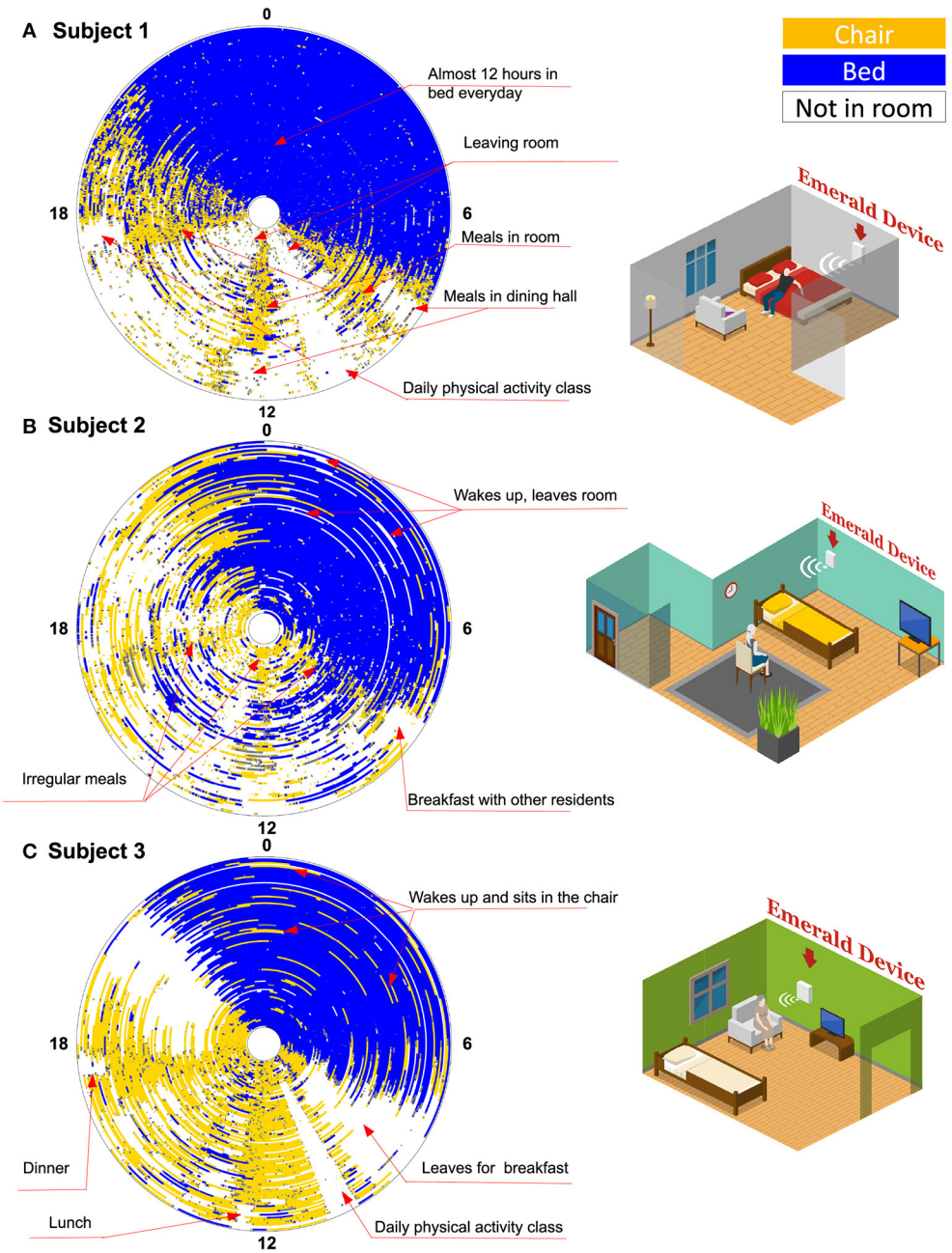


**FIGURE 3 |** Activity graph. Each circle is 1 day. Angle around each circle represents the hour in the day (Zero at the top refers to midnight, and 12 refers to noon time). The innermost circle is the first day the patient became COVID negative, while the outermost circle is the last monitoring day for that patient. The different colors refer to different locations: bed in blue, chair in yellow, and outside the room in white. In total, there are 87 days for subject 1, 87 days for subject 2, and 99 days for subject 3 (Missing days are not visualized).

virus itself and environmental strain from implementing social distancing within an assisted living facility may trigger changes in behavior, such as where to eat and routine on activities.

The study also sheds lights on the specific utility of the studied metrics. It shows that respiration rate may serve as a biomarker in tracking recovery status during the acute phase of the disease. For symptomatic subjects (i.e., subjects 1 and 2), the RR during the acute phase was elevated from its baseline, even when patients did not report breathing issues, and decreased to its baseline as patients became COVID negative. Further, an abnormally high RR is detected for subject 2 before her hospitalization, which indicates that a sudden elevation of RR could serve as a precursor to symptom escalation in COVID-19. In contrast, sleep and gait speed highlighted problems during the post-acute phase.

Specifically, subject 1 developed movement difficulties and took several months to recover his mobility after becoming COVID negative. All subjects exhibited signs of sleep issues that persisted for a long time.

The activity graph is a novel matric that can objectively quantify subjects' daily behavior, and potential changes in their habits. In this graph, each day is abstracted and summarized in a single colored circle, allowing us to visualize daily and repetitive behavior. This provides novel insights into patients' quality of life and daily functions. For example, in our study, the activity graphs have revealed that, while all three subjects live in the same ALF and in principle follow the same ALF schedule, subjects 1 and 3 have regular routines, while subject 2 does not. This could be a sign of subject's 2 agitation and exacerbated cognitive impairment, which was repeatedly noted by the staff. The ability to capture subjects' routine provides a new metric that can help in behavioral phenotyping, an area that currently heavily depends on subjective questionnaires.

The study has several limitations. The sample size is small and may not be representative of the broader COVID-19 population. In addition, we were not able to collect objective behavioral data using standardized scales to compare sensor data with. Also, the monitored physiological signals are limited to respiration, sleep, walking speed and activities, and the monitored space is limited by the radio coverage area. Despite these limitations, we believe that the results demonstrate the feasibility of passive and contactless phenotyping, and its potential for studying the short- and long-term progression of COVID-related symptoms among older adults recovering at home. Our team aims to address these limitations, particularly the absence of comparison measures in ongoing and future work.

At the time of writing, there remains a lack of clarity on how the neuropsychiatric symptoms of COVID may play out in the long term. Our findings indicate that passive digital phenotyping can be a powerful tool to facilitate understanding this issue. With the ability to closely track intra-individual changes in respiration, gait, sleep, and activities, such an approach holds the potential to unlock relationships between different behavioral phenomena associated with COVID-19 infection and recovery. Additionally, contactless passive monitoring technologies can uniquely facilitate detailed longitudinal studies of symptoms' trajectories in older adults, without burdening patients or caregivers.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the MIT institutional review board and the Mass General Brigham Human Research Protections Program ceded review to the MIT IRB. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

IV and DK: initial study concept and design. GZ, IV, YL, WM, and DK: analysis and interpretation of data. WM: acquisition of data and device installation. RM, HC, and WM: human subject management. GZ, YL, and YY: Software/Code development. GZ and YL: statistical analysis. All authors contributed to writing the manuscript.

## FUNDING

The project was funded by the MIT J-Clinic for Machine Learning in Health, a research center at the Massachusetts Institute of Technology. IV was supported by the McLean Division for Geriatric Psychiatry.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.754169/full#supplementary-material

## REFERENCES

1. Weerahandi H, Hochman KA, Simon E, Blaum C, Chodosh J, Duan E, et al. Post-discharge health status and symptoms in patients with severe COVID-19. *J Gen Intern Med*. (2021) 36:738–45. doi: 10.1007/s11606-020-06338-4
2. Puntmann VO, Carerj ML, Wieters I, Fahim M, Arendt C, Hoffmann J, et al. Outcomes of cardiovascular magnetic resonance imaging in patients recently recovered from coronavirus disease 2019 (COVID-19). *JAMA Cardiol*. (2020) 5:1265. doi: 10.1001/jamacardio.2020.3557
3. Carfì A, Bernabei R, Landi F, Group for the GAC-19 P-ACS. Persistent symptoms in patients after acute COVID-19. *JAMA*. (2020) 324:603–5. doi: 10.1001/jama.2020.12603
4. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. (2020) 584:430–6. doi: 10.1038/s41586-020-2521-4
5. Hao F, Tan W, Jiang L, Zhang L, Zhao X, Zou Y, et al. Do psychiatric patients experience more psychiatric symptoms during COVID-19 pandemic and lockdown? A case-control study with service and research implications for immunopsychiatry. *Brain Behavior Immunity*. (2020) 87:100–6. doi: 10.1016/j.bbi.2020.04.069
6. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. (2020) 579:265–9. doi: 10.1038/s41586-020-2008-3
7. Kunz J, Propper C. *Hospital Quality and Deaths From COVID-19 in US Counties*. VoxEU.org (2020). Available online at: https://voxeu.org/article/hospital-quality-and-deaths-covid-19-us-counties (accessed April 12, 2021).
8. Kim EJ, Marrast L, Conigliaro J. COVID-19: magnifying the effect of health disparities. *J Gen Intern Med*. (2020) 35:2441–2. doi: 10.1007/s11606-020-05881-4
9. Sudre CH, Murray B, Varsavsky T, Graham MS, Penfold RS, Bowyer RC, et al. Attributes and predictors of long COVID. *Nature Med*. (2021) 27:626–31. doi: 10.1038/s41591-021-01292-y
10. Jordan RE, Adab P, Cheng KK. Covid-19: risk factors for severe disease and death. *BMJ*. (2020) 368:m1198. doi: 10.1136/bmj.m1198
11. Wang C, Pan R, Wan X, Tan Y, Xu L, Ho CS, et al. Immediate psychological responses and associated factors during the initial stage

of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China. *Int J Environ Res Public Health*. (2020) 17:1729. doi: 10.3390/ijerph17051729
12. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, china, of novel coronavirus-infected pneumonia. *N Engl J*. (2020) 382:1199–207. doi: 10.1056/NEJMoa2001316
13. Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, Ganesh S, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Med*. (2020) 26:1037–40. doi: 10.1038/s41591-020-0916-2
14. Vindegaard N, Benros ME. COVID-19 pandemic and mental health consequences: systematic review of the current evidence. *Brain Behav Immun*. (2020) 89:531–42. doi: 10.1016/j.bbi.2020.05.048
15. Meyer J, McDowell C, Lansing J, Brower C, Smith L, Tully M, et al. Changes in physical activity and sedentary behavior in response to COVID-19 and their associations with mental health in 3052 US adults. *Int J Environ Res Public Health*. (2020) 17:6469. doi: 10.3390/ijerph17186469
16. CDC. *Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19)*. Centers for Disease Control and Prevention (2020). Available online at: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html (accessed April 12, 2021).
17. Orri M, Lipset CH, Jacobs BP, Costello AJ, Cummings SR. Web-based trial to evaluate the efficacy and safety of tolterodine ER 4mg in participants with overactive bladder: REMOTE trial. *Contempor Clin Trials*. (2014) 38:190–7. doi: 10.1016/j.cct.2014.04.009
18. Marcum ZA, Hanlon JT, Murray MD. Improving medication adherence and health outcomes in older adults: an evidence-based review of randomized controlled trials. *Drugs Aging*. (2017) 34:191–201. doi: 10.1007/s40266-016-0433-7
19. Vahia IV, Kabelac Z, Hsu C-Y, Forester BP, Monette P, May R, et al. Radio signal sensing and signal processing to monitor behavioral symptoms in dementia: a case study. *Am J Geriatr Psychiatry*. (2020) 28:820–5. doi: 10.1016/j.jagp.2020.02.012
20. Adib F, Kabelac Z, Katabi D. Multi-person localization via {RF} body reflections. In: *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, 279–292.
21. Adib F, Mao H, Kabelac Z, Katabi D, Miller RC. Smart homes that monitor breathing and heart rate. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul Republic of Korea: ACM), 837–46. doi: 10.1145/2702123.2702200
22. Hsu C-Y, Liu Y, Kabelac Z, Hristov R, Katabi D, Liu C. Extracting gait velocity and stride length from surrounding radio signals. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI'17* (Denver: ACM Press). p. 2116–26. doi: 10.1145/3025453.3025937
23. Zhao M, Yue S, Katabi D, Jaakkola TS, Bianchi MT. Learning sleep stages from radio signals: a conditional adversarial architecture. In: *International Conference on Machine Learning* (PMLR). p. 4100–9. Available online at: http://proceedings.mlr.press/v70/zhao17d.html (accessed April 12, 2021).
24. Vahia IV, May R, Kabelac Z, Hoti K, Munir U, Owoyemi P, et al. AI-based digital phenotyping of behavioral and psychiatric symptoms in dementia: clinical implications. *Alzheimer's Dement*. (2019) 15:P163. doi: 10.1016/j.jalz.2019.06.4335
25. Kabelac Z, Tarolli CG, Snyder C, Feldman B, Glidden A, Hsu C-Y, et al. Passive monitoring at home: a pilot study in Parkinson disease. *DIB*. (2019) 3:22–30. doi: 10.1159/000498922
26. Lichstein K, Stone KC, Donaldson J, Nau SD, Soeffing JP, Murray D, et al. Actigraphy validation with insomnia. *Sleep*. (2006) 29:232–9. doi: 10.1093/sleep/29.2.232
27. Spira AP, Stone K, Beaudreau SA, Ancoli-Israel S, Yaffe K. Anxiety symptoms and objectively measured sleep quality in older women. *Am J Geriatr Psychiatry*. (2009) 17:136–43. doi: 10.1097/JGP.0b013e3181871345

**Conflict of Interest:** DK is a co-founder of Emerald Innovations, which uses wireless sensors similar to the one used in the study to analyze physiological signals.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychiatry

PERSPECTIVE
published: 28 October 2021
doi: 10.3389/fpsyt.2021.732773

# Use of Passive Sensing in Psychotherapy Studies in Late Life: A Pilot Example, Opportunities and Challenges

*Jihui Lee*[1], *Nili Solomonov*[2], *Samprit Banerjee*[1], *George S. Alexopoulos*[2] and *Jo Anne Sirey*[2]*

[1] Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, United States, [2] Weill Cornell Institute of Geriatric Psychiatry, Weill Cornell Medicine, White Plains, NY, United States

Late-life depression is heterogenous and patients vary in disease course over time. Most psychotherapy studies measure activity levels and symptoms solely using self-report scales, administered periodically. These scales may not capture granular changes during treatment. We introduce the potential utility of passive sensing data collected with smartphone to assess fluctuations in daily functioning in real time during psychotherapy for late life depression in elder abuse victims. To our knowledge, this is the first investigation of passive sensing among depressed elder abuse victims. We present data from three victims who received a 9-week intervention as part of a pilot randomized controlled trial and showed a significant decrease in depressive symptoms (50% reduction). Using a smartphone, we tracked participants' daily number of smartphone unlocks, time spent at home, time spent in conversation, and step count over treatment. Independent assessment of depressive symptoms and behavioral activation were collected at intake, Weeks 6 and 9. Data revealed patient-level fluctuations in activity level over treatment, corresponding with self-reported behavioral activation. We demonstrate how passive sensing data could expand our understanding of heterogenous presentations of late-life depression among elder abuse. We illustrate how trajectories of change in activity levels as measured with passive sensing and subjective measures can be tracked concurrently over time. We outline challenges and potential solutions for application of passive sensing data collection in future studies with larger samples using novel advanced statistical modeling, such as artificial intelligence algorithms.

**Keywords: depression, psychotherapy, mobile health, passive sensing, late life**

OPEN ACCESS

***Edited by:***
*Andrea Iaboni,*
*University Health Network, Canada*

***Reviewed by:***
*Ashley Hagaman,*
*Yale University, United States*
*Sanjeev Kumar,*
*University of Toronto, Canada*
*Alastair Van Heerden,*
*Human Sciences Research Council, South Africa*

***Correspondence:**
*Jo Anne Sirey*
*jsirey@med.cornell.edu*

***Specialty section:***
*This article was submitted to Aging Psychiatry, a section of the journal Frontiers in Psychiatry*

***Received:*** *29 June 2021*
***Accepted:*** *30 September 2021*
***Published:*** *28 October 2021*

***Citation:***
*Lee J, Solomonov N, Banerjee S, Alexopoulos GS and Sirey JA (2021) Use of Passive Sensing in Psychotherapy Studies in Late Life: A Pilot Example, Opportunities and Challenges. Front. Psychiatry 12:732773. doi: 10.3389/fpsyt.2021.732773*

## INTRODUCTION

Major Depressive Disorder (MDD) in later life is a heterogenous condition characterized by high variability in biological and clinical features (1, 2). Individuals with MDD vary in their disease course with fluctuations in activity levels and mood during treatment (3). Most depression studies use rating scales administered once weekly to track change

and these assessments do not capture granular time-sensitive changes (2, 4). Passive sensing data collection using smartphone sensors, such as pedometer, accelerometer, gyroscope, GPS, and microphone can capture fluctuations in daily functioning in real time (5, 6). The granularity and multimodal nature of passive sensing data can inform behaviors associated with outcomes and predict response more precisely (7).

While passive sensing has gained its popularity in mental health studies in youth and adult populations (8, 9), few studies examined its applicability on studying mental disorders in late life (10, 11). Even less is known about the use of passive sensing among older adults suffering from trauma and coping with chronic stress and high rates of depression and anxiety. This population is historically underserved and suffers from high rates of depression, anxiety, and post-traumatic stress. Insights from passive sensing data could help understand the heterogenous pattern of treatment response for each patient and thus guide personalization of these therapies to older adults' specific needs and circumstances.

Real-time routine tracking of movement and activity levels in depressed older adults—especially those suffering from trauma—can inform the study of engagement in behavioral activation (BA) psychotherapies that target increasing activity levels to reduce depression severity (12, 13). In these therapies patients are encouraged to engage in meaningful, rewarding activities, including increased time away from home, physical activity, and social interactions (13–15). We developed PROTECT, a behavioral activation and goal directed intervention for late life depression in elder abuse victims. PROTECT is intervention designed to reduce depression among elder abuse victims seeking elder mistreatment reduction services. It targets depressive symptoms by reducing victims' social isolation and increasing behavioral activation leading to a sense of agency and empowerment.

In this paper, we use case study examples from the PROTECT study (16) to present the potential utility of smartphone as a data collection tool in studies of psychotherapy for late-life depression. We examined the individual fluctuations in behavioral activation levels as well as trajectories of passive sensing measures during treatment course. We discuss opportunities and challenges and provide potential solutions and recommendations for future research.

## METHODS

PROTECT psychotherapy includes 9 weekly sessions, where the therapist and the client work collaboratively toward realistic goals by implementing step-by-step action plans. PROTECT has shown to reduce depression severity and increase behavioral activation among elder abuse victims [See (17); See (16) for details]. Patients' reported levels of activity during the study was measured using the Behavioral Activation for Depression Scale (BADS) (18). During the 9-week treatment, BADS were measured at three time points; at baseline, weeks 6 and 9 (treatment end).

At recruitment, the participants consented to carry their smartphones during 9 weeks of intervention for passive sensing data collection and were informed of the types of data collected. They were given an iPhone if they did not own a smartphone. Participants received an instruction step-by-step booklet accompanied by technological training by research assistants on how to operate and use the smartphone. Therapists and research assistants provided ongoing technological support as needed. Participants were informed of the extent of passive sensing data collected from their smartphones and the data were securely stored and managed by using a server-based tracking program.

We focused on four passive sensing measures including step count, time spent at home, time in conversation and the number of times the phone was turned on (screen unlocks), and all measures were recorded daily. We utilized passive sensing data to infer an individual's daily living patterns. For example, higher daily step count reflects increased physical activity levels while more time spent at home may reflect greater isolation and lack of outside activity. More time in conversation may represent more social interaction with others. Finally, the number of screen unlocks is used as a utility measure, reflecting the level of engagement with the smartphone over time. The number of screen unlocks is also used to evaluate the granularity of passive sensing data. More screen unlocks is thought to indicate greater use of the phone and may increase data reliability and granularity (19).

One or more sensors were involved to define each passive sensing measure. For example, a pedometer was used to count the number of steps. Longitude and latitude coordinates were derived via Wi-Fi, cell phone towers and GPS. This location information was used to identify "home" and calculate time spent at home on a day. To protect participants' privacy, we did not record participants' actual geographic location but rather traveling patterns—moving east/west and north/south from an arbitrary reference location. Audio from a smartphone's microphone was sampled periodically to capture the participant's voiced signal. To protect user privacy, proprietary algorithms processed audio data in real time, destroying all contents and only capturing if and how long the participant was engaging in a conversation.

We preprocessed passive sensing data by removing unreliably low (or high) observations to prevent potential bias. Passive sensing recording is intrinsically dependent on the participant's level of engagement with their smartphone. Participants were asked to carry their phone at all times, but their level of engagement with their devices varied because participants might have not carried their phone during the day or their phones either was off/charging, was left at home, or had trouble authorizing the data collection. Heterogeneous levels of engagement across days within a single participant and across participants may result in different degrees of underestimation in passive sensing data and introduce biases if analyzed without addressing this issue. We implemented a 2-stage preprocessing algorithm for mobile health data. The first stage involved principal component analysis on the utility measures from the smartphone such as variability in the battery level and the number of raw observations of each passive measure within a day to quantify each participant's level

of engagement. Days with extremely low level of engagement (the composite engagement score lower than the 30th quantile) were considered unreliable and labeled missing. The second stage used k-nearest neighbors algorithm to classify all unlabeled days. We did not impute the missing data; instead, we presented a smoothed curve to depict the overall trajectories of passive sensing data.

We explored the relation between individual reported activity on the behavioral activation measure (BADS) and passive data collected during the treatment. Overall, we examined the fluctuations in BADS scores and passive sensing measures on a within-person level. We visually inspected whether BADS scores and/or passive sensing measures increased or decreased compared to the individual's average levels and reported how the change in one coincided with the change in the other measures. Fluctuations of daily recorded passive sensing data were captured using a smooth local polynomial regression (LOESS) curve. This trajectory was visually compared with changes in BADS scores from baseline to weeks 6 and 9. As a result, we created individual-specific narratives to link observations from passive sensing data with their potential clinical implications.

## PILOT EXAMPLES

We selected three patients from a small pilot study comparing PROTECT with a referral control condition. These were representative pilot cases to illustrate the potential utility of passive sensing data among depressed older adults. The three patients were most compliant with study protocols of carrying around the smartphones during 9 weeks of treatment and thus produced the most granular passive sensing data. All three patients showed a clinical response, defined as 50% reduction in depressive symptoms on the Montgomery-Asberg Depression Rating Scale (MADRS) (20) score by the end of treatment (week 9). The study was approved by Weill Cornell Medicine's Institutional Review Board and all participants provided written consent for collection and processing of deidentified passive sensing data. **Table 1** shows demographic and clinical characteristics of the subsample of three patients. **Figure 1** shows the clinical and passive sensing data from these three patients.

Patient A is a 62-year-old Black man whose baseline MADRS score was 27. He experienced financial, verbal/emotional, and physical abuse. He showed significant improvement in depression during treatment with a MADRS score of 6 at treatment end. We were not able to obtain passive sensing data from his smartphone in the first 2 weeks of the study, likely due to technical difficulties or need of reminders to carry the phone at all times. However, following this initial period, his consistently high screen unlocks suggested high engagement with his smartphone throughout the intervention after the first 2 weeks. Patient A began and ended therapy with relatively low levels of behavioral activation (Intake BADS = 102; Week 9 BADS = 109). These low activity levels correspond with less time spent in conversation at the beginning and end of treatment and decrease in step count by Week 9. At Week 6, Patient A reported the highest activity levels (BADS = 127). This increase corresponded with more time spent at home and more time in conversation, which may indicate this patient engaged in helpful activities and conversations with others at home during mid-treatment.

Patient B is a 65-year-old White Hispanic woman who was divorced and lived alone with a history of verbal/emotional and physical abuse. She reported severe major depression (MADRS score of 33 at baseline) and extremely low levels of behavioral activation (BADS = 39) at the start of treatment. Patient B was consistently engaged with her phone, as reflected by high screen unlocks throughout treatment. By Week 6, the patient reported an increase in behavioral activation levels (BADS = 49). Patient B showed consistently high step count (averaging 4,000 steps a day), spent time in conversation and was away from home throughout treatment. This high engagement in multiple activities may have contributed to her significant reduction in depression (MADRS at Week 9 = 15),

**TABLE 1 |** Demographic and clinical characteristics of the sample.

| | | Patient A | Patient B | Patient C |
|---|---|---|---|---|
| Treatment group | | PROTECT | PROTECT | Referral |
| Age (years) | | 62 | 65 | 69 |
| Gender | | Male | Female | Female |
| Marital status | | Separated | Divorced | Married |
| Living situation | | Lives with Others | Lives Alone | Lives with Others |
| Ethnicity | | Non-Hispanic | Hispanic | Non-Hispanic |
| Race | | African American | White | African American |
| Religion | | Other | Catholic | Catholic |
| Education (years) | | 14 | 12 | 14 |
| Financial situation | Perception of financial status | Has just enough | Has just enough | Has just enough |
| | Annual Income | <9K | 13K–16K | 13K–16K |
| Abuse | Financial | Y | | |
| | Verbal / Emotional | Y | Y | |
| | Physical | Y | Y | Y |

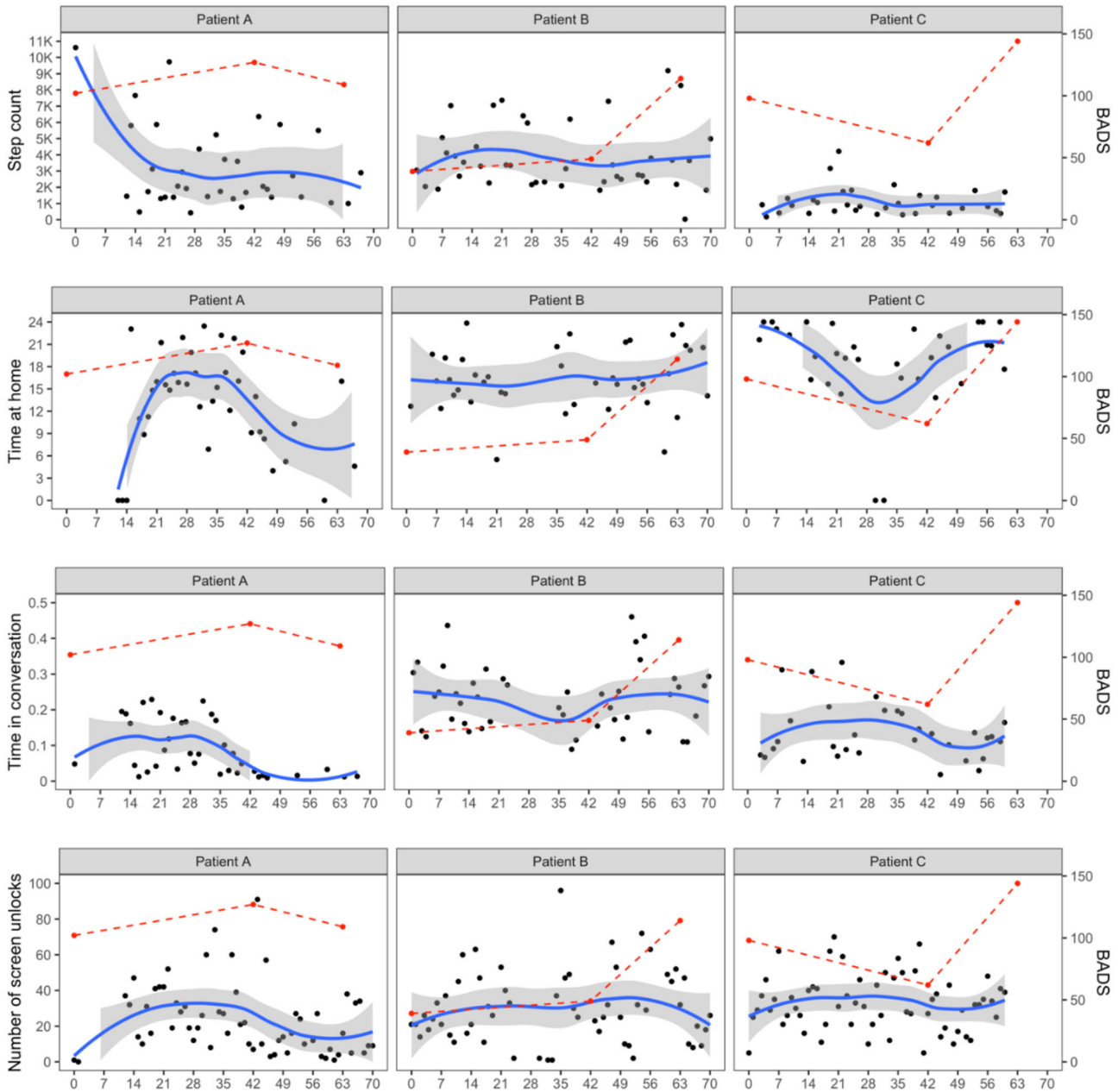


**FIGURE 1 |** Passive sensing data fluctuations over treatment. BADS, Behavioral Activation for Depression Scale. Columns represent patients and rows represent different smartphone data (step count, time spent at home, time in conversation, and the number of screen unlocks). For all panels, the x-axis shows time in days. Points in black represent daily recorded smartphone data and a blue line with a shaded band is a smooth local polynomial regression (LOESS) curve with its 95% confidence interval. Points and a dashed line in red show the fluctuations in BADS scores from baseline, weeks 6 to 9 (end of treatment).

as well as meaningful increase in levels of behavioral activation (BADS = 114).

Patient C is a 69-year-old Black woman who reported physical abuse with moderate depression (MADRS = 23) and relatively high levels of behavioral activation at intake (BADS = 98). Patient C maintained high levels of screen unlocks reflecting consistent engagement with her phone. However, low step count and time spent in conversation throughout treatment reflected low outside and social activity levels. Nevertheless, her BADS score increased to 114 by treatment end. For this patient, more time at home corresponded with higher levels of behavioral activation at the beginning and end of treatment. During Week 6, the patient did not spend much time at home, suggesting potential disruption to her usual routine. It may have contributed to the lower behavioral activation (BADS = 62). This also suggests she may have engaged in pleasurable activities at home.

## OPPORTUNITIES AND CHALLENGES

Passive sensing data offer the potential to observe daily activity between standardized assessments of behavioral activation and changes in depression severity (21). Our data illustrate the individual level variation observed among three patients who showed improved depression during PROTECT treatment. All three patients showed clinically significant response and were engaged with the smartphone during the study. However, the figures illustrate variability in passive sensing data and behavioral activity level reports within-patient over time as well as between-patient differences. Real-time assessment of individuals in their natural environments maximizes ecological validity and the granularity of smartphone data can capture detailed fluctuations of behavior over the study period (4). The variability observed may also reflect the different ways that patients become activated as part of the therapy.

Multimodal data can provide a more nuanced understanding of behavioral patterns for each individual. The passive sensing data provide an opportunity for digital phenotyping, i.e., moment-by-moment quantification of the individual-level human phenotype *in-situ* (22, 23). Passive sensing can reflect changes in physical activity and time spent outside (24), which may correlate with mental health outcomes, such as loneliness and social isolation (25), as well as anxiety, stress and depression (26). In our project, step count and time spent at home were used to quantify participants' daily activity. By considering these measures simultaneously we discerned days with high activity level at or around home from those days with a greater travel diameter. Additional information regarding planned goals and the types of activities patients engaged in while at home or outside could expand our understanding of clinical meaning of passive sensing data on an individual level. Further, future work could investigate associations between activity levels and loneliness and social isolation, which is prevalent among elder abuse victims. Collection of these data in a large sample may contribute to the understanding of behavioral patterns associated with treatment response and guide development of personalized treatments (27, 28).

Unique characteristics of passive sensing introduce a new area of data analyses methods. Although we did not apply advanced analytic methods in this small, classic statistical methods such as mixed effects models and generalized estimation equation as well as pre vs. post hypothesis tests can be used to analyze the temporal changes in passive sensing data (29–32). Creating a platform that streamlines the passive sensing data collection, management and analysis allows to collect a bigger sample (23), and passive sensing data in a large sample provides ample opportunities to develop and implement sophisticated statistical methods and machine learning algorithms (32) for suicide prediction, for example. Types of machine learning and artificial intelligence algorithms for passive sensing data range from feature extraction and selection (33), gradient boosting (6), to artificial neural networks (34). Large sample data might help ferret out what activities are most frequently associated with increase in activity levels, behavioral activation reports and improvement in depression.

A challenge of implementation of passive data is reliance on the engagement with the smartphone. The patients we presented consistently used their smartphone. However, many elder abuse victims may struggle to maintain high levels of engagement. These individuals often struggle with chronic and acute trauma and are more likely to be members of marginalized minorities, come from lower socioeconomic background, and experience medical burden and disability (35). Previous studies have documented a digital divide within the older adult population, with those from lower socioeconomic status and less resources least likely to adapt to technology (36–38). However, studies have shown that technology use among older adults has increased dramatically over the past two decades (37). In our case study examples, we observed from Patient A that it may take a while for older adults to get used to using the mobile devices, but they could adapt to use the new technology and provide useful passive sensing data (39). Nonetheless, tailoring technology to the older adults' specific needs and circumstances can significantly enhance passive sensing data quality, validity and accuracy (40–42). To protect participants' privacy, we did not collect the content of conversations or the specific locations visited when participants left their home. Data on content could elucidate the affective valence of conversations and their potential effect on outcome. Social interactions with supportive others are especially therapeutic (15). However, it is also possible that elder abuse victims spent time speaking with supportive others, or alternatively with the identified abuser. Similarly, we do not know whether participants who left their home engaged in pleasurable activities aligned with their treatment goals, or activities that may have increased distress. Further research is needed to examine the qualitative nature of passive sensing data collected to increase clinical interpretability.

In summary, passive data tracking can provide nuanced granular data on activity and engagement patterns over time. Despite substantially growing interest in incorporating mobile technology to mental health studies in recent years, the extent of technology used for continuous monitoring of older adults has been relatively limited to environmental such as in-home sensors (11). To our knowledge, this is the first study which used passive sensing data from a population of elder abuse victims. If integrated with clinical response trajectories, passive sensing data can improve identification of personalized interventions leading to increased activity and well-being among older adults (10). However, the reliability of smartphone data is dependent on the participant's active and sustained engagement with smartphones (43). Challenges include low perceived ease of smartphone use and the lack of technological support tailored to older adults' needs. Potential solutions include implementing changes in mobile technology based on older adults' needs and preferences and use of wearable devices. Future work will investigate relationships between activity levels measured by passive sensing and treatment outcomes in larger samples using advanced statistical approaches.

## DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because data could be made available by the authors in compliance with the funder's data sharing policy. Requests to access the datasets should be directed to jsirey@med.cornell.edu.

## ETHICS STATEMENT

The study was approved by Weill Cornell Medicine's Institutional Review Board. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

## AUTHOR CONTRIBUTIONS

JL, NS, and SB contributed to the analyses. GA, JS, and SB contributed to study design. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was supported by P50 MH113838 (GA) and K23 MH123864 (NS).

