DxA SPEECH CONSORTIUM (SpeechDx)

Harmonized Repository for the Exploration of Speech and Language Biomarkers for Neurodegenerative Disease

DATA COLLECTION STUDY PROCEDURES

Table of Contents

1.0 List of abbreviations

Term	Description
AD	Alzheimer's Disease
ADDI	Alzheimer's Disease Data Initiative
AE	Adverse Event
AES	Advanced Encryption Standard
API	Application Programming Interface
CN	Cognitively Normal
DAC	Data Access Committee
FTD	Frontotemporal Degeneration
GDPR	General Data Protection Regulation
ICF	Informed Consent Form
IRB	Institutional Review Board
LAR	Legally Authorized Representative
LBD	Lewy Body Dementia
MCI	Mild Cognitive Impairment
MRI	Magnetic Resonance Imaging (Brain)
NP	Neuropsychiatric
PET	Positron Emission Tomography
PD	Parkinson's Disease
PI	Principal Investigator
PII	Personal Identifiable Information
SUID	Site Unique Participant ID (PI/site assigned)
SCIUD	Speech Consortium Study Unique ID (MMS assigned)
SCD	Subjective Cognitive Decline
SSH	Secure Shell
UDS	Unified Data Set

2.0 Summary

Table 1. Summary of the project

Project Title	DxA Speech Consortium: Harmonized Repository for the Exploration of Speech and Language Biomarkers for Neurodegenerative Disease.
Study Design	Longitudinal, observational study for collection of at-home speech and language tasks
Duration of Study Participation	Each participant will contribute data for at least 3 years upon enrolment.
Sample Size	Final number of participants is expected to be approximately N=2000, Ages >45
Objectives	To collect participant speech and language data together with cognitive and participant characterization to create a database of speech collected over 3 years that will be offered for the development of new cognitive symptom tracking and quantification by qualified researchers.
Purpose	The overall purpose of this study is to generate a comprehensive repository of speech samples collected via harmonized procedures from a large, diverse cohort of subjects representing different accents, languages, speech and language components, and AD disease stages.
Patient Population	Participants currently enrolled in longitudinal, observational studies will be eligible to participate in the SpeechDx observational study. Participants will range from cognitively normal to AD.

3.0 Introduction

Background

Existing Need

Successful drug development for Alzheimer's disease (AD) depends on a clinician's ability to diagnose and quantify disease severity and progression—especially via clear, measurable biomarkers that can detect subtle changes in patients' pathologic cognitive capacities and neuronal decline long before they show more serious symptoms. While blood biomarkers show great promise in detecting Alzheimer's related pathology early, many individuals who are amyloid positive do not go on to develop symptomatic AD [14]. Alterations of speech and language show promise as possible early biomarkers of AD [1], and may be used alongside AD-related blood biomarkers to predict who will go on to experience cognitive decline. Such changes hold additional promise as potential diagnostic or prognostic biomarkers for AD [1, 6, 8].

