




The need for a harmonized speech dataset for Alzheimer's disease biomarker development

Nicole L. Bjorklund¹, Howard Fillit¹, Kristina Malzbender², Shobha Purushothama¹, Lampros Kourtis^{2,3,4*}

¹Alzheimer's Drug Discovery Foundation, New York, NY 10019, USA

²Gates Ventures, Kirkland, WA 98033, USA

³Circadic, Arlington, MA 02476, USA

⁴Clinical & Translational Science Institute, Tufts University Medical Center, Boston, MA 02111, USA

***Correspondence:** Lampros Kourtis, Circadic, Arlington, MA 02476, USA. lampros@circadic.io

Academic Editor: Lindsay A. Farrer, Boston University School of Medicine, USA

Received: August 24, 2020 **Accepted:** October 20, 2020 **Published:** December 31, 2020

Cite this article: Bjorklund NL, Fillit H, Malzbender K, Purushothama S, Kourtis L. The need for a harmonized speech dataset for Alzheimer's disease biomarker development. *Explor Med.* 2020;1:359-63. <https://doi.org/10.37349/emed.2020.00024>

Abstract

This commentary is the product of a concerted effort to understand the needs, barriers, and gaps in the field of speech and language biomarkers for Alzheimer's disease (AD). It distills interviews, surveys, and extensive correspondence with global leaders in the areas of dementia research, clinical trials, linguistics, and data analytics into an idealized clinical-study design for the harmonized collection of voice recordings. The ultimate goal of the effort is to democratize the ongoing speech and language analytics efforts by making such rich datasets available to the wider research ecosystem.

Keywords

Voice data collection, speech and language biomarkers, acoustic and linguistic features, cohort characterization, open-access database, Alzheimer's disease, neurodegenerative disease

Introduction

Successful drug development for Alzheimer's disease (AD) depends on clinicians' ability to diagnose and monitor the disease's progression—especially via clear, measurable biomarkers that can detect subtle changes in patients' pathologic neuronal decline long before they show other, more serious symptoms. Alterations of speech and language are showing promise as possible early biomarkers of AD [1].

Researchers can collect and analyze speech and language information using new and improved technology, hardware and data analytics. Likewise, ubiquitous use of smart devices enables remote data collection, both active (prompted by the user) and passive (without user prompts). These tools can measure acoustic features such as pitch and amplitude, as well as lexical and syntactic aspects of speech and features of written language such as text contextual or semantic information—all of which are associated with early AD and its progression [2, 3].

© The Author(s) 2020. This is an Open Access article licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



Yet researchers have not been able to fully take advantage of the opportunities these tools can offer. To optimize speech and language biomarker discovery, researchers need a comprehensive speech-sample repository that covers a large, diverse cohort of subjects representing different accents, languages, speech and language components, and disease stages. They also need state-of-the-art participant characterization along with harmonized protocols and standards that cover the types of speech and language samples. These activities are nearly impossible for most research groups or startups to achieve on their own due to the costs associated with participant characterization [such as repeated positron emission tomography (PET) scans, magnetic resonance imaging (MRI), and blood-based biomarkers in large longitudinal cohorts].

We believe a global partnership between clinicians, researchers, and data scientists can meet these challenges, facilitating further identification, development and validation of speech-based biomarkers to enable researchers to apply artificial-intelligence algorithms for AD screening, detection, prediction, diagnosis, and monitoring. Existing consortia in related fields demonstrate that global collaboration and data sharing can indeed produce meaningful results (Table 1).

Table 1. Selected examples of productive consortia efforts

Consortium type	Weblink	Goals	Outcome
Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA) Network	http://enigma.ini.usc.edu/	Fifty active working groups dedicated to sharing ideas, algorithms, data, and information	Replicating promising findings among a network of researchers in the field of imaging genetics
Institutional Neuroimaging Data-Sharing Initiative	http://fcon_1000.projects.nitrc.org/	Access to thousands of functional MRI datasets for their analysis	Standardized imaging data
Critical Path for Parkinson's Consortium	https://c-path.org/programs/cpp/	Links academic researchers with scientists from the pharmaceutical industry, government agencies, and patient-advocacy organizations	Facilitate the development of therapies with improved clinical endpoints
Linguistic Data Consortium	https://www ldc.upenn.edu/	Open network of universities, libraries, corporations, and government-research laboratories that supports language-related education, research and technology development	Creating and sharing linguistic resources, such as data, tools and standards
AphasiaBank	https://aphasia.talkbank.org/	The development of standardized evaluation methods to guide the development and evaluation of effective methods for improving language usage in people with aphasia	Improvement of patient-oriented treatment of aphasia

This manuscript summarizes interviews, surveys, and extensive correspondence with global leaders in the areas of dementia research, clinical trials, linguistics, and data analytics and outlines an ideal approach to generating a comprehensive, gold-standard set of speech- and language-based data. The end product of such an approach would be: 1) a rich, diverse, longitudinal, repeatedly measured, high-quality set of speech samples and 2) participant-characterization labels (such as imaging, blood-based biomarkers, or neuropsychological testing and clinical diagnosis) that researchers around the world can use to generate new diagnostic and prognostic algorithms. Here we focus on three broad areas: cohort selection, study design, and data collection and dissemination.