## REFERENCES


1. Beijers L, Wardenaar KJ, van Loo HM, Schoevers RA. Data-driven biological subtypes of depression: systematic review of biological approaches to depression subtyping. *In Mol Psychiatry*. (2019) 24:888–900. doi: 10.1038/s41380-019-0385-5
2. Fried E. Moving forward: how depression heterogeneity hinders progress in treatment and research. *In Exp Rev Neurother.* (2017) 17: 423–5. doi: 10.1080/14737175.2017.1307737
3. Schoevers RA, Van Borkulo CD, Lamers F, Servaas MN, Bastiaansen JA, Beekman ATF, et al. Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. *Psychol Med.* (2021) 51:1906–15. doi: 10.1017/S0033291720000689
4. Bos FM, Schoevers RA, aan het Rot M. Experience sampling and ecological momentary assessment studies in psychopharmacology: a systematic review. *In Eur Neuropsychopharmacol.* (2015) 25:1853–64. doi: 10.1016/j.euroneuro.2015.08.008
5. Fortuna KL, Torous J, Depp CA, Jimenez DE, Areán PA, Walker R, et al. A future research agenda for digital geriatric mental healthcare. *In Am J Geriatric Psychiatry.* (2019) 27:1277. doi: 10.1016/j.jagp.2019.05.013
6. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. *Sensors.* (2020) 20:1–16. doi: 10.3390/s2012 3572
7. Zulueta J, Leow AD, Ajilore O. Real-time monitoring: a key element in personalized health and precision health. *FOCUS.* (2020) 18:175–80. doi: 10.1176/appi.focus.20190042
8. Kerst A, Zielasek J, Wolfgang G. Smartphone applications for depression: a systematic literature review and a survey of health care professionals' attitudes towards their use in clinical practice. *Eur Arch Psychiatry Clin Neurosci.* (2020) 2:139–52. doi: 10.1007/s00406-018-0974-3
9. Liang Y, Zheng X, Zeng DD. A survey on big data-driven digital phenotyping of mental health. *Inform Fusion.* (2019) 52:290–307. doi: 10.1016/j.inffus.2019.04.001
10. Grossman JT, Frumkin MR, Rodebaugh TL, Lenze EJ. MHealth assessment and intervention of depression and anxiety in older adults. *In Harvard Rev Psychiatry.* (2020) 28:203–14. doi: 10.1097/HRP.0000000000000255
11. Kang HG, Mahoney DF, Hoenig H, Hirth VA, Bonato P, Hajjar I. In situ monitoring of health in older adults: technologies and issues. *J Am Geriatr Soc.* (2010) 58:1579–86. doi: 10.1111/j.1532-5415.2010.02959.x
12. Alexopoulos GS, Areán PA. A model for streamlining psychotherapy in the RDoC era: the example of "Engage". *Mol Psychiatry.* (2014) 19:14–9. doi: 10.1038/mp.2013.150
13. Jacobson NS, Martell CR, Dimidjian S. Behavioral activation treatment for depression: returning to contextual roots. *Clin Psychol Sci Practice.* (2001) 8:255–70. doi: 10.1093/clipsy.8.3.255
14. Alexopoulos GS, Raue PJ, Banerjee S, Marino P, Renn BN, Solomonov N, et al. Comparing the streamlined psychotherapy "Engage" with problem-solving therapy in late-life major depression. *A randomized clinical trial. Mol Psychiatry*. (2020). doi: 10.1038/s41380-020-0832-3. [Epub ahead of print].
15. Solomonov N, Bress JN, Anne Sirey JA, Gunning FM, Flückiger C, Raue PJ, et al. Engagement in socially and interpersonally rewarding activities as a predictor of outcome in 'Engage' behavioral activation therapy for late-life depression. *American J Geriatr Psychiatry.* (2019) 27:571–8. doi: 10.1016/j.jagp.2018.12.033
16. Sirey JA, Solomonov N, Guillod A, Zanotti P, Lee J, Soliman M, et al. PROTECT: a novel psychotherapy for late-life depression in elder abuse victims. *Int Psychog.* (2021) 33:521–5. doi: 10.1017/S1041610221000430
17. Sirey JA, Halkett A, Chambers S, Salamone A, Bruce ML, Raue PJ, et al. PROTECT: a pilot program to integrate mental health treatment into elder abuse services for older women. *J Elder Abuse Neglect.* (2015) 27:438–53. doi: 10.1080/08946566.2015.1088422
18. Kanter JW, Mulick PS, Busch AM, Berlin KS, Martell CR. The behavioral activation for depression scale (BADS): psychometric properties and factor structure. *J Psychopathol Behav Assess.* (2007) 29:191–202. doi: 10.1007/s10862-006-9038-5
19. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. *Evidence-Based Mental Health.* (2020) 23:161–6. doi: 10.1136/ebmental-2020-300180
20. Montgomery SA, Åsberg MA. A new depression scale designed to be sensitive to change. *Br J Psychiatry.* (1979) 134:382–9. doi: 10.1192/bjp.134.4.382
21. Onnela JP. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology.* (2021) 46:45–54. doi: 10.1038/s41386-020-0771-3
22. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *In Neuropsychopharmacol.* (2016) 41:1691–6. doi: 10.1038/npp.2016.7
23. Torous J, Kiang MV, Lorme J, Onnela J-P. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health.* (2016) 3:e16. doi: 10.2196/mental.5165
24. Trifan A, Oliveira M, Oliveira JL. Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. *JMIR MHealth UHealth.* (2019) 7:e12649. doi: 10.2196/12649
25. Afsaneh Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, Cohen S, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data. *JMIR Mhealth Uhealth.* (2019) 2019:7E13209. doi: 10.2196/13209
26. Müller SR, Peters H, Matz SC, Wang W, Harari GM. Investigating the relationships between mobility behaviours and indicators of subjective well–being using smartphone–based experience sampling and GPS tracking. *Eur J Pers.* (2020) 34:714–32. doi: 10.1002/per.2262
27. Aung MH, Matthews M, Choudhury T. Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies. *In Dep Anxiety.* (2017) 34:603–9. doi: 10.1002/da.22646

28. Seppälä J, De Vita I, Jämsä T, Miettunen J, Isohanni M, Rubinstein K, et al. Mobile phone and wearable sensor-based mHealth approaches for psychiatric disorders and symptoms: systematic review. *JMIR Mental Health.* (2019) 6:e9819. doi: 10.2196/mental.9819
29. Barnett I, Torous J, Staples P, Keshavan M, Onnela JP. Beyond smartphones and sensors: choosing appropriate statistical methods for the analysis of longitudinal data. *In J Am Med Inform Assoc.* (2018) 25:1669–74. doi: 10.1093/jamia/ocy121
30. Hebden L, Cook A, van der Ploeg HP, King L, Bauman A, Allman-Farinelli M. A mobile health intervention for weight management among young adults: a pilot randomised controlled trial. *J Human Nutr Diet.* (2014) 27:322–32. doi: 10.1111/jhn.12155
31. Proudfoot J, Clarke J, Birch MR, Whitton AE, Parker G, Manicavasagar V, et al. Impact of a mobile phone and web program on symptom and functional outcomes for people with mild-to-moderate depression, anxiety and stress: a randomised controlled trial. *BMC Psychiatry.* (2013) 13:1–12. doi: 10.1186/1471-244X-13-312
32. Torous J, Staples P, Barnett I, Sandoval LR, Keshavan M, Onnela, et al. -P. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *Npj Dig Med.* (2018) 1:15. doi: 10.1038/s41746-018-0022-8
33. Chikersal P, Doryab A, Tumminia M, Villalba DK, Dutcher JM, Liu X, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM Trans Comp Human Int.* (2021) 28:1–41. doi: 10.1145/3422821
34. Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. *In Mol Psychiatry.* (2019) 24:1583–98. doi: 10.1038/s41380-019-0365-9
35. Dong XQ. Elder abuse: systematic review and implications for practice. *In J Am Geriatr Soc.* (2015) 63:1214–38. doi: 10.1111/jgs.13454
36. Hong YA, Cho J. Has the digital health divide widened? Trends of health-related internet use among older adults from 2003 to 2011. *J Gerontol Series B Psychol Sci Soc Sci.* (2017) 72:856–63. doi: 10.1093/geronb/gbw100
37. Hung LY, Lyons JG, Wu CH. Health information technology use among older adults in the United States, 2009–2018. *Curr Med Res Opinion.* (2020) 36:789–97. doi: 10.1080/03007995.2020.1734782
38. Levy H, Janke AT, Langa KM. Health literacy and the digital divide among older americans. *J Gen Int Med.* (2015) 30:284–9. doi: 10.1007/s11606-014-3069-5
39. Benda NC, Alexopoulos GS, Marino P, Sirey JA, Kiosses DN, Ancker JS. The age limit does not exist: a pilot usability assessment of a SMS-messaging and smartwatch-based intervention for older adults with depression. *AMIA Ann Symp Proc Arch.* (2020) 213–22.
40. Jayasinghe N, Moallem BI, Kakoullis M, Ojie MJ, Sar-Graycar L, Wyka K, et al. Establishing the feasibility of a tablet-based consent process with older adults: a mixed-methods study. *Gerontologist.* (2019) 59:124–34. doi: 10.1093/geront/gny045
41. Ramsey AT, Wetherell JL, Depp C, Dixon D, Lenze E. Feasibility and acceptability of smartphone assessment in older adults with cognitive and emotional difficulties. *J Technol Human Serv.* (2016) 34:209–23. doi: 10.1080/15228835.2016.1170649
42. Reading Turchioe M, Grossman LV, Baik D, Lee CS, Maurer MS, Goyal P, et al. Older adults can successfully monitor symptoms using an inclusively designed mobile application. *J Am Geriatr Soc.* (2020) 68:1313–8. doi: 10.1111/jgs.16403
43. Berrouiguet S, Perez-Rodriguez MM, Larsen M, Baca-García E, Courtet P, Oquendo M. From eHealth to iHealth: transition to participatory and personalized medicine in mental health. *J Med Int Res.* (2018) 20:e7412. doi: 10.2196/jmir.7412


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

PERSPECTIVE
published: 15 November 2021
doi: 10.3389/fpsyt.2021.734909

# Artificial Intelligence: An Interprofessional Perspective on Implications for Geriatric Mental Health Research and Care

*Brenna N. Renn*[1]*, *Matthew Schurr*[1], *Oleg Zaslavsky*[2] and *Abhishek Pratap*[3,4,5,6]

[1] Department of Psychology, University of Nevada, Las Vegas, NV, United States, [2] Department of Biobehavioral Nursing and Health Informatics, University of Washington, Seattle, WA, United States, [3] Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, ON, Canada, [4] Vector Institute for Artificial Intelligence, Toronto, ON, Canada, [5] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States, [6] Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom

Artificial intelligence (AI) in healthcare aims to learn patterns in large multimodal datasets within and across individuals. These patterns may either improve understanding of current clinical status or predict a future outcome. AI holds the potential to revolutionize geriatric mental health care and research by supporting diagnosis, treatment, and clinical decision-making. However, much of this momentum is driven by data and computer scientists and engineers and runs the risk of being disconnected from pragmatic issues in clinical practice. This interprofessional perspective bridges the experiences of clinical scientists and data science. We provide a brief overview of AI with the main focus on possible applications and challenges of using AI-based approaches for research and clinical care in geriatric mental health. We suggest future AI applications in geriatric mental health consider pragmatic considerations of clinical practice, methodological differences between data and clinical science, and address issues of ethics, privacy, and trust.

**Keywords: machine learning, deep learning, psychotherapy, older adults, technology, depression, natural language processing, personalized medicine/personalized health care**

OPEN ACCESS

***Edited by:***
*Ellen E. Lee,*
*University of California, San Diego,*
*United States*

***Reviewed by:***
*Helmet Karim,*
*University of Pittsburgh, United States*
*Huali Wang,*
*Peking University Sixth Hospital, China*

***Correspondence:**
*Brenna N. Renn*
*brenna.renn@unlv.edu*

***Specialty section:***
*This article was submitted to Aging Psychiatry, a section of the journal Frontiers in Psychiatry*

***Received:*** *01 July 2021*
***Accepted:*** *07 October 2021*
***Published:*** *15 November 2021*

***Citation:***
*Renn BN, Schurr M, Zaslavsky O and Pratap A (2021) Artificial Intelligence: An Interprofessional Perspective on Implications for Geriatric Mental Health Research and Care. Front. Psychiatry 12:734909. doi: 10.3389/fpsyt.2021.734909*

## INTRODUCTION

Artificial intelligence (AI) learns patterns in large multimodal datasets both within and across individuals (1) to help improve understanding of current clinical status [e.g., calculating a risk score for heart disease (2)] or predict a future outcome [e.g., predicting daily mood fluctuations (3)]. Such technology is increasingly critical and opportune in our digital healthcare revolution. Advances in technology, such as the ubiquity of smartphones, other wearables, and embedded sensors, in addition to the emergence of large datasets (e.g., electronic health records) have altered the landscape of clinical care and research. AI approaches can dynamically interpret such complex data and generate incredible insight to potentially improve clinical methods and results. AI holds the potential to revolutionize geriatric mental health care and research by learning and applying such individualized predictions to guide clinical decision-making. Specifically, AI can contribute to the proactive and objective assessment of mental health symptoms to aid in diagnosis and treatment delivery to suit individual needs, including long-term monitoring and care management.

The big promise for AI in mental health care and research—largely due to its reliance on big data—is to facilitate understanding of what works for whom, and when. However, much of this momentum is driven by machine learning experts (e.g., data and computer scientists and engineers)

and runs the risk of being disconnected from pragmatic issues in clinical practice. In this piece, we bring the perspectives of clinician-scientists in clinical geropsychology (BNR and MS) and geriatric nursing (OZ) to bear on expertise in AI and data science (AP). We provide a brief overview of AI in mental health with the main focus on possible applications and challenges of using AI-based approaches for research and clinical care in geriatric mental health.

## CLINICAL APPLICATIONS OF AI

The field of geriatric mental health focuses on both normal and pathological aging from a biological and psychological perspective; this encompasses acute and chronic physical illness, neurodegeneration and cognitive impairment, and mental disorders in people aged 65 and older. Research and clinical applications within geriatric mental health focus on both care delivery and the evaluation, diagnosis, prevention, and treatment of such disorders. The appeal of such AI-enabled technology to advance geriatric mental health care is 2-fold. First, AI technologies hold the potential to develop precision models that are both personalized and conceivably more accurate than traditional clinical care using vast amounts of real-world multimodal data about patients, including the influence of environmental and other risk and protective factors. Secondly, technology in general has long been heralded as a means to overcome traditional access barriers of cost, time, distance, and stigma, all of which are relevant for older adults. While a thorough review of AI is beyond the scope of this Perspective, relevant machine learning (ML) and deep learning applications [including natural language processing (NLP)] of AI are presented in **Table 1**. Interested readers are directed to other reviews for more in-depth descriptions of AI in mental health (10–13). We briefly review three clinical domains relevant to geriatric mental health care below and subsequently suggest specific areas where AI can assist clinical care (see **Figure 1**).

### Assessment, Symptom Recognition, and Diagnosis

A major issue in geriatric mental health care and research is accurate classification of a disorder. Many mental health conditions, including late-life depression, go undetected and

**TABLE 1 |** Overview of artificial intelligence (AI) technologies with relevance to geriatric mental health.

| Type of AI technology | Definition | Clinical example |
|---|---|---|
| Machine learning (ML) | A family of statistical techniques that allow computer programs to make predictions and decisions based on past data. | |
| Supervised | A type of machine learning that uses labeled datasets to "train" algorithms. For example, a dataset includes a label for cognitive impairment (cognitively impaired or not). The model learns on a set of training data, then the algorithm is tested on unlabeled data to ensure its accuracy in classifying the target variable. | Modeling a variety of clinical, lifestyle, and sociodemographic factors to help predict cognitive function in older people; clinicians could use this non-invasive screening method to decide whether or not a patient warrants further in-depth cognitive assessment (4). |
| Unsupervised | A type of machine learning based on analyzing unlabeled data to discover hidden patterns or data groups. The algorithm is not provided with a label thus, subject-matter experts must evaluate the data output to ensure its usefulness. Unlabeled data are sorted into groups or patterns to identify the underlying structure of the data. | Identifying high likelihood of dementia in population-based surveys (5). |
| Deep learning (DL) | A subfield of machine learning; deep learning models use computer programs called artificial neural networks to discover latent relationships in complex, raw data. DL algorithms develop multiple hierarchical layers of data representation and learn complex underlying patterns. | A trial in India used deep learning to predict depression among older adults and had a high prediction accuracy (97.2%) based on sociodemographic variables and morbidity (sleep difficulties, mobility difficulty, hearing, and visual impairment) (6). |
| Natural language processing (NLP) | Natural language processing (NLP) aims to comprehend human language by extracting word features (such as syntax, grammar, and semantic meaning) from text and transcribed speech. It holds much potential in mental health research and care, where text (e.g., electronic health records [EHRs]) and speech (e.g., psychotherapy session content) are key real-world data sources. | Using speech features (e.g., speech fluency, prosody, duration) to detect late-life depression (7). |
| Computer vision | Computer vision is used to detect and classify objects. The model imposes a grid-like structure on images and learns key features, such as edges and curves, to build a unique model to recognize similar objects. | Extracting gait features from video recordings of older adults with dementia ("human pose estimation") to predict future falls (8). |
| Reinforcement learning | Deep reinforcement learning (RL) is a form of adaptive learning that rewards desired outcomes (behaviors) and penalizes undesirable or unwanted outcomes. Such algorithms learn to sense and interpret the right and wrong actions in an environment and train through trial and error. | Helping providers by editing written therapeutic exchanges to increase the level of expressed empathy, a critical component of therapeutic conversations (9). |

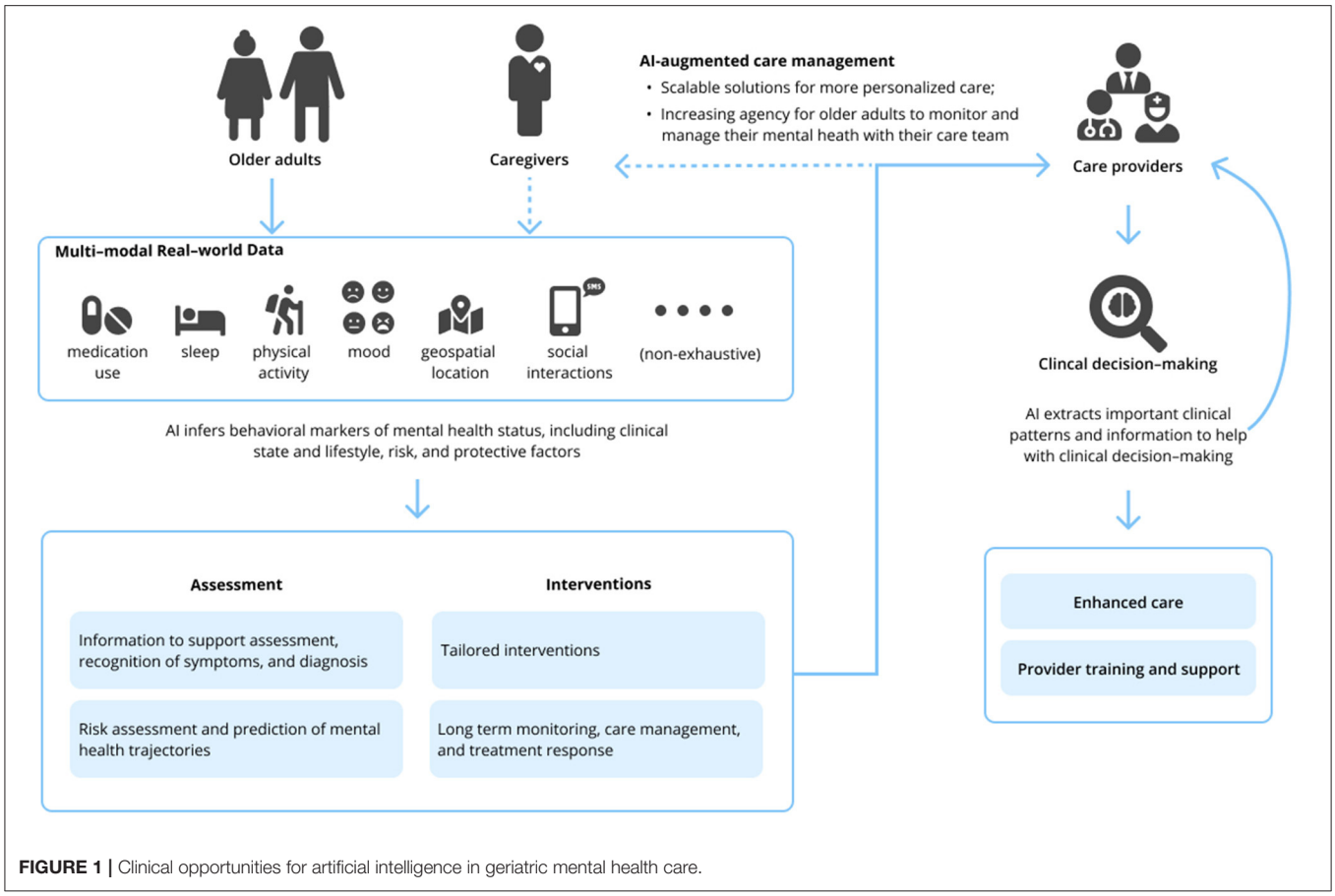


**FIGURE 1 |** Clinical opportunities for artificial intelligence in geriatric mental health care.

untreated (14). When symptoms are recognized, diagnosis primarily relies on subjective recollections of symptoms, which leads to a considerable amount of diagnostic variability and may be subject to patient recall bias (15). Moreover, differential diagnosis can be particularly challenging in older adult patients with multimorbidities or when considering conditions with overlapping symptoms. A compelling application of AI is accurately predicting who needs mental health treatment before someone realizes they need it—or, before symptoms become too burdensome—by tracking early cues related to a change in an individual's daily behavior. One of the most ubiquitous opportunities is personal sensing, which converts the huge amount of sensor data collected by our phones (or other wearable devices) into clinically meaningful information about behaviors, thoughts, and emotions to make inferences about clinical status and/or disorders (16). These data sources can be rich and multimodal, encapsulating sleep, social interaction, and physical activity, to name a few features. Such data may serve as objective measures for hallmark symptoms (e.g., fatigue and sleep disturbances) in the diagnosis of depression in older adults (17).

A small but growing body of literature has begun to apply AI approaches to geriatric mental health assessment and diagnosis, largely in the context of depression (10) and neurocognitive impairment (11). For example, language ability and processing—including spontaneous speech—is often an early affected cognitive domain in the course of dementia, especially Alzheimer's disease, and has been proposed as a target for early recognition and diagnosis (18). However, traditional methods of early recognition and diagnosis often produce significant overlap with "normal" cognitive functioning among older adults, and thus have reduced clinical utility in early detection (19). AI techniques such as NLP may detect speech features (e.g., acoustic features such as pause duration and emotion) that are sensitive to cognitive decline and may better differentiate those with early impairment than traditional neuropsychological assessment (19, 20).

## Treatment and Treatment Monitoring

The shortage of geriatric mental health specialists (21) and barriers to treatment seeking among older adults (22–24) mean that patients with mental health needs are often delayed in obtaining treatment, if they receive treatment at all. As it stands in current clinical practice, access to evidence-based treatment is often limited (25), and when implemented, treatment decisions are often guided by trial and error. Ongoing assessment is also crucial to assess effectiveness of treatment, but may be overlooked or untenable in routine practice, rendering ineffective treatment decisions. As AI aims to predict who needs mental health services, the next compelling application of such technology will be to answer the question of "What works for whom, and when?"

A promising application of AI for mental health, inspired by precision medicine, is to identify subgroups of patients with similar symptom expressions and outcomes to guide treatment decisions, commonly referred to as "subtyping" (26). Once treatment is initiated, AI could also help clinicians monitor response to treatment and symptom trajectory, such as through passive recording of behavioral data using wearable sensors.

Another example to optimize treatment is to quickly mobilize tailored supports using just-in-time adaptive interventions. These adapt the type, timing, and intensity of treatment based on the individual's need at the moment and context they most need the support (27). Such efforts in geriatrics include computational modeling based on smartphone data to target health behavior change (i.e., low physical activity and sedentary behavior) in older adults (28). These models use sensor data to monitor health states with the goal of delivering personalized interventions to mitigate behavioral and psychological factors that contribute to health risk.

Intelligent voice assistants, virtual health agents, and conversational agents (e.g., chatbots) are designed to reduce health care system burden (29) and improve patient autonomy and self-management (30). While mainstream conversational agents have yet to be tested with older adults, preliminary evidence suggests that older adults are comfortable self-disclosing with other conversational agents (19). It is conceivable that such AI may one day be used to support "aging in place," such as allowing older adults to complete remote assessments for routine monitoring. Researchers are also prototyping AI-based "smart homes" to support safety and independence among older adults and individuals with disabilities and chronic conditions (31). However, ongoing engagement is required for AI to assist with long-term monitoring or treatment delivery. For example, while intelligent voice assistants such as Amazon's Echo have the potential to support independence among older adults, users may discontinue such products if they do not realize benefits or experience challenges using such devices in shared spaces (32).

## Clinical Decision-Making, Provider Training and Support

AI may free up time for the clinician to implement treatment decisions and focus on other therapeutic targets (e.g., client rapport) where the application of current AI technologies has been ineffective (33). AI-based data collection and harmonization may streamline patient flow, automate assessments, monitor longitudinal trajectories and outcomes, reduce paperwork, and monitor medication(s) and potential contraindications (34), thus freeing providers to practice the "human" elements of mental health care. AI may also be used to train mental health professionals. This is particularly relevant to the current geriatric workforce shortage (21). Examples include virtual patient simulations to train and evaluate clinical skills (e.g., asking proper diagnostic questions) (35) and NLP to analyze the quality of engagement between a therapist and a patient in a psychotherapy session (36). However, limitations to this technology remain; this work found models only modestly predicted patient-rated alliance from psychotherapy session content (36).

# CHALLENGES AND OPPORTUNITIES

Now we want to highlight some challenges and propose how AI solutions can be applied to real-world problems in geriatric mental health care and research. We suggest the AI community partners with clinician-researchers and care teams (including nursing staff, providers, and caregivers), and vice versa, in order to make most relevant the potential of such technology. This is particularly germane to issues of geriatric mental health care.

## Unique Challenges in Geriatrics

Aging is a complex process that involves interconnected changes spanning cellular to psychological to sociocultural processes, the results of which present unique challenges when working in geriatrics. First, older adults are less likely than younger adults to receive accurate diagnosis and treatment for mental health issues (37), and barriers are greater among racially and ethnically diverse older adults compared to their non-Latino White counterparts (23). Workforce shortages, specifically lack of providers with competencies in the specific needs of older adults, contribute to these issues (21). Older adults also present with greater comorbidity, chronicity, and complexity than their younger adult counterparts; acute and chronic physical health conditions, medication use, and cognitive, sensory, or functional impairments can all complicate the detection and diagnosis of a mental health condition. Additionally, the variation in manifestation of mental health symptoms and treatment responses in older adults affects timely and accurate diagnosis. For example, an enduring finding in geriatric mental health care is that older adults with depressive symptoms are less likely than younger adults to present with sadness and are more apt to endorse anhedonia (loss of interest or pleasure), apathy, and somatic symptoms such as fatigue, diffuse aches and pain, or malaise (38). Somatic symptoms of late-life depression also overlap with symptoms of chronic disease, potentially obscuring or complicating diagnosis of mental health conditions. Moreover, older adults may be poor utilizers of mental health services if they are uncertain whether their symptoms are due to psychological problems or normal aging (39). Thus, AI holds promise to capture real-world behavioral data to aid in the recognition and diagnosis of mental health conditions in older adults. The majority of the literature points to applications of AI among younger adults (often college-aged convenience samples). Next steps are to prototype, train, and validate AI approaches on data from diverse respondents, including older adults, to capture the specific clinical needs and heterogeneity in the population.

Social, environmental, and familial contexts are important considerations in geriatric mental health. Caregiving is one such relevant factor. Persons with chronic or life-limiting disease–often older adults—require progressively extensive attention and assistance with activities of daily living. This care is often provided by family members or other unpaid caregivers. AI technologies may better prepare and support caregivers in their

tasks. A systematic review of 30 studies (40) described a range of assistive AI devices designed to facilitate caregiving, such as support with dressing or handwashing or detecting falls. However, the review noted that most studies were descriptive or exploratory, offering very limited evidence of such technology to date.

Social factors such as social isolation and loneliness may also exacerbate mental health issues; indeed, a recent federal report highlighted the epidemic of social isolation and loneliness among older adults (41). AI could be used to both assess and offer supports for loneliness. For example, a proof-of-concept study used NLP to identify loneliness among U.S. community-dwelling older adults based on speech from qualitative interviews (42). Importantly, this study attempted to understand sex differences in the reporting of such a complex psychological construct—something with which clinicians may struggle. As with much of the AI applications to date in geriatric mental health, the authors note that future work will need larger, more diverse samples and to incorporate multimodal data streams to improve the predictions. In any case, AI supports designed for older adults will need to address not only psychological and biological/medical factors, but social and environmental factors to be most relevant.

The term "older adult" encompasses a wide range of the lifespan and includes diverse individuals from various birth cohorts; racial/ethnic, cultural and socioeconomic backgrounds; and functional abilities. As healthcare in general, and AI opportunities specifically, relies on technology, there is concern that older adults will be left out of such a digital health revolution. Even though many members of the "young-old" (65–74 years) and older cohorts may be accustomed to smart devices and other technologies, older adults are often left out of the design and marketing of such innovations (43). Sensory issues, ranging from tremors to limited vision, may also impede the use of conventional technological devices designed for users with normative abilities. When innovations are marketed toward older people, they often reflect a pathological view of aging and are limited to support for emergency monitoring (e.g., fall detection). Our call to action is that AI developers leverage a user-centered perspective, including diverse older adults with a range of health-related quality of life, during the design and evaluation (44) to uncover such technology's viability and fit-for-purpose in the target population.

## Methodological, Practical, and Other Challenges

Given the pursuit of such rapid and novel innovation, not all AI developments will readily translate to clinical or other real-world settings. While not exhaustive, we outline a few key challenges in an attempt to bridge data science with clinical science in geriatric mental health and suggest next steps in addressing such challenges.

First, there has been a paradigm shift away from traditional experimental studies that typify mental health research to rapid innovations in AI (13). The empirical approaches familiar to clinicians—namely hypothesis testing and reliance on evidence-based practice—are potentially at odds with the proof-of-concept, hypothesis-generating demonstrations that characterize much AI research to date. The innovations propelling AI forward are often tested on small samples to demonstrate proof-of-concept (40); however, this runs the risk that ML models will be overfit, leading to spurious findings and lacking generalizability to new data sources. External validation of the model (that is, testing in new datasets) is essential to improve prediction, yet only three of 51 studies in a recent review of ML in psychotherapy research did so (45). When large datasets are available, they are often prone to bias arising from differential recruitment, attrition, and engagement over time (46). Importantly, adults over the age of 60 are those least represented in digital health study samples, and such studies rarely reflect the racial/ethnic and geographical diversity of the U.S., limiting the validity of findings (46). Moreover, researchers from non-health science fields may use different reporting norms than clinician scientists, resulting in missing key pieces of information, including participant demographics and other aspects of methods (e.g., location of data collection) (40), which limit inferences and generalizability.

When it comes to implementation of AI, clinicians may override algorithm-based recommendations, or patients may be wary of algorithm-recommended treatment. Although computational modeling is a powerful tool to sift through predictors to develop complex algorithms, the "black box" of such computations may be off-putting to clinicians who have long relied on their own clinical reasoning to drive decision-making, or who may not fully understand the statistical models (47). Moreover, algorithm recommendations may not fully incorporate all clinical considerations, including patient restrictions or preferences. A major pitfall of using AI for mental health care—geriatric or otherwise—is that such systems will sometimes be wrong, resulting in patient harm. For example, a patient with a depressive disorder may be misclassified and not treated. While such error happens in human-based decision making, it will be important to build in safeguards when implementing such AI systems at scale (e.g., transparency around computational inferences and classification; routine clinician assessment to augment such AI classification for greater reliability; development of other safety nets in healthcare). Finally, even if we could use AI to accurately predict clinical state or worsening of a patient *via* sensor data or other algorithmic prediction, what would a clinic or individual clinicians actually do with such data? A clear bridge between developing and implementing such predication-based models is developing appropriate clinical workflows and interventions to address such predictions.

Data scientists must also partner with clinicians and clinical scientists to ensure that data features are meaningful and valid for older adults (16). In our own work using ML to model daily variation in depressive symptoms based on mobility data, we were unable to access raw mobility data from the proprietary sensor software and translate such data into meaningful variables (48). We also ran into issues with intra- and interindividual variation in phone usage patterns—data are only as robust as

the degree to which users use the device (3, 48), which may vary between older and younger adults. More work is needed to understand older adults as unique users of devices, such as smartphones, rather than simply extrapolating assumptions from younger users. Finally, sensors and other multimodal sources of data may detect incredible variability in clinical states and behavioral markers. However, for practical utility, AI models need to be trained to differentiate features that are clinically relevant—that is, diagnostic—from transient mood states. This will again require models based on large and diverse samples of older adults to ascertain features associated with geriatric mental health conditions.

One cannot tread into the topic of AI without running into discussion of ethics, structural inequalities, privacy, and trust issues. A full discussion of these topics is beyond the scope of this paper but has been discussed elsewhere (49–51). Briefly, these will be critical issues to consider as the innovation of data science meets the practical applications of clinical work. For example, what are the bioethical considerations if an AI algorithm recommends a particular intervention, which the clinician decides against it, and the patient decompensates? Or, conversely, where does liability lie if a patient dies after a clinician deploys an algorithm-recommended treatment (52)? It is also crucial to acknowledge that racial, gender, and ageist biases and discrimination are deeply embedded in healthcare—and as a result, in the AI systems that learn from such data sources. When unchecked, the inferences drawn from such technologies are likely to perpetuate systemic injustices in healthcare. These may result from bias and a lack of transparency in developing algorithms, such as using training data from a preponderance of young White men or using flawed proxy variables to calculate risk scores (53). Such bias is then further maintained in how providers respond to such algorithmic predictions. Thus, understanding and preventing the root causes for bias in AI systems must be a priority to monitor and mitigate such consequences. Privacy concerns among users of various technology-based assessments and interventions has also been a central theme arising in research from our group (54–56). Trust may vary as a function of who is conducting the research—for example, trust in internet-based research is higher (and participants more likely to share their data) when the research is conducted by university researchers compared to private companies (55). Building trustworthiness of AI in geriatric mental health care and research will rely on reconciling some of the issues discussed above—namely, *explainability* (the ability to understand or describe how a model arrived at its prediction), *transparency* (clear and transparent methodology), and *generalizability* (related to methodology; exhaustive testing and validation of models) (57).

## CONCLUSION

AI holds promise for more accurate diagnosis and personalized treatment recommendations, yet the field is nascent with no established pathway for integration into routine clinical care. A recent market research survey found that healthcare providers remain highly skeptical of consumer technology, remote data collection, and the integrity of such data (58). Moreover, development and implementation of such technology must incorporate clinicians, patients, and caregivers as key stakeholder groups to build trust and adopt user-centered approaches that address privacy and usability issues. We may be on the cusp of a new era that will allow the full potential of AI to take hold in mental health care broadly, and geriatrics specifically. However, until clinicians join forces with data scientists, engineers, and developers—and until such technology addresses the pragmatic concerns that clinicians and patients face—we will only scratch the surface of such potential for these technologies.

## AUTHOR CONTRIBUTIONS

BNR and AP conceptualized and designed the work. BNR, MS, and AP drafted the manuscript. All authors critically read and revised the manuscript for important intellectual content and approved the submitted version.

## FUNDING

This project was supported in part by the National Institute of Mental Health [grant P50 MH115837] and the National Institute on Aging [grant K23 AG059912]. AP's effort was supported by a grant from the Krembil Foundation, Toronto, Canada. Open access publication fees were provided by BNR's startup funds from the University of Nevada, Las Vegas. The sponsors played no role in the design of this manuscript, nor did they have any role during its execution or decision to submit.

## ACKNOWLEDGMENTS

We thank Alejandra Lopez from the University of Washington ALACRITY Center for her assistance in formatting and designing our figure.