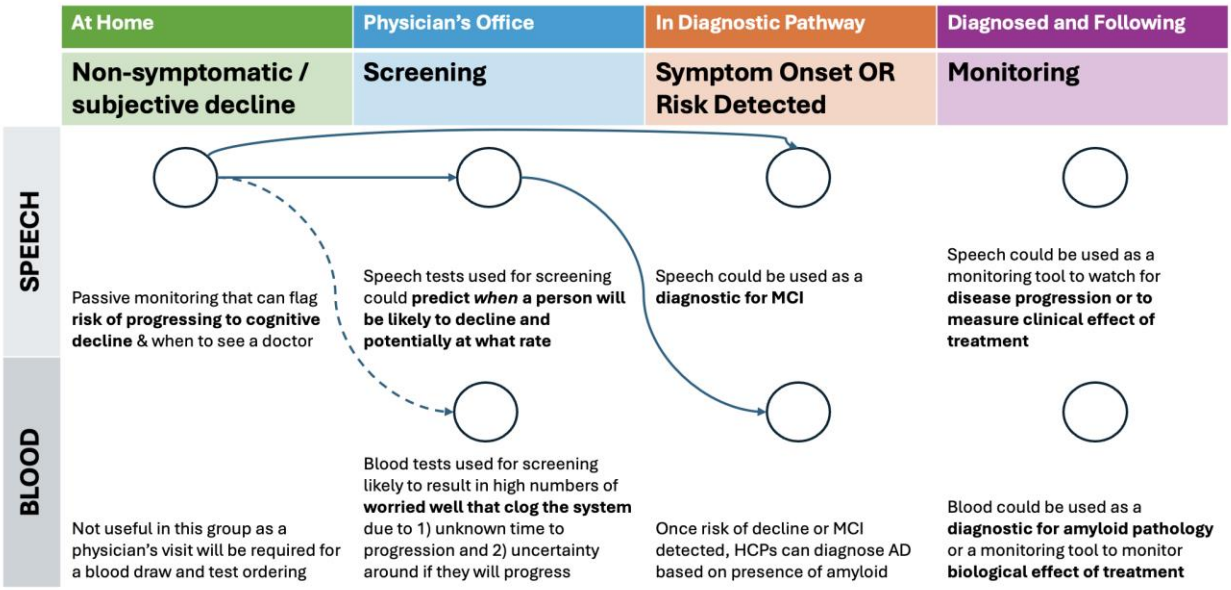


Figure 1. *An example of a diagnostic ecosystem to combine functional (speech) and biology (blood) biomarkers to forecast, diagnose and monitor Alzheimer's Disease. Speech and language biomarkers can work hand-in-hand with blood biomarkers to allow for early detection of the disease. Once the disease is diagnosed, speech and languagee biomarkers can be used to monitor its progression and the efficacy of potential treatments.*

Speech and language data is widespread and can be collected with low burden to patients. 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, lexical, and syntactic aspects of speech, as well as features of written language —all of which are associated with early AD and its progression [2,3]. These include aspects of pitch, amplitude, contextual information, semantic information, and more.

Thus far, 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, longitudinal speech-sample

repository that covers a large, diverse cohort of subjects representing different AD disease stages. They also need state-of-the-art participant characterization, with harmonized clinical and AD biomarker data that can be paired with these speech and language samples. Due to the costs associated with participant characterization (such as repeated PET scans, MRIs, and blood-based biomarkers in large longitudinal cohorts), these activities are nearly impossible for most research groups or startups to achieve on their own [4]. SpeechDx engages a global partnership between clinicians, researchers, and data scientists to meet these challenges. By facilitating further identification, development, and validation of speech-based biomarkers, this partnership will enable researchers to develop artificial-intelligence algorithms for AD screening, detection, prediction, diagnosis, and monitoring.

Use of speech and language in Alzheimer's disease prediction

Speech and language analysis shows great promise in aiding early Alzheimer's disease detection and prediction, with several recent publications highlighting this promise. Amini et al [6] used natural language processing (NLP) and machine learning to predict the progression to AD within 6 years using speech data. They found that linguistic features extracted from speech, along with demographic information, could predict AD progression with an accuracy of 78.5% and a sensitivity of 81.1%. Notably, the model using text features alone outperformed models using demographic features alone and even surpassed the predictive power of traditional neuropsychological test scores. This highlights the potential for using structured interviews to capture language deficits indicative of cognitive decline. In another study, Ezzati et al [11] used various machine learning methods and features from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database to predict the clinical outcome of MCI participants, achieving an accuracy of 77% in predicting conversion to AD over 48 months. Eyigoz et al [8] focused on predicting AD onset in cognitively normal individuals using automated linguistic analysis of written responses to the cookie-theft picture-description task. The authors found that linguistic markers were significantly associated with future AD onset and achieved an AUC of 0.74 and an accuracy of 0.70 in predicting AD onset, outperforming models based solely on non-linguistic variables like APOE genotype, demographics, and traditional neuropsychological test scores.