Cohort/patient selection

To obtain a set of speech samples that has the greatest utility for researchers, patients should range from healthy controls (HC) with no risk factors and HC with high risk factors [such as having the apolipoprotein E (*APOE*) 4 allele] to preclinical/suspected to prodromal/mild cognitive impairment (MCI) to mild AD and eventually to AD. Including disease controls, such as Parkinson's or frontotemporal degeneration, is also important.

To use speech and language biomarkers as a measure of disease progression, the cohorts selected should allow for repeated, longitudinal, preferably high-frequency measurements. The cohorts should also include characterization using digital or traditional neuropsychological tests, genetic testing, MRI or PET imaging,

and blood-based or cerebrospinal fluid (CSF) biomarkers. Finally, researchers should try to mitigate the burdens—of cost, time, and effort—on patients.

Protocol and study design and data collection

Given the diversity of potential approaches to collecting, processing, and analyzing speech- and language-based data, study design for a gold-standard dataset must carefully consider the attributes outlined in [Table 2](#).

Table 2. Key considerations for the collection and development of a harmonized speech and language dataset

Cohort/patient selection	Study design	Data collection and dissemination
<ul style="list-style-type: none"> Which patients should be included (e.g., pre-clinical, SCD, mild AD, MCI)? Should disease controls be included (e.g., PD, FTD)? Should patients with known risk factors (e.g., APOE positive) be included? What is the appropriate balance of ethnicities, geographic diversity, and genders? What is the appropriate cohort size? What is the minimum level of characterization required (e.g., neuropsychological tests, PET imaging, blood, CSF biomarkers)? How should a diversity of languages and accents be incorporated? 	<ul style="list-style-type: none"> Which types of speech samples should be collected? Consider spanning cognitive domains and cognitive load levels. Are the tests active or passive? How are the tests categorized (e.g., constrained, non-constrained)? Which speech sample collection tests will be best to characterize a patient's disease progression? Per disease stage? Which tests will be most applicable to real-world settings? How can speech sample collection be harmonized? How can researchers ensure that data coming from different cohorts can be aggregated to one database? 	<ul style="list-style-type: none"> What is the appropriate frequency and duration of test administration? Will the setting of data collection (in-clinic or remote) impact patient compliance? Can tests be refined/adjusted over time if needed? How can annotation and collection be consistently ensured? How can broad data sharing and access be facilitated while ensuring patient privacy?

SCD: subjective cognitive decline; PD: Parkinson's disease; FTD: frontotemporal dementia

Voice recordings can be constrained, in which the subject is prompted to perform a clearly defined task such as recalling a list of words; unconstrained, in which speech samples are collected while the user is performing basic communication tasks such as talking with someone on the telephone; or somewhere in between ([Figure 1](#)).