## REFERENCES

1. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis.* (2018) 66:149–53. doi: 10.1093/cid/cix731
2. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS ONE.* (2019) 14:e0213653. doi: 10.1371/journal.pone.0213653
3. Pratap A, Atkins DC, Renn BN, Tanana MJ, Mooney SD, Anguera JA, et al. The accuracy of passive phone sensors in predicting daily mood. *Depress Anxiety.* (2019) 36:72–81. doi: 10.1002/da.22822
4. Rankin D, Black M, Flanagan B, Hughes CF, Moore A, Hoey L, et al. Identifying key predictors of cognitive dysfunction in older people using supervised machine learning techniques: observational study. *JMIR Med Inform.* (2020) 8:e20995. doi: 10.2196/20995
5. Cleret de Langavant L, Bayen E, Yaffe K. Unsupervised machine learning to identify high likelihood of dementia in population-based surveys:

development and validation study. *J Med Internet Res.* (2018) 20:e10493. doi: 10.2196/10493
6. Sau A, Bhakta I. Artificial neural network (ANN) model to predict depression among geriatric population at a Slum in Kolkata, India. *J Clin Diagn Res.* (2017) 11:VC01–4. doi: 10.7860/JCDR/2017/23656.9762
7. DeSouza DD, Robin J, Gumus M, Yeung A. Natural language processing as an emerging tool to detect late-life depression. *Front Psychiatry.* (2021) 12:719125. doi: 10.3389/fpsyt.2021.719125
8. Ng K-D, Mehdizadeh S, Iaboni A, Mansfield A, Flint A, Taati B. Measuring gait variables using computer vision to assess mobility and fall risk in older adults with dementia. *IEEE J Transl Eng Health Med.* (2020) 8:1–9. doi: 10.1109/JTEHM.2020.2998326
9. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Towards facilitating empathic conversations in online mental health support: a reinforcement learning approach. In: *Proceedings of the Web Conference 2021*. Ljubljana: ACM (2021). p. 194–205. doi: 10.1145/3442381.3450097
10. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim H-C, et al. Artificial intelligence for mental health and mental illnesses: an overview. *Curr Psychiatry Rep.* (2019) 21:116. doi: 10.1007/s11920-019-1094-0
11. Graham SA, Lee EE, Jeste DV, Van Patten R, Twamley EW, Nebeker C, et al. Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: a conceptual review. *Psychiatry Res.* (2020) 284:112732. doi: 10.1016/j.psychres.2019.112732
12. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim H-C, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biol Psychiatry Cogn Neurosci Neuroimaging.* (2021) 6:856–64. doi: 10.1016/j.bpsc.2021.02.001
13. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry.* (2021) 20:154–70. doi: 10.1002/wps.20882
14. Unützer J. Late-life depression. *N Engl J Med.* (2007) 357:2269–76. doi: 10.1056/NEJMcp073754
15. Folsom DP, Lindamer L, Montross LP, Hawthorne W, Golshan S, Hough R, et al. Diagnostic variability for schizophrenia and major depression in a large public mental health care system dataset. *Psychiatry Res.* (2006) 144:167–75. doi: 10.1016/j.psychres.2005.12.002
16. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu Rev Clin Psychol.* (2017) 13:23–47. doi: 10.1146/annurev-clinpsy-032816-044949
17. Berke EM, Choudhury T, Ali S, Rabbi M. Objective measurement of sociability and activity: mobile sensing in the community. *Ann Fam Med.* (2011) 9:344–50. doi: 10.1370/afm.1266
18. Taler V, Phillips NA. Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *J Clin Exp Neuropsychol.* (2008) 30:501–56. doi: 10.1080/13803390701550128
19. Beltrami D, Gagliardi G, Rossini Favretti R, Ghidoni E, Tamburini F, Calzà L. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Front Aging Neurosci.* (2018) 10:369. doi: 10.3389/fnagi.2018.00369
20. Gil D, Johnsson M. Diagnosing Parkinson by using artificial neural networks and support vector machines. *Glob J Comput Sci Technol.* (2009) 9:63–71.
21. Committee on the Mental Health Workforce for Geriatric Populations, Board on Health Care Services, Institute of Medicine. The mental health and substance use workforce for older adults: in whose hands? In: Eden J, Maslow K, Le M, Blazer D, editors, Washington, DC: National Academies Press (US) (2012). Available online at: http://www.ncbi.nlm.nih.gov/books/NBK201410/ (accessed June 25, 2021).
22. Brenes GA, Danhauer SC, Lyles MF, Hogan PE, Miller ME. Barriers to mental health treatment in rural older adults. *Am J Geriatr Psychiatry.* (2015) 23:1172–8. doi: 10.1016/j.jagp.2015.06.002
23. Jimenez DE, Bartels SJ, Cardenas V, Dhaliwal SS, Alegría M. Cultural beliefs and mental health treatment preferences of ethnically diverse older adult consumers in primary care. *Am J Geriatr Psychiatry.* (2012) 20:533–42. doi: 10.1097/JGP.0b013e318227f876
24. Mojtabai R, Olfson M, Sampson NA, Jin R, Druss B, Wang PS, et al. Barriers to mental health treatment: results from the National Comorbidity Survey Replication. *Psychol Med.* (2011) 41:1751–61. doi: 10.1017/S0033291710002291
25. Areán PA, Renn BN, Ratzliff A. Making psychotherapy available in the United States: implementation challenges and solutions. *Psychiatr Serv.* (2021) 72:222–4. doi: 10.1176/appi.ps.202000220
26. Saria S, Goldenberg A. Subtyping: what it is and its role in precision medicine. *IEEE Intell Syst.* (2015) 30:70–5. doi: 10.1109/MIS.2015.60
27. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med Publ Soc Behav Med.* (2017) 52:446–62. doi: 10.1007/s12160-016-9830-8
28. Müller AM, Blandford A, Yardley L. The conceptualization of a Just-In-Time Adaptive Intervention (JITAI) for the reduction of sedentary behavior in older adults. *mHealth.* (2017) 3:37–37. doi: 10.21037/mhealth.2017.08.05
29. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J.* (2019) 6:94–8. doi: 10.7861/futurehosp.6-2-94
30. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR MHealth UHealth.* (2018) 6:e12106. doi: 10.2196/12106
31. Euronews. *Smarter Than the Average Home—Technology and Assisted Living*. euronews (2020). Available online at: https://www.euronews.com/next/2020/09/25/smarter-than-the-average-home-how-technology-is-aiding-assisted-living (accessed June 25, 2021).
32. Trajkova M, Martin-Hammond A. "Alexa is a toy": exploring older adults' reasons for using, limiting, and abandoning echo. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery (2020). p. 1–13. doi: 10.1145/3313831.3376760
33. Luxton DD. Recommendations for the ethical use and design of artificial intelligent care providers. *Artif Intell Med.* (2014) 62:1–10. doi: 10.1016/j.artmed.2014.06.004
34. Bindoff I, Stafford A, Peterson G, Kang BH, Tenni P. The potential for intelligent decision support systems to improve the quality and consistency of medication reviews. *J Clin Pharm Ther.* (2012) 37:452–8. doi: 10.1111/j.1365-2710.2011.01327.x
35. Bond WF, Lynch TJ, Mischler MJ, Fish JL, McGarvey JS, Taylor JT, et al. Virtual standardized patient simulation: case development and pilot application to high-value care. *Simul Healthc.* (2019) 14:241–50. doi: 10.1097/SIH.0000000000000373
36. Goldberg SB, Flemotomos N, Martinez VR, Tanana MJ, Kuo PB, Pace BT, et al. Machine learning and natural language processing in psychotherapy research: alliance as example use case. *J Couns Psychol.* (2020) 67:438–48. doi: 10.1037/cou0000382
37. Bor JS. Among the elderly, many mental illnesses go undiagnosed. *Health Aff.* (2015) 34:727–31. doi: 10.1377/hlthaff.2015.0314
38. Gallo JJ, Rabins PV, Anthony JC. Sadness in older persons: 13-year follow-up of a community sample in Baltimore, Maryland. *Psychol Med.* (1999) 29:341–50. doi: 10.1017/S0033291798008083
39. Knight BG, Winterbotham S. Rural and urban older adults' perceptions of mental health services accessibility. *Aging Ment Health.* (2020) 24:978–84. doi: 10.1080/13607863.2019.1576159
40. Xie B, Tao C, Li J, Hilsabeck RC, Aguirre A. Artificial intelligence for caregivers of persons with Alzheimer's disease and related dementias: systematic literature review. *JMIR Med Inform.* (2020) 8:e18189. doi: 10.2196/18189
41. National Academies of Sciences, Engineering, and Medicine. *Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System*. Washington, DC: National Academies Press (2020). Available online at: https://www.nap.edu/catalog/25663 (accessed May 16, 2020).
42. Badal VD, Graham SA, Depp CA, Shinkawa K, Yamada Y, Palinkas LA, et al. Prediction of loneliness in older adults using natural language processing: exploring sex differences in speech. *Am J Geriatr Psychiatry.* (2021) 29:853–66. doi: 10.1016/j.jagp.2020.09.009
43. Seifert A, Cotten SR, Xie BA. Double burden of exclusion? Digital and social exclusion of older adults in times of COVID-19. *J Gerontol Ser B.* (2021) 76:e99–103. doi: 10.1093/geronb/gbaa098

44. Kolovson S, Pratap A, Duffy J, Allred R, Munson SA, Areán PA. Understanding participant needs for engagement and attitudes towards passive sensing in remote digital health studies. In: *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*. Atlanta, GA: ACM (2020). p. 347–62. doi: 10.1145/3421937.3422025
45. Aafjes-van Doorn K, Kamsteeg C, Bate J, Aafjes M. A scoping review of machine learning in psychotherapy research. *Psychother Res.* (2021) 31:92–116. doi: 10.1080/10503307.2020.1808729
46. Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. *Npj Digit Med.* (2020) 3:21. doi: 10.1038/s41746-020-0224-8
47. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. *J Am Med Assoc.* (2007) 298:1010–22. doi: 10.1001/jama.298.9.1010
48. Renn BN, Pratap A, Atkins DC, Mooney SD, Areán PA. Smartphone-based passive assessment of mobility in depression: challenges and opportunities. *Ment Health Phys Act.* (2018) 14:136–9. doi: 10.1016/j.mhpa.2018.04.003
49. Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell.* (2019) 1:501–7. doi: 10.1038/s42256-019-0114-4
50. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. *Annu Rev Public Health.* (2018) 39:95–112. doi: 10.1146/annurev-publhealth-040617-014208
51. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *Npj Digit Med.* (2020) 3:81. doi: 10.1038/s41746-020-0288-5
52. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *J Am Med Assoc.* (2019) 322:1765. doi: 10.1001/jama.2019.15064
53. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* (2019) 366:447–53. doi: 10.1126/science.aax2342
54. Areán PA, Pratap A, Hsin H, Huppert TK, Hendricks KE, Heagerty PJ, et al. Perceived utility and characterization of personal google search histories to detect data patterns proximal to a suicide attempt in individuals who previously attempted suicide: pilot cohort study. *J Med Internet Res.* (2021) 23:e27918. doi: 10.2196/27918
55. Pratap A, Allred R, Duffy J, Rivera D, Lee HS, Renn BN, et al. Contemporary views of research participant willingness to participate and share digital data in biomedical research. *J Am Med Assoc Netw Open.* (2019) 2:e1915717. doi: 10.1001/jamanetworkopen.2019.15717
56. Renn BN, Hoeft TJ, Lee HS, Bauer AM, Areán PA. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the US. *Npj Digit Med.* (2019) 2:6. doi: 10.1038/s41746-019-0077-1
57. Chandler C, Foltz PW, Elvevåg B. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophr Bull.* (2019) 2019:sbz105. doi: 10.1093/schbul/sbz105
58. Shinkman R. *Survey Casts Doubt on Utility of Wearable Devices in Healthcare*. Healthcare Dive (2021). Available online at: https://www.healthcaredive.com/news/survey-casts-doubt-on-utility-of-wearable-devices-in-healthcare/598846/ (accessed May 10, 2021).


**Conflict of Interest:** BR receives unrelated research support from Sanvello.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 16 November 2021
doi: 10.3389/fpsyt.2021.728732

# Do Words Matter? Detecting Social Isolation and Loneliness in Older Adults Using Natural Language Processing

*Varsha D. Badal*[1,2], *Camille Nebeker*[3], *Kaoru Shinkawa*[4], *Yasunori Yamada*[4], *Kelly E. Rentscher*[5], *Ho-Cheol Kim*[6] *and Ellen E. Lee*[1,2,7]*

[1] Department of Psychiatry, University of California San Diego, La Jolla, CA, United States, [2] Sam and Rose Stein Institute for Research on Aging, University of California San Diego, La Jolla, CA, United States, [3] Herbert Wertheim School of Public Health and Longevity Science, University of California San Diego, La Jolla, CA, United States, [4] Digital Health, IBM Research-Tokyo, Tokyo, Japan, [5] Cousins Center for Psychoneuroimmunology, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, CA, United States, [6] AI and Cognitive Software, IBM Research-Almaden, San Jose, CA, United States, [7] VA San Diego Healthcare System, La Jolla, CA, United States

OPEN ACCESS

**Edited by:**
Ruth Asch,
Yale University, United States

**Reviewed by:**
Sylvester Orimaye,
East Tennessee State University, United States
Migita Michael D'Cruz,
National Institute of Mental Health and Neurosciences (NIMHANS), India
André Carlos Ponce De Leon Ferreira De Carvalho,
University of São Paulo, Brazil
Lars Meyer,
Max-Planck-Gesellschaft (MPG), Germany

***Correspondence:**
Ellen E. Lee
eel013@health.ucsd.edu

**Specialty section:**
This article was submitted to Aging Psychiatry, a section of the journal Frontiers in Psychiatry

**Received:** 22 June 2021
**Accepted:** 08 October 2021
**Published:** 16 November 2021

**Citation:**
Badal VD, Nebeker C, Shinkawa K, Yamada Y, Rentscher KE, Kim H-C and Lee EE (2021) Do Words Matter? Detecting Social Isolation and Loneliness in Older Adults Using Natural Language Processing. Front. Psychiatry 12:728732. doi: 10.3389/fpsyt.2021.728732

**Introduction:** Social isolation and loneliness (SI/L) are growing problems with serious health implications for older adults, especially in light of the COVID-19 pandemic. We examined transcripts from semi-structured interviews with 97 older adults (mean age 83 years) to identify linguistic features of SI/L.

**Methods:** Natural Language Processing (NLP) methods were used to identify relevant interview segments (responses to specific questions), extract the type and number of social contacts and linguistic features such as sentiment, parts-of-speech, and syntactic complexity. We examined: (1) associations of NLP-derived assessments of social relationships and linguistic features with validated self-report assessments of social support and loneliness; and (2) important linguistic features for detecting individuals with higher level of SI/L by using machine learning (ML) models.

**Results:** NLP-derived assessments of social relationships were associated with self-reported assessments of social support and loneliness, though these associations were stronger in women than in men. Usage of first-person plural pronouns was negatively associated with loneliness in women and positively associated with emotional support in men. ML analysis using leave-one-out methodology showed good performance (F1 = 0.73, AUC = 0.75, specificity = 0.76, and sensitivity = 0.69) of the binary classification models in detecting individuals with higher level of SI/L. Comparable performance were also observed when classifying social and emotional support measures. Using ML models, we identified several linguistic features (including use of first-person plural pronouns, sentiment, sentence complexity, and sentence similarity) that most strongly predicted scores on scales for loneliness and social support.

**Discussion:** Linguistic data can provide unique insights into SI/L among older adults beyond scale-based assessments, though there are consistent gender differences. Future research studies that incorporate diverse linguistic features as well as other

behavioral data-streams may be better able to capture the complexity of social functioning in older adults and identification of target subpopulations for future interventions. Given the novelty, use of NLP should include prospective consideration of bias, fairness, accountability, and related ethical and social implications.

**Keywords: artificial intelligence, social connectedness, gender, loneliness, NLP, Social support, linguistic features**

## INTRODUCTION

"*No man is an island entire of itself; every man is a piece of the continent, a part of the main…* "—John Donne.

Rates of social isolation and loneliness (SI/L) have increased over the past few decades among older adults, impacting both mental and physical health (1, 2). SI/L is associated with increased alcohol and drug abuse (3), cognitive decline (4), development of depressive and anxiety symptoms (5, 6), poor physical functioning (7–9), as well as increased mortality (10, 11). Furthermore, the adoption of physical distancing guidelines during the COVID-19 pandemic has further isolated seniors from relationships and meaningful activities, impacting health and well-being (12).

While current studies rely on self-report measures of social interactions and subjective experiences to assess SI/L, these approaches may not fully capture the nature or quality of the social connections. Our previous work has used Natural Language Processing (NLP) approaches to identify subtle speech-based linguistic features that reflect loneliness in older adults. We found strong gender differences in the acknowledgment of loneliness and expressed sentiment among older adults (13). These findings provided foundational support that unstructured text data can provide unique insights into internal subjective experiences, including for the detection and understanding of SI/L. Building upon our previous work, the current study examined how older men and women describe relationships and social supports during a semi-structured interview. This NLP analysis was applied to interview segments that focused on social relationships (where loneliness was not specifically mentioned), successful aging, and technology.

We hypothesize that linguistic features may be reflective of SI/L due to the social nature of language, which often reflects how an individual relates to others. Linguistic data may provide a novel data source for understanding and assessing SI/L and may be particularly useful to social media companies, as such data is widely accessible unlike clinical and psychological measures including depression. For example, an individual who is lonely may have higher usage of first-person singular pronouns ("I") than that of first-person plural pronouns ("we"). This may reflect a lack of social contacts, close family members, or significant others, as well as signal a lack of closeness or commonality with social contacts. Similarly, an individual who is lonely may feel distressed over a lack of social relationships and would use more negative language to describe them to an interviewer. In addition to pronoun usage, we also explored use of other parts-of-speech and syntactic complexity in relationship to SI/L. While few NLP studies have explored this previously, there have been links between socioeconomic status and education with loneliness—which could be indirectly reflected by syntactic complexity (longer and more complex sentence structures) or diversity of language used.

In this proof-of-concept study, we explored the interplay of gender and SI/L on transcribed speech data, using validated self-report scales for SI/L and NLP techniques, to provide a qualitative assessment of relationships. We examined the association between a scale-based measure of social support with the number and type of described relationships. We also examined how textual features, in particular pronoun usage, reflected self-report ratings of SI/L. Last we created machine learning (ML) models to predict SI/L based on sociodemographic and linguistic -based features, comparing the top-ranking features for different aspects of social support and loneliness.

## RESEARCH DESIGN AND METHODS

### Participants and Procedures

For this study, we utilized data collected *via* interviews with residents living independently at a continuing care senior housing community (CCSHC) in southern California. Cohort characteristics and study procedures have been previously published (14, 15). The study was approved by the University of California San Diego Human Research Protections Program (HRPP). Study inclusion criteria were: (1) English speaking individuals 65+ years old, (2) Ability to complete study assessments and engage in a qualitative interview, and (3) No known diagnosis of dementia or any other disabling illness.

### Sociodemographic and Clinical Measures

Sociodemographic data including age, gender, racial background, years of education and marital status were collected along with scales to measure depression (Patient Health Questionnaire, 9-item) (16) and anxiety (Brief Symptom Inventory—Anxiety subscale).

### Measures of Social Functioning

Social support was assessed using scales from the MacArthur Studies of Successful Aging that included measures of Emotional Support (6-item scale, including "How often does your spouse make you feel loved and cared for?"), Instrumental Support (6-item scale, including "How often does your spouse help with daily tasks like shopping, giving you a ride, or helping with household tasks?") and Negative aspects of Social Relationships (6-item scale, including "How often does your spouse make too many demands on you?") (17).

Loneliness was assessed with the UCLA Loneliness scale (Version 3) or UCLA-3, a validated and commonly used research instrument. The UCLA-3 has high internal consistency, validity, and test-retest reliability (18). Unlike single-item assessments of loneliness, the UCLA-3 does not explicitly use the word "lonely." The 20 items inquire about specific experiences, e.g., "How often do you feel in tune with others around you?" using a 4-point Likert scale (1 = "I never feel this way" to 4 = "I often feel this way"). The cut-offs for loneliness severity on the UCLA-3 scale were adapted from Doryab et al. (19), such that total scores $\leq 40$ are categorized as not lonely and total scores $>40$ are categorized as lonely. Q2 (What makes those meaningful to you?) was included in extraction of linguistic features. However, due to the open-ended scope of the question and lack of concrete or objective information for further analyses, we only included linguistic features from those responses. Some of the commonly used social scales are henceforth referred to using acronyms: ESS-E, Emotional Support Scale—Emotional Support score; ESS-I, Emotional Support Scale—Instrumental Support; ESS-NI, Emotional Support Scale—Negative Interaction Score; SSI, Social Support Index.

## Qualitative Interviews

In addition to the aforementioned data collection, semi-structured interviews were conducted with participants covering a variety of topics (loneliness, relationships, and wisdom). Interviews were conducted by research staff trained in qualitative methods (Patton 2002) and occurred between April 2018 and August 2019. The interview protocol included six questions on the topic of relationships: (Q1) "So, this first section is about family, friendships and relationships. Do you have important relationships in your life? Please describe them." (Q2) "What makes those relationships meaningful to you?" (Q3) "Do you feel that there are people in your life who fully understand you?" (Q4) "How often do you spend time with or connect (*via* phone, email, or social media) with others?" (Q5) "Do you feel you are part of a larger community? Please explain." (Q6) "When you are feeling disconnected or isolated what do you do?" Each interview was audio-taped and subsequently transcribed by a commercial company (MModal). The interviews were manually transcribed verbatim and distinguished between the interviewer and interviewee. The same interviewer conducted all the interviews. This study focused upon Q1 responses to extract the number of important relationships, Q3 responses to extract the number of relationships in which one felt understood, and Q4 responses to extract frequency and mode of communication. The relationship section of the interview was used to extract linguistic features since these questions were consistent between the self-reported lonely and not-lonely, whereas for pronoun usage, we used the entire interview text in addition to the relationship section, given that focusing the conversation upon relationships could bias the pronoun usage (e.g., increased use overall of pronouns to describe their social network).

## Analytic Procedures

NLP techniques allow us to isolate relevant pieces of information within a response and suitably encode the information into numerical values or "features." Some of these features are derived from the entire transcript, while others are derived from responses to specific questions or an entire thematic section. Many of these features are present in varying strengths, commonly referred to as "impurity" levels in NLP analysis, among classes based on user-defined criteria (e.g., gender, loneliness levels). This impurity of features (probability of incorrectly classifying) is exploited by ML techniques to discriminate among the classes even if the impurity is not significant, or the association is non-linear, or if several features must be composed together for the ML analyses. The following subsections discuss the steps involved and implementation details.

## Text Processing to Localize Responses

Term Frequency—Inverse Document Frequency (TF-IDF) techniques (20, 21) were used to identify specific questions and subsequent responses. These TF-IDF techniques are commonly used in document retrieval and data mining approaches (22). Briefly, within this method, the transcript of the interview is akin to a "corpus," the entirety of text to be searched. Each question in the actual interview is analogous to a "document," which must be matched (and its location retrieved) to a template question of interest, or a "query."

Matching the query with the document uses vector algebra. First, the corpus (or collection of documents) is converted into vectors to capture the frequency (TF component) and the uniqueness (IDF component) of words (henceforth referred to as "terms"). Next, the queries are also vectorized. Finally, the query vectors can be matched with document vectors (using cosine-similarity) to identify best matches. The procedure is repeated for each transcript.

The transcribed interviews identified the interviewer's utterances with a new line preceded by the character "Q," while the interviewee's answers were preceded by the character "A." TF-IDF implementation queries were used with the actual questions in the transcripts. The TF-IDF approaches allowed text to be identified within each transcript that best matched the template query. After identifying the location of the question, we extracted the subsequent response (several lines following the "A" in the transcribed interview text).

## Text Processing to Extract Information

### Linguistic Features

Linguistic features include frequency and ratio of parts of speech, vocabulary richness (Brunét's index, Honore's statistic, type token ratio), filled pauses (dysfluency in speech), syntactic complexity (complex and compound phrase structure within a sentence), sentence similarity (similarity between all pairs of sentences), and sentiment (23). For sentiment analysis, we used VADER (Valence Aware Dictionary for Sentiment Reasoning) a highly regarded and freely available tool. VADER is sensitive to polarity (positive/negative) as well as the strength of conveyed emotions. VADER is based on a dictionary which maps words into sentiment values (covering the positive to negative range), and also rates text based on capitalization and punctuation. VADER is ranked as one of the best in a 2016 benchmark study

of commonly used sentiment analyzers (24). Once the location of relevant text in the transcripts was identified, a variety of techniques were used to quantify the represented information. As previously mentioned, all linguistic features, aside from pronoun usage, were extracted from only the relationship section of the interview. Specific details on these features are available in the **Supplementary Appendix A.**

### Pronoun Usage-Based Features

We computed the density of first-person singular (I, me, my, and mine), first-person plural (we, our, us, and ours) and the third-person plural (he, she, they, them, and their) pronouns, but excluded the second-person pronouns (you, your, and yours) because they were primarily used to address the interviewer in the transcripts. Although these features are also linguistic in nature, they are mentioned in a separate category due to the nuanced semantics conveyed about relationships with others. Due to the focus on relationships with others in the relationship section, the section had higher pronoun usage and effect sizes were small to very small ($<0.20$). We used the transcript from the entire interview for pronoun-related analyses, which provided higher discrimination.

### Relationship Word-Based Features

A dictionary of words was manually created to identify relationships mentioned by participants in their responses. These relationship words were further mapped into categories, e.g., "husband" and "wife" were categorized into "spouse." **Supplementary Table 1** outlines the mapping between relationship words and assigned categories. We also created a dictionary of predefined phrases that are often used in American English to identify modes and frequency of communication. To assess communication frequency, the phrases were mapped to approximate frequency as shown in **Supplementary Table 2**.

## ML Classification

Socio-demographic features (education, age, race, marital status, etc.), linguistic features, and all pronoun density features ($N = 97$) were used to classify participants into objective categories for loneliness (UCLA-3 severity, cutoff score of 40) and social support (median cutoff) using Artificial neural network (ANN) with 200 internal units in Orange version 3.27.1, scikit-learn version 0.24.2. (25) was used. Various ML models such as Artificial neural network (ANN) with activation functions (Logistic, ReLu, and tanh), support vector machine (SVM), k-nearest neighbors (kNN), Tree and random forest were used (25). **Figure 1** depicts the overall procedure along with features and sources used for our processing. Performances of binary-classification models were evaluated by using F1 score and the area under receiver operating characteristic curve (AUC) with leave-one-subject-out cross validation.

## Feature Ranking

Classifiers usually benefit from a large feature set, however, as the size of the feature set grows, at some point, error rates begin to increase (26). This phenomenon becomes even more relevant as the size of the feature set becomes comparable to the sample size, as in our case. There is a strong possibility of overfitting, and many features may be a source of noise. The approach usually (27, 28) is to rank features and then use top features incrementally to find the best performing set. This usually results in improved performance.

To determine the top-ranking NLP features that contribute to SI/L, we assessed how differently the feature is distributed across classes (previously referred to as impurity). GINI is a popular impurity-based feature ranking technique (29) that states the probability that the feature is wrongly classified (0 = "pure," 0.5 = equal distribution across all classes, 1 = random distribution across classes) (29, 30). GINI was used to rank the features that were most strongly associated with the SI/L classification.

## A Caveat on Anaphora and Overestimation

Using NLP to extract information about relationships from unstructured text has a few notable challenges. For example, a response may mention "I have children. A son and a daughter." Such responses require establishing correspondences between nouns (and pronouns), possibly separated by long spans of text. Anaphoric resolution (establishing correspondence among nouns and pronouns that refer to the same entity within and across sentences) is difficult, hence we acknowledge the possibility of overestimation in this process (31, 32). Our analysis relies upon counting words from our dictionary of relationship terms. Thus, our NLP-guided count of relationships may overestimate the intended number of relationships in the response due to possible anaphoric references.

# RESULTS

Of the 101 interviews, 97 participants also completed other baseline assessments and were included in the analyses for this study.

## Description of the Study Sample

Participants ranged between 66 and 94 years of age (**Table 1**). Men were older (Mean age = 86.2 vs. 81.7 years for women, Cohen's $d = -0.68$) and had more years of education (Cohen's $d = -0.40$) than women. Racial background, marital status, mean UCLA-3 scores, instrumental support, negative interactions, anxiety, and depression scores were similar by gender.

## Comparison of Self-Report and NLP-Based Measures of Social Support

The location of responses corresponding to Q1 and Q3 in the transcripts were identified correctly for all 97 interviewees, and more than 97% of responses were captured for the analyses. **Figures 2A,B** show the relationship type and distribution of important relationship terms by gender, in response to Q1. The identified relations were mapped into relationship categories. Children were most commonly reported as important relationships (63.5% women, average 1.5 mentions per interviewee for women overall, 52.9% men averaged 1.76 mentions per interviewee overall for men), followed by siblings and spouses. **Figures 2C,D** show, by gender, the relationship type and distribution of relationships in which the participant

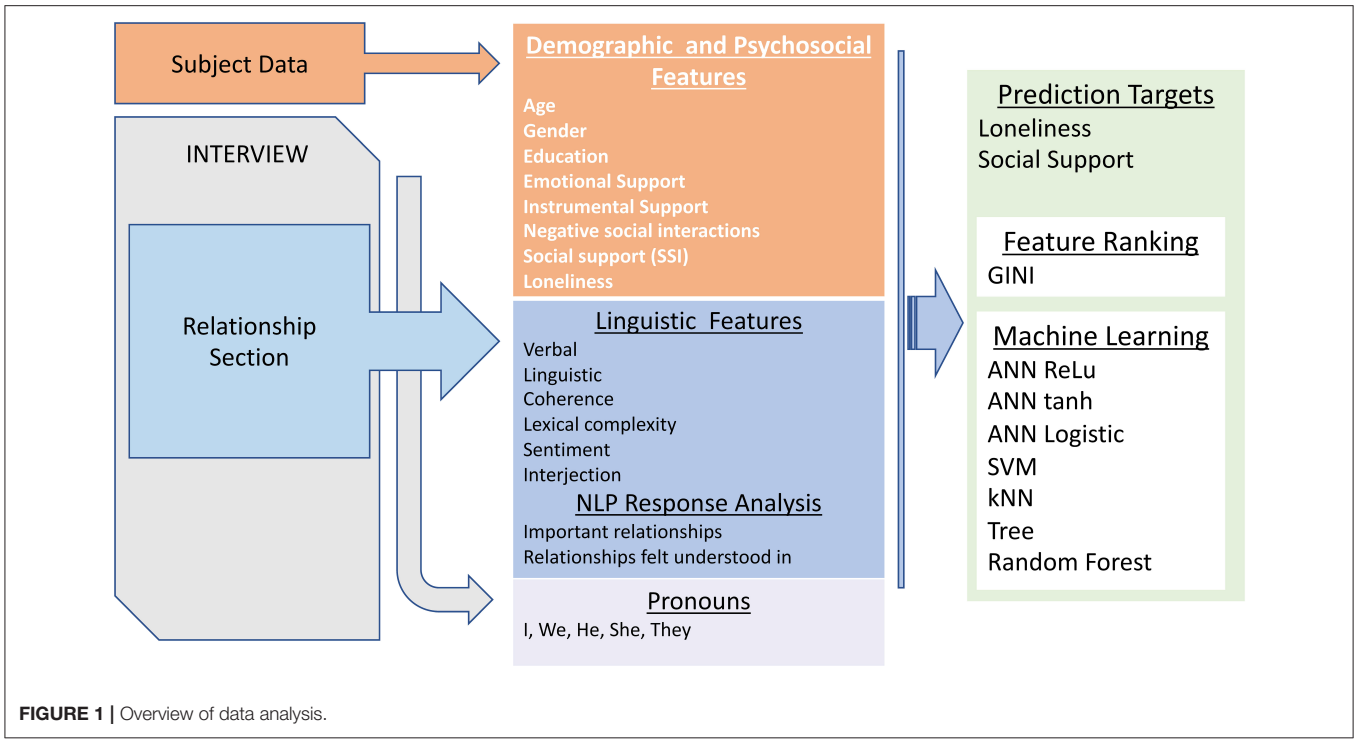


**FIGURE 1 |** Overview of data analysis.

**TABLE 1 |** Demographics information.

| | Women | | | | Men | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N* | MEAN | MEDIAN | SD | N* | MEAN | MEDIAN | SD | t or $\chi^2$ | df | p | Cohen's d |
| Age at visit (years) | 63 | 81.7 | 81.5 | 6.94 | 34 | 86.2 | 86.5 | 5.90 | −3.36 | 96 | <0.001 | −0.68 |
| Education (years) | 63 | 15.4 | 16.0 | 2.42 | 34 | 16.4 | 16.0 | 2.23 | −1.95 | 96 | 0.06 | −0.40 |
| Race (% Caucasian) | 63 | 90.5 | | | 34 | 94.1 | | | 0.06 | 1 | 0.81 | |
| Marital Status (% not single) | 63 | 34.9 | | | 34 | 52.9 | | | 2.26 | 1 | 0.13 | |
| Loneliness (UCLA-3 score) | 54 | 36.2 | 35.0 | 9.35 | 30 | 39.3 | 38.5 | 11.54 | −1.24 | 83 | 0.22 | −0.30 |
| Emotional support (ESS-E) | 60 | 2.8 | 3.0 | 0.41 | 33 | 2.6 | 2.5 | 0.47 | 1.97 | 92 | 0.05 | 0.44 |
| Instrumental support (ESS-I) | 60 | 2.0 | 2.0 | 0.85 | 33 | 1.9 | 2.0 | 0.79 | 0.59 | 92 | 0.55 | 0.13 |
| Negative social interactions (ESS-NI) | 60 | 0.7 | 0.5 | 0.72 | 33 | 0.8 | 0.5 | 0.65 | −0.98 | 92 | 0.33 | −0.21 |
| Social support (SSI) | 53 | 52.0 | 52.0 | 7.41 | 31 | 49.6 | 50.0 | 7.24 | 1.41 | 83 | 0.16 | 0.32 |
| PHQ-9 | 57 | 2.7 | 2.0 | 3.55 | 31 | 3.5 | 2.0 | 3.89 | −0.91 | 87 | 0.37 | −0.21 |

*N refers to number of available observations at baseline. Some information was incomplete (unavailable).*
*ESS-E, Emotional Support Scale—Emotional Support score; ESS-I, Emotional Support Scale—Instrumental Support; ESS-NI, Emotional Support Scale—Negative Interaction Score; PHQ-9, Patient Health Questionnaire 9-item; SSI, Social Support Index; UCLA-3, UCLA Loneliness Scale (Version 3).*

feels understood. In terms of feeling understood, participants most commonly noted children, spouses, and parents. Women and men reported similar numbers of important relationships (Mann–Whitney $U = 993.0$, $p = 0.18$, Cohen's $d = -0.084$) and relationships in which they felt understood (Mann–Whitney $U = 989.5$, $p = 0.26$, Cohen's $d = -0.085$). A sizable fraction of men (35.2%) and women (46.0%) reported they were not understood by anyone.

Women communicated with their social network more frequently than men based upon key phrases in response to Q4 mapped to frequency (23.5 times a month vs. 8.0 times a month, Mann–Whitney $U = 131.5$ $p < 0.001$, Cohen's $d = 0.76$; **Figure 3A**). The most frequently mentioned mode of communication was phone ($N = 26$), followed by email ($N = 22$), and social media ($N = 17$), which included Facebook and Instagram **(Figure 3B)**.

Emotional and instrumental support were associated with the NLP-derived assessments of relationships. The number of important relationships was correlated with greater emotional support in women, but not men (Spearman's $\rho = 0.28$, $p = 0.03$ and Spearman's $\rho = -0.06$, $p = 0.73$, respectively). Furthermore, the number of important relationships was negatively correlated

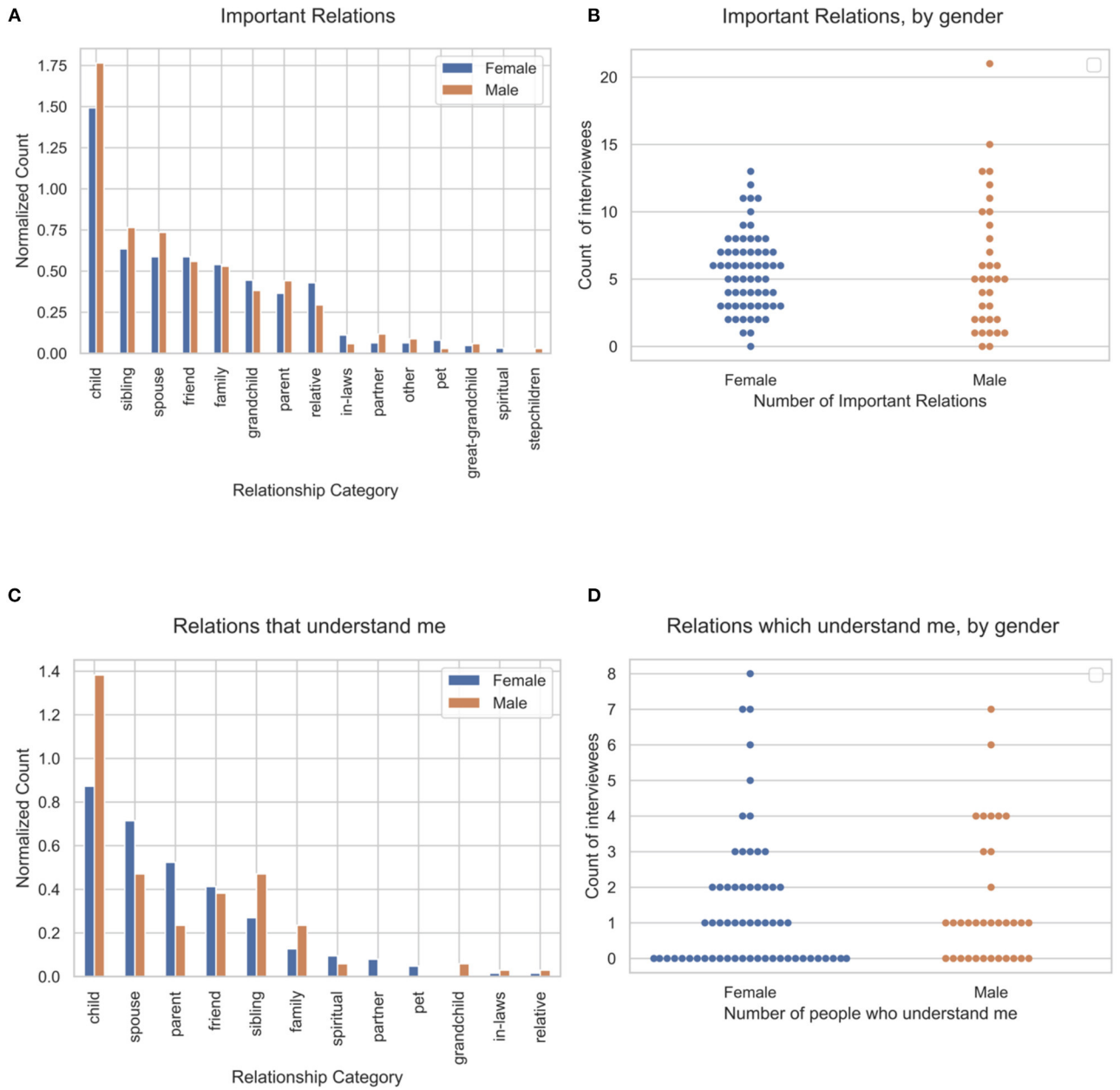


**FIGURE 2 |** Number and type of important relationships and relationships in which one feels understood. **(A)** Relationship category, **(B)** no significant difference in the distribution of the counts by gender, **(C)** relations that understand in response to Q3, and **(D)** no significant difference in the distribution of the counts by gender.

with negative social interactions in women, but not in men (Spearman's $\rho = -0.34$, $p = 0.009$ and Spearman's $\rho = 0.11$, $p = 0.55$, respectively). The numbers of important relationships were not significantly correlated with UCLA-3 loneliness scale scores (Spearman's $\rho = -0.15$, $p = 0.16$) in either gender.