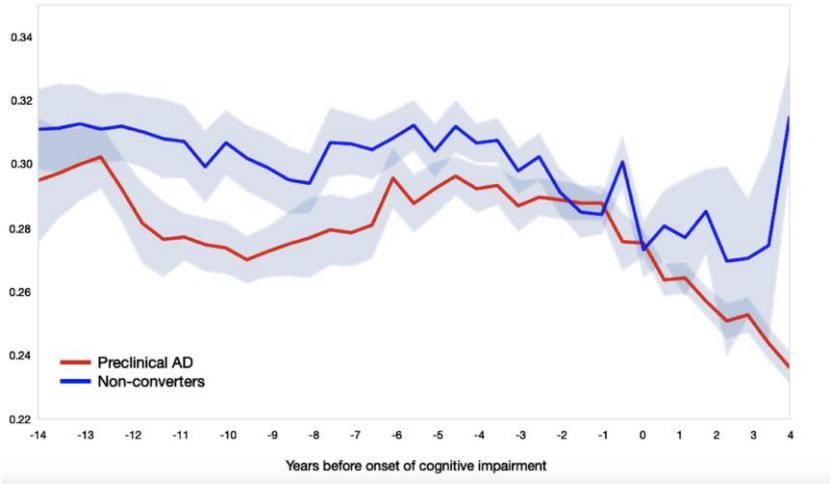


Figure 2. Feature extraction and Machine Learning using speech recordings from the Framingham Heart Study was able to predict conversion to AD long before other modalities (from Eyigoz et al 2020).

The literature emphasizes the potential of speech and language analysis to develop cost-effective, accessible, and easy-to-administer screening tools and discusses the advantages of using speech data, which can be collected remotely and inexpensively. With the average person spending significant daily time communicating through a digital device, this presents a unique opportunity to investigate *functional* symptoms in a manner that is passive, unobtrusive, continuous, globally applicable, cost effective, and scalable.

Use of speech and language in Alzheimer's disease screening and diagnosis.

Several sources report high diagnostic performance of speech and language biomarkers in detecting AD and mild cognitive impairment (MCI). Hajjar et al [10] used lexical-semantic and acoustic features derived from audio recordings of mixed populations and achieved an AUC of 0.80 for lexical-semantic scores and 0.77 for acoustic scores in detecting MCI. Chou et al [12] found that a combination of linguistic features and biomarkers could achieve up to 88% accuracy in distinguishing early-AD patients from normal controls. Furthermore, a correlation between speech and language biomarkers and neuroanatomy has been established. For example, García-Gutiérrez et al [7] used various machine learning (ML) models to classify amyloid status based on these acoustic features and achieved an AUC of 0.79 surpassing that of conventional neuropsychological tests typically used to evaluate cognitive functions. Similarly, Konig et al [9] report similar results, with an accuracy of 87% ± 3% in distinguishing between healthy controls and AD patients. Hajjar et al [10] found that acoustic scores derived from speech analysis were significantly associated with hippocampal volume, a well-established hallmark of AD progression. In conclusion, speech and language analysis holds significant potential as a valuable tool for the detection, diagnosis, and monitoring of AD.

3.1 Study Approach

These procedures have been designed in collaboration with global leaders in the areas of dementia research, clinical trials, linguistics, and data analytics and aim to generate a comprehensive, gold-standard set of speech- and language-based data. The end product of this data-gathering exercise will be:

- A PII-free*, diverse, longitudinal, multi-lingual, high-quality set of speech sample audio files, along with their annotated transcripts and meta-information. All audio files will be matched to:
- Pseudonymized ** participant characterization labels (such as imaging, blood-based biomarkers, and neuropsychological testing and clinical diagnosis), harmonized across all participating sites.