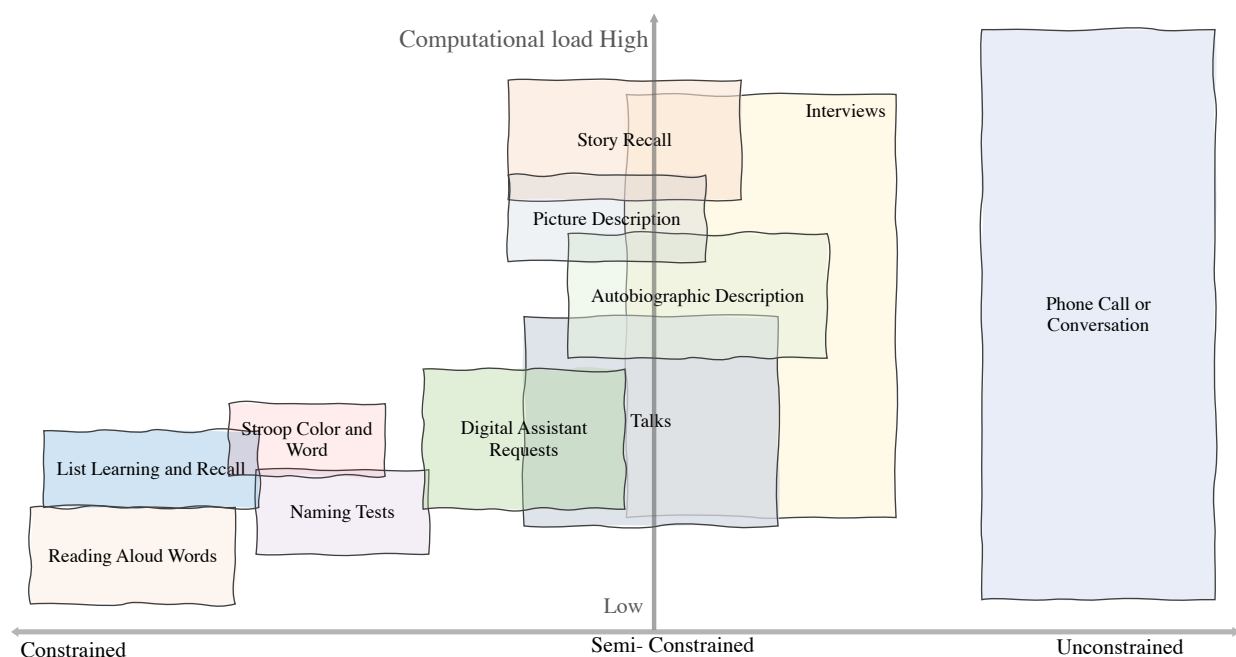


Figure 1. Speech and language tests—diversity in terms of type and patient burden

Each of these approaches carries a different cognitive load and highlights different aspects of speech or language, and likewise provides the ability to reveal changes in speech, language and interaction patterns

in addition to changes across multiple cognitive domains. A dataset that combines the raw data from these assessments will provide the largest variety of speech and language features for analysis.

Researchers must consider which aspects of speech and communication they can reliably and consistently collect across different cohorts using different technology platforms. They should also develop standardized protocols for administering, recording, labeling, and annotating (where applicable) the voice samples. These standardized protocols will truly permit meaningful comparisons.

Data dissemination and privacy

The utility of the speech and language dataset depends on researchers' ability to access and analyze it while still maintaining patient privacy and data security. An ideal data sharing platform should address aspects of access (open, limited, nested) and enable virtual processing of datasets within the repository to maintain patient privacy. Possible approaches include allowing researchers to process raw data, run their algorithms, and extract features on a remote privacy-maintaining server *versus* downloading onto individual computers. Moreover, different levels of processing could be allowed for each interested party, such as limiting access to phonetic and acoustic features, thereby preserving subjects' privacy as much as possible. Approaches to maintaining privacy are evolving and best practices should be implemented and updated when appropriate. Existing voice repositories, such as the Linguistic Data Consortium and DementiaBank, can serve as an example [4, 5].

Conclusion

A comprehensive, harmonized, open-access speech-sample repository covering well characterized, large, diverse cohort(s) of subjects can enable the development of better biomarkers that characterize the onset and progression of AD (and other neurodegenerative diseases) in a minimally invasive, low-cost way. At the same time, democratizing speech and language analytics must be a joint effort: at every step along the way, collaboration and cooperation are key. Together, these can facilitate truly seismic shifts in neurodegeneration research.

Abbreviations

AD: Alzheimer's disease

MRI: magnetic resonance imaging

PET: positron emission tomography

Declarations

Acknowledgments

The authors would like to thank the companies and investigators who provided their insightful expertise for the development of this commentary. The authors would also like to thank Dr. Niranjana Bose, Health and Life Sciences, Gates Ventures for thoughtful discussions and critical review of the manuscript. Also, the authors would like to thank Visar Berisha, Julie Liss, Shira Hahn, and Jessica Robin for their help in creating Figure 1. Writing assistance was provided by Emily Lieb.

Author contributions

NLB, LK, KM, SP and HF contributed to the conception and design of the commentary. NLB, LK and SP wrote the first draft of the manuscript. SP created the tables. All authors contributed to manuscript revision, read and approved the submitted version.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and materials

Not applicable.

Funding

Funding was provided through the Diagnostics Accelerator, an initiative funded by a coalition of funders including the Alzheimer's Drug Discovery Foundation and Gates Ventures. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright

© The Author(s) 2020.

References

1. Kourtis LC, Regele OB, Wright JM, Jones GB. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *npj Digital Med.* 2019;2:9.
2. Boschi V, Catricalà E, Consonni M, Chesi C, Moro A, Cappa SF. Connected speech in neurodegenerative language disorders: a review. *Front Psychol.* 2017;8:269.
3. Mueller KD, Hermann B, Mecollari J, Turkstra LS. Connected speech and language in mild cognitive impairment and Alzheimer's disease: a review of picture description tasks. *J Clin Exp Neuropsychol.* 2018;40:917-39.
4. Linguistics Data Consortium [Internet]. Linguistic Data Consortium, The Trustees of the University of Pennsylvania. c1992-2020 - [cited 2020 Nov 9]. Available from: <https://www ldc.upenn.edu/>
5. DementiaBank [Internet]. DementiaBank Consortium, Carnegie Mellon University. c1999 - [cited 2020 Nov 9]. Available from: <https://dementia.talkbank.org/>