## Text Features Related to SI/L: Pronoun Usage

The density of types of pronouns, computed as a ratio of their occurrence counts divided by total number of words uncovered several interesting associations. First-person plural pronoun usage negatively correlated with loneliness in women ($\rho = -0.31$, $p = 0.025$). Emotional support in women was directly related to third-person pronoun density (Spearman's $\rho = 0.30$, $p = 0.008$).

## Binary Classification Models and GINI Based Feature Ranking

The ANN using Logistic activation function outperformed the others (F1 = 0.73, AUC = 0.75, specificity = 0.76, and sensitivity = 0.69) in predicting loneliness. This approach also performed

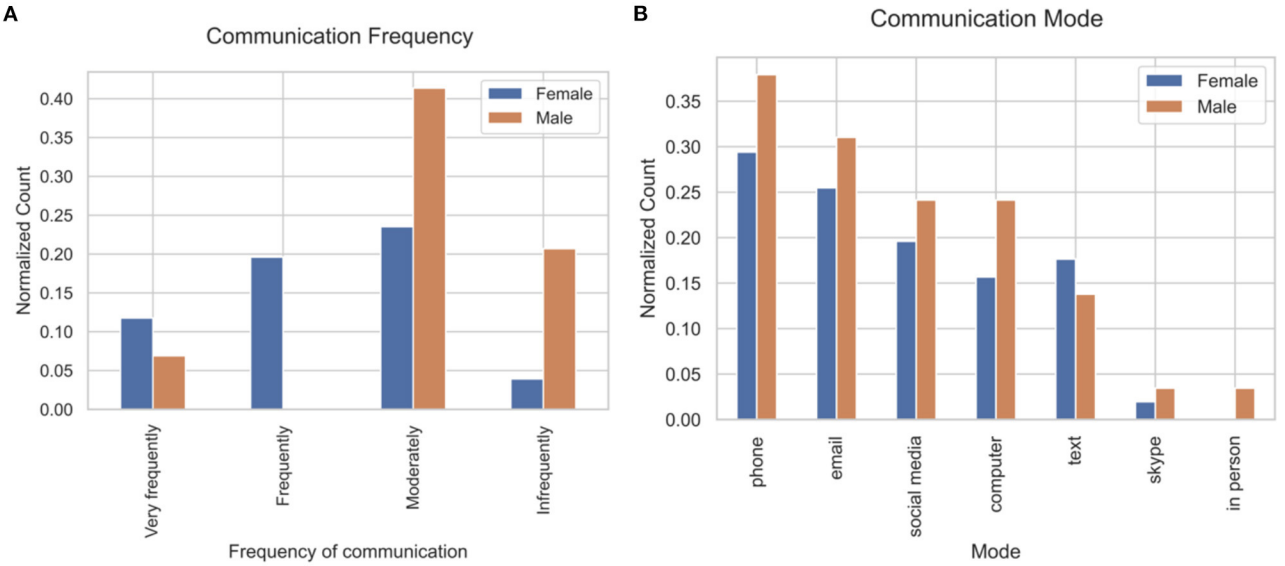


**FIGURE 3 |** Frequency and mode of communication with social network. **(A)** Communication frequency and **(B)** communication mode. Details are shown in **Supplementary Table 2**.

**TABLE 2 |** Binary-classification performance with loneliness (Leave one out)*$.

| | | | | Confusion matrix | | | |
|---|---|---|---|---|---|---|---|
| **Models** | **Hyper parameters** | **AUC** | **F1 Score** | **TP** | **FP** | **TN** | **FN** |
| ANN Logistic | Scipy implementation, Number of hidden layers = 1, Number of neurons in hidden layer = 200, solver = Adam | 0.75 | 0.73 | 24 | 12 | 37 | 11 |
| SVM | Cost (C) = 1.00, Numerical Tolerance = 0.001, Epsilon = 0.10, g = auto, kernel = RBF | 0.74 | 0.73 | 23 | 11 | 38 | 12 |
| ANN tanh | Scipy implementation, Number of hidden layers = 1, Number of neurons in hidden layer = 200, solver = Adam | 0.67 | 0.65 | 20 | 14 | 35 | 15 |
| ANN ReLu | Scipy implementation, Number of hidden layers = 1, Number of neurons in hidden layer = 100, solver = Adam | 0.70 | 0.64 | 21 | 16 | 33 | 14 |
| Tree | Max depth = 100, Min number of instance in leaves = 1 | 0.59 | 0.57 | 19 | 20 | 29 | 16 |
| Random Forest | Number of Trees = 8, Number of attributes for split = 4, Limit depth = 7, Don't split subsets smaller than 2 | 0.57 | 0.55 | 12 | 14 | 35 | 23 |
| kNN | Number of neighbors k = 9, Metric = Chebyshev, weight = distance | 0.54 | 0.54 | 16 | 20 | 29 | 19 |

*Features comprising socio-demographic features, language features, and pronoun features. $top 10 features.*
*ANN, artificial neural network; SVM, support vector machine; kNN, k-nearest neighbors algorithm; TP, true positive; FP, false positive; TN, true negative; FN, false negative.*

similar to our previous approach for quantitative loneliness (13). Performance for ML for various measures such as loneliness and social support are shown in **Tables 2, 3**. Social support classification using ML showed acceptable performance (ESSES: F1 score = 0.67, AUC = 0.72; ESSIS: F1 score = 0.66, AUC = 0.71; ESSNS: F1 score = 0.67, AUC = 0.62) for median split.

We used GINI to rank features for the classification task, the top 10 features results are shown in **Table 4**. Description of specific features is categorized and grouped in **Supplementary Table 3**. Several of the top-ranked features were consistently related to loneliness and social support. Lower usage of first-person plural pronouns was linked to higher loneliness, while higher usage of pronouns in general was associated with better emotional and instrumental support, as well as with fewer negative social interactions. Similarly, greater sentence similarity was associated with lower instrumental support while lower sentence complexity was associated with higher loneliness and lower emotional support. Shorter response length in the relationship section was associated with higher loneliness, while shorter responses throughout the interview were associated with lower emotional and instrumental support. Higher education levels were linked to greater loneliness. Lower positive sentiment and higher negative sentiment were consistently linked to less emotional support, less instrumental support, and more negative social interactions.

Feature rankings suggest greater role of age than gender in SI/L, with relative information gains of 0.02 vs. 0.01.

## DISCUSSION

Our study explored how text features were associated with SI/L in older community-dwelling adults. Older women's responses to questions about important relationships were more strongly correlated with their ratings on social support scales than older men's. Pronoun density was associated with loneliness and social support in both men and women and were consistently a

**TABLE 3 |** Binary-classification performance for social and emotional support at median and 75 percentile (Leave one out)*.

| Target | Cutoff | AUC | F1 Score | Top model | Number of features included* | Confusion matrix | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | TP | FP | TN | FN |
| A:Binary-classification performance for social and emotional support at median (Leave one out)*. | | | | | | | | | |
| Emotional support | 3.0 | 0.72 | 0.67 | ANN ReLu | 60 | 44 | 18 | 19 | 12 |
| Instrumental support | 2.0 | 0.71 | 0.66 | ANN tanh | 20 | 42 | 18 | 20 | 13 |
| Negative social interactions | 0.5 | 0.62 | 0.67 | Tree | 30 | 51 | 17 | 12 | 13 |
| B:Binary-classification performance for social and emotional support at 75 percentiles (Leave one out)*. | | | | | | | | | |
| Emotional support | 3.0 | 0.72 | 0.67 | ANN ReLu | 60 | 44 | 18 | 19 | 12 |
| Instrumental support | 2.5 | 0.63 | 0.63 | ANN ReLu | 5 | 15 | 15 | 44 | 19 |
| Negative social interactions | 1.0 | 0.63 | 0.71 | SVM | 15 | 20 | 8 | 47 | 18 |

*Features comprising socio-demographic features, linguistic features, and pronoun features.*
*TP, true positive; FP, false positive; TN, true negative; FN, false negative.*

**TABLE 4 |** Top GINI-ranked predictors in machine learning models for loneliness and social isolation*#.

| Loneliness | Emotional support | Instrumental support | Negative social interactions |
|---|---|---|---|
| First-person plural pronoun (We) (Density, entire transcript) | Response length (Word minimum response) | Sentence similarity (frequency) | Noun usage (frequency) |
| –* | + | – | – |
| Compound (positive and negative) sentiment (SD) | Positive sentiment (mean) | Pronoun usage (frequency) | Negative sentiment (SD) |
| –* | +* | + | + |
| Interjection (ratio) | Pronoun usage (Ratio of pronoun to noun) | Negative sentiment (maximum) | Negative sentiment (mean) |
| – | +* | + | +* |
| Sentence complexity (average yngve depth, Median) | Positive sentiment (median) | Response length (Total number words) | Verb usage (frequency) |
| –* | +* | + | – |
| Response length (Total words, mean) | Compound sentiment (median) | Neutral sentiment (median) | Sentence similarity (median) |
| –* | | – | – |
| Sentence complexity (yngve depth, Total) | Sentiment neutral (median) | Sentence similarity | Pronoun usage (frequency) |
| –* | –* | – | – |
| Response length (Total characters, median) | Sentence complexity (Average yngve depth, median) | Frequency of adjectives | Ratio of nouns |
| –* | + | + | + |
| Education | Gender (female) | Response length (Total number characters) | Filler frequency |
| +* | +* | + | – |
| Total words (Relationship section) | Pronoun usage (ratio) | Verb usage (frequency) | Number of important relationships |
| – | +* | + | – |
| Adverb usage (frequency) | Compound sentiment (mean) | Vocabulary Brunett index | Positive sentiment (SD) |
| –* | + | + | + |

$Features comprising socio-demographic features, language features, and pronoun features. *Significant.*
#*Description of linguistic features is mentioned in* ***Supplementary Appendix A*** *and* ***Supplementary Table 3****.*
*+Associated with higher scores on loneliness, emotional support etc.*
*–Associated with lower scores on loneliness, emotional support, etc.*
*SD, standard deviation.*
***Color coding****:*

***Pronouns***, ***Socio-demographic***, ***Response length***, ***Parts of speech (non-pronouns)***, ***Sentiment***, ***Sentence complexity and similarity***, ***Vocabulary richness***, ***Relationship words.***

top feature in models of loneliness and social support. Other top linguistic features included sentence similarity/complexity, response length, and sentiment.

The current finding that usage of first-person plural pronouns was linked to lower loneliness among the women and higher social support among men is consistent with previous research on first-person plural pronoun use as a linguistic indicator of interdependence that has been consistently associated with higher quality relationships and better physical and mental health functioning (33). Studies have also described the links between first-person plural pronoun usage and better perceived support, an expanded sense of self (34), and better conflict resolution in couples (35–38). First-person plural pronoun use also reflects social support within couples, exhibited in how dyads cope together with challenges such as a cancer diagnosis (39–42). Conversely, usage of first-person singular pronoun has been linked with depressive symptoms (43) and negative affective states, noting these associations to be stronger in women (44–46). Language may also influence mood states. Subjects who recalled a depressing incident from a self-distanced perspective (using fewer first-person pronouns) had less depressed affect for up to a week, compared to those who used a more self-immersive stance (47).

The current study illustrates how diverse sets of linguistic features can be used to predict SI/L with good accuracy. The linguistic models presented here (which included a broader variety of linguistic features and sociodemographic information) slightly outperformed our previous models (13), which were limited to NLU-based emotions, sentiment, and question stems from the structured interview template. The current models found that in addition to sentiment, sentence complexity and similarity, usage of pronouns and other parts of speech, and response length were top-ranked features in predicting SI/L. This suggests that a broader variety of linguistic features may outperform purely emotion and sentiment-based models, though more comprehensive models should also include auditory features (e.g., tone, response latency), semantic features (word usage), and longitudinal follow-up. A 2017 study by Mehl et al. (48) reported that lonely individuals used fewer propositions and less time spent talking with others. One study reported that linguistic features such as tentativeness and non-fluencies are associated with depression and anxiety symptoms (49), while another study of Twitter messages found that posts that used "lonely" or "alone" had consistent themes of anger, anxiety, difficulties in interpersonal relationships, substance use, unhealthy eating and sleep (50). One novel study of professional football players and their coaches found longitudinal decline in language complexity in the players (who were at high risk for head trauma) relative to their coaches (51). NLP approaches can capture the breadth of information conveyed through language, augmenting our ability to assess an individual's internal emotional state and social functioning.

All participants were assessed on a wide range of socio-demographic and psychological factors including depression. Previously published studies have shown the overlapping prevalence of depression and loneliness, however due to the low prevalence of depression in this cohort (7.2% had a PHQ-9 score of 10 or greater, 2.1% had a PHQ-9 score of 15 or greater and 0.0% had 20 or greater) and due to a lack of depressive symptoms beyond the mild level of severity, we did not include depression as a confound. For the purposes of this study, only socio-demographic factors and linguistic features were used to predict loneliness and social support.

Our findings included a sizeable number of Facebook users in this age group; it is not very surprising given previous studies that have found older adults to be capable users of technology (52) and, increasingly, social media users—with Facebook use reaching 50% (53) even as younger adults ceased using the social media platform (54).

Several studies have attempted novel techniques to remedy the lack of interpretability of ML models (or their black-box nature). A recent review on the topic, which details the advantages and major drawbacks can be found here (55). Many of these methods have short histories, or, are not widely and openly accepted and/or understood. This is in contrast to ANN models, which are often not only powerful, but they also have a long history, are well and widely understood, studied, and are relatable by most in the field. Most professionals can find a common ground in ANN.

## Study Limitations

Properties of speech (e.g., pitch, prosody, meaningless sounds, amplitude, and modulations) are meaningful features with clinically relevant implications, however, in the current study, we did not assess speech acoustics and relied solely upon the transcribed text.

Our study was cross-sectional and limited to a small sample of independent-living older adults and may not be representative of nor generalizable to the broader class of individuals in the same age group. Our statistical analysis showed a significant age difference between the two genders (Cohen's $d = -0.68$, $p < 0.001$) which potentially confounds age and gender. Follow-up analyses examined the confounding effects of age and gender. Machine learning models exploit combining features in complex non-linear ways to predict the target variables; however, they are difficult to interpret. Linguistic features, by definition, are influenced by language proficiency. Thus, NLP features in non-native English speakers may manifest differently (56, 57). In the current study, we did not control for English proficiency. The models were derived from participants who are fluent in English and may not be applicable to other older adult populations. Pronoun usage may depend on variety of factors such as the number of siblings and size of family when growing up, the choice of profession, and involvement in leadership roles (58). The current study did not control for these factors. Mental health status and momentary emotional state of both, the interviewer and interviewee and their interpersonal dynamics, can influence the interview. Due to a large variety of factors that shape conversations, predictions using these approaches are difficult to perfect.

Character and personality play an important role in verbal expression and are worthy of independent investigation, however this is beyond the scope of the present study. Despite promising initial findings, commonly used sentiment analyzers may be susceptible to bias, due to highly variable assessments, large

breadth of applications, or specificity to a particular test case (24).

In this analysis, we have trimmed the least important features, stopping when performance of the model is reduced. While this method of selecting features based on information gain or impurity rank, may result in including features that could be inter-correlated, this does not adversely affect the performance or the results in contrast to traditional statistical methods. This method may not provide the minimal feature set, which is very difficult to identify (59), but roughly identified sets such as ours work well in practice.

For this project, we aimed to compare language usage differences between people with and without SI/L. Transformers, despite being very useful in certain cases that require extraction of meaning, have limited applicability in our study. First, they are intended to process text, not linguistic features. Second, transformers are uniquely equipped for tasks such as translation and summarizing as they are designed to retain meaningful concepts using attention (60). But this has an effect of deemphasizing less important details, which have less to do with the meaning but more to do with expression e.g., vocabulary richness, filler words, and pronouns. Third, recent studies have reported that Bidirectional Encoder Representations from Transformers (BERT—a well-known architecture that first introduced the idea of attention and was quickly embraced by the community) often cannot outperform some common classification and other simpler baselines (61–63). Crafting an appropriate transformer for the task may not be straightforward, and advantages may translate into just a few percentage points in performance.

### Future Directions/Overall Conclusions

The application of NLP for the purpose of facilitating understanding of human health is exciting. The fact that myriad factors can influence conversations, more research is needed to refine the predictive accuracy of these models. NLP assessments of unstructured language may be integrated with self-report and behavioral assessments to provide nuanced and sensitive evaluations of SI/L. Moreover, the narrative data that forms the basis of the NLP training data must be evaluated to ensure that it is representative of people for whom the results may be applied. Given its novelty, those exploring NLP applications, including researchers and clinicians, should become knowledgeable about how to approach its use and consider issues of bias, fairness, accountability, and related ethical and social implications early and often during the study. While this study was limited to common architectures used in ML, newer attention-based models, such as transformers, may provide additional improvements.

Due to low rates of depression in this cohort, we were not able to assess language features that were reflective of depressive symptoms. However, future NLP studies of lonely cohorts with higher rates of depression should consider how the impact of depression on language, both independent of SI/L as well as through effects on social functioning.

## DATA AVAILABILITY STATEMENT

The study/data is governed by University of California San Diego Human Research Protections Program (HRPP) rules and other contract. It is not publicly available due to privacy concerns, may include HIPAA regulations. For access, qualified researchers may contact the corresponding author.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of California San Diego Human Research Protections Program. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

VB and EL contributed to the conception and design of the study and had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. VB wrote the first draft of the manuscript and conducted the data analyses. VB, YY, and KS developed the NLP tools used in the analyses. EL, VB, CN, YY, KS, KR, and H-CK were involved in the data interpretation. VB, EL, and CN wrote sections of the manuscript and were involved in data interpretation. All authors contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

This study received funding from IBM. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.728732/full#supplementary-material

## REFERENCES

1. Jeste DV, Lee EE. Loneliness vs. wisdom amid two concurrent pandemics of loneliness. *Am J Geriatr Psychiatry*. (2020) 28:1245–7. doi: 10.1016/j.jagp.2020.08.017

2. Lee EE, Depp C, Palmer BW, Glorioso D, Daly R, Liu J, et al. High prevalence and adverse health effects of loneliness in community-dwelling adults across the lifespan: role of wisdom as a protective factor. *Int Psychogeriatr*. (2019) 31:1447–62. doi: 10.1017/S1041610218002120

3. Brown BB, Chiang CP. Drug and alcohol abuse among the elderly: is being alone the key? *Int J Aging Hum Dev.* (1983) 18:1–12. doi: 10.2190/VRMK-T7UV-QKRT-KN1R
4. Shankar A, Hamer M, McMunn A, Steptoe A. Social isolation and loneliness: relationships with cognitive function during 4 years of follow-up in the English Longitudinal Study of Ageing. *Psychosom Med.* (2013) 75:161–70. doi: 10.1097/PSY.0b013e31827f09cd
5. Ellwardt L, van Tilburg T, Aartsen M, Wittek R, Steverink N. Personal networks and mortality risk in older adults: a twenty-year longitudinal study. *PLoS One.* (2015) 10:e0116731. doi: 10.1371/journal.pone.0116731
6. van der Veen DC, van Zelst WH, Schoevers RA, Comijs HC, Voshaar RC. Comorbid anxiety disorders in late-life depression: results of a cohort study. *Int Psychogeriatr.* (2015) 27:1157–65. doi: 10.1017/S1041610214002312
7. Crewdson JA. The effect of loneliness in the elderly population: a review. *Healthy Aging Clin Care Elderly.* (2016) 8:1–8. doi: 10.4137/HACCE.S35890
8. Momtaz YA, Hamid TA, Yusoff S, Ibrahim R, Chai ST, Yahaya N, et al. Loneliness as a risk factor for hypertension in later life. *J Aging Health.* (2012) 24:696–710. doi: 10.1177/0898264311431305
9. Buchman AS, Boyle PA, Wilson RS, James BD, Leurgans SE, Arnold SE, et al. Loneliness and the rate of motor decline in old age: the Rush Memory and Aging Project, a community-based cohort study. *BMC Geriatr.* (2010) 10:77. doi: 10.1186/1471-2318-10-77
10. Penninx BW, van Tilburg T, Kriegsman DM, Deeg DJ, Boeke AJ, van Eijk JT. Effects of social support and personal coping resources on mortality in older age: the Longitudinal Aging Study Amsterdam. *Am J Epidemiol.* (1997) 146:510–9. doi: 10.1093/oxfordjournals.aje.a009305
11. Sugisawa H, Liang J, Liu X. Social networks, social support, and mortality among older people in Japan. *J Gerontol.* (1994) 49:S3–S13. doi: 10.1093/geronj/49.1.S3
12. Daly J, Depp C, Graham SA, Jeste DV, Kim HC, Lee EE, et al. Health impacts of the "Stay at Home" order on community dwelling older adults and how technologies may help: a focus group study. *JMIR Aging.* (2021) 4:e25779. doi: 10.2196/preprints.25779
13. Badal VD, Graham SA, Depp CA, Shinkawa K, Yamada Y, Palinkas LA, et al. Prediction of loneliness in older adults using natural language processing: exploring sex differences in speech. *Am J Geriatr Psychiatry.* (2021) 29:853–66. doi: 10.1016/j.jagp.2020.09.009
14. Jeste DV, Glorioso D, Lee EE, Daly R, Graham S, Liu J, et al. Study of independent living residents of a continuing care senior housing community: sociodemographic and clinical associations of cognitive, physical, mental health. *Am J Geriatr Psychiatry.* (2019) 27:895–907. doi: 10.1016/j.jagp.2019.04.002
15. Morlett Paredes A, Lee EE, Chik L, Gupta S, Palmer BW, Palinkas LA, et al. Qualitative study of loneliness in a senior housing community: the importance of wisdom and other coping strategies. *Aging Ment Health.* (2021) 25:559–66. doi: 10.1080/13607863.2019.1699022
16. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* (2001) 16:606–13. doi: 10.1046/j.1525-1497.2001.016009606.x
17. Seeman TE, Lusignolo TM, Albert M, Berkman L. Social relationships, social support, and patterns of cognitive aging in healthy, high-functioning older adults: MacArthur studies of successful aging. *Health Psychol.* (2001) 20:243. doi: 10.1037/0278-6133.20.4.243
18. Russell DW. UCLA Loneliness Scale (Version 3): reliability, validity, factor structure. *J Pers Assess.* (1996) 66:20–40. doi: 10.1207/s15327752jpa6601_2
19. Doryab A, Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data. *JMIR Mhealth Uhealth.* (2019) 7:e13209. doi: 10.2196/13209
20. Joachims T. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* Carnegie-Mellon Univ Pittsburgh Pa Dept of Computer Science (1996).
21. Aizawa A. An information-theoretic perspective of tf-idf measures. *Inf Process Manag.* (2003) 39:45–65. doi: 10.1016/S0306-4573(02)00021-3
22. Rajaraman A, Ullman JD. *Mining of Massive Datasets*. New York, NY: Cambridge University Press (2011).
23. Yamada Y, Shinkawa K, Shimmei K. Atypical repetition in daily conversation on different days for detecting Alzheimer's disease: evaluation of phone-call data from a regular monitoring service. *JMIR Mental Health.* (2020) 7:e16790. doi: 10.2196/16790
24. Ribeiro FN, Araújo M, Gonçalves P, Gonçalves MA, Benevenuto F. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.* (2016) 5:1–29. doi: 10.1140/epjds/s13688-016-0085-1
25. Demsar J, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, et al. Orange: data mining toolbox in Python. *J Mach Learn Res.* (2013) 14:2349–53.
26. Hughes G. On the mean accuracy of statistical pattern recognizers. *IEEE Trans Inf Theory.* (1968) 14:55–63. doi: 10.1109/TIT.1968.1054102
27. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics.* (2007) 23:2507–17. doi: 10.1093/bioinformatics/btm344
28. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. *Comput Biol Med.* (2019) 112:103375. doi: 10.1016/j.compbiomed.2019.103375
29. Brown AJM. *Comprehensive Chemometrics, Vol. 3.* GINI Index (2009). Available online at: https://www.sciencedirect.com/referencework/9780444527011/comprehensive-chemometrics (accessed March, 2021).
30. Tyagi N. *Understanding the Gini Index and Information Gain in Decision Trees*. Available online at: https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8 (accessed March, 2021).
31. Von Heusinger K, Egli U. Introduction: reference and the semantics of anaphora. In: von Heusinger K, Egli U, editors. *Reference and Anaphoric Relations*. Dordrecht: Springer (2000). p. 1–13.
32. Smit J, Steglich-Petersen A. Anaphora and semantic innocence. *J Seman.* (2010) 27:119–24. doi: 10.1093/jos/ffp012
33. Karan A, Rosenthal R, Robbins ML. Meta-analytic evidence that we-talk predicts relationship and personal functioning in romantic couples. *J Soc Pers Relationsh.* (2019) 36:2624–51. doi: 10.1177/0265407518795336
34. Aron A, Aron EN, Norman C. Self-expansion model of motivation and cognition in close relationships and beyond. In: Fletcher GJO, Clark MS, editors. *Blackwell Handbook of Social Psychology.* Malden, MA: Interpersonal Processes Blackwell Publishers (2001). p. 478–501.
35. Simmons RA, Gordon PC, Chambless DL. Pronouns in marital interaction: what do "You" and "I" say about marital health? *Psychol Sci.* (2005) 16:932–6. doi: 10.1111/j.1467-9280.2005.01639.x
36. Agnew CR, Van Lange PAM, Rusbult CE, Langston CA. Cognitive interdependence: commitment and the mental representation of close relationships. *J Pers Soc Psychol.* (1998) 74:939–54. doi: 10.1037/0022-3514.74.4.939
37. Seider BH, Hirschberger G, Nelson KL, Levenson RW. We can work it out: age differences in relational pronouns, physiology, and behavior in marital conflict. *Psychol Aging.* (2009) 24:604–13. doi: 10.1037/a0016950
38. Williams-Baucom KJ, Atkins DC, Sevier M, Eldridge KA, Christensen A. "You" and "I" need to talk about "us": linguistic patterns in marital interactions. *Pers Relationsh.* (2010) 17:41–56. doi: 10.1111/j.1475-6811.2010.01251.x
39. Hagedoorn M, Sanderman R, Bolks HN, Tuinstra J, Coyne JC. Distress in couples coping with cancer: a meta-analysis and critical review of role and gender effects. *Psychol Bull.* (2008) 134:1–30. doi: 10.1037/0033-2909.134.1.1
40. Rohrbaugh MJ, Mehl MR, Shoham V, Reilly ES, Ewy GA. Prognostic significance of spouse "we" talk in couples coping with heart failure. *J Consult Clin Psychol.* (2008) 76:781–9. doi: 10.1037/a0013238
41. Helgeson VS, Jakubiak B, Seltman H, Hausmann L, Korytkowski M. Implicit and explicit communal coping in couples with recently diagnosed type 2 diabetes. *J Soc Pers Relationsh.* (2017) 34:1099–121. doi: 10.1177/0265407516669604
42. Rentscher KE, Rohrbaugh MJ, Shoham V, Mehl MR. Asymmetric partner pronoun use and demand-withdraw interaction in couples coping with health problems. *J Fam Psychol.* (2013) 27:691–701. doi: 10.1037/a0034184
43. Rude S, Gortner E-M, Pennebaker J. Language use of depressed and depression-vulnerable college students. *Cogn Emot.* (2004) 18:1121–33. doi: 10.1080/02699930441000030

44. Nolen-Hoeksema S. Sex differences in unipolar depression: evidence and theory. *Psychol Bull.* (1987) 101:259. doi: 10.1037/0033-2909.101.2.259
45. Lyubomirsky S, Layous K, Chancellor J, Nelson SK. Thinking about rumination: the scholarly contributions and intellectual legacy of Susan Nolen-Hoeksema. *Annu Rev Clin Psychol.* (2015) 11:1–22. doi: 10.1146/annurev-clinpsy-032814-112733
46. Nolen-Hoeksema S, Larson J, Grayson C. Explaining the gender difference in depressive symptoms. *J Pers Soc Psychol.* (1999) 77:1061. doi: 10.1037/0022-3514.77.5.1061
47. Kross E, Ayduk O. Facilitating adaptive emotional analysis: distinguishing distanced-analysis of depressive experiences from immersed-analysis and distraction. *Pers Soc Psychol Bull.* (2008) 34:924–38. doi: 10.1177/0146167208315938
48. Mehl MR, Raison CL, Pace TWW, Arevalo JMG, Cole SW. Natural language indicators of differential gene regulation in the human immune system. *Proc Natl Acad Sci U S A.* (2017) 114:12554–9. doi: 10.1073/pnas.1707373114
49. ODea B, Boonstra TW, Larsen ME, Nguyen T, Venkatesh S, Christensen H. The relationship between linguistic expression and symptoms of depression, anxiety, and suicidal thoughts: a longitudinal study of blog content. *PLoS ONE.* (2021) 16:e0251787.
50. Guntuku SC, Schneider R, Pelullo A, Young J, Wong V, Ungar L, et al. Studying expressions of loneliness in individuals using twitter: an observational study. *BMJ Open.* (2019) 9:e030355. doi: 10.1136/bmjopen-2019-030355
51. Berisha V, Wang S, LaCross A, Liss J, Garcia-Filion P. Longitudinal changes in linguistic complexity among professional football players. *Brain Lang.* (2017) 169:57–63. doi: 10.1016/j.bandl.2017.02.003
52. Smith A. *Older Adults and Technology Use.* Pew Research Center (2014).
53. Statista Research Department. *Percentage of U.S. Adults Who Use Facebook as of February 2021 by Age Group* (2021).
54. Bosak K, Park SH. Characteristics of adults' use of facebook and the potential impact on health behavior: secondary data analysis. *Interact J Med Res.* (2018) 7:e9554. doi: 10.2196/ijmr.9554
55. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: a review of machine learning interpretability methods. *Entropy.* (2021) 23:18. doi: 10.3390/e23010018
56. Cuskley C, Colaiori F, Castellano C, Loreto V, Pugliese M, Tria F. The adoption of linguistic rules in native and non-native speakers: evidence from a Wug task. *J Mem Lang.* (2015) 84:205–23. doi: 10.1016/j.jml.2015.06.005
57. Bloem J, Wieling M, Nerbonne J. Automatically identifying characteristic features of non-native English accents. *Future Dialects.* (2016) 155–73. doi: 10.17169/LANGSCI.B81.148
58. Brown R, Gilman A. The pronouns of power and solidarity. In: Fishman JA, editor. *Readings in the Sociology of Language*. Berlin; Boston: De Gruyter Mouton (2012). p. 252–75. doi: 10.1515/9783110805376.252
59. Sitnikov D, Titova O, Romanenko O, Ryabov O. A method for finding minimal sets of features adequately describing discrete information objects. *WIT Trans Inf Commun Technol.* (2009) 42:143–53. doi: 10.2495/DATA090141
60. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst.* (2017) 5998–6008. Available online at: https://arxiv.org/abs/1706.
61. Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon H-J, et al. Limitations of transformers on clinical text classification. *IEEE J Biomed Health Inf.* (2021) 25:3596–607. doi: 10.1109/JBHI.2021.3062322
62. Data ScienceStack Exchange. *Why do BERT Classification Do Worse With Longer Sequence Length*?.
63. Liu B, Huang J, Cai X, Church K. Better than BERT but worse than baseline. *arXiv:2105.05915 [arXiv preprint]* (2021).


**Conflict of Interest:** KS, YY, and H-CK are employees of IBM.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 13 December 2021
doi: 10.3389/fpsyt.2021.712251

# Automatic Assessment of Loneliness in Older Adults Using Speech Analysis on Responses to Daily Life Questions

*Yasunori Yamada*[1]*, *Kaoru Shinkawa*[1]*, *Miyuki Nemoto*[2] and *Tetsuaki Arai*[3]

[1] IBM Research, Tokyo, Japan, [2] Dementia Medical Center, University of Tsukuba Hospital, Tsukuba, Japan, [3] Division of Clinical Medicine, Department of Psychiatry, Faculty of Medicine, University of Tsukuba, Tsukuba, Japan

Loneliness is a perceived state of social and emotional isolation that has been associated with a wide range of adverse health effects in older adults. Automatically assessing loneliness by passively monitoring daily behaviors could potentially contribute to early detection and intervention for mitigating loneliness. Speech data has been successfully used for inferring changes in emotional states and mental health conditions, but its association with loneliness in older adults remains unexplored. In this study, we developed a tablet-based application and collected speech responses of 57 older adults to daily life questions regarding, for example, one's feelings and future travel plans. From audio data of these speech responses, we automatically extracted speech features characterizing acoustic, prosodic, and linguistic aspects, and investigated their associations with self-rated scores of the UCLA Loneliness Scale. Consequently, we found that with increasing loneliness scores, speech responses tended to have less inflections, longer pauses, reduced second formant frequencies, reduced variances of the speech spectrum, more filler words, and fewer positive words. The cross-validation results showed that regression and binary-classification models using speech features could estimate loneliness scores with an $R^2$ of 0.57 and detect individuals with high loneliness scores with 95.6% accuracy, respectively. Our study provides the first empirical results suggesting the possibility of using speech data that can be collected in everyday life for the automatic assessments of loneliness in older adults, which could help develop monitoring technologies for early detection and intervention for mitigating loneliness.

**Keywords: health-monitoring, speech analysis and processing, mental health, voice, social connectedness**

OPEN ACCESS

***Edited by:***
*Helmet Karim,*
*University of Pittsburgh, United States*

***Reviewed by:***
*Manuela Barreto,*
*University of Exeter, United Kingdom*
*Ipsit Vahia,*
*McLean Hospital, United States*

***Correspondence:**
*Yasunori Yamada*
*ysnr@jp.ibm.com*
*Kaoru Shinkawa*
*kaoruma@jp.ibm.com*

***Specialty section:***
*This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry*

***Received:*** *20 May 2021*
***Accepted:*** *19 November 2021*
***Published:*** *13 December 2021*

***Citation:***
*Yamada Y, Shinkawa K, Nemoto M and Arai T (2021) Automatic Assessment of Loneliness in Older Adults Using Speech Analysis on Responses to Daily Life Questions. Front. Psychiatry 12:712251. doi: 10.3389/fpsyt.2021.712251*

## 1. INTRODUCTION

Loneliness is a subjective and perceived state of social and emotional isolation. Importantly, loneliness is a specific construct that is associated with but distinguished from depression, anxiety, and objective social isolation. As the world's elderly population increases, loneliness in older adults is becoming a serious health problem. In older adults, loneliness has been prospectively associated with a wide range of adverse health outcomes including morbidity and mortality (1, 2), function decline (3), depression (4, 5), cognitive decline (6, 7), and incidents of dementia, especially Alzheimer's disease (8, 9). A meta-analysis has shown that loneliness increases the risk of mortality

comparable with other well-known risk factors, such as smoking, obesity, and physical inactivity (10). Moreover, the increases in the aging population and prevalence of loneliness make loneliness a more serious social and health problem (11, 12). In fact, the prevalence of loneliness has increased from an estimated 11–17% in the 1970s (11, 13) to about 20–40% for older adults (1, 14, 15). From these perspectives, a growing body of research have actively investigated possible interventions to reduce the prevalence of loneliness and its harmful consequences (11, 12, 16), and early detection of loneliness is urgently needed. One of the simplest ways is to use a direct question such as "Do you feel lonely?" but it has been reported to lead to underreporting due to the stigma associated with loneliness (17–19). Instead, multidimensional scales without explicitly using the word "lonely" [e.g., the 20-item UCLA Loneness Scale (20)] have been widely used for measuring loneliness in older adults (17). If loneliness measured by such a multidimensional scale can be automatically estimated by using passively collected data without requiring individuals to perform any task, this would help early detection of lonely individuals through frequent assessments with less burden on older adults.

Several studies have reported the possibility of automatic assessment of loneliness by using daily behavioral data (21–23). For example, one study collected behavioral data using in-home sensors such as time out-of-home and number of calls from 16 older adults for 8 months, and reported that a regression model using them could estimate scores of the UCLA Loneliness Scale with a correlation of 0.48 (21). Another study collected behavioral data including mobility, social interactions, and sleep from the smartphones and Fitbits of 160 college students and reported that a binary-classification model using these behavioral data could detect individuals with high loneliness scores at 80.2% accuracy (22). Although they suggested that the loneliness may produce measurable changes in daily behaviors and be automatically assessed by using these behavioral data, the behavioral types investigated in previous studies as well as studies researching these behavioral types still remains limited. Being capable of assessing loneliness using various types of daily behaviors would help improve performance and extend the application scope.