* Personally Identifiable Information (PII) is defined as any information that can distinguish or trace an individual's identity. This includes direct identifiers like names, Social Security numbers, or biometric records, as well as indirect identifiers, such as date of birth or mother's maiden name, or unique achievements and traits, when combined with other linked data. The U.S. National Institute of Standards and Technology (NIST) describes PII as "information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual."

**According to the European General Data Protection Regulation (GDPR), pseudonymization is "..processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept seperately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable nature person."

The SpeechDx dataset will first be made available to a limited group of program partners and later to the wider community of researchers around the world to help them create, train, and validate speech and language biomarker algorithms. To share the dataset with any given entity, the entity must prove that it complies with national or local regulations, such as GDPR or the AI Act. Due to privacy restrictions and obligations, researchers are required to undertake to protect the privacy of the individuals participating in this study. The speech samples/data are still treated as personal information under European/UK privacy laws.

3.1.1 Cohort/Patient Selection Strategy

To obtain a set of speech samples that has the greatest utility for researchers, participants from ongoing cohorts (see section 5.2 for detailed descriptions) range from healthy controls (HC) with no risk factors and HC with high risk factors (such as APOE4 allele), to preclinical/suspected, to prodromal/MCI, to mild AD, and eventually to early AD. In addition, a small number of voice recordings from other neurodegenerative diseases, such as Parkinson's or frontotemporal degeneration, may be collected. To enable the development of speech biomarkers with potential for earliest possible disease detection and prediction of future cognitive decline, the SpeechDx participant population is biased towards earlier stages of the cognitive spectrum.

Selected cohorts will participate in longitudinal speech data collection, with quarterly speech data collected performed remotely (at home). Participants will be characterized using digital or traditional neuropsychological tests, genetic testing, MRI imaging, and blood-based AD biomarkers, while a subset of participants will also be characterized via amyloid PET imaging and AD biomarkers in CSF. The addition of the proposed measurements has been designed to minimize cost, time, and effort burdenfor both study staff and participants.

3.1.2 **Data to be collected and respective test types.**

Speech and language stimuli 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 (See Figure 1).

Each of these approaches carries a different cognitive load, highlights different aspects of speech or language, and provides the ability to reveal changes in speech and language patterns, in addition to changes across multiple cognitive domains. This proposal includes assessments that range from unconstrained to semi-constrained, aiming to generate an optimal dataset that offers the greatest variety of speech and language features for analysis.

Standardized procedures for administering, recording, labeling, and annotating (where applicable) the speech and language samples are under development. Detailed descriptions of the proposed assessments can be found in Section 5.2 and the Appendix.