Speech is an attractive candidate for automatically assessing loneliness. There is growing interest in using speech data for healthcare applications (24, 25), due to the improvement in audio quality recorded by portable devices and the popularity of voice-based interaction systems such as voice assistants in smart speakers and smartphones. For example, a number of studies used phone conversations passively recorded (26, 27) and others used speech responses to tasks with mobile devices (28–32). If automatic assessment of loneliness is possible using speech responses collected in either way (i.e., conversations with other people or speech responses collected through voice interfaces), it would greatly increase the opportunity and accessibility of assessment for early detection of loneliness.

Speech data has been used for capturing changes in various types of emotional states and mental health conditions including depression (33–41), suicidality (35, 42), and bipolar disorder (27, 43). As a result of a complexity of the speech production process involving motor, cognitive, and physiological factors, speech has been thought to be a sensitive output system such that changes in individuals' emotional states and mental health conditions can produce measurable acoustic, prosodic, and linguistic changes (35, 44, 45). Studies have shown the promise in using speech as an objective biomarker for detecting/predicting mental illness (46, 47) and monitoring a patient's symptoms (48, 49). For example, previous studies on depressive speech reported substantial changes in acoustic, prosodic, and linguistic features including reduced formant frequencies (35, 36, 40), reduced pitch variation (less inflections) (41, 50), more pauses (34, 38), more negative words, and fewer positive words (33, 37). Although there has been no study investigating the relationship between loneliness and speech data that can be collected in everyday situations, it is reasonable to explore the possibility that speech data could be used for assessing loneliness in older adults.

We aimed to investigate whether speech features associated with loneliness levels in older adults can be found in speech data that can be collected in everyday life and whether these speech data can be used for estimating loneliness levels and detecting individuals with higher levels of loneliness. To this end, we developed a tablet-based application and collected speech responses to daily life questions regarding, for example, one's feelings and future travel plans. We also collected self-rated scores of the UCLA Loneliness Scale from the same participants. From audio data of the speech responses, we automatically extracted speech features characterizing acoustic, prosodic, and linguistic aspects, and investigated the association between these speech features and loneliness scores using correlation analysis and machine learning models.

## 2. METHODS

### 2.1. Participants

We recruited healthy older adults through local recruiting agencies or advertisements in the local community in Ibaraki, Japan. Participants were excluded if they had self-reported diagnoses of mental illness at the time of recruitment (e.g., major depression, bipolar disorder, and schizophrenia), had self-reported prior diagnoses of neurodegenerative diseases (e.g., Parkinson's disease, dementia, and mild cognitive impairment), or had other serious diseases or disabilities that would interfere with the assessments of this study. All examinations were conducted in Japanese. This study was conducted under the approval of the Ethics Committee, University of Tsukuba Hospital (H29-065). All participants provided written informed consent after the procedures of the study had been fully explained.

In addition to the speech data collection and loneliness survey, all participants underwent the Mini-Mental State Examination to assess global cognition and Geriatric Depression Scale to assess depressive symptoms conducted by neuropsychologists. They also answered self-report instruments about their education level and marital status.

A total of 57 older individuals completed the speech data collection and loneliness survey [30 women (52.6%); 62–81 years; mean (SD) age, 73.2 (4.5) years; **Table 1**]. **Table 1** summarizes the information about participant characteristics.

**TABLE 1 |** Characteristics of study participants (*N*=57).

| Characteristics | | |
|---|---|---|
| Age [years], mean (SD) | 73.2 | (4.5) |
| Sex, n (%) | | |
| Men | 27 | (47.4) |
| Women | 30 | (52.6) |
| Education [years], mean (SD) | 13.8 | (2.2) |
| Marital status, n (%) | | |
| Never married | 0 | (0) |
| Divorced | 2 | (3.5) |
| Widowed | 8 | (14.0) |
| Married | 47 | (82.5) |
| Mini-Mental State Examination[a], mean (SD) | 27.4 | (1.9) |
| Geriatric Depression Scale[b], mean (SD) | 2.9 | (2.5) |
| UCLA Loneliness Score[c], mean (SD) | 37 | (8.6) |

[a]*The total possible score ranges from 0 to 30.*
[b]*The total possible score ranges from 0 to 15.*
[c]*The total possible score ranges from 20 to 80.*


**FIGURE 1 |** Overview of experimental setup for collecting speech data. **(A)** Participant's turn and **(B)** tablet's turn.

## 2.2. Loneliness Survey

Loneliness levels for our participants were measured by the Japanese version of the UCLA Loneliness Scale version 3 (20, 51). This scale is a validated, self-rated instrument designed to measure feelings of emotional loneliness in a wide group of respondents, including older adults, and implemented in numerous epidemiologic studies of aging (3, 52, 53). It consists of 20 Likert-type questions on a four-point scale from "never" to "always". The total score ranges from 20 to 80 with a higher score indicating greater loneliness, and there is no identified cut-off score that defines loneliness (54). We used this total score for the analysis.

## 2.3. Speech Data Collection

Participants sat down in front of the tablet and answered questions presented by a voice-based application on the tablet in a quiet room with low reverberation. The participants were asked to speak as naturally as possible. The tablet indicated whether it was speaking or listening (**Figure 1**). We used an iPad Air 2 and recorded voice responses by using the iPad's internal microphone (core audio format, 44,100 Hz, 16-bit).

The participants were asked eight daily life questions. The first two questions were frequently-used ones in daily conversations, that is, how one feels today and one's sleep quality last night. The next three questions were related to past experiences in terms of recalling old memories about a fun childhood activity as well as recent memories related to what was eaten for dinner yesterday and the day before yesterday. The next two questions were related to future expectations in terms of risk planning, such as one's response plans for an earthquake, and travel planning where participants chose one option from among two regarding future travel destinations and gave three reasons for their choice. The final question was related to general knowledge where participants explained a Japanese traditional event. For the actual sentences of the daily life questions, please see **Supplementary Table 1**.

## 2.4. Speech Data Analysis

From the speech responses of each participant to the eight questions, we automatically extracted a total of 160 speech features consisting of 128 acoustic features, 16 prosodic features, and 16 linguistic features. These features were determined on the basis of previous studies on inferring changes in emotional states and mental health conditions such as depression and suicidality (27, 33–43, 55–57). Full list of speech features is available in **Supplementary Table 2**.

As a preprocessing step, we first converted the audio data of each response into text data (i.e., automatic speech recognition) and divided the audio signals into voice and silence segments (i.e., voice activity detection) by using the IBM Watson Speech to Text service. All acoustic and prosodic features were extracted from the audio signals of voice segments, except for pause-related features, which were calculated by using the time duration of silence segments. Linguistic features were extracted from the text data after word tokenization, part-of-speech tagging, and word lemmatization using the Japanese morphological analyzer Janome (version 0.3.10 [1]) in Python (version 3.8).

The acoustic features consisted of two feature types related to formant frequencies and Mel-frequency cepstral coefficients (MFCCs). Formant frequencies contain information related to

[1]https://mocobeta.github.io/janome/en/

acoustic resonances of the vocal tract and thus are thought to be able to capture changes in vocal tract properties affected by both an increase in muscle tension and changes in salivation and mucus secretion due to mental state changes (35). For example, the first and second formant (F1 and F2) are mainly associated with the tongue position: the F1 frequency is inversely related to the height of the tongue, while the F2 frequency is directly related to the frontness of the tongue position (58, 59). Limited movements of the articulators and particularly of the tongue, for example due to increased muscle tension, lead to inadequate vowel formation characterized by a lowering of normally high frequency formants and by an elevation of normally low frequency formants (60). Decreased formant frequencies were reported with increasing levels of speaker depression (35, 36, 40), and formant-based features have been frequently used for detecting depressive speech (46, 47). In addition, MFCCs are spectral features characterizing the frequency distribution of a speech signal at specific time instance information and designed to take into account the response properties of the human auditory system (61). As with the formant-based features, MFCCs have consistently been observed to change with individuals' mental states (35), and have been successfully used for various speech tasks including emotion recognition (62, 63), mood detection (64), and detection of depression (49, 65). In particular, the variances of the derivative of MFCCs were reported to show a consist trend of negative correlations with depression severity (49, 50). These decreased temporal variations in MFCCs with increasing depressive severity are thought to capture monotony and dullness of speech in clinical descriptions (35, 49). We thus used the first two formant frequencies (F1 and F2) and the variances of the first order derivatives ($\Delta$) of the first 14 MFCCs. Because these features were extracted from each response to the eight questions, we obtained $(2 + 14) \times 8 = 128$ acoustic features for each participant. To extract them, we used the Python-based (version 3.8) audio processing library librosa [version 0.8.0 (66)].

Prosodic features such as rhythm, stress, and intonation in speech conveys important information regarding individual's mental states. Commonly-used examples include pitch variation (i.e., inflection) and pause duration. Multiple studies reported a reduced pitch variation and an increased pause duration in accordance with increasing levels of depression severity (34, 38) as well as brief emotion induction of sadness in normal participants (39), although a number of studies showed no substantial change (34, 67). We thus used pitch variation and pause duration for prosodic features. Specifically, we calculated the pitch variation and pause duration in all eight responses of each participants and used total $2 \times 8 = 16$ features as the prosodic features. For estimating pitch, we used fundamental frequency calculated with the Python-based audio processing library Signal_Analysis (version 0.1.26 [2]).

The linguistic features consisted of three feature types related to positive words, negative words and filler words (e.g., "umm," "hmm," "uh"). Sentiment analysis has been one of most representative approaches to detect changes in mental health conditions from linguistic cues. For example, several studies reported that depressed individuals tended to use more negative words and fewer positive words than non-depressed individuals (33, 37). Filler words are commonly found in spontaneous speech and have been suggested as important signatures for detecting depression (55–57). We thus used the number of positive and negative words and the proportion of filler words as linguistic features. Specifically, we counted the number of positive and negative words, respectively, in each speech response to the four questions expected to include positive or negative words: questions about a fun childhood activity, response plan for an earthquake, future travel plans, and a Japanese traditional event. Each word was determined to be positive (or negative) by using the Japanese Sentiment Polarity Dictionary (68, 69). The number of filler words was obtained by counting two kinds of words: those estimated as hesitation by automatic speech recognition using the IBM Watson Speech to Text service and those defined as fillers in the Japanese IPA dictionary[3]. We thus used $2 \times 4 + 1 \times 8 = 16$ linguistic features for each participant.

## 2.5. Statistical Analysis

Spearman's nonparametric rank correlation coefficient was computed to test the null hypothesis that there is no correlation between each speech feature and scores of the UCLA Loneliness Scale. We did not adjust for multiple comparisons, and $P$ values below 0.05 were considered to be statistically correlated. The statistical analyses were conducted using the Statistics and Machine Learning Toolbox (version 11.1) for the MATLAB (version R2017a, The MathWorks Inc) environment.

## 2.6. Regression and Classification Models

The regression and binary-classification models were built to investigate whether speech features can be used for estimating scores of the UCLA Loneliness Scale and for detecting individuals with high scores, respectively (**Figure 2**). For the cut-off score for building the binary-classification model, because there is no designated cut-off score, we used the mean + 1SD of our participants' scores in the same manner as a previous study on characteristics of lonely older adults using the same UCLA Loneliness Scale (54). To facilitate interpretations and compare model performance with those of previous studies on automatic assessment of loneliness by using daily behavioral data (21–23), we focused on developing models using only speech features without other demographic information such as gender.

The regression and binary-classification models were built by using multiple types of machine learning models by combining them with automatic feature selection using a sequential forward selection algorithm. Model performances were evaluated by 20 iterations of 10-fold cross-validation methods. In the ten-fold cross-validation, the model was trained using 90% of the data (the "training set") while the remaining 10% was used for testing. The process was repeated ten times to cover the entire span of the data, and the average model performance was calculated. Regression model performances were evaluated by using $R^2$, explained variance (EV), mean absolute error (MAE),

[2]https://brookemosby.github.io/Signal_Analysis/

[3]https://ja.osdn.net/projects/ipadic/

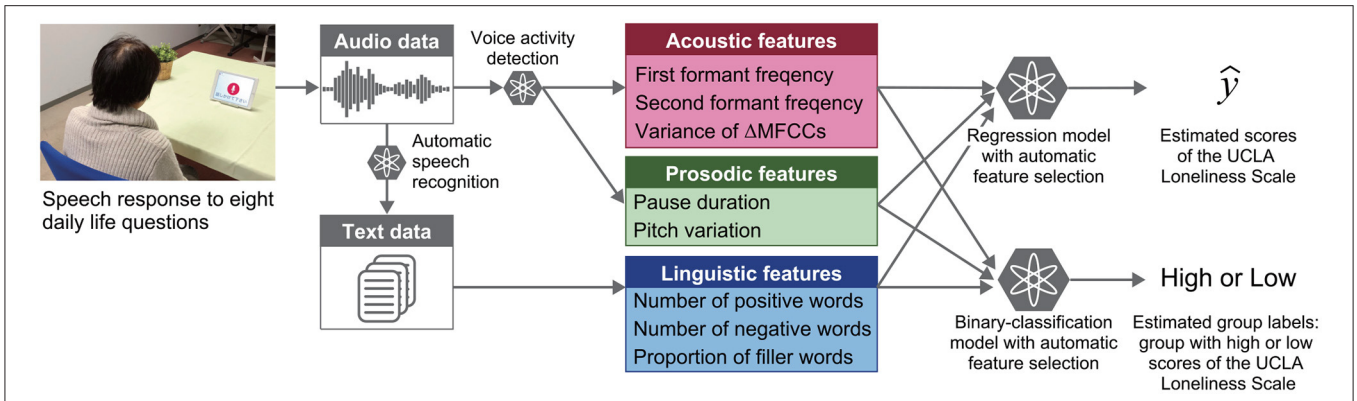


**FIGURE 2 |** Overview of automatic analysis pipeline for estimating loneliness scores and for detecting individuals with high loneliness scores from speech responses to daily life questions.

and root mean square error (RMSE). MAE and RMSE were calculated by the following equations: $\text{MAE} = 1/n \sum_{i=1}^{n} |y_i - \hat{y}_i|$ and $\text{RMSE} = \sqrt{1/n \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$, where $y_i$ and $\hat{y}_i$ are actual and estimated scores of the UCLA Loneliness Scale for the $i$-th participant, respectively. Binary-classification model performances were evaluated by accuracy, sensitivity, specificity, and F1 score. The total number of input features to the regression and binary-classification models was set to 48 so that the number of acoustic, prosodic, and linguistic features would be the same (i.e., $16 \times 3 = 48$). The inputs of acoustic features were selected on the basis of absolute values of Spearman correlation coefficients with scores of the UCLA Loneliness Scale in the training set.

The machine learning models included k-nearest neighbors (70), random forest (RF) (71), and support vector machine (SVM) (72). The parameters that we studied were as follows: the number of neighbors for the k-nearest neighbors; the number and the maximum depth of trees for RF; kernel functions, penalty parameter, the parameter associated with the width of the radial basis function (RBF) kernel, class weights for the classification model, and the parameter of the regression model related to the loss function for the SVM. We used algorithms implemented using the Python package scikit-learn (version 0.23.2) and all other parameters were kept at their default values. We performed a grid search and determined the aforementioned parameters.

## 3. RESULTS

The mean score for the UCLA Loneliness Scale was 37.0 (SD: 8.6; range for participants, 20–63; possible range, 20–80; **Figure 3**). The Cronbach's alpha coefficient was 0.89. The cut-off score for dividing participants into two groups with low and high loneliness scores for building a binary-classification model was 46 points, which was determined by the mean + 1SD of our participants' scores in the same manner as that of a previous study (54). In our sample, ten older adults (6 males, 4 females; 18% of the participants) scored equal to or greater than the cut-off score. They were similar values reported in the previous study investigating 173 older adults: cut-off score was 48 points, and

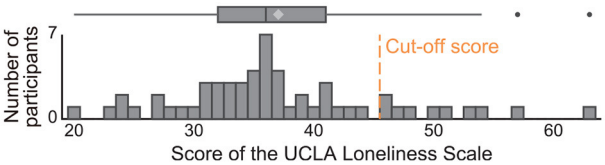


**FIGURE 3 |** Histogram of scores of the UCLA Loneliness Scale for study participants. Cut-off score was determined by using the mean + 1SD of our participants' scores and 46 points. In our sample, 10 older adults (18% of the participants) scored equal to or greater than the cut-off score.

19% of their participants scored equal to or greater than their cut-off score (54). In regard to the speech data, we obtained an average of 319.7 sec (SD: 108.5) of speech responses to the eight daily life questions. The average duration of responses to each question varied between 4.2 and 75.4 sec.

We first investigated associations of loneliness scores with each speech feature. Consequently, we found 21 speech features weakly correlated with loneliness scores (Spearman correlation $\rho$; $0.26 < |\rho| < 0.41$; $P < 0.05$; **Supplementary Table 3**): 15 acoustic features (13 features related to variance of ΔMFCCs and 2 features related to F2), 3 prosodic features (pitch variation and two features related to pause duration), 3 linguistic features (positive word frequency and two features related to filler words). With increasing loneliness scores, the acoustic features showed decreased F2 and reduced the variance of ΔMFCCs, the prosodic features showed decreased pitch variation and increased pause duration, and the linguistic features showed a decrease in the number of positive words and an increase in the proportion of filler words (**Figure 4A** and **Supplementary Table 3**). After controlling for age and sex as potential confounding factors (35, 73, 74), 15 of the 21 speech features remain correlated with loneliness scores (**Supplementary Table 3**).

We next built regression models using speech features to investigate whether speech response to daily life questions could be used for estimating scores of the UCLA Loneliness Scale. The result of iterative ten-fold cross validations showed that the

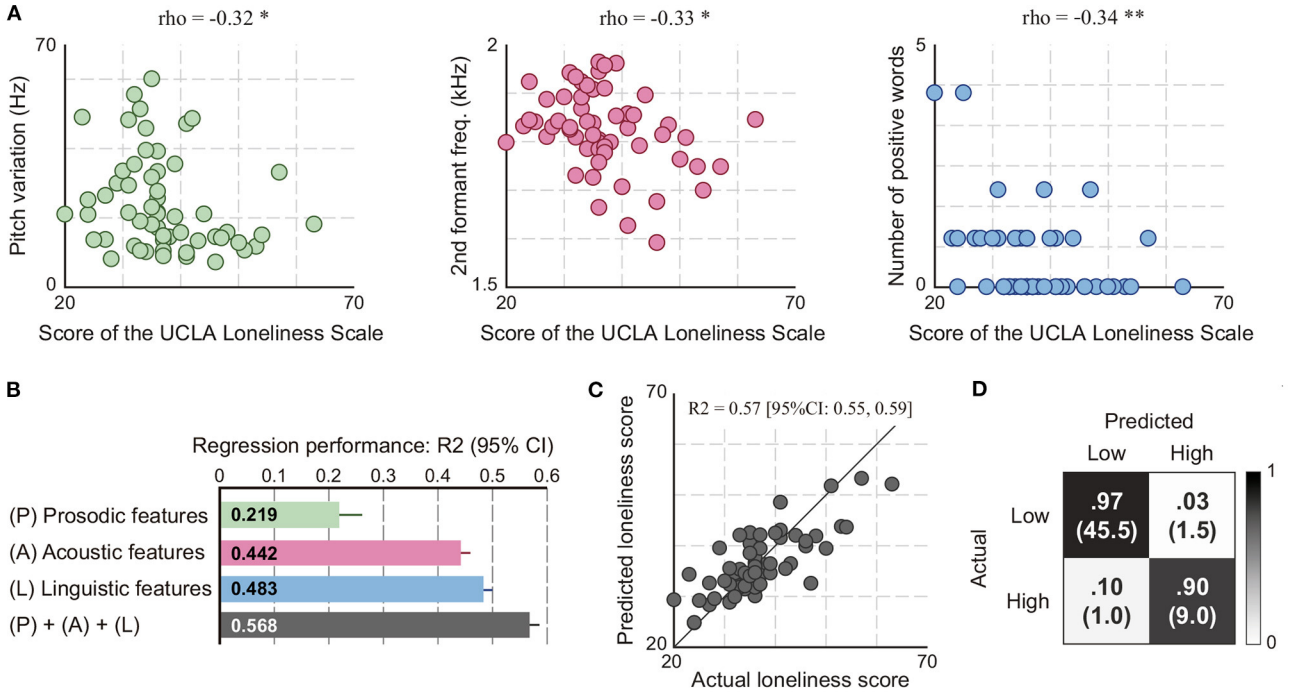


**FIGURE 4 |** Analysis results of the associations of speech responses to eight daily life questions with scores of the UCLA Loneliness Scale. **(A)** Examples of speech features correlated with loneliness scores (Spearman correlation; $^{*}P < 0.05$ and $^{**}P < 0.01$). **(B)** Regression performances of the models using speech features for estimating loneliness scores. **(C)** Actual and predicted loneliness scores by the regression model using acoustic, prosodic, and linguistic features. **(D)** Confusion matrix of the binary-classification model using acoustic, prosodic, and linguistic features for detecting individuals with high loneliness scores. It was obtained using 20 iterations of 10-fold cross-validation. The number in parentheses indicates the mean number of participants among 20 iterations.

model using speech features consisting of acoustic, prosodic, and linguistic features could estimate loneliness scores with an $R^2$ of 0.568 (EV of 0.570, MAE of 4.46, and RMSE of 5.63) (**Figures 4B,C** and **Table 2**). This model was based on an SVM with an RBF kernel using 4 acoustic feature, 1 prosodic feature, and 3 linguistic features selected by the automatic feature selection procedure. The performances of this model calculated separately by sex were $R^2$ of 0.599 (95% CI: 0.579 to 0.620) for women and $R^2$ of 0.511 (95% CI: 0.487–0.535) for men. When building regression models separately by sex, the performances of the model for women and men were $R^2$ of 0.648 (95% CI: 0.628–0.668) and $R^2$ of 0.764 (95% CI: 0.744–0.784), respectively. We also built regression models separately using each acoustic, prosodic, and linguistic feature sets and compared their performances. Consequently, the model using linguistic features had the highest performance with an $R^2$ of 0.483 (EV of 0.484, MAE of 4.75, and RMSE of 6.16) followed by that using acoustic features with an $R^2$ of 0.442 (EV of 0.442, MAE of 4.86, and RMSE of 6.40), and that using prosodic features with an $R^2$ of 0.219 (EV of 0.227, MAE of 5.96, and RMSE of 7.56) (**Figure 4B** and **Table 2**).

We finally investigated whether speech data could be used for detecting individuals with high loneliness scores by building a binary-classification model with speech features. The results of iterative ten-fold cross validations showed that the model using acoustic, prosodic, and linguistic features could detect individuals with high loneliness scores at 95.6% accuracy (90.0% sensitivity, 96.8% specificity, and 87.9% F1 score) (**Figure 4D** and **Table 3**). This model was based on an SVM with an RBF kernel using 5 acoustic features, 2 prosodic features, and 3 linguistic features. The performances of this model calculated separately by sex were 98.3% accuracy (95% CI: 97.1–99.5) for women and 92.6% accuracy (95% CI: 92.6–92.6) for men. When building binary-classification models separately by sex, the performances of the model for women and men were 100.0% accuracy (95% CI: 100.0–100.0) and 98.9% accuracy (95% CI: 98.1–99.7), respectively. For the models using the acoustic, prosodic, and linguistic feature sets separately, the results showed similar trends with those of the regression models: the model using acoustic features had the highest accuracy at 92.7% (95% CI: 92.2–93.3), followed by that using linguistic features with 91.0% accuracy (95% CI: 90.7–91.3), and that using prosodic features with 87.7% accuracy (95% CI: 87.7–87.7) (**Table 3**).

## 4. DISCUSSION

We collected speech responses to eight daily life questions with our tablet-based application and investigated the associations of speech features automatically extracted from audio data of these speech responses with scores the UCLA Loneliness Scale. Our first main finding was that acoustic, prosodic, and linguistic characteristics each may have features affected by loneliness levels in older adults. Through correlation analysis, we could

**TABLE 2 |** Regression model performance of speech features predicting loneliness scores resulting from 20 iterations of 10-fold cross validation.

| Input variables | $R^2$ | EV | MAE | RMSE |
|---|---|---|---|---|
| (P) Prosodic | 0.219 [0.177, 0.261] | 0.227 [0.185, 0.268] | 5.96 [5.79, 6.12] | 7.56 [7.36, 7.76] |
| (A) Acoustic | 0.442 [0.424, 0.459] | 0.442 [0.425, 0.460] | 4.86 [4.78, 4.93] | 6.40 [6.30, 6.50] |
| (L) Linguistic | 0.483 [0.467, 0.500] | 0.484 [0.468, 0.501] | 4.75 [4.66, 4.83] | 6.16 [6.06, 6.26] |
| (P) + (A) + (L) | 0.568 [0.550, 0.586] | 0.570 [0.553, 0.587] | 4.46 [4.36, 4.57] | 5.63 [5.51, 5.74] |

*Each value indicates the average value across 20 iterations with 95% confidence interval. EV, explain variance; MAE, mean absolute error; RMSE, root mean square error.*

**TABLE 3 |** Classification model performance of speech features detecting individuals with high loneliness level resulting from 20 iterations of 10-fold cross validation.

| Input variables | Accuracy (%) | Sensitivity (%) | Specificity (%) | F1 score (%) |
|---|---|---|---|---|
| (P) Prosodic | 87.7 [87.7, 87.7] | 30.0 [30.0, 30.0] | 100.0 [100.0, 100.0] | 46.2 [46.2, 46.2] |
| (L) Linguistic | 91.0 [90.7, 91.3] | 60.0 [60.0, 60.0] | 97.6 [97.2, 97.9] | 70.0 [69.3, 70.7] |
| (A) Acoustic | 92.7 [92.2, 93.3] | 70.0 [70.0, 70.0] | 97.6 [96.9, 98.2] | 77.2 [75.9, 78.5] |
| (P) + (L) + (A) | 95.6 [95.0, 96.2] | 90.0 [90.0, 90.0] | 96.8 [96.1, 97.6] | 87.9 [86.4, 89.4] |

*Each value indicates the average value across 20 iterations with 95% confidence interval.*

find acoustic, prosodic, and linguistic features correlated with loneliness scores. Our second finding was that the combination of acoustic, prosodic, and linguistic features could achieve high performances both for estimating loneliness scores and for detecting individuals with high loneliness scores. These findings showed the possibility of the use of speech responses usually observed in daily conversations (e.g., responses regarding today's feeling and future travel plans) for automatically assessing loneliness in older adults, which can help to promote future efforts toward developing applications for assessing and monitoring loneliness in older adults.

We found speech features correlated with loneliness scores in acoustic, prosodic, and linguistic characteristics in speech response to daily life questions. With increasing loneliness scores, speech responses tended to have less inflections and longer pauses in prosodic features; reduced second formant frequencies and variances of the speech spectrum (ΔMFCCs) in acoustic features; and fewer positive words and more filler words in linguistic features. All these trends in their changes were consistent with those observed in individuals with changes in emotional states and mental health conditions, especially those reported in previous studies on depressed speech [for F2 (35, 36, 40); for the variance of ΔMFCCs (49, 50); for pitch variation (41, 50); for pauses (34, 38); for positive words (33, 37); for filler words (55–57)]. This result may be reasonable because loneliness and depression are different constructs but closely correlated with each other (11). Considering similarities between loneliness and depression, including in their effects on speech characteristics, further studies including longitudinal data collection are required to ensure that the speech changes are due to either loneliness or depression and to identify changes in speech features particularly sensitive to loneliness rather than depression or mood. The potential mechanisms underlying the effects of loneliness on speech characteristics are poorly understood (75–77), but we may be able to explain them from the perspective of the associations of chronic psychological stress. Lonely individuals reported experiencing a great number of chronic stressors (78) and were more likely to perceive daily events as stressful (79, 80). Further, empirical studies suggested the associations of loneliness with exaggerated stress responses (75). These changes may potentially affect processes involved in the phonation and articulation muscular systems and speech production via changes to the somatic and autonomic nervous systems, which may result in producing measurable acoustic, prosodic, and linguistic changes (35). In this study, we observed the effects of loneliness on speech responses to daily life questions that were not designed to induce emotional responses, although we did not test effects of these questions on mood. This result suggest that loneliness may affect even daily speech through chronic psychological stress, although further research is needed. In addition, due to a complexity of loneliness, there are multiple scales for measuring loneliness from different viewpoints. For example, the UCLA Loneliness Scale is used in an attempt to measure loneliness as a global, unidimensional construct, while the de Jong Gierveld Loneliness Scale (81) is used to attempt to measure it as multifaceted phenomenon with separate emotional and social components (17). Therefore, investigating speech changes related to different loneliness scales may provide useful insights to deepening our understanding of the wide and complex profiles of loneliness.

The cross-validation results showed that the regression and binary-classification models using speech features could estimate loneliness scores with an $R^2$ of 0.57 (Pearson correlation of 0.76) and detect individuals with high loneliness scores with 95.6% accuracy, respectively. Previous studies on assessments of loneliness using behavioral data focused on behavioral patterns such as phone usage, time out-of-home, step counts, and sleep duration, and they reported a regression performance with a correlation of 0.48 (21) and classification accuracy ranging from 80.2 to 91.7% accuracy (22, 23). Compared with their performance, both regression and classification models in our study showed better performances. Although there are differences in the methodology such as target population, cut-off scores, and number of samples, this improvement of model performance might come from the use of speech data instead

of the behavioral patterns investigated in previous studies. Aligning with previous studies on the associations of speech with depression and suicidality, our results suggest that speech may be one of the key behavioral markers for automatically detecting and predicting changes in mental health conditions including loneliness in older adults. One of our contributions lies in providing the first empirical evidence showing the feasibility of using the automatic analysis of speech for detecting changes due to loneliness in older adults. In addition, many recent studies have explored the use of speech data for healthcare applications for monitoring various types of health statuses in older adults, for example, for detecting cognitive impairments (31, 82–84) and Alzheimer's disease (26, 28, 29, 32, 85–90), for detecting depression (38, 91, 92), and for predicting driving risks (30). Together with these previous studies, our results may help future efforts toward developing applications using speech data for automatically and simultaneously monitoring various types of health statuses including loneliness. On the other hand, these applications have raised numerous ethical concerns including informed consent, especially when using passive data, i.e., data generated without the active participation of the individual (e.g., GPS, accelerometer data, phone call) (93). Thus, the ethical implications need to be considered parallel to the development of these healthcare applications.

Comparing the model performances among speech feature types showed that acoustic features could achieve high accuracies comparable with linguistic features. In particular, for detecting individuals with high loneliness scores, the binary-classification model using acoustic features achieved the best accuracy. Although user-interface studies reported that voice input was effective and was preferable as an input modality for older adults (94–96), other studies reported that the performance of automatic speech recognition tended to be worse in older adults than in other age groups (97, 98). Because we analyzed only speech data collected in a lab setting, we may need to consider the possibility that there would be a situation where automatic speech recognition would be difficult to use for extracting linguistic features from speech data collected in living situations. In that case, our results may suggest that an approach focusing on developing a model for detecting individuals with high loneliness scores using paralinguistic features, especially acoustic features, would be useful and effective.

There were several limitations in this study. First, the number of questions was small and limited. Although our study provided the first empirical evidence of the usefulness of daily life questions for assessing loneliness in older adults, it still remains uninvestigated what kinds of daily conversations could particularly elicit changes associated with loneliness. To investigate this, data collection at home would be a good way to collect many speech responses by having participants using applications on a daily basis. Second, in terms of statistical analysis of correlation coefficients, we did not adjust for multiple comparisons across speech features due to the exploratory nature of this investigation. In addition, the results of a *post hoc* power analysis revealed that speech features except for the variance of ΔMFCC14 did not reach a power of 0.8 with a significance level of 0.05 (two-sided). A future study on larger samples should confirm our result about the effects of loneliness on speech characteristics. Third, residual confounding such as medication can still exist in addition to age and sex considered in the analysis (35). We also excluded individuals with diagnoses of mental illness such as major depression, because they may affect speech. Therefore, a further study using large samples with these confounding factors is required to further confirm our results about the usefulness of speech analysis for assessing loneliness. Fourth, the number of participants with higher loneliness scores was small and limited. This might affect the generalizability of our results. Finally, the results were obtained by analyzing speech data in Japanese. Thus, we need to investigate speech data in other languages to confirm our results regarding the usefulness of speech responses to daily life questions for assessing loneliness.

In summary, we provide the first empirical results suggesting the possibility of using the automatic analysis of speech responses to daily life questions for estimating loneliness scores and detecting individuals with high loneliness scores. The results presented in this work indicate that it could be feasible to automatically assess loneliness in older adults from daily conversational data, which can help promote future efforts toward the early detection and intervention for mitigating loneliness.

## DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available but derived and supporting data may be available from the corresponding author on reasonable request and with permission from the Ethics Committee, University of Tsukuba Hospital. Requests to access the datasets should be directed to Yasunori Yamada, ysnr@jp.ibm.com; Kaoru Shinkawa, kaoruma@jp.ibm.com.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee, University of Tsukuba Hospital. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

YY and KS contributed to conception and design of the study and performed analysis and wrote the manuscript. YY, KS, and MN conducted the experiments. All authors have approved the final version.

## FUNDING

This work was supported by JSPS KAKENHI grant no. 19H01084.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.712251/full#supplementary-material

## REFERENCES


1. Luo Y, Hawkley LC, Waite LJ, Cacioppo JT. Loneliness, health, and mortality in old age: a national longitudinal study. *Soc Sci Med*. (2012) 74:907–14. doi: 10.1016/j.socscimed.2011.11.028
2. Rico-Uribe LA, Caballero FF, Martín-María N, Cabello M, Ayuso-Mateos JL, Miret M. Association of loneliness with all-cause mortality: a meta-analysis. *PLoS ONE*. (2018) 13:e0190033. doi: 10.1371/journal.pone.0190033
3. Perissinotto CM, Cenzer IS, Covinsky KE. Loneliness in older persons: a predictor of functional decline and death. *Arch Intern Med*. (2012) 172:1078–84. doi: 10.1001/archinternmed.2012.1993
4. Cacioppo JT, Hughes ME, Waite LJ, Hawkley LC, Thisted RA. Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses. *Psychol Aging*. (2006) 21:140. doi: 10.1037/0882-7974.21.1.140
5. Heinrich LM, Gullone E. The clinical significance of loneliness: a literature review. *Clin Psychol Rev*. (2006) 26:695–718. doi: 10.1016/j.cpr.2006.04.002
6. Tilvis RS, Kähönen-Väre MH, Jolkkonen J, Valvanne J, Pitkala KH, Strandberg TE. Predictors of cognitive decline and mortality of aged people over a 10-year period. *J Gerontol A Biol Sci Med Sci*. (2004) 59:M268–74. doi: 10.1093/gerona/59.3.M268
7. Donovan NJ, Wu Q, Rentz DM, Sperling RA, Marshall GA, Glymour MM. Loneliness, depression and cognitive function in older US adults. *Int J Geriatr Psychiatry*. (2017) 32:564–73. doi: 10.1002/gps.4495
8. Wilson RS, Krueger KR, Arnold SE, Schneider JA, Kelly JF, Barnes LL, et al. Loneliness and risk of Alzheimer disease. *Arch Gen Psychiatry*. (2007) 64:234–40. doi: 10.1001/archpsyc.64.2.234
9. Sundström A, Adolfsson AN, Nordin M, Adolfsson R. Loneliness increases the risk of all-cause dementia and Alzheimer's disease. *J Gerontol B Psychol Sci Soc Sci*. (2020) 75:919–26. doi: 10.1093/geronb/gbz139
10. Holt-Lunstad J, Smith TB, Baker M, Harris T, Stephenson D. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspect Psychol Sci*. (2015) 10:227–37. doi: 10.1177/1745691614568352
11. Cacioppo S, Grippo AJ, London S, Goossens L, Cacioppo JT. Loneliness: clinical import and interventions. *Perspect Psychol Sci*. (2015) 10:238–49. doi: 10.1177/1745691615570616
12. Jeste DV, Lee EE, Cacioppo S. Battling the modern behavioral epidemic of loneliness: suggestions for research and interventions. *JAMA Psychiatry*. (2020) 77:553–4. doi: 10.1001/jamapsychiatry.2020.0027
13. Peplau L, Russell D, Heim M. The experience of loneliness. In: Frieze IH, Bar-Tal D, Carroll JS, editors. *New Approaches to Social Problems: Applications of Attribution Theory*. San Francisco, CA (1979).
14. Savikko N, Routasalo P, Tilvis RS, Strandberg TE, Pitkälä KH. Predictors and subjective causes of loneliness in an aged population. *Arch Gerontol Geriatr*. (2005) 41:223–33. doi: 10.1016/j.archger.2005.03.002
15. Theeke LA. Predictors of loneliness in US adults over age sixty-five. *Arch Psychiatr Nurs*. (2009) 23:387–96. doi: 10.1016/j.apnu.2008.11.002
16. Hawkley LC, Cacioppo JT. Loneliness matters: a theoretical and empirical review of consequences and mechanisms. *Ann Behav Med*. (2010) 40:218–27. doi: 10.1007/s12160-010-9210-8
17. Ong AD, Uchino BN, Wethington E. Loneliness and health in older adults: a mini-review and synthesis. *Gerontology*. (2016) 62:443–449. doi: 10.1159/000441651
18. Pinquart M, Sorensen S. Influences on loneliness in older adults: a meta-analysis. *Basic Appl Soc Psychol*. (2001) 23:245–66. doi: 10.1207/S15324834BASP2304_2
19. Shiovitz-Ezra S, Ayalon L. Use of direct versus indirect approaches to measure loneliness in later life. *Res Aging*. (2012) 34:572–91. doi: 10.1177/0164027511423258
20. Russell DW. UCLA loneliness scale (Version 3): reliability, validity, and factor structure. *J Pers Assess*. (1996) 66:20–40. doi: 10.1207/s15327752jpa6601_2
21. Austin J, Dodge HH, Riley T, Jacobs PG, Thielke S, Kaye J. A smart-home system to unobtrusively and continuously assess loneliness in older adults. *IEEE J Transl Eng Health Med*. (2016) 4:1–11. doi: 10.1109/JTEHM.2016.2579638
22. Doryab A, Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data. *JMIR Mhealth Uhealth*. (2019) 7:e13209. doi: 10.2196/13209
23. Sanchez W, Martinez A, Campos W, Estrada H, Pelechano V. Inferring loneliness levels in older adults from smartphones. *J Ambient Intell Smart Environ*. (2015) 7:85–98. doi: 10.3233/AIS-140297
24. Manfredi C, Lebacq J, Cantarella G, Schoentgen J, Orlandi S, Bandini A, et al. Smartphones offer new opportunities in clinical voice research. *J Voice*. (2017) 31:111.e1-111.e7. doi: 10.1016/j.jvoice.2015.12.020
25. Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *NPJ Digit Med*. (2019) 2:1–9. doi: 10.1038/s41746-019-0084-2
26. Yamada Y, Shinkawa K, Shimmei K. Atypical repetition in daily conversation on different days for detecting alzheimer disease: evaluation of phone-call data from a regular monitoring service. *JMIR Mental Health*. (2020) 7:e16790. doi: 10.2196/16790
27. Faurholt-Jepsen M, Busk J, Frost M, Vinberg M, Christensen EM, Winther O, et al. Voice analysis as an objective state marker in bipolar disorder. *Transl Psychiatry*. (2016) 6:e856. doi: 10.1038/tp.2016.123
28. König A, Satt A, Sorin A, Hoory R, Derreumaux A, David R, et al. Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Curr Alzheimer Res*. (2018) 15:120–9. doi: 10.2174/1567205014666170829111942
29. Yamada Y, Shinkawa K, Kobayashi M, Nishimura M, Nemoto M, Tsukada E, et al. Tablet-based automatic assessment for early detection of alzheimer's disease using speech responses to daily life questions. *Front Digit Health*. (2021) 3:30. doi: 10.3389/fdgth.2021.653904
30. Yamada Y, Shinkawa K, Kobayashi M, Takagi H, Nemoto M, Nemoto K, et al. Using speech data from interactions with a voice assistant to predict the risk of future accidents for older drivers: prospective cohort study. *J Med Internet Res*. (2021) 23:e27667. doi: 10.2196/27667
31. Kobayashi M, Kosugi A, Takagi H, Nemoto M, Nemoto K, Arai T, et al. Effects of age-related cognitive decline on elderly user interactions with voice-based dialogue systems. In: *IFIP Conference on Human-Computer Interaction*. Cyprus (2019). p. 53–74.
32. Hall AO, Shinkawa K, Kosugi A, Takase T, Kobayashi M, Nishimura M, et al. Using tablet-based assessment to characterize speech for individuals with dementia and mild cognitive impairment: preliminary results. *AMIA Jt Summits Transl Sci Proc*. (2019) 2019:34.
33. Scibelli F. *Detection of Verbal and Nonverbal Speech Features as Markers of Depression: Results of Manual Analysis and Automatic Classification*. Napoli: Università degli Studi di Napoli Federico II (2019).
34. Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. *Biol Psychiatry*. (2012) 72:580–7. doi: 10.1016/j.biopsych.2012.03.015
35. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Commun*. (2015) 71:10–49. doi: 10.1016/j.specom.2015.03.004
36. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J Neurolinguistics*. (2007) 20:50–64. doi: 10.1016/j.jneuroling.2006.04.001
37. Rude S, Gortner EM, Pennebaker J. Language use of depressed and depression-vulnerable college students. *Cogn Emot*. (2004) 18:1121–33. doi: 10.1080/02699930441000030
38. Alpert M, Pouget ER, Silva RR. Reflections of depression in acoustic measures of the patient's speech. *J Affect Disord*. (2001) 66:59–69. doi: 10.1016/S0165-0327(00)00335-9
39. Sobin C, Alpert M. Emotion in speech: the acoustic attributes of fear, anger, sadness, and joy. *J Psycholinguist Res*. (1999) 28:347–65.
40. Flint AJ, Black SE, Campbell-Taylor I, Gailey GF, Levinton C. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *J Psychiatr Res*. (1993) 27:309–19. doi: 10.1016/0022-3956(93)90041-Y
41. Kuny S, Stassen H. Speaking behavior and voice sound characteristics in depressive patients during recovery. *J Psychiatr Res*. (1993) 27:289–307. doi: 10.1016/0022-3956(93)90040-9

42. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes M. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans Biomed Eng*. (2000) 47:829–37. doi: 10.1109/10.846676
43. Guidi A, Vanello N, Bertschy G, Gentili C, Landini L, Scilingo EP. Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients. *Biomed Signal Process Control*. (2015) 17:29–37. doi: 10.1016/j.bspc.2014.10.011
44. Scherer KR. Vocal affect expression: a review and a model for future research. *Psychol Bull*. (1986) 99:143–65. doi: 10.1037/0033-2909.99.2.143
45. Ozdas A, Shiavi RG, Silverman SE, Silverman MK, Wilkes DM. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Trans Biomed Eng*. (2004) 51:1530–40. doi: 10.1109/TBME.2004.827544
46. Low LSA, Maddage NC, Lech M, Sheeber LB, Allen NB. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Trans Biomed Eng*. (2010) 58:574–86. doi: 10.1109/TBME.2010.2091640
47. Helfer BS, Quatieri TF, Williamson JR, Mehta DD, Horwitz R, Yu B. Classification of depression state based on articulatory precision. In: *Interspeech*. Lyon (2013). p. 2172–6.
48. Place S, Blanch-Hartigan D, Rubin C, Gorrostieta C, Mead C, Kane J, et al. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *J Med Internet Res*. (2017) 19:e75. doi: 10.2196/jmir.6678
49. Cummins N, Epps J, Sethu V, Breakspear M, Goecke R. Modeling spectral variability for the classification of depressed speech. *In: Interspeech*. Lyon (2013). p. 857–61.
50. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J. Analysis of acoustic space variability in speech affected by depression. *Speech Commun*. (2015) 75:27–49. doi: 10.1016/j.specom.2015.09.003
51. Masuda Y, Tadaka E, Dai Y. Reliability and validity of the Japanese version of the UCLA loneliness scale version 3 among the older population. *J Jpn Acad Commun Health Nurs*. (2012) 15:25–32. doi: 10.1186/s12905-019-0792-4
52. Donovan NJ, Okereke OI, Vannini P, Amariglio RE, Rentz DM, Marshall GA, et al. Association of higher cortical amyloid burden with loneliness in cognitively normal older adults. *JAMA Psychiatry*. (2016) 73:1230–7. doi: 10.1001/jamapsychiatry.2016.2657
53. Lee EE, Depp C, Palmer BW, Glorioso D, Daly R, Liu J, et al. High prevalence and adverse health effects of loneliness in community-dwelling adults across the lifespan: role of wisdom as a protective factor. *Int Psychogeriatr*. (2019) 31:1447. doi: 10.1017/S1041610218002120
54. Adams KB, Sanders S, Auth E. Loneliness and depression in independent living retirement communities: risk and resilience factors. *Aging Ment Health*. (2004) 8:475–85. doi: 10.1080/13607860410001725054
55. Stasak B. *An Investigation of Acoustic, Linguistic, and Affect Based Methods for Speech Depression Assessment*. Sydney, NSW: UNSW Sydney (2018).
56. Stasak B, Epps J, Cummins N. Depression prediction via acoustic analysis of formulaic word fillers. *Polar*. (2016) 77:230.
57. Morales M, Scherer S, Levitan R. A linguistically-informed fusion approach for multimodal depression detection. In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. New Orleans, LA (2018). p. 13–24.
58. Skodda S, Grönheit W, Schlegel U. Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease. *PLoS ONE*. (2012) 7:e32132. doi: 10.1371/journal.pone.0032132
59. Titze IR. *Principles of Voice Production*. Allyn & Bacon (1994).
60. Roy N, Nissen SL, Dromey C, Sapir S. Articulatory changes in muscle tension dysphonia: evidence of vowel space expansion following manual circumlaryngeal therapy. *J Commun Disord*. (2009) 42:124–35. doi: 10.1016/j.jcomdis.2008.10.001
61. Mermelstein P. Distance measures for speech recognition, psychological and instrumental. *Pattern Recogn Artif Intell*. (1976) 116:374–88.
62. Koolagudi SG, Rao KS. Emotion recognition from speech: a review. *Int J Speech Technol*. (2012) 15:99–117. doi: 10.1007/s10772-011-9125-1
63. Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. *Int J Speech Technol*. (2018) 21:93–120. doi: 10.1007/s10772-018-9491-z
64. Pan Z, Gui C, Zhang J, Zhu J, Cui D. Detecting manic state of bipolar disorder based on support vector machine and Gaussian mixture model using spontaneous speech. *Psychiatry Investig*. (2018) 15:695. doi: 10.30773/pi.2017.12.15
65. Sturim D, Torres-Carrasquillo PA, Quatieri TF, Malyska N, McCree A. Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In: *Twelfth Annual Conference of the International Speech Communication Association*. Florence (2011). doi: 10.21437/Interspeech.2011-746
66. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, et al. librosa: Audio and music signal analysis in python. In: *Proceedings of the 14th Python in Science Conference. Vol. 8*. Austin, TX (2015). p. 18–25.
67. Yang Y, Fairbairn C, Cohn JF. Detecting depression severity from vocal prosody. *IEEE Trans Affect Comput*. (2012) 4:142–50. doi: 10.1109/T-AFFC.2012.38
68. Kobayashi N, Inui K, Matsumoto Y, Tateishi K, Fukushima T. Collecting evaluative expressions for opinion extraction. In: *International Conference on Natural Language Processing*. Hainan Island: Springer (2004). p. 596–605.
69. Higashiyama M, Inui K, Matsumoto Y. Learning sentiment of nouns from selectional preferences of verbs and adjectives. In: *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*. Tokyo (2008). p. 584–7.
70. Goldberger J, Hinton GE, Roweis S, Salakhutdinov RR. Neighbourhood components analysis. In: *Proceedings of the 17th International Conference on Neural Information Processing Systems*. Vancouver, BC (2004). p. 513–20.
71. Breiman L. Random forests. *Mach Learn*. (2001) 45:5–32. doi: 10.1023/A:1010933404324
72. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY (1992). p. 144–52.
73. Badal VD, Graham SA, Depp CA, Shinkawa K, Yamada Y, Palinkas LA, et al. Prediction of loneliness in older adults using natural language processing: exploring sex differences in speech. *Am J Geriatr Psychiatry*. (2021) 29:853–66. doi: 10.1016/j.jagp.2020.09.009
74. Badal VD, Nebeker C, Shinkawa K, Yamada Y, Rentscher KE, Kim HC, et al. Do Words matter? Detecting social isolation and loneliness in older adults using natural language processing. *Front Psychiatry*. (2021) 12:728732. doi: 10.3389/fpsyt.2021.728732
75. Brown EG, Gallagher S, Creaven AM. Loneliness and acute stress reactivity: a systematic review of psychophysiological studies. *Psychophysiology*. (2018) 55:e13031. doi: 10.1111/psyp.13031
76. Boss L, Kang DH, Branson S. Loneliness and cognitive function in the older adult: a systematic review. *Int Psychogeriatr*. (2015) 27:541–53. doi: 10.1017/S1041610214002749
77. National Academies of Sciences, Engineering, Medicine. *Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System*. Washington, DC: National Academies Press (2020).
78. Hawkley LC, Cacioppo JT. Aging and loneliness: downhill quickly? *Curr Direct Psychol Sci*. (2007) 16:187–91. doi: 10.1111/j.1467-8721.2007.00501.x
79. Turner JR. Individual differences in heart rate response during behavioral challenge. *Psychophysiology*. (1989) 26:497–505. doi: 10.1111/j.1469-8986.1989.tb00701.x
80. Cacioppo JT. Social neuroscience: autonomic, neuroendocrine, and immune responses to stress. *Psychophysiology*. (1994) 31:113–28. doi: 10.1111/j.1469-8986.1994.tb01032.x
81. De Jong-Gierveld J, Kamphuls F. The development of a Rasch-type loneliness scale. *Appl Psychol Meas*. (1985) 9:289–99. doi: 10.1177/014662168500900307
82. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Alzheimer's dementia recognition through spontaneous speech: the ADReSS Challenge. *arXiv preprint* arXiv:200406833. (2020) doi: 10.21437/Interspeech.2020-2571
83. Mueller KD, Koscik RL, Hermann BP, Johnson SC, Turkstra LS. Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. *Front Aging Neurosci*. (2018) 9:437. doi: 10.3389/fnagi.2017.00437
84. Roark B, Mitchell M, Hosom JP, Hollingshead K, Kaye J. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Trans Audio Speech Lang Process*. (2011) 19:2081–2090. doi: 10.1109/TASL.2011.2112351
85. Gosztolya G, Vincze V, Tóth L, Pákáski M, Kálmán J, Hoffmann I. Identifying mild cognitive impairment and mild Alzheimer's disease based on

spontaneous speech using ASR and linguistic features. *Comput Speech Lang*. (2019) 53:181–97. doi: 10.1016/j.csl.2018.07.007
86. Hernández-Domínguez L, Ratté S, Sierra-Martínez G, Roche-Bergua A. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimers Dement*. (2018) 10:260–8. doi: 10.1016/j.dadm.2018.02.004
87. Beltrami D, Gagliardi G, Rossini Favretti R, Ghidoni E, Tamburini F, Calzà L. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Front Aging Neurosci*. (2018) 10:369. doi: 10.3389/fnagi.2018.00369
88. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis*. (2016) 49:407–422. doi: 10.3233/JAD-150520
89. König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement (Amst)*. (2015) 1:112–24. doi: 10.1016/j.dadm.2014.11.012
90. Yamada Y, Shinkawa K, Kobayashi M, Caggiano V, Nemoto M, Nemoto K, et al. Combining multimodal behavioral data of gait, speech, and drawing for classification of Alzheimer's disease and mild cognitive impairment. *J Alzheimers Dis*. (2021) 84:315–27. doi: 10.3233/JAD-210684
91. Sanchez MH, Vergyri D, Ferrer L, Richey C, Garcia P, Knoth B, et al. Using prosodic and spectral features in detecting depression in elderly males. In: *Twelfth Annual Conference of the International Speech Communication Association*. Florence (2011).
92. Stasak B, Epps J, Goecke R. Automatic depression classification based on affective read sentences: opportunities for text-dependent analysis. *Speech Commun*. (2019) 115:1–14. doi: 10.1016/j.specom.2019.10.003
93. Maher NA, Senders JT, Hulsbergen AF, Lamba N, Parker M, Onnela JP, et al. Passive data collection and use in healthcare: a systematic review of ethical issues. *Int J Med Inform*. (2019) 129:242–7. doi: 10.1016/j.ijmedinf.2019.06.015
94. Smith AL, Chaparro BS. Smartphone text input method performance, usability, and preference with younger and older adults. *Hum Factors*. (2015) 57:1015–28. doi: 10.1177/0018720815575644
95. Liu YC, Chen CH, Lin YS, Chen HY, Irianti D, Jen TN, et al. Design and usability evaluation of mobile voice-added food reporting for elderly people: randomized controlled trial. *JMIR mHealth uHealth*. (2020) 8:e20317. doi: 10.2196/20317
96. Stigall B, Waycott J, Baker S, Caine K. Older adults' perception and use of voice user interfaces: a preliminary review of the computing literature. In: *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. Fremantle, WA (2019). p. 423–7.
97. Werner L, Huang G, Pitts BJ. automated speech recognition systems and older adults: a literature review and synthesis. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 63*. Seattle, WA (2019). p. 42–6.
98. Portet F, Vacher M, Golanski C, Roux C, Meillon B. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. *Pers Ubiquit Comput*. (2013) 17:127–4. doi: 10.1007/s00779-011-0470-5

**Conflict of Interest:** YY and KS are employed by the IBM Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 22 December 2021
doi: 10.3389/fpsyt.2021.739022

# The Effects of Bipolar Disorder Risk on a Mobile Phone Keystroke Dynamics Based Biomarker of Brain Age

*John Zulueta[1], Alexander Pantelis Demos[2], Claudia Vesel[3], Mindy Ross[4], Andrea Piscitello[1], Faraz Hussain[1], Scott A. Langenecker[5], Melvin McInnis[6], Peter Nelson[7], Kelly Ryan[6], Alex Leow[1,3] and Olusola Ajilore[1]**

[1] Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, United States, [2] Department of Psychology, University of Illinois at Chicago, Chicago, IL, United States, [3] Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, United States, [4] Graduate College, University of Illinois at Chicago, Chicago, IL, United States, [5] Department of Psychiatry, The University of Utah, Salt Lake City, UT, United States, [6] Department of Psychiatry, University of Michigan, Ann Arbor, MI, United States, [7] College of Engineering, University of Illinois at Chicago, Chicago, IL, United States

OPEN ACCESS

**Edited by:**
Ellen E. Lee,
University of California, San Diego,
United States

**Reviewed by:**
Rao Kosagisharaf,
Instituto de Investigaciones Científicas y Servicios de Alta Tecnología, Panama
Andrea Iaboni,
University Health Network, Canada

***Correspondence:**
Olusola Ajilore
oajilore@uic.edu

**Specialty section:**
This article was submitted to Aging Psychiatry, a section of the journal Frontiers in Psychiatry

**Received:** 09 July 2021
**Accepted:** 19 November 2021
**Published:** 22 December 2021

**Citation:**
Zulueta J, Demos AP, Vesel C, Ross M, Piscitello A, Hussain F, Langenecker SA, McInnis M, Nelson P, Ryan K, Leow A and Ajilore O (2021) The Effects of Bipolar Disorder Risk on a Mobile Phone Keystroke Dynamics Based Biomarker of Brain Age. Front. Psychiatry 12:739022. doi: 10.3389/fpsyt.2021.739022

**Background:** Research by our group and others have demonstrated the feasibility of using mobile phone derived metadata to model mood and cognition. Given the effects of age and mood on cognitive performance, it was hypothesized that using such data a model could be built to predict chronological age and that differences between predicted age and actual age could be a marker of pathology.

**Methods:** These data were collected via the ongoing BiAffect study. Participants complete the Mood Disorders Questionnaire (MDQ), a screening questionnaire for bipolar disorder, and self-reported their birth year. Data were split into training and validation sets. Features derived from the smartphone kinematics were used to train random forest regression models to predict age. Prediction errors were compared between participants screening positive and negative on the MDQ.

**Results:** Three hundred forty-four participants had analyzable data of which 227 had positive screens for bipolar disorder and 117 had negative screens. The absolute prediction error tended to be lower for participants with positive screens (median 4.50 years) than those with negative screens (median 7.92 years) ($W = 508$, $p = 0.0049$). The raw prediction error tended to be lower for participants with negative screens (median $= -5.95$ years) than those with positive screens (median $= 0.55$ years) ($W = 1{,}037$, $p = 0.037$).

**Conclusions:** The tendency to underestimate the chronological age of participants screening negative for bipolar disorder compared to those screening positive is consistent with the finding that bipolar disorder may be associated with brain changes that could reflect pathological aging. This interesting result could also reflect that those who screen negative for bipolar disorder and who engaged in the study were more likely to have higher premorbid functioning. This work demonstrates that age-related changes may be detected *via* a passive smartphone kinematics based digital biomarker.

**Keywords: digital biomarkers, bipolar disorder, brain age estimation, smartphone, digital phenotyping**

## INTRODUCTION

The development of biomarkers has long been a goal for psychiatry with the hope that these biomarkers may be able to facilitate early detection, diagnosis, and treatment selection—moving the field closer to a paradigm of precision medicine (1, 2). Aging is a heterogenous process associated with increased risk of morbidity and mortality. It has been proposed that differences between biological age and chronological age may be indicative of pathology, and various phenomena have been investigated as potential aging biomarkers including telomere length (3), DNA methylation (4), and features derived from neuroimaging (5, 6).

In previous work, our group identified age associated effects on smartphone typing kinematics—specifically enhancement of the difference between midday typing speed and typing speed at the beginning and end of the day (7). These kinematic data were collected *via* the BiAffect platform which collects such data passively as participants use their smartphones in their day-to-day routines thus enabling the creation of ecologically valid and temporally associated markers of cognitive performance (8).

Bipolar disorder is a psychiatric disorder characterized by recurrent episodes of mood disturbances. It is associated with cognitive deficits during mood episodes, some of which remain during euthymia (9, 10). It has also been proposed that bipolar disorder may exacerbate age associated neuropathologic processes in a phenomenon-termed neuroprogression (11, 12). In this study we investigated the hypothesis that cognitive changes associated with the disorder would be detectable *via* changes in typing kinematics. To do this, we leveraged the BiAffect platform taking advantage of the open enrollment of the project to obtain a large, heterogenous sample. Rather than utilizing binary self-report of diagnosis to distinguish between healthy controls and participants with a bipolar spectrum disorder, we categorized participants based on screening status on the Mood Disorders Questionnaire (MDQ), a screening instrument for bipolar disorder (13) using standard cut-off scores with sensitivity of 61% and specificity of 88% (14). We then examined differences in the performance characteristics of age prediction between the groups to investigate smartphone kinematic based age prediction's utility as a digital biomarker.

## METHODS

Data for this study was collected as part of the open science BiAffect project. This study began in March 2018 with enrollment open to all adults in the United States with an iOS based smartphone that supports the BiAffect app. As of the time of the writing of this manuscript, the study is ongoing. Its protocol has been approved by the University of Illinois at Chicago Institutional Review Board.

The data for this study was collected from March 2018 to February 2021. Subject enrollment and data collection were all performed within the BiAffect app. The app includes modules for participants to complete questionnaires and perform tasks designed to measure aspects of cognitive performance such as response inhibition, set shifting, and reaction time. Its core technology is a custom built keyboard designed to replace the default keyboard. This keyboard collects typing related metadata including the type of keypress event (alphanumeric, backspace, autocorrection, etc.) and the associated timestamp. It does not collect the actual alphanumeric content. These data are then securely uploaded to the study server.

Participants completed the mood disorders questionnaire (MDQ) a screening instrument for bipolar disorder (13). The MDQs were typically completed at study entry. The performance characteristics of this instrument vary based on the setting but in general has been estimated to have relatively high specificity of 88% and adequate sensitivity of 61% (14) with a cut score of $\geq$7. Participants also provided self-reports of whether they have a diagnosis of a bipolar spectrum disorder, their birth year, and their gender. Given the fact that MDQ performance characteristics are better characterized than the reliability of self-reported bipolar disorder diagnosis and given the high rates of participants not disclosing their diagnosis status, it was decided to utilize MDQ status as a feature of interest in our analysis.

Analysis was restricted to participants who had provided at least 12 weeks worth of typing data. This was determined by calculating the median number of keystrokes per day across the entire sample and then filtering accordingly.

### Data Processing and Feature Engineering

Each subject's typing data was tokenized into sessions by grouping together consecutive keystroke events which have differences in timestamps of <5 seconds. Metrics were calculated for each session and then summarized for each subject. For sample entropy calculations the following parameters were used: $m = 2$, $r = 0.2$ * the standard deviation and tau = 1. **Table 1** includes a description of these features. This data processing was performed via the pandas package, version 1.2.4 (15) for Python (Version 3.8.3).

### Model Training and Assessment

Data were split into training and validation sets (75:25). Because of the collinearity among many of the features and the relative robustness of random forest models to collinearity (16), random forest models were used. Random forest models consist of a collection of decision trees whose individual predictions are aggregated to make a single prediction. They are a popular analytic tool in bioinformatics given their ability to model complex interactions (17). Two random forest regression models were trained using the caret and randomForest packages for R (18, 19). The mtry parameter determines the number of features that will be available for use when splitting nodes during the training of the model's decision trees. The mtry parameter was selected via a grid search of values ranging from 1 to 30 features using 10-fold cross-validation with 3 repeats. The mtry value which minimized the Root Mean Square Error (RMSE) was selected as the value used in the final models. The models were constructed in a stepwise fashion with the first model including only typing related features, and the second model included all features from the first with the addition of gender and MDQ screening status.

Each model's performance was assessed using the validation set to calculate RMSE, Breiman's pseudo R-squared, and median

**TABLE 1 |** Model features.

| Feature | Description |
|---|---|
| Mean_keypresses_per_session | Mean number of keypresses per session |
| Median_keypresses_per_session | Median number of keypresses per session |
| Standard_deviation_keypress_per_session | Standard deviation of keypresses per session |
| Median_absolute_deviation_keypress_per_session | Median absolute deviation of keypresses per session |
| Mean_interkey_time_mean | Mean of mean of interkey times per session |
| Median_interkey_time_mean | Median of mean interkey times per session |
| Standard_deviation_interkey_time_mean | Standard deviation of mean interkey times per session |
| Median_absolute_deviation_interkey_time_mean | Median absolute deviation of mean interkey times per session |
| Mean_interkey_time_median | Mean of median of interkey times per session |
| Median_interkey_time_median | Median of median interkey times per session |
| Standard_deviation_interkey_time_median | Standard deviation of median interkey times per session |
| Median_absolute_deviation_interkey_time_median | Median absolute deviation of median interkey times per session |
| Mean_autocorrect_rate | Mean autocorrect rate per session (# of autocorrect events / total # of keystrokes per session) |
| Median_autocorrect_rate | Median autocorrect rate per session (# of autocorrect events / total # of keystrokes per session) |
| Standard_deviation_autocorrect_rate | Standard deviation of autocorrect rate per session (# of autocorrect events / total # of keystrokes per session) |
| Median_absolute_deviation_autocorrect_rate | Median absolute deviation of autocorrect rate per session (# of autocorrect events / total # of keystrokes per session) |
| Mean_backspace_rate | Mean backspace rate per session (# of backspace events / total # of keystrokes per session) |
| Median_backspace_rate | Median backspace rate per session (# of backspace events / total # of keystrokes per session) |
| Standard_deviation_backspace_rate | Standard deviation of backspace rate per session (# of backspace events / total # of keystrokes per session) |
| Median_absolute_deviation_backspace_rate | Median absolute deviation of backspace rate per session (# of backspace events / total # of keystrokes per session) |
| Mean_session_length | Mean length of sessions in seconds |
| Median_session_length | Median length of sessions in seconds |
| Standard_deviation_session_length | Standard deviation of length of sessions in seconds |
| Median_absolute_deviation_session_length | Median absolute deviation of length of sessions in seconds |
| Sample_entropy_keypress | Sample entropy of # of keypresses per sessions |
| Sample_entropy_interkey_time_mean | Sample entropy of mean interkey times per session |
| Sample_entropy_interkey_time_median | Sample entropy of median interkey times per session |
| Sample_entropy_autocorrect_rate | Sample entropy of autocorrect rate per session |
| Sample_entropy_backspace_rate | Sample entropy of backspace rate per session |
| Sample_entropy_session_length | Sample entropy of session length in seconds |

absolute error. Differences in model performance testing were assessed using paired Wilcoxon tests of their absolute errors. Feature importance was assessed using out-of-bag changes in Mean Square Error (MSE). Accumulated Local Effects plots (ALE Plots) (20) were constructed for features which appeared important or interesting. These plots allow the visualization of the effect of individual features and the interaction of two features on the model's prediction. They are especially useful when features may be correlated. Differences within model performance between participants based on MDQ screen status were assessed using Wilcoxon tests comparing raw prediction error scores and absolute prediction error scores. All tests were two sided with a significance level of 0.05. Family-wise error rates were controlled using the Holm-Bonferroni method. All statistical testing was performed in R (Version 4.0.0).

## RESULTS

A total of 344 participants met criteria for inclusion in this analysis: 117 with negative MDQ screens and 227 with positive screens. As summarized in **Table 2**, the group with positive screens tended to have fewer males (Fisher's Exact $p = 0.0042$) and be younger than the positive screen group ($W = 15{,}887$, $p = 0.0028$. Compared to participants with positive MDQ screens, participants with negative screens had a lower rate of reporting a diagnosis of bipolar disorder, a higher rate of reporting no history of bipolar disorder, and also provided no diagnosis history at a lower rate (all comparisons Fisher's Exact $p < 0.001$). The participants with negative screens tended to have lower MDQ scores comparted to those with positive screens ($W = 23{,}322$, $p < 0.001$) and have a greater total number of keypresses ($W = 15{,}098$, $p = 0.037$).

Using the criterion of minimizing RMSE for tuning the mtry parameter of the models, Model 1 which only included the typing metrics had an mtry = 15, and Model 2 which included the features of Model 1 as well as gender and MDQ status had an mtry = 10. **Figure 1** depicts the tuning results.

Using the training set, both Model 1 and Model 2 had an RMSE of 8.7 years. The Breiman Pseudo R-squared values were 0.56 and 0.57 for Models 1 and 2, respectively. The performance

**TABLE 2 |** Subject characteristics.

| | MDQ negative | MDQ positive | |
|---|---|---|---|
| # of participants | 117 | 227 | |
| % not male | 60% | 75% | $p = 0.0042$ |
| Age in years, mean (sd) | 41 (16) | 35 (11) | |
| Age in years, median (mad) | 39 (16) | 33 (12) | W = 15,887, $p = 0.0028$ |
| Age (min, max) | (20, 88) | (18, 70) | |
| Self-reports history of diagnosis with bipolar spectrum disorder | 24 (21%) | 115 (51%) | $p < 0.001$ |
| Self-reports no history of diagnosis with bipolar spectrum disorder | 66 (56%) | 29 (13%) | $p < 0.001$ |
| Does not provide any information regarding diagnosis of bipolar spectrum disorder | 27 (23%) | 83 (37%) | $p < 0.001$ |
| MDQ score, mean (sd) | 6 (4) | 12 (1) | $W = 23{,}322$, $p < 0.001$ |
| Total keypresses, mean (sd) | 37,027 (87,464) | 36,381 (71,262) | |
| Total keypresses, median (mad) | 7,600 (7,465) | 12,043 (12,682) | $W = 15{,}098$, $p = 0.037$ |

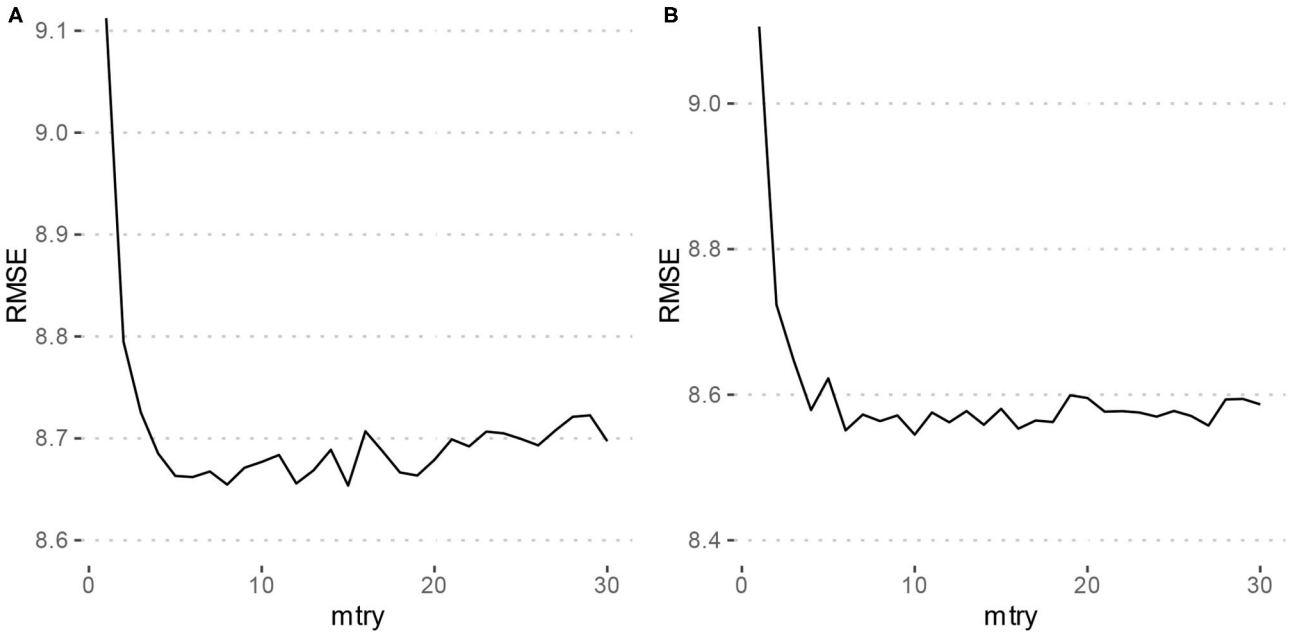


**FIGURE 1 |** Parameter tuning results for random forest models. RMSE, Root mean squared error. **(A)** Depicts the grid search results for Model 1 (typing metrics only) which achieved a minimum RMSE at mtry = 15. **(B)** Depicts the grid search results for Model 2 (typing metrics with gender and MDQ status) which achieved a minimum RMSE at mtry = 10.

of these models using the validation set are described in **Table 3**. Using the validation set, Model 1 had an RMSE of 9.7, and Model 2 had an RMSE of 9.5. Breiman's Pseudo R-squared was 0.42 and 0.44 for Models 1 and 2, respectively. Model 1 had a median absolute error of 5.9 and Model 2 had a median absolute error of 5.5. This difference was not statistically significant ($V = 2{,}109$, $p = 0.21$).

Given the trend toward improved performance with the inclusion of gender and MDQ status as model features, analysis of feature importance and differences in prediction by MDQ status are presented only for Model 2.