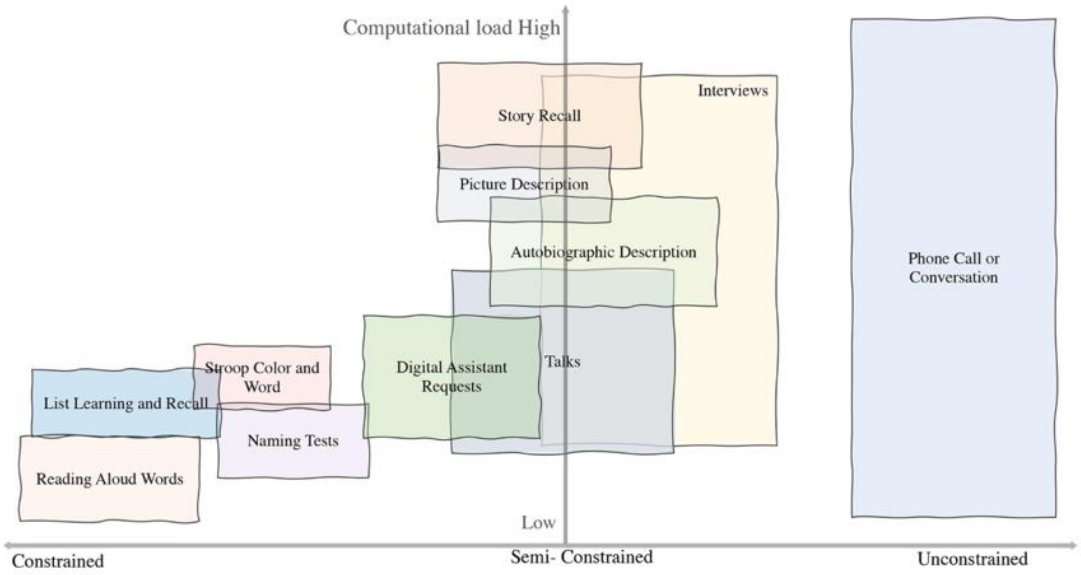


Figure 3. A classification of reported tests that have at least an element of speech or language (from Bjorklund et al, 2020)

Study Objectives

SpeechDx is a data-collection study with the following objectives:

- Enable the development of speech-related AD biomarker products that can effectively predict future cognitive decline
- Enable the development of speech-related AD biomarker products that accurately detect cognitive impairment
- Enable the development of speech-related AD biomarker products that be used for monitoring of cognitive impairment and/or treatment efficacy

SpeechDx aims to do this via:

- Collection of a unified, longitudinal dataset of recorded speech-based tests across multiple sites, leveraging well-characterized subjects from existing cohort studies. Subject characterization includes demographics, neuropsychological tests results with component scores, biometrics, clinical diagnosis, raw MRI (T1 and T2 FLAIR) imaging files and analysis results, raw PET imaging files when available, blood biomarkers (i.e. plasma tau217), medication history and other biomarker and clinical data where available.
- Sharing this unified dataset with researchers around the world while maintaining data security and privacy for the purpose of generating and validating speech-based AD biomarkers.

4.0 Study Design

This is a data-collection study from approximately 2000 study participants that will unify data coming from ongoing studies at Study Centers. Study participants will be recruited from each Site's Principal Investigator's (PI) existing patient populations and will agree to amendments to the existing ICF that they have consented to or will agree to new ICFs accounting for the activities in these procedures.

4.1 Power Calculations

As this is an exploratory study conducted for the purpose of data collection, a formal power calculation is not possible. However, we aim to capture the data of at least 100 converters (participants that convert from one diagnostic state to another throughout the duration of the study (e.g. from Cognitively Normal to MCI)) in both English and Spanish. A statistical model was created based on the demographics of the participating population, disease statistics, and a conventional dropout rate, in order to estimate the number of participants who will convert from one state to the next during the three years of SpeechDx data collection. This model has been implemented into a dashboard that allows us to track predicted conversion rates as participants enroll into the study at various dropout rates.

Statistical Model: We created a model using NACC* (**National Alzheimer's Coordinating Center)** data. The goal of this analysis is to use this model to estimate the number of conversions (CN to MCI, MCI to AD) within the SpeechDx population. This set of NACC data contains individuals diagnosed as cognitively normal (CN), subjective cognitive decline (SCD), mild cognitive impairment (MCI), and Alzheimer's Disease (AD). The current variables allow us to build two models (one for CN conversions, and one for conversions to AD) to predict cognitive decline as a categorical measure.

Predicting Transitions from CN: The first model identifies patients who are likely to transition from CN to another clinical diagnosis. The key variables in the model include age, years smoking, baseline MMSE, amyloid levels in CSF, and tau levels in CSF. As anticipated, patients who are older, have smoked for longer, have lower MMSE scores, lower amyloid-42 levels in CSF, and higher tau levels in CSF are predicted to have higher probability of transitioning from CN.