Feature importance is depicted in **Figure 2**. Features whose exclusion from the model results in a larger increase in Mean Squared Error are considered more important. While this allows us to understand the relative importance of the features, it does not provide information on how each feature's value is associated with age. One method that allows for the examination of these relationships in random forest models is an ALE plot. Given that many of the most important features are different summaries of the same essential feature (e.g., interkey time), in plots A–D of **Figure 3** we present the ALE plots of four of the most important features: the median of mean interkey times, the mean session

**TABLE 3 |** Model performance comparison using the validation dataset.

| | Model 1 (only typing metrics) | Model 2 (typing metrics with MDQ status and gender) | |
|---|---|---|---|
| RMSE | 9.7 | 9.5 | |
| Breiman's Pseudo R-Squared | 0.42 | 0.44 | |
| Median absolute error | 5.9 | 5.5 | $V = 2{,}109$, $p = 0.21$ |

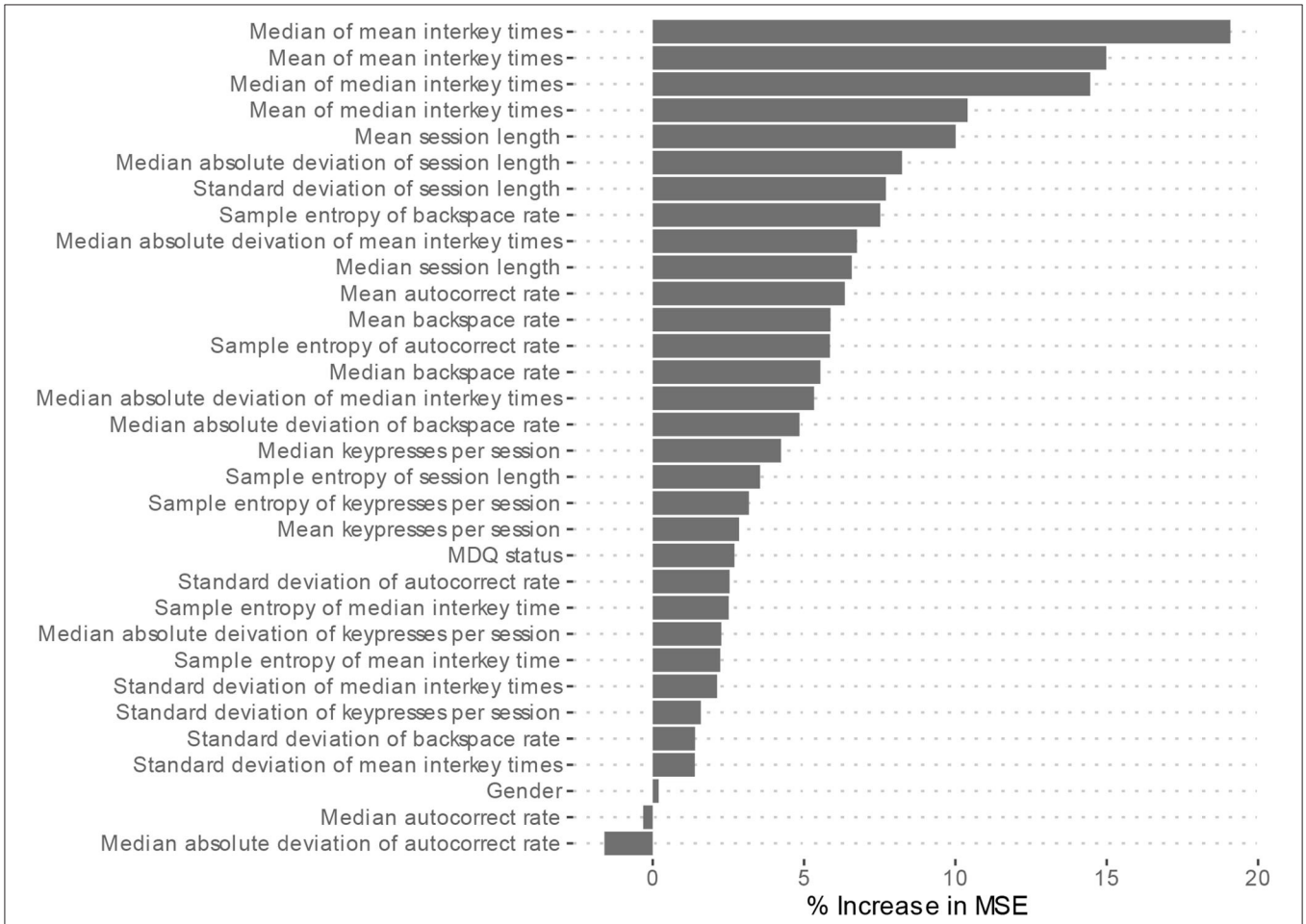


**FIGURE 2 |** Model 2 Feature importance. MSE, Mean square error. Higher increases in MSE indicated increased importance of the feature in predicting age but do not indicate directionality of the relationship.

length, the sample entropy of the backspace rate, and the mean backspace rate. Based on these plots, increased interkey time and session length are both generally associated with increased age; whereas, increased sample entropy of the backspace rate is associated with younger age, and the association between age and the mean backspace rate is not monotonic. Plots E and F of **Figure 3** depict the interaction between the median of mean interkey times and the mean session length and between the mean backspace rate and the sample entropy of the backspace rate, respectively. In these plots we see that the existence and directionality of linear trends between the predicted age and these features depend on the range of a second associated feature highlighting the complexity of the relationship between typing behaviors and predicted age.

The raw prediction error for age, i.e., how many years over or under the model predicted from the correct age, tended to be lower for participants with a negative screen (median $= -5.95$) than those with positive screens (median $= 0.55$) ($W = 1{,}037$, $p = 0.037$). The absolute prediction error, which measures the absolute deviation from the correct age, tended to be lower for participants with a positive screen (median $= 4.50$) than those with negative screens (median $= 7.92$) ($W = 508$, $p = 0.0049$). These comparisons are depicted with boxplots in **Figure 4**. The significant difference in absolute prediction error between the groups suggests the existence of an intrinsic difference between the groups in terms of how each group's typing behaviors relate to age, and the significant difference in raw errors specifically demonstrates that participants with a negative MDQ screen tend

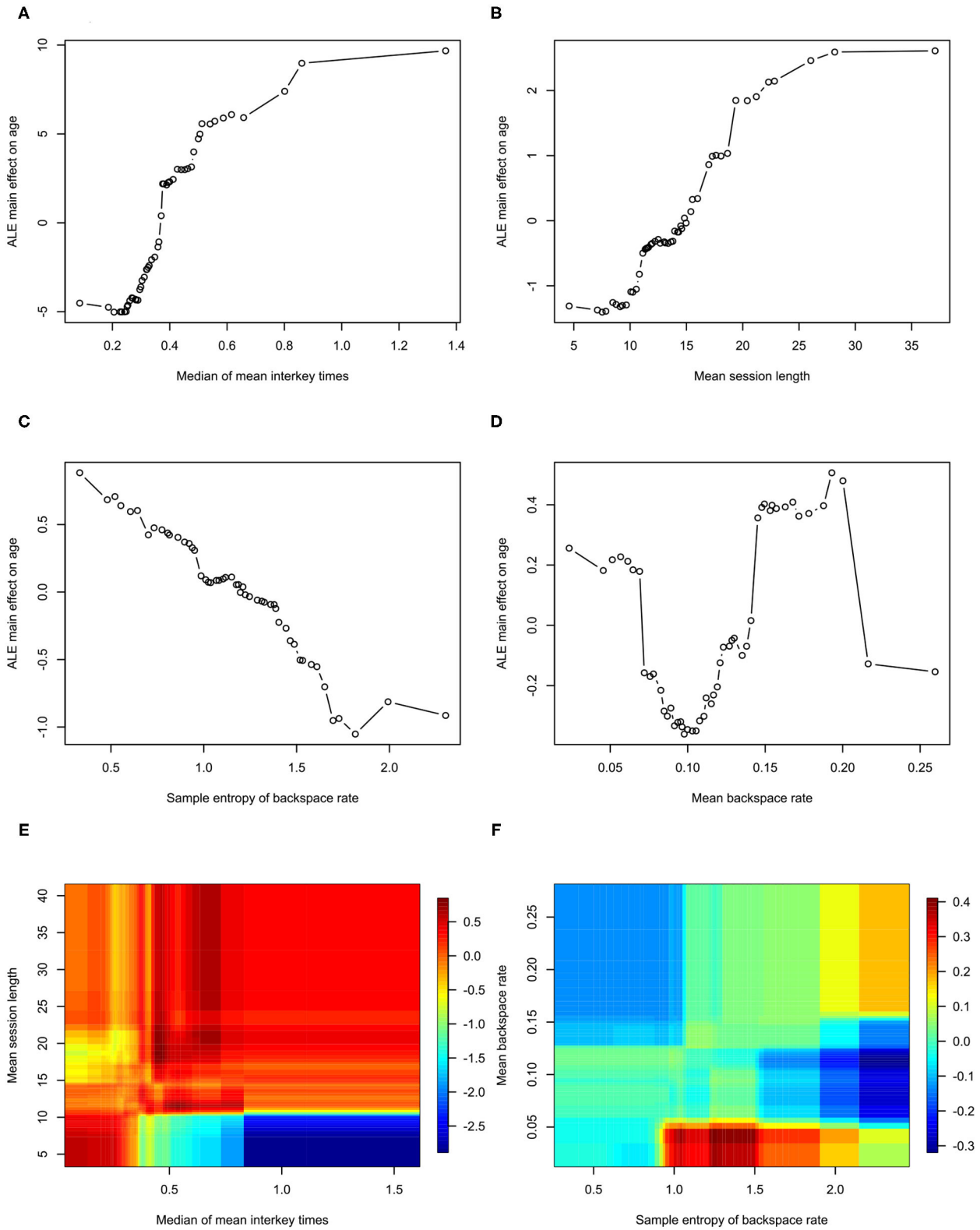


**FIGURE 3 |** Accumulated Local Effects plots for Model 2. ALE, Accumulated Local Effects. **(A–D)** Depict the effects of individual features on age prediction. **(E,F)** Depict the interaction of the two indicated effects on age.

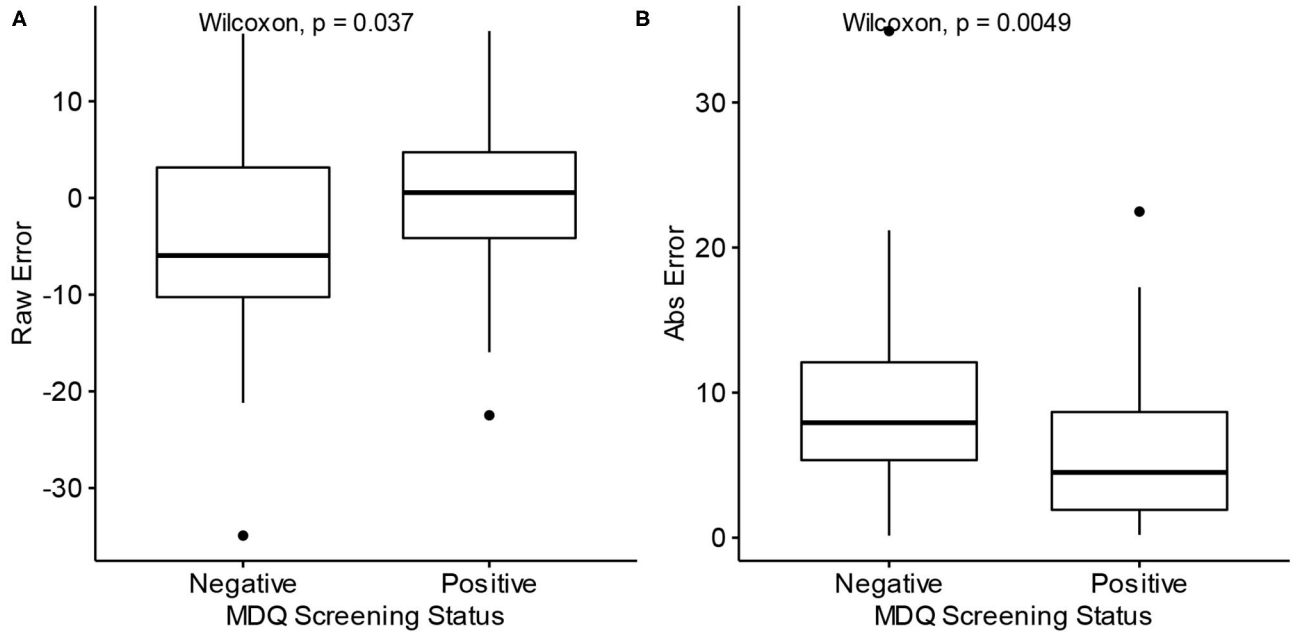


**FIGURE 4 |** Differences in prediction error by MDQ status. MDQ, Mood disorder questionnaire. Abs Error, Absolute error. **(A)** Raw prediction errors of Model 2 by MDQ Screening status. **(B)** Absolute prediction errors of Model 2 by MDQ Screening status.

to be predicted as younger may be consistent with the theory of bipolar disorder's association with neuroprogression.

## DISCUSSION

Objective biomarkers of psychiatric pathology have the potential to transform the practice of psychiatry by providing clinicians more precise and reliable data to inform treatment decisions. In this study, we investigate the possibility of creating such a biomarker by using passively collected keyboard dynamics metadata derived from smartphone usage. This method of collection has the advantages of enabling high frequency sampling and perhaps even more importantly enabling such sampling to take place in people's normal day-to-day lives.

While there have been several studies which have used mobile phone derived metadata to predict demographic features such as age, these studies have tended to use features such as the number of calls and length of calls, the number of text messages, time of day of usage, and metrics derived from the networks of interactions (i.e., calls/text messages) between the users and other people (21–23). Further, these studies appear to have been focused on the utility of such methods for marketing applications, and in keeping with that aim, they used binned age groups and measured model performance based on correct classification of users to those groups. This makes comparing performance between these models and our regression oriented models difficult.

Although models of biological age are typically constructed by using a cohort of "healthy" participants to train a model which is then applied to participants with pathology, for this study we trained our model on both healthy and non-healthy participants in order to make maximum use of the available data. This is a limitation that we plan to address in future studies via larger, more well-characterized samples. The significant difference in prediction accuracy between participants with positive and negative MDQ screens suggests that there may be some intrinsic difference in the pattern of typing between participants who are likely at elevated risk of having bipolar disorder compared to those without such risk. That participants with positive screens have a lower absolute prediction error may be a consequence of the fact that the training sample consisted mostly of participants with positive screens; however, it may also be consistent with the emerging finding that psychiatric disorders may be characterized by a decrease in complexity and variability of behavior, which makes the brain less adaptable to a constantly changing environment (24). Even if the difference is primarily driven by the imbalance between participants with positive and negative MDQ screens, the very fact that such a difference exists is notable in that it suggests that the psychiatric pathology associated with a positive screen produces detectable changes in mobile phone typing kinematics. That there was also a tendency for the model to underpredict the age of participants with negative MDQ screens is consistent with the concept that pathology is associated with the phenomenon of biological age exceeding chronological age (5).

With this study design, we were not able to include in our models other potential covariates of cognitive and motor performance which could affect age prediction errors. Such factors could include co-morbidities, medication status, specific psychiatric diagnoses and severity, and general facility with

phone usage. Given the tendency for bipolar disorder to be associated with a host of other co-morbidities (25), the between group differences found between participants with positive and negative MDQ screens could be due to bipolar disorder itself or some other combination of co-morbidities that is associated with people who are at increased risk of bipolar disorder that alters typing behaviors. Without such information, we cannot rule out the possibility that the differences in prediction accuracy between the participants with positive and negative MDQ screens are due to a disproportionate allocation of these characteristics which are not associated with the pathology associated with a positive MDQ screen. In future studies, we plan to collect data such as diagnoses, severity, and treatment status in order to create models which will properly attribute and quantify the effect of these variables on brain age biomarkers.

Examining which features are most important in predicting age, the most important features are measures of typing speed and the length of typing session. This is consistent with previous studies that have found correlations between age and typing performance (26–28). Plots A and B of **Figure 3** demonstrate that both interkey time and session length tend to be positively correlated with age – older age is associated with slower speed and longer session; however, in examining the interaction of these two features depicted in plot D of **Figure 3**, we see that for sessions under 10 s in length, there is actually a negative correlation between interkey time and age. One possible explanation is that in these short sessions the interkey time represents the pauses that occur in a rapid exchange of text messages with another person, and that for longer sessions the interkey time represents the pauses that occur in the composition of a longer body of text.

The relatively high importance of the sample entropy of backspace rates across sessions is an intriguing finding. This feature theoretically measures the complexity of participant backspace use. Based on its ALE plot, plot E of **Figure 3**, it appears to generally negatively correlated with age; however, notably if we examine the interaction of the sample entropy of the rate with the overall mean rate depicted in plot D of **Figure 3**, we see that at relatively low overall usage of backspace increased entropy is associated with younger age, but that at relatively high usage increased entropy is associated with older age. Several studies have examined how measures of complexity like entropy can be applied to functional imaging derived brain networks and how complexity changes with aging (29–31). With data we are collecting in one of our current studies we will be able to examine how neuroimaging derived measures of brain complexity are associated with the complexity of signals derived from the kinematics of smartphone usage (32).

### Limitations

This study is limited by the fact that all subjective data were provided via self-report. A sample which consists of participants with clinically confirmed diagnoses would yield greater insight into the utility of digital biomarkers such as those described here in characterizing psychiatric disorders. Along those same lines, objective neuropsychological performance data would also have helped better contextualize our findings with existing literature in bipolar disorder. Although we included age and gender as features in our model, a sample in which age and gender distributions are equivalent across the case and control groups would likely yield more robust findings.

### Future Directions

Future directions of this research include attempting to replicate these findings in a more well-characterized sample and determining how this marker compares to other biomarkers such as neuroimaging based markers and other digital biomarkers. We also aim to investigate how differences in handedness (one-handed vs. two-handed typing) and distinguishing between different types of keystroke transitions (e.g., alphanumeric to alphanumeric vs. alphanumeric to backspace) might yield better performing models. Another potential line of investigation is determining whether differences between predicted age and chronological age by our model are associated with state level phenomena such as the severity of mood episodes.

## CONCLUSION

Passively collected typing kinematics can be used to estimate biomarkers of brain age. The differences we found in this study between the performance of such a biomarker in participants with and without a positive screen for bipolar disorder—i.e., the tendency to underestimate the age of healthy participants—suggest that this biomarker may also be a useful marker of pathology. Further investigation to refine the model and determine its relation to other markers of pathology such as neuroimaging and neuropsychological testing is warranted.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Illinois at Chicago Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

JZ did the analysis and wrote the manuscript. All other authors contributed data and edited the manuscript.

## FUNDING

This study was partially funded by Mood Challenge for Research kit (AL and PN) and 1R01MH120168 (OA and AL).

## ACKNOWLEDGMENTS

The BiAffect team would like to thank all the citizen scientists who have enrolled in our study and contributed their data.

## REFERENCES


1. McGorry PD. Early clinical phenotypes, clinical staging, and strategic biomarker research: building blocks for personalized psychiatry. *Biol Psychiatry*. (2013) 74:394–5. doi: 10.1016/j.biopsych.2013.07.004
2. Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF. Berk M. The new field of "precision psychiatry." *BMC Med*. (2017) 15:1–7. doi: 10.1186/s12916-017-0849-x
3. Mather KA, Jorm AF, Parslow RA, Christensen H. Is telomere length a biomarker of aging? A review. *J Gerontol A Biol Sci Med Sci*. (2011) 66:202–13. doi: 10.1093/gerona/glq180
4. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks: challenges and recommendations. *Genome Biol*. (2019) 20:1–24. doi: 10.1186/s13059-019-1824-y
5. Cole JH, Marioni RE, Harris SE, Deary IJ. Brain age and other bodily 'ages': implications for neuropsychiatry. *Molecular Psychiatry*. (2019) 24:266–81. Available from: doi: 10.1038/s41380-018-0098-1
6. Cole JH, Franke K. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends Neurosci*. (2017) 40:681–90. doi: 10.1016/j.tins.2017.10.001
7. Vesel C, Rashidisabet H, Zulueta J, Stange JP, Duffecy J, Hussain F, et al. Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: a biaffect iOS study. *J Am Med Informatics Assoc*. (2020) 27:1007–18. doi: 10.1093/jamia/ocaa057
8. BiAffect. *The First Study on Mood and Cognition Using Mobile Typing Kinematic*. Available online at: https://www.biaffect.com/ (accessed June 26, 2021).
9. Ryan KA, Vederman AC, McFadden EM, Weldon AL, Kamali M, Langenecker SA, et al. Differential executive functioning performance by phase of bipolar disorder. *Bipolar Disord*. (2012) 14:527–36. doi: 10.1111/j.1399-5618.2012.01032.x
10. Langenecker SA, Saunders EFH, Kade AM, Ransom MT, McInnis MG. Intermediate: cognitive phenotypes in bipolar disorder. *J Affect Disord*. (2010) 122:285–93. doi: 10.1016/j.jad.2009.08.018
11. Eclarinal E, Ajilore O. Neurobiology of older age bipolar disorder. In: *Bipolar Disorder in Older Age Patients*. Cham: Springer International Publishing (2017) p. 43–56. doi: 10.1007/978-3-319-48912-4_3
12. Berk M, Kapczinski F, Andreazza AC, Dean OM, Giorlando F, Maes M, et al. Pathways underlying neuroprogression in bipolar disorder: focus on inflammation, oxidative stress and neurotrophic factors. *Neurosci Biobehav Rev*. (2011) 35:804–17. doi: 10.1016/j.neubiorev.2010.10.001
13. Hirschfeld RMA, Williams JBW, Spitzer RL, Calabrese JR, Flynn L, Keck J, et al. Development and validation of a screening instrument for bipolar spectrum disorder: the mood disorder questionnaire. *Am J Psychiatry*. (2000) 157:1873–5. doi: 10.1176/appi.ajp.157.11.1873
14. Zimmerman M, Galione JN. Screening for bipolar disorder with the mood disorders questionnaire: a review. *Harv Rev Psychiatry*. (2011) 19:219–28. doi: 10.3109/10673229.2011.614101
15. McKinney W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*. Austin, TX (2010). p. 56–61. doi: 10.25080/Majora-92bf1922-00a
16. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography (Cop)*. (2013) 36:27–46. doi: 10.1111/j.1600-0587.2012.07348.x
17. Boulesteix A-L, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov*. (2012) 2:493–507. doi: 10.1002/widm.1072
18. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, et al. *Caret: Classification and Regression Training*. R package version 6.0–21. CRAN Wien, Austria (2014)
19. Liaw A, Wiener M. Classification and regression by random forest. *R News*. (2002) 3:18–22.
20. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. *J R Stat Soc Ser B Stat Methodol*. (2020) 82:1059–86. doi: 10.1111/rssb.12377
21. Dong Y, Yang Y, Tang J, Chawla N V. Inferring user demographics and social strategies in mobile social networks. *Proc ACM SIGKDD Int Conf Knowl Discov Data Min*. (2014) 14:15–24. doi: 10.1145/2623330.2623703
22. Jahani E, Sundsøy P, Bjelland J, Bengtsson L. Sandy PA, De Montjoye Y-A. Improving official statistics in emerging markets using machine learning and mobile phone data. *EPJ Data Sci*. (2017) 6:3. doi: 10.1140/epjds/s13688-017-0099-3
23. Al-Zuabi IM, Jafar A, Aljoumaa K. Predicting customer's gender and age depending on mobile phone data. *J Big Data*. (2019) 6:1–16. doi: 10.1186/s40537-019-0180-9
24. Yang AC, Tsai SJ. Is mental illness complex? From behavior to brain. *Prog Neuro-Psychopharmacol Biol Psychiatry*. (2013) 45:253–7. doi: 10.1016/j.pnpbp.2012.09.015
25. Crump C, Sundquist K, Winkleby MA, Sundquist J. Comorbidities and mortality in bipolar disorder: a swedish national cohort study. *JAMA Psychiatry*. (2013) 70:931–9. doi: 10.1001/jamapsychiatry.2013.1394
26. Czaja SJ, Sharit J. Age differences in the performance of computer-based work. *Psychol Aging*. (1993) 8:59–67. doi: 10.1037/0882-7974.8.1.59
27. Kalman YM, Kavé G, Umanski D. Writing in a digital world: self-correction while typing in younger and older adults. *Int J Environ Res Public Health*. (2015) 12:12723–34. doi: 10.3390/ijerph121012723
28. Chen R, Jankovic F, Marinsek N, Foschini L, Kourtis L, Signorini A, et al. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams world from consumer-grade multimodal sensor streams. In*: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage, AK (2019). p. 2145–55. doi: 10.1145/3292500.3330690
29. Zhou J, Lo OY, Halko MA, Harrison R, Lipsitz LA, Manor B. The functional implications and modifiability of resting-state brain network complexity in older adults. *Neurosci Lett*. (2020) 720:134775. doi: 10.1016/j.neulet.2020.134775
30. Sokunbi MO, Cameron GG, Ahearn TS, Murray AD, Staff RT. Fuzzy approximate entropy analysis of resting state fMRI signal complexity across the adult life span. *Med Eng Phys*. (2015) 37:1082–90. doi: 10.1016/j.medengphy.2015.09.001
31. Liu CY, Krishnan AP, Yan L, Smith RX, Kilroy E, Alger JR, et al. Complexity and synchronicity of resting state blood oxygenation level-dependent (BOLD) functional MRI in normal aging and cognitive decline. *J Magn Reson Imaging*. (2013) 38:36–45. doi: 10.1002/jmri.23961
32. *Unobtrusive Monitoring of Affective Symptoms and Cognition Using Keyboard Dynamics (UnMASCK)*. Available online at: https://clinicaltrials.gov/ct2/show/NCT04358900 (accessed May 20, 2021).


**Conflict of Interest:** OA is a co-founder of KeyWIse AI. He also serves on the advisory boards of Embodied Labs and Blueprint Health. AL is an advisor for Buoy health and a consultant for Otsuka USA and ATAI Life Sciences, in addition to being a Co-founder of KeyWise AI.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychiatry

REVIEW
published: 23 December 2021
doi: 10.3389/fpsyt.2021.782183

# Affective Computing for Late-Life Mood and Cognitive Disorders

*Erin Smith*[1,2,3,4,5]*, *Eric A. Storch*[6], *Ipsit Vahia*[7,8], *Stephen T. C. Wong*[9], *Helen Lavretsky*[10], *Jeffrey L. Cummings*[11] and *Harris A. Eyre*[1,2,4,5,6,12]

[1] The PRODEO Institute, San Francisco, CA, United States, [2] Organisation for Economic Co-operation and Development (OECD), Paris, France, [3] Department of Neurology & Neurological Sciences, Stanford University, Stanford, CA, United States, [4] Global Brain Health Institute, University of California, San Francisco, San Francisco, CA, United States, [5] Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland, [6] Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, TX, United States, [7] Division of Geriatric Psychiatry, McLean Hospital, Boston, MA, United States, [8] Division of Geriatric Psychiatry, Harvard Medical School, Boston, MA, United States, [9] Systems Medicine and Biomedical Engineering Houston Methodist, Houston, TX, United States, [10] Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, CA, United States, [11] Chambers-Grundy Center for Transformative Neuroscience, Department of Brain Health, School of Integrated Health Sciences, University of Nevada, Las Vegas (UNLV), Las Vegas, NV, United States, [12] IMPACT, The Institute for Mental and Physical Health and Clinical Translation, Deakin University, Geelong, VIC, Australia

Affective computing (also referred to as artificial emotion intelligence or emotion AI) is the study and development of systems and devices that can recognize, interpret, process, and simulate emotion or other affective phenomena. With the rapid growth in the aging population around the world, affective computing has immense potential to benefit the treatment and care of late-life mood and cognitive disorders. For late-life depression, affective computing ranging from vocal biomarkers to facial expressions to social media behavioral analysis can be used to address inadequacies of current screening and diagnostic approaches, mitigate loneliness and isolation, provide more personalized treatment approaches, and detect risk of suicide. Similarly, for Alzheimer's disease, eye movement analysis, vocal biomarkers, and driving and behavior can provide objective biomarkers for early identification and monitoring, allow more comprehensive understanding of daily life and disease fluctuations, and facilitate an understanding of behavioral and psychological symptoms such as agitation. To optimize the utility of affective computing while mitigating potential risks and ensure responsible development, ethical development of affective computing applications for late-life mood and cognitive disorders is needed.

**Keywords: affective computing, late-life depression, dementia, Alzheimer's disease, digital phenotyping**

OPEN ACCESS

***Edited by:***
*Valeria Manera,*
*Université Côte d'Azur, France*

***Reviewed by:***
*Maribel Pino,*
*Broca Hospital (APHP), France*
*Hatice Kose,*
*Istanbul Technical University, Turkey*

***Correspondence:**
*Erin Smith*
*erin.smith@gbhi.org*

***Specialty section:***
*This article was submitted to Aging Psychiatry, a section of the journal Frontiers in Psychiatry*

***Received:*** *23 September 2021*
***Accepted:*** *29 November 2021*
***Published:*** *23 December 2021*

***Citation:***
*Smith E, Storch EA, Vahia I, Wong STC, Lavretsky H, Cummings JL and Eyre HA (2021) Affective Computing for Late-Life Mood and Cognitive Disorders. Front. Psychiatry 12:782183. doi: 10.3389/fpsyt.2021.782183*

## INTRODUCTION

Between 2019 and 2050, the number of people aged 65 years or over in the world will increase from 703 million to 1.5 billion people (1). With the rapid growth in the aging population around the world, improving the standard of care in late-life mood and cognitive conditions is of the utmost importance. Late-life mood and cognitive conditions are characterized by their complexity, multisystemic nature and broad societal impact, hence making them poorly suited to siloed approaches of thinking and innovation (2). Issues such as overlapping symptoms, comorbidities, and misdiagnosis among mental health and neurological disorders represents only a small subset of the challenges facing late-life mood and cognitive conditions (2). For example,

psychiatric symptoms often occur during prodromal stages of neurodegenerative diseases (3–5). Neurodegenerative diseases are often misclassified as psychiatric disease, which can cause the patient to receive delayed, inappropriate treatment and experience more distress (4). Relatedly, many neurodegenerative conditions have clinical and neuropathological overlap, which can cause difficulty with accurate diagnosis and treatment (6). Comorbidities also present further challenges. As an example, 61% of people with dementia are estimated to have three or more comorbid diagnoses (7), which often remain over or under-treated and negatively affect the physical and psychological well-being of people with dementia (8). As the burden of late-life mood and cognitive disorders continues to rise, improved prevention, diagnosis, and treatment options are urgently needed (9, 10). To solve the unprecedented complexities and challenges associated with late-life mood and cognitive conditions, new technologies and approaches to care are needed. Affective computing has immense potential to benefit the treatment and care of late-life mood and cognitive disorders. Affective computing (also referred to as artificial emotion intelligence or emotion AI) is the study and development of systems and devices that can recognize, interpret, process, and simulate emotion or other affective phenomena (11, 12). It is a transdisciplinary field that combines engineering and computer science with psychology, cognitive science, neuroscience, sociology, education, psychophysiology, value-centered design, ethics, and more.

In this paper, we aim to provide an overview of common affective computing modalities with applications for late-life mood and cognitive disorders. We additionally explore specific applications for late-life depression and Alzheimer's disease (AD). Lastly, we discuss ethical implications and analyze key challenges that must be overcome to ensure ethical development of affective computing for late-life mood and cognitive conditions.

## METHODS

Publications were collected in September 2021 from four databases: PubMed, PsycINFO, OvidSP, and Web of Science. To construct the search protocol, the research question was structured in terms of the following topics: Affective Computing, digital phenotyping, late-life depression, dementia, Alzheimer's disease, and ethical issues. Synonyms and main terms for these topics were selected to construct the search codes. After an initial search, publication titles and abstracts were screened according to year of publication, publication in English language, and of peer-reviewed type. A total of 150 articles were utilized in this review.

## RESULTS

### Affective Computing and Late-Life Depression

Depression is the leading cause of disability worldwide and a major contributor to the overall global burden of disease (13). The scope and burden of late-life depression is significant and expected to rise in the 21$^{st}$ century (14). Depression is estimated to affect 29% of elderly Europeans (15) and 30.6% of elderly Chinese (16). When compared to younger patients with depression, older adults with depression typically have more medical and neurologic comorbidities and display more cognitive impairment (17). Data suggest that 1 in 10 cases of dementia world-wide can be attributed to depression (18). Neurotoxicity due amyloid and tau protein aggregation may represent a pathophysiological cascade which, along with vascular compromise, may predispose individuals to late-life depression (17). **Table 1** highlights clinical challenges in late-life depression and ways in which affective computing may be beneficial.

Williamson et al. (23) examined changes in motor output in people with depression from vocal acoustics and facial movements (23). Using the 4th International Audio/Video Emotion Challenge (AVEC), which consists of a read passage and free-response speech segment from subjects with varying depression levels according to their self-reported Beck depression inventory assessment, they developed a multimodal analysis pipeline that leverages complementary information in audio and video signals including structure and timing features for estimating depression severity. Using the identified features of changes in coordination, movement, and timing of vocal and facial movements, the developed algorithm was able to predict the Beck depression inventory ratings from the AVEC test set with a root-mean-square error of 8.12 and mean absolute error of 6.31 (23).

De Choudhury et al. (36) used behavioral attributes from social media to characterize severity of depression at a population level by developing a social media depression index (SMDI) (73). Using crowdsourcing techniques, they built a corpus of over 69K Twitter postings shared by individuals diagnosed with clinical depression that was measured using the Center for Epidemiologic Studies Depression Scale (CES-D) screening test. By analyzing behavioral features including emotional expression, linguistic style, user engagement, and egocentric social network properties, they built a model that can predict if a post is indicative of depression with an accuracy of more than 70% and precision of 0.82. Lastly, they developed the SMDI metric. The SMDI metric leverages the prediction model to predict posts indicative of depression on Twitter and helps characterize the levels of depression in populations. The geographical, demographic, and seasonal patterns of depression given by SMDI confirm known clinical characteristics of depression and are highly correlated with depression statistics reported by the Centers for Disease Control and Prevention (CDC) (73).

Mundt et al. (42) used vocal acoustic biomarkers to predict depression severity and treatment response (42). One hundred five adults with depression were recruited into a 4-week, randomized, double-blind, place-controlled clinical trial. Speech samples were collected at baseline and study end point using an automated telephone system. Clinician-rated and patient-reported measures of depression severity and treatment response were collected. Results from the study replicated and supported findings from prior studies. More severe depression produced

**TABLE 1 |** Affective computing applications for clinical challenges in late-life depression.

| | Vocal biomarkers | Facial expression biomarkers | Body movements | Eye movements | Keystroke dynamics | Social media behavior | Socially assistive robots (SARs) |
|---|---|---|---|---|---|---|---|
| **APPROACHES FOR AFFECTIVE COMPUTING IN CLINICAL CHALLENGES OF LATE-LIFE DEPRESSION** | | | | | | | |
| Inadequacies of current screening and diagnostic approaches | Use vocal biomarkers to detect depression (19–23) | Use facial expression biomarkers to detect depression (23–30) | Use head movements and pose to detect depression (26, 30–32) | Use gaze and eye movement to detect depression to detect behavioral consistent with depression (33) | Use keystroke dynamics to detect typing behavior associated with depression (34) | Use behavioral attributes related to social engagement, emotion, language, linguistic style, and writing aspects to detect depression (35–40) | Use SARs to administer screening and diagnostic approaches that leverage affective computing biomarkers (41) |
| Trial-and-error treatment approaches | Monitor depression severity and treatment response to determine optimal treatment using vocal biomarkers (42–45) | Monitor depression severity and treatment response to determine optimal treatment using facial expression biomarkers (46) | Leverage body and head movement analysis to measure depression severity throughout the course of treatment to determine optimal approach (46, 47) | Measure depression treatment response using anti-saccade eye movement tasks (48) | Determine optimal treatment by measuring depression severity via touchscreen typing (49) | Assess social media behavior and develop treatment strategy for social media usage that reinforces depressive beliefs and symptoms to improve overall treatment outcomes (50) | Use SARs to provide in-home therapeutic approach and collect real-time data on affective computing biomarkers to determine optimal treatment strategies (41) |
| Loneliness and social isolation | Detect loneliness and social isolation and better identify behavioral phenotypes of loneliness and social isolation through vocal biomarkers (51) | Assess spontaneous smile mimicry to detect and monitor loneliness (52) | Assess body movement coordination, which may be impaired during loneliness due to changes in the left posterior superior temporal sulcus (53) | Assess eye movement, which may be impaired during loneliness due to changes in the left posterior superior temporal sulcus (53) | Assess keystroke dynamics and hand action, which may be impaired during loneliness due to changes in the left posterior superior temporal sulcus (53) | Monitor social media behavior for early detection of loneliness (54) | Use socially assistive robots (SARs) or social companion robots that have affective computing capabilities to help older adults with depression (55–57) |
| Poor treatment follow-up | Monitor daily fluctuations using vocal biomarkers during time outside of the clinic and receive alerts if problems with treatment (19, 23) | Monitor daily fluctuations using facial expression biomarkers during time outside of the clinic and receive alerts if problems with treatment (23) | Assess body movement during time outside of the clinic to better understand symptom fluctuations (46, 47) | Track eye movement to understand depression symptoms between in-clinic visits (48) | Monitor keystroke behavior to assess treatment efficacy (34) | Monitor behavioral attributes related to social engagement, emotion, language, linguistic style, and writing aspects to monitor depression symptom severity between treatment sessions or clinic visits (35–40) | Use SARs to interact with older adults and understand depression symptom progression and severity (55–57) |
| Co-occurrence with anxiety disorders | Differentiate between depression and anxiety disorders using vocal biomarkers | Differentiate between depression and anxiety disorders using facial expression biomarkers | Identify and monitor anxiety and depression severity scores using digital gait movement (58) | Assess anxiety and depression severity using eye movement (59) | Monitor and differentiate between anxiety and depression via touchscreen typing (34) | Detect anxiety and depression via social media behavior (40) | Capture affective computing biomarkers to differentiate between anxiety and depression via SARs (41) |
| Co-occurrence with Alzheimer's disease | Monitor symptoms associated with depression and Alzheimer's disease using vocal biomarkers (60) | Monitor symptoms associated with depression and Alzheimer's disease using facial expression biomarkers (61) | Kinematic analysis can detect co-morbid Alzheimer's disease for patients with depression (62) | Detect and differentiate between Alzheimer's and depression via eye movement tracking (63, 64) | Monitor early stages of Alzheimer's disease and depression via touchscreen typing (34, 65) | Detect late-life depression in Alzheimer's disease patients via speech and language analysis on social media (66) | Use SARs to analyze speech and language of people with late-life depression and Alzheimer's disease (66) |
| Risk of suicide | Detect suicidal ideation and risk using vocal biomarkers (67) | Evaluate risk of suicide using facial expression biomarkers (68) | Monitor body movement to detect risk of suicide (69) | Assess eye movement to identify attention bias for suicide related stimuli (70) | Use digital phenotyping from smartphone typing to detect risk of suicide (69) | Detect suicidal ideation and risk using social media behavioral analysis (71, 72) | For people at risk of suicide, SARs can be leveraged to track affective computing biomarkers that indicate risk of suicide (69) |

longer recordings with more pause time, more variable pause lengths, a greater percentage of pause time, smaller speech/pause ratios, and slower speaking rates. Speech pause times were found to shorten with clinical improvement following treatment, and depressed patients who did not improve clinically were found to have smaller vocal acoustic changes and/or changes that were directionally opposite to treatment responders.

## Affective Computing and Alzheimer's Disease

Alzheimer's disease (AD) is the most common neurodegenerative disorder and largest cause of dementia in the world with rapidly growing personal, societal, economic, and medical implications. In the United States alone, over 6.2 million people suffer from AD that costs the healthcare system more than $355 billion annually, not including the value of informal caregiving (74). At a global level, there are more than 35 million people currently living with AD, and by 2050 the number is expected to more than triple, exceeding 115 million people (75). There is growing interest in detecting AD during prodromal stages because (a) the likelihood of reversing anatomic and physiologic changes (such as neuronal death) likely decreased dramatically as the disease advances (61, 76), (b) there is a growing body of evidence that cognitive, sensory, and motor changes may precede clinical manifestation of AD by 10–20 years (74, 76), and (c) aducanumab, the only approved disease-modifying therapy is recommended only for patients with early AD (77). **Table 2** highlights clinical challenges in AD and ways in which affective computing may be beneficial.

Ahmed et al. (83) examined connected speech as a marker of disease progression in autopsy-proven AD (83). Samples of connected speech were obtained from 15 patients who were part of a longitudinal cohort study in whom AD was diagnosed during life and later confirmed at post-mortem. The study analyzed spoken discourse over the course from MCI to mild AD dementia to moderate AD dementia. Samples were analyzed using measures of syntactic complexity, lexical content, speech production, fluency, and semantic content. Subtle changes were found in spoken language that were detectable in MCI stages and enabled monitoring progression through successive clinical stages of AD. Language biomarkers could help identify prodromal AD and provide a way to monitor disease in therapeutic trials (83).

Bayat et al. (113) evaluated the ability of in-vehicle GPS data loggers and driving behavior to distinguish cognitively normal older drivers with preclinical AD from those without preclinical AD using machine learning algorithms (113). For 1 year, 139 subjects (64 with preclinical AD; 75 without preclinical AD, as determined by cerebrospinal fluid biomarkers) were monitored while they drove with a commercial in-vehicle GPS data logger. Random Forest models were trained on the GPS data. The receiver operating curve (ROC) area under the curve (AUC) for predicting preclinical AD from driving features alone was 0.82, with the addition of age alone increased to 0.94, and with the additions of age and APOE ε4 status increased to 0.96 (113).

Gills et al. (95) developed and validated a short digital eye-tracking assessment that predicts cognitive status among adults (95). Fifty-five adults (11 with MCI and 44 cognitively normal) were tested on two occasions. During the first visit, participants underwent a brief eye-tracking based visual paired-comparison (VPC), Montreal Cognitive Assessment (MoCA), Digital Symbol Coding test (DSC), and NIH Toolbox Cognitive Battery (NIHTB-CB). During the second visit, participants underwent VPC, DSC, NIHTB-CB, and dual-task (DT). VPC reliably predicted cognitive status while demonstrating high test-retest reliability and displayed significant associations with gold standard cognitive assessments (95). Eye-tracking based VPC may provide a useful, brief, and scalable screening tool for cognitive impairment (95).

## Ethical Implications of Affective Computing in Healthcare

New approaches are needed to address the technical, scientific, philosophical, and ethical challenges associated with affective computing applications in healthcare (114). One key challenge is ensuring that the tools account for sex, gender, racial, ethnic, and culture-based differences. For example, two-thirds of AD patients worldwide are women (115, 116). Women have a higher lifetime risk of stroke than men, and women are twice as likely to be diagnosed with depression and anxiety disorders, and migraines (116–118). These are all specific risks factors for developing dementia.

Given the higher rate of depression and AD among women research is needed to understand if and how sex and gender-based differences affect disease manifestations that may result in the need for different digital biomarkers and machine learning affective computing approaches for males vs. females. It is critical to ensure that the machine learning algorithms capture symptoms that may be more common or different in females.

We must ensure that algorithmic fairness with affective computing does not stop at merely accounting for sex and gender-based differences. Data sets and algorithms used for affective computing must also include bias reduction measures that account for ethnic, racial, geographical, cultural, and other human biases (119). As one example, facial analysis algorithms are often trained on datasets that are predominately comprised of lighter-skinned males and may fail to detect female faces and people of different races and ethnicities (120). Algorithmic and human bias must be addressed to ensure greater fairness, transparency, and accountability in the development of affective computing applications.

The various ways affective computing can be leveraged in healthcare also have different ethical implications. For example, current affective computing technologies typically leverage passive or active data collection. In the context of affective computing for healthcare, passive data collection may entail information continuously collected from smartphone usage, driving, or social media, whereas active data collection may involve specific vocal or facial expression assessments within a clinical setting. Different ethical and practical considerations arise for passive vs. active data collection and the setting in which data is conducted (e.g., at home using everyday technological

**TABLE 2 |** Affective computing applications for clinical challenges in AD.

| | Vocal biomarkers | Facial expression biomarkers | Body movements | Eye movements | Keystroke dynamics |
|---|---|---|---|---|---|
| **APPROACHES FOR AFFECTIVE COMPUTING IN CLINICAL CHALLENGES OF ALZHEIMER'S DISEASE** | | | | | |
| Lack of early, objective screening and diagnostic approaches | Use vocal biomarkers to detect and access mild cognitive impairment (MCI) and prodromal stages of AD (61, 78–83) | Assess cognitive and neuropsychiatric symptoms of AD using facial expression impairments (84, 85) | Capture motor impairments that precede signs of cognitive impairment by over a decade in people with AD through measuring gait speed, stride length, and gait symmetry (61, 86–90) | Use eye movement to detect MCI and prodromal stages of AD (91–95) | Capture differences in reaction speed and movement that have been found in early stages of AD using an active finger tapping test or passive data collection from daily computer, tablet, or smartphone keyboard use (61, 96–98) |
| Lack of objective biomarkers for monitoring disease progression and comprehensive, daily fluctuations | Use vocal biomarkers to monitor disease progression (83) | Monitor diseases progression and daily fluctuations of symptoms using facial expressions (84, 85) | Assess gait and balance throughout AD to monitor disease progression (99–101) | Monitor eye movement to track AD progression (102, 103) | Monitor progression of cognitive impairment including MCI to AD using touchscreen typing (61, 65) |
| Understanding and addressing the behavioral and psychological symptoms, such as agitation and pain, experienced by patients with AD | Capture vocal biomarkers using sensing technology to monitor behavioral and psychological symptoms of AD (104) | Digitize facial expressions and movements to monitor behavioral and psychological symptoms of AD using sensing technology (104) | Use body movements, such as number of transitions between spaces, to detect and better understand different behavioral and psychological symptoms of AD (105) | Use eye movements to understand and monitor behavioral and psychological symptoms of AD via sensing technology (104) | Analyze typing to identify subtypes of AD based on the presence and intensity of behavioral and psychological symptoms of AD |
| Co-occurrence with depression | Monitor symptoms associated with depression and AD using vocal biomarkers (60) | Monitor symptoms associated with depression and AD using facial expression biomarkers (61) | Kinematic analysis can detect co-morbid AD for patients with depression (62) | Detect and differentiate between AD and depression via eye movement tracking (63, 64) | Monitor early stages of AD and depression via touchscreen typing (34, 65) |
| Misdiagnosis between AD and other neurodegenerative disorders during early stages of disease progression | Differentiate between AD, Parkinson's disease (AD), and Lewy Body Disease (LBDs) using vocal biomarkers (60, 106–109) | Differentiate between AD and PD using facial expression analysis | Differentiate between AD and PD using digital gait analysis (99–101, 110) | Assess PD, AD, and Lewy body dementia via eye movement analysis (111) | Detect and differentiate early stages of PD and AD using typing and keyboard dynamics (61, 65, 112) |

devices or via a specific assessment in-clinic). These challenges must be addressed.

Other ethical considerations that will need to be addressed and mitigated with the advent of affective computing applications in healthcare are (a) addressing privacy and data security concerns; b) determining what it means for people and society to know they have a disorder 10–20 years before clinical manifestations (such as with AD or PD); (c) navigating how to update affective computing models after deployment to make them optimally adaptive and effective; and (d) considering what privacy means when a doctor or friend or family member could tell if you have a certain health condition or are suicidal.

The 2021 "Ethics and governance of artificial intelligence for health: WHO guidance" report by the World Health Organization (WHO) highlights key ethical principles for the use of artificial intelligence in healthcare, which include protecting autonomy; promoting human well-being, human safety and the public interest; ensuring transparency, explainability and intelligibility; fostering responsibility and accountability; ensuring inclusiveness and equity; and promoting artificial intelligence that is responsive and sustainable (121). These ethical principles must guide the development of affective computing applications for late-life mood and cognitive conditions to ensure human rights are upheld and that patient and community interests do not become subordinate to the powerful commercial interest of technology companies or the interests of governments in surveillance and social control (121). Questions related to data security, privacy, and ownership must also be addressed; how CAN these concerns be navigated on a global scale? Initiatives such as the National Institute of Health Bridge to Artificial Intelligence (Bridge2AI) program brings together technologists, biomedical experts, social scientists, and humanists to develop ethical data sets and tools and will help navigate these pressing questions related to affective computing applications in brain health (122).

By addressing the ethical implications and challenges during early stages of technology development, the introduction of affective computing technology in healthcare can bring a new era of health that is marked by a proactive, personalized, and preventative approach to care. Affective computing has the

possibility to be applied globally, such as through the ubiquity of smartphones, and can help address challenges related to health equity. Affective computing technology has the potential to improve early detection and screening, disease severity and progression monitoring, treatment efficacy monitoring, and the quality of life for people around the world with a myriad of different neurological health conditions.

## CONCLUSION

Affective computing can address challenges associated with late-life mood and cognitive conditions, including depression and AD. Affective computing technologies- ranging from vocal dynamics to facial expressions to social media usage to driving behavior- can provide objective biomarkers and tools for early detection, monitoring treatment response, tracking disease progression, and more comprehensively understanding the daily life of patients. To leverage affective computing to increase global brain health equity and a precision medicine approach to care, efforts are needed to ensure ethical development of affective computing for late-life mood and cognitive conditions that account for algorithmic and human bias. With these safeguards affective computing can become a major tool of care of late life affective and cognitive disorders.

## AUTHOR CONTRIBUTIONS

ES led the manuscript development and wrote the initial review. HE, JC, HL, SW, IV, and EAS contributed edits, additional examples, and ideas. All authors contributed to the idea development, writing, editing of this manuscript, and approved the submitted version.

## REFERENCES


1. UN. *World Population Ageing*. (2019). United Nations Department of Economic and Social Affairs (2019).
2. Smith E, Au R, Mossé M, Lavretsky H, Forbes M, Eyre HA. Rebooting late-life mental health innovation and entrepreneurship with convergence science. *Am J Geriatr Psychiatry*. (2020) 28:591–6. doi: 10.1016/j.jagp.2020.03.003
3. Fujishiro H, Nakamura S, Sato K, Iseki E. Prodromal dementia with L ewy bodies. *Geriatr Gerontol Int.* (2015) 15:817–26. doi: 10.1111/ggi.12466
4. Woolley JD, Khan BK, Murthy NK, Miller BL, Rankin KP. The diagnostic challenge of psychiatric symptoms in neurodegenerative disease: rates of and risk factors for prior psychiatric diagnosis in patients with early neurodegenerative disease. *J Clin Psychiatry*. (2011) 72: 126–33. doi: 10.4088/JCP.10m06382oli
5. Pellicano C, Benincasa D, Pisani V, Buttarelli FR, Giovannelli M, Pontieri FE. Prodromal non-motor symptoms of Parkinson's disease. *Neuropsychiatr Dis Treat.* (2007) 3:145. doi: 10.2147/nedt.2007.3.1.145
6. Das S, Zhang Z, Ang LC. Clinicopathological overlap of neurodegenerative diseases: a comprehensive review. *J Clin Neurosci.* (2020) 78:30–3. doi: 10.1016/j.jocn.2020.04.088
7. Fillit HM. The pharmacoeconomics of Alzheimer's disease. *Am J Manag Care.* (2000) 6:S1139–48.
8. Fox C, Smith T, Maidment I, Hebding J, Madzima T, Cheater F, et al. The importance of detecting and managing comorbidities in people with dementia? *Age Ageing.* (2014) 43:741–3. doi: 10.1093/ageing/afu101
9. Eyre H, Baune B, Lavretsky H. Clinical advances in geriatric psychiatry: a focus on prevention of mood and cognitive disorders. *Psychiatr Clin North Am.* (2015) 38:495–514. doi: 10.1016/j.psc.2015.05.002
10. Richardson S, Sinha A, Vahia I, Dawson W, Kaye J, Reynolds III CF, et al. Brain health living labs. *Am J Geriatr Psychiatry*. (2020) 29:698–703. doi: 10.1016/j.jagp.2020.11.010
11. el Kaliouby R. *We Need Computers With Empathy*. Cambridge, MA: MIT Technology Review (2017).
12. Picard RW. *Affective Computing*. Cambridge, MA: MIT press (2000).
13. World Health Organization: Depression. World Health Organization (2020).
14. Forbes M, Rego T, Lavretsky H. *III CFR. Convergence Mental Health Across the Life Span: Advances in Precision Geriatric Psychiatry, Convergence MEntal Health: A Transdisciplinary Approach to Innovation*. Oxford: Oxford University Press (2020).
15. Horackova K, Kopecek M, Machu V, Kagstrom A, Aarsland D, Motlova LB, et al. Prevalence of late-life depression and gap in mental health service use across European regions. *Eur Psychiatry*. (2019) 57:19–25. doi: 10.1016/j.eurpsy.2018.12.002
16. Zhong BL, Xu YM, Xie WX, Liu XJ, Huang ZW. Depressive symptoms in elderly chinese primary care patients: prevalence and sociodemographic and clinical correlates. *J Geriatr Psychiatry Neurol.* (2019) 32:312–8. doi: 10.1177/0891988719862620
17. Eyre HA. P. Siddarth, van Dyk K, N St Cyr, Baune BT, Barrio JR, Small GW, Lavretsky H. Neural correlates of apathy in late-life depression: a pilot [(18) F]FDDNP positron emission tomography study. *Psychogeriatrics.* (2017) 17:186–93. doi: 10.1111/psyg.12213
18. Norton S, Matthews FE, Barnes DE, Yaffe K, Brayne C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. *Lancet Neurol.* (2014) 13:788–94. doi: 10.1016/S1474-4422(14)70136-X
19. Williamson JR, Quatieri TF, Helfer BS, Horwitz R, Yu B, Mehta DD. Vocal biomarkers of depression based on motor incoordination. In: *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge* (Barcelona). (2013). p. 41–8.
20. Cummins N, Epps J, Breakspear M, Goecke R. An investigation of depressed speech detection: Features and normalization. In: *Twelfth Annual Conference of the International Speech Communication Association* (Florence). (2011).
21. Trevino AC, Quatieri TF, Malyska N. Phonologically-based biomarkers for major depressive disorder. *EURASIP J Adv Signal Process.* (2011) 2011:1–18. doi: 10.1186/1687-6180-2011-42
22. Scherer S, Stratou G, Gratch J, Morency LP. *Investigating Voice Quality as a Speaker-Independent Indicator of Depression and PTSD*. Lyon: Interspeech (2013). p. 847–51.
23. Williamson JR, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD. Vocal and facial biomarkers of depression based on motor incoordination and timing. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (Orlando, FL). (2014). p. 65–72.
24. Cohn JF, Kruez TS, Matthews I, Yang Y, Nguyen MH, Padilla MT, et al. Detecting depression from facial actions and vocal prosody (2009). In: *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE* (Amsterdam). (2009). p. 1–7. doi: 10.1109/ACII.2009.5349358
25. Maddage NC, Senaratne R, Low LS, Lech M, Allen N. Video-based detection of the clinical depression in adolescents. *Annu Int Conf IEEE Eng Med Biol Soc.* (2009) 2009:3723–6. doi: 10.1109/IEMBS.2009.5334815
26. Stratou G, Scherer S, Gratch J, Morency LP. Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences, (2013). In: *Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE* (Geneva). (2013). p. 147–52.
27. Joshi J, Dhall A, Goecke R, Breakspear M, Parker G. Neural-net classification for spatio-temporal descriptor based depression analysis. In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE* (Tsukuba). (2012). p. 2634–8.
28. Pampouchidou A, Simos PG, Marias K, Meriaudeau F, Yang F, Pediaditis M, et al. Automatic assessment of depression based on visual cues: a systematic review. *IEEE Trans Affect Comput.* (2017) 10:445–70. doi: 10.1109/TAFFC.2017.2724035

29. Cohn JF, Cummins N, Epps J, Goecke R, Joshi J, Scherer S. Multimodal assessment of depression from behavioral signals. *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2* (New York, NY: ACM Books). (2018). p. 375–417.
30. Dibeklioglu H, Hammal Z, Cohn JF. Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding. *IEEE J Biomed Health Inform.* (2018) 22:525–36. doi: 10.1109/JBHI.2017.2676878
31. Joshi J, Goecke R, Parker G, Breakspear M. Can body expressions contribute to automatic depression analysis? (2013). In: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE* (Shanghai). (2013). p. 1–7. doi: 10.1109/FG.2013.6553796
32. Alghowinem S, Goecke R, Wagner M, Parkerx G, Breakspear M. Head pose and movement analysis as an indicator of depression, (2013). In: *Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE* (Geneva). (2013). p. 283–288.
33. Alghowinem S, Goecke R, Wagner M, Parker G, Breakspear M. Eye movement analysis for depression detection, (2013). In: *IEEE International Conference on Image Processing, IEEE* (Beijing). (2013). p. 4220–4.
34. Vesel C, Rashidisabet H, Zulueta J, Stange JP, Duffecy J, Hussain F, et al. Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: a BiAffect iOS study. *J Am Med Inf Assoc.* (2020) 27:1007–18. doi: 10.1093/jamia/ocaa057
35. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. *Health Inf Sci Syst.* (2018) 6:8. doi: 10.1007/s13755-018-0046-0
36. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In: *Proceedings of the International AAAI Conference on Web and Social Media* (Montreal, QC). (2013).
37. Cacheda F, Fernandez D, Novoa FJ, Carneiro V. Early detection of depression: social network analysis and random forest techniques. *J Med Internet Res.* (2019) 21:e12554. doi: 10.2196/12554
38. Aalbers G, McNally RJ, Heeren A, De Wit S, Fried EI. Social media and depression symptoms: a network perspective. *J Exp Psychol: Gen.* (2019) 148:1454. doi: 10.1037/xge0000528
39. Song H, You J, Chung JW, Park JC. "Feature Attention Network: Interpretable Depression Detection from Social Media," In: *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, China* (2018).
40. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Curr Opin Behav Sci.* (2017) 18:43–9. doi: 10.1016/j.cobeha.2017.07.005
41. Randall N, Bennett CC, Šabanović S, Nagata S, Eldridge L, Collins S, et al. More than just friends: in-home use and design recommendations for sensing socially assistive robots (SARs) by older adults with depression. *Paladyn, J Behav Rob.* (2019) 10:237–55. doi: 10.1515/pjbr-2019-0020
42. Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. *Biol Psychiatry.* (2012) 72:580–7. doi: 10.1016/j.biopsych.2012.03.015
43. Yang Y, Fairbairn C, Cohn JF. Detecting depression severity from vocal prosody. *IEEE Trans Affect Comput.* (2013) 4:142–50. doi: 10.1109/T-AFFC.2012.38
44. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J. Analysis of acoustic space variability in speech affected by depression. *Speech Commun.* (2015) 75:27–49. doi: 10.1016/j.specom.2015.09.003
45. Scherer S, Hammal Z, Yang Y, Morency LP, Cohn JF. Dyadic behavior analysis in depression severity assessment interviews. *Proc ACM Int Conf Multimodal Interact.* (2014) 2014:112–9. doi: 10.1145/2663204.2663238
46. Dibeklioglu H, Hammal Z, Yang Y, Cohn JF. Multimodal detection of depression in clinical interviews. In: *Proceedings of the (2015). ACM on International Conference on Multimodal Interaction.* (2015). p. 307–10. doi: 10.1145/2818346.2820776
47. Joshi J, Dhall A, Goecke R, Cohn JF. Relative body parts movement for automatic depression analysis, (2013). In: *Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE* (Geneva). (2013). p. 492–7.
48. Sweeney JA, Strojwas MH, Mann JJ, Thase ME. Prefrontal and cerebellar abnormalities in major depression: evidence from oculomotor studies. *Biol Psychiatry.* (1998) 43:584–94. doi: 10.1016/S0006-3223(97)00485-X
49. Mastoras RE, Iakovakis D, Hadjidimitriou S, Charisis V, Kassie S, Alsaadi T, et al. Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Sci Rep.* (2019) 9:1–12. doi: 10.1038/s41598-019-50002-9
50. Bettmann JE, Anstadt G, Casselman B, Ganesh K. Young adult depression and anxiety linked to social media use: assessment and treatment. *Clin Soc Work J.* (2021) 49:368–79. doi: 10.1007/s10615-020-00752-1
51. Doryab A, Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and fitbit data. *JMIR Mhealth Uhealth.* (2019) 7:e13209. doi: 10.2196/13209
52. Arnold AJ, Winkielman P. Smile (but only deliberately) though your heart is aching: Loneliness is associated with impaired spontaneous smile mimicry. *Soc Neurosci.* (2021) 16:26–38. doi: 10.1080/17470919.2020.1809516
53. Kanai R, Bahrami B, Duchaine B, Janik A, Banissy MJ, Rees G. Brain structure links loneliness to social perception. *Curr Biol.* (2012) 22:1975–9. doi: 10.1016/j.cub.2012.08.045
54. Hunt MG, Marx R, Lipson C, Young J. No more FOMO: Limiting social media decreases loneliness and depression. *J Soc Clin Psychol.* (2018) 37:751–68. doi: 10.1521/jscp.2018.37.10.751
55. Scoglio AA, Reilly ED, Gorman JA, Drebing CE. Use of social robots in mental health and well-being research: systematic review. *J Med Internet Res.* (2019) 21:e13322. doi: 10.2196/13322
56. Chen SC, Jones C, Moyle W. Social robots for depression in older adults: a systematic review. *J Nurs Scholarsh.* (2018) 50:612–22. doi: 10.1111/jnu.12423
57. Beuscher LM, Fan J, Sarkar N, Dietrich MS, Newhouse PA, Miller KF, et al. Socially assistive robots: measuring older adults' perceptions. *J Gerontol Nurs.* (2017) 43:35–43. doi: 10.3928/00989134-20170707-04
58. Zhao N, Zhang Z, Wang Y, Wang J, Li B, Zhu T, et al. See your mental state from your walk: Recognizing anxiety and depression through Kinect-recorded gait data. *PLoS ONE.* (2019) 14:e0216591. doi: 10.1371/journal.pone.0216591
59. Armstrong T, Olatunji BO. Eye tracking of attention in the affective disorders: a meta-analytic review and synthesis. *Clin Psychol Rev.* (2012) 32:704–23. doi: 10.1016/j.cpr.2012.09.004
60. Yu B, Quatieri TF, Williamson JR, Mundt JC. Cognitive impairment prediction in the elderly based on vocal biomarkers. In: *Sixteenth Annual Conference of the International Speech Communication Association* (Dresden). (2015).
61. Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *NPJ Digit Med.* (2019) 2:1–9. doi: 10.1038/s41746-019-0084-2
62. Schröter A, Mergl R, Bürger K, Hampel H, Möller H-J, Hegerl U. Kinematic analysis of handwriting movements in patients with Alzheimer's disease, mild cognitive impairment, depression and healthy subjects. *Dement Geriatr Cogn Disord.* (2003) 15:132–42. doi: 10.1159/000068484
63. Hutton JT, Nagel J, Loewenson RB. Eye tracking dysfunction in Alzheimer-type dementia. *Neurology.* (1984) 34:99–99. doi: 10.1212/WNL.34.1.99
64. Zhu J, Wang Z, Gong T, Zeng S, Li X, Hu B, et al. An improved classification model for depression detection using EEG and eye tracking data. *IEEE Trans Nanobioscience.* (2020) 19:527–37. doi: 10.1109/TNB.2020.2990690
65. Ntracha A, Iakovakis D, Hadjidimitriou S, Charisis VS, Tsolaki M, Hadjileontiadis LJ. Detection of mild cognitive impairment through natural language and touchscreen typing processing. *Front Digit Health.* (2020) 19. doi: 10.3389/fdgth.2020.567158
66. Fraser KC, Rudzicz F, Hirst G. Detecting late-life depression in Alzheimer's disease through analysis of speech and language. In: *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology* (San Diego, CA). (2016). p. 1–11.
67. Stasak B, Epps J, Schatten HT, Miller IW, Provost EM, Armey MF. Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. *Speech Commun.* (2021) 132:10–20. doi: 10.1016/j.specom.2021.05.004

68. Amico F, Healy G, Arvaneh M, Kearney D, Mohedano E, Roddy D, et al. Multimodal validation of facial expression detection software for real-time monitoring of affect in patients with suicidal intent. *Eur Psychiatry.* (2016) 33:S596. doi: 10.1016/j.eurpsy.2016.01.2225
69. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digit Med.* (2019) 2:1–11. doi: 10.1038/s41746-019-0166-1
70. Vannoy S, Gable S, Brodt M, Cadet M, Andrews B, Maloney M. An application of eye tracking technology to detect attention bias for suicide related stimuli (2016).
71. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering shifts to suicidal ideation from mental health content in social media. In: Proceedings of the (2016). *CHI conference on human factors in computing systems* (San Jose, CA). (2016). p. 2098–110.
72. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. *Biomed Inform Insights.* (2018) 10:1–11. doi: 10.1177/1178222618792860
73. De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: *Proceedings of the 5th annual ACM web science conference* (New York, NY). (2013). p. 47–56.
74. Alzheimer's Assocation: (2021). Alzheimer's Disease Facts and Figures (2021).
75. Population Reference Bureau: Dementia Cases Expected to Triple by 2050 as World Population Ages (2012).
76. Sperling R, Mormino E, Johnson K. The evolution of preclinical Alzheimer's disease: implications for prevention trials. *Neuron.* (2014) 84:608–22. doi: 10.1016/j.neuron.2014.10.038
77. Cummings J, Aisen P, Apostolova L, Atri A, Salloway S, Weiner M. Aducanumab: Appropriate use recommendations. *J Prev Alzheimers Dis.* (2021) 1–13. doi: 10.14283/jpad.2021.41
78. König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement (Amst).* (2015) 1:112–24. doi: 10.1016/j.dadm.2014.11.012
79. Martínez-Sánchez F, Meilán JJG, Carro J, Ivanova OA. Prototype for the voice analysis diagnosis of Alzheimer's disease. *J Alzheimers Dis.* (2018) 64:473–81. doi: 10.3233/JAD-180037
80. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. *Digit Biomarkers.* (2021) 5:78–88. doi: 10.1159/000515346
81. Pistono A, Jucla M, Barbeau EJ, Saint-Aubert L, Lemesle B, Calvet B, et al. Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. *J Alzheimers Dis.* (2016) 50:687–98. doi: 10.3233/JAD-150408
82. Yeung A, Iaboni A, Rochon E, Lavoie M, Santiago C, Yancheva M, et al. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimers Res Ther.* (2021) 13:1–10. doi: 10.1186/s13195-021-00848-x
83. Ahmed SA. Haigh MF, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain.* (2013) 136:3727–37. doi: 10.1093/brain/awt269
84. Seidl U, Lueken U, Thomann PA, Kruse A, Schröder J. Facial expression in Alzheimer's disease: impact of cognitive deficits and neuropsychiatric symptoms. *Am J Alzheimers Dis Other Demen.* (2012) 27:100–6. doi: 10.1177/1533317512440495
85. Burton KW, Kaszniak AW. Emotional experience and facial expression in Alzheimer's disease. *Aging Neuropsychol Cogn.* (2006) 13:636–51. doi: 10.1080/13825580600735085
86. Verghese J, Robbins M, Holtzer R, Zimmerman M, Wang C, Xue X, et al. Gait dysfunction in mild cognitive impairment syndromes. *J Am Geriatr Soc.* (2008) 56:1244–51. doi: 10.1111/j.1532-5415.2008.01758.x
87. Lyons BE, Austin D, Seelye A, Petersen J, Yeargers J, Riley T, et al. Pervasive computing technologies to continuously assess Alzheimer's disease progression and intervention efficacy. *Front Aging Neurosci.* (2015) 7:102. doi: 10.3389/fnagi.2015.00232
88. Ellis RJ, Ng YS, Zhu S, Tan DM, Anderson B, Schlaug G, et al. A validated smartphone-based assessment of gait and gait variability in Parkinson's disease. *PLoS ONE.* (2015) 10:e0141694. doi: 10.1371/journal.pone.0141694
89. Buracchio T, Dodge HH, Howieson D, Wasserman D, Kaye J. The trajectory of gait speed preceding mild cognitive impairment. *Arch Neurol.* (2010) 67:980–6. doi: 10.1001/archneurol.2010.159
90. Albers MW, Gilmore GC, Kaye J, Murphy C, Wingfield A, Bennett DA, et al. At the interface of sensory and motor dysfunctions and Alzheimer's disease. *Alzheimers Demen.* (2015) 11:70–98. doi: 10.1016/j.jalz.2014.04.514
91. Crutcher MD, Calhoun-Haney R, Manzanares CM, Lah JJ, Levey AI, Zola SM. Eye tracking during a visual paired comparison task as a predictor of early dementia. *Am J Alzheimers Dis Other Demen.* (2009) 24:258–66. doi: 10.1177/1533317509332093
92. Zola SM, Manzanares C, Clopton P, Lah J, Levey A, A. behavioral task predicts conversion to mild cognitive impairment and Alzheimer's disease. *Am J Alzheimers Dis Other Demen.* (2013) 28:179–84. doi: 10.1177/1533317512470484
93. Fernández G, Manes F, Politi LE, Orozco D, Schumacher M, Castro L, et al. Patients with mild Alzheimer's disease fail when using their working memory: evidence from the eye tracking technique. *J Alzheimer's Dis.* (2016) 50:827–38. doi: 10.3233/JAD-150265
94. Ladas A, Frantzidis C, Bamidis P, Vivas AB. Eye blink rate as a biological marker of mild cognitive impairment. *Int J Psychophysiol.* (2014) 93:12–6. doi: 10.1016/j.ijpsycho.2013.07.010
95. Gills JL, Bott NT, Madero EN, Glenn JM, Gray MA. short digital eye-tracking assessment predicts cognitive status among adults. *GeroScience.* (2021) 43:297–308. doi: 10.1007/s11357-020-00254-5
96. Rabinowitz I, Lavner Y. Association between finger tapping, attention, memory, and cognitive diagnosis in elderly patients. *Percept Mot Skills.* (2014) 119:259–78. doi: 10.2466/10.22.PMS.119c12z3
97. Austin D, Jimison H, Hayes T, Mattek N, Kaye J, Pavel M. Measuring motor speed through typing: a surrogate for the finger tapping test. *Behav Res Methods.* (2011) 43:903–9. doi: 10.3758/s13428-011-0100-1
98. Stringer G, Couth S, Brown L, Montaldi D, Gledson A, Mellor J, et al. Can you detect early dementia from an email? A proof of principle study of daily computer use to detect cognitive and functional decline. *Int J Geriat Psychiatry.* (2018) 33:867–74. doi: 10.1002/gps.4863
99. Gras LZ, Kanaan SF, McDowd JM, Colgrove YM, Burns J, Pohl PS. Balance and gait of adults with very mild Alzheimer disease. *J Geriatr Phys Ther.* (2015) 38:1–7. doi: 10.1519/JPT.0000000000000020
100. Wittwer JE, Webster KE, Menz HB, A. longitudinal study of measures of walking in people with Alzheimer's disease. *Gait Posture.* (2010) 32:113–7. doi: 10.1016/j.gaitpost.2010.04.001
101. Maquet D, Lekeu F, Warzee E, Gillain S, Wojtasik V, Salmon E, et al. Gait analysis in elderly adult patients with mild cognitive impairment and patients with mild Alzheimer's disease: simple versus dual task: a preliminary report. *Clin Physiol Funct Imaging.* (2010) 30:51–6. doi: 10.1111/j.1475-097X.2009.00903.x
102. Freitas Pereira ML, Camargo ZA, Aprahamian I, Forlenza OV. Eye movement analysis and cognitive processing: Detecting indicators of conversion to Alzheimer's disease. *Neuropsychiatr Dis Treat. (*2014) 10:1273–85. doi: 10.2147/NDT.S55371
103. Coubard OA. What do we know about eye movements in Alzheimer's disease? The past 37 years and future directions. *Fut Med.* (2016) 10:677–80. doi: 10.2217/bmm-2016-0095
104. Husebo BS, Heintz HL, Berge LI, Owoyemi P, Rahman AT, Vahia IV. Sensing technology to monitor behavioral and psychological symptoms and to assess treatment response in people with dementia. A systematic review. *Front Pharmacol.* (2020) 10:1699. doi: 10.3389/fphar.2019.01699
105. Au-Yeung TM, Miller L, Beattie Z, May R, Cray HV, Kabelac Z, et al. Monitoring behaviors of patients with late-stage dementia using passive environmental sensing approaches: a case series. *Am J Geriat Psychiatry.* (2021) 30:1–11. doi: 10.1016/j.jagp.2021.04.008
106. Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JJG. Ten years of research on automatic voice and speech analysis of people with Alzheimer's Disease and mild cognitive impairment: a systematic review article. *Front Psychol.* (2021) 12:645. doi: 10.3389/fpsyg.2021.620251
107. Martínez F. Acoustic analysis of speech and voice disorders in patients with Lewy Body Diseases (2019).

108. Smith KM, Williamson JR, Quatieri TF. Vocal markers of motor, cognitive, and depressive symptoms in Parkinson's disease, (2017). In: *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE* (San Antonio, TX). (2017). p. 71–78.
109. Williamson JR, Quatieri TF, Smith KM. *Vocal Markers of Motor, Cognitive, and Depressive Symptoms in Parkinson's Disease*. Lexington, MA: MIT Lincoln Laboratory Lexington United States (2017).
110. Coates L, Shi J, Rochester L, Del Din S, Pantall A. Entropy of real-world gait in Parkinson's disease determined from wearable sensors as a digital marker of altered ambulatory behavior. *Sensors.* (2020) 20:2631. doi: 10.3390/s20092631
111. Mosimann UP, Müri RM, Burn DJ, Felblinger J, O'Brien JT, McKeith IG. Saccadic eye movement changes in Parkinson's disease dementia and dementia with Lewy bodies. *Brain.* (2005) 128:1267–76. doi: 10.1093/brain/awh484
112. Adams WR. High-accuracy detection of early Parkinson's Disease using multiple characteristics of finger movement while typing. *PLoS ONE.* (2017) 12:e0188226. doi: 10.1371/journal.pone.0188226
113. Bayat S, Babulal GM, Schindler SE, Fagan AM, Morris JC, Mihailidis A, et al. GPS driving: a digital biomarker for preclinical Alzheimer disease. *Alzheimer's Res Ther.* (2021) 13:1–9. doi: 10.1186/s13195-021-00852-1
114. Picard RW. Affective computing: from laughter to IEEE. *IEEE Trans Affect Comput.* (2010) 1:11–7. doi: 10.1109/T-AFFC.2010.10
115. Alzheimer's disease facts and figures. *Alzheimers Dement*. (2016) 12:459–509. doi: 10.1016/j.jalz.2016.03.001
116. Women's Brain Project (2021).
117. Organization WH. *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. World Health Organization (2016).
118. Chuang CS, Lin CL, Lin MC, Sung FC, Kao CH. Migraine and risk of dementia: a nationwide retrospective cohort study. *Neuroepidemiology*. (2013) 41:139–45. doi: 10.1159/000353559
119. Lee NT, Resnick P, Barton G. *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms. Washington, DC, USA:Brookings Institute (*2019).
120. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Conference on fairness, accountability and transparency, PMLR* (New York, NY). (2018). p. 77–91.
121. *Ethics and Governance of Artificial Intelligence for Health: World Health Organization Guidance*. Geneva: World Health Organization (2021).
122. National Institute of Health: Bridge to Artificial Intelligence (Bridge2AI) (2021).


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

# Advantages of publishing in Frontiers

**OPEN ACCESS**
Articles are free to read for greatest visibility and readership

**FAST PUBLICATION**
Around 90 days from submission to decision

**HIGH QUALITY PEER-REVIEW**
Rigorous, collaborative, and constructive peer-review

**TRANSPARENT PEER-REVIEW**
Editors and reviewers acknowledged by name on published articles

**Frontiers**
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

**Visit us:** www.frontiersin.org
**Contact us:** frontiersin.org/about/contact

**REPRODUCIBILITY OF RESEARCH**
Support open data and methods to enhance research reproducibility

**DIGITAL PUBLISHING**
Articles designed for optimal readership across devices

**FOLLOW US**
@frontiersin

**IMPACT METRICS**
Advanced article metrics track visibility across digital media

**EXTENSIVE PROMOTION**
Marketing and promotion of impactful research

**LOOP RESEARCH NETWORK**
Our network increases your article's readership