Predicting Transitions to AD: The second model identifies patients who are likely to transition to AD from any other clinical diagnosis. The key variables in the model included age, ApoE4 carrier status, baseline MMSE, baseline CDR-sb, amyloid in CSF, total tau in CSF, and baseline diagnosis. As anticipated, patients who are older, are ApoE4 carriers, have lower MMSE scores, higher CDR-sb scores, lower amyloid-42 levels in the CSF, and higher tau levels in the CSF are predicted to have higher probability of transitioning to AD. Patients who are diagnosed with MCI at baseline are also predicted to have a higher probability of transitioning.

*NACC Data: The NACC data contains Alzheimer's disease research information collected from 37 Alzheimer's Disease and Research Centers (ADRC) across 26 states. Various biomarkers, clinical outcomes, and demographic information are collected from people at or above the age of 36, with most subjects falling between the ages of 60 and 100. For the purpose of this analysis, we limit the population to individuals who had at least one visit three or four years after their baseline visit.

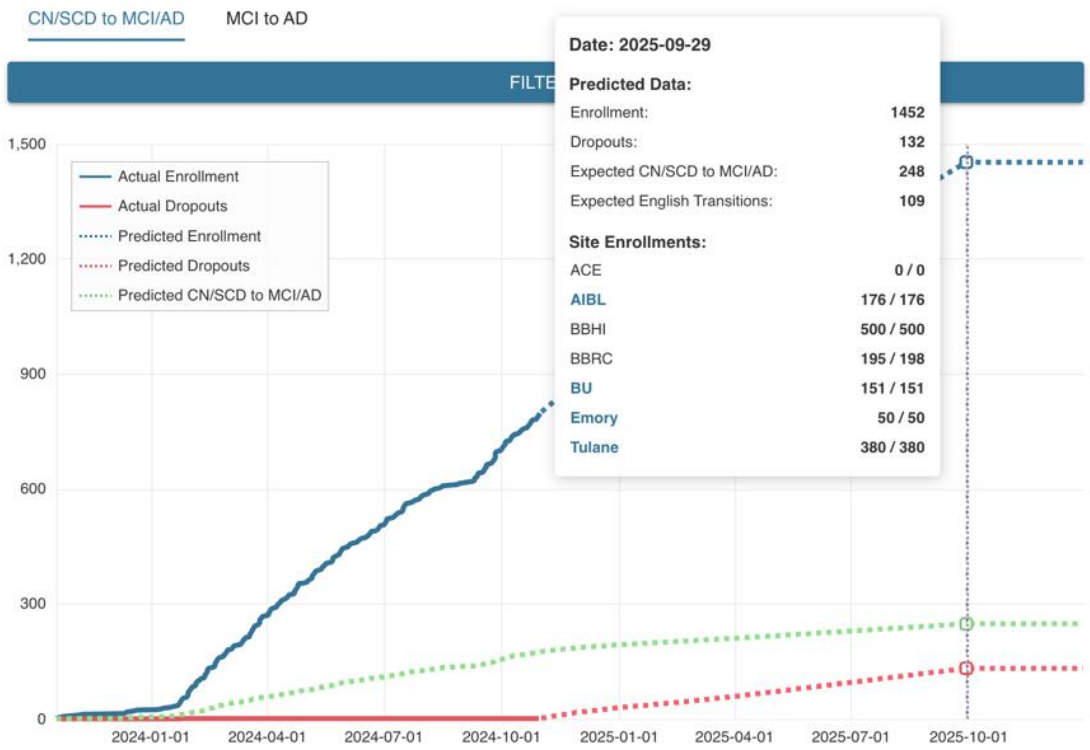


Figure 4. Internal dashboard modeling projected disease state conversions over the 3-year duration of data active SpeechDx collection, assuming a 20% participant dropout rate. Numbers represented are projections and are subject to change

4.2 Participating Sites

Data collection will include voice recordings from a varied population in terms of accent, language, and stage of disease. We do not yet know how speech and language features may change as neurodegeneration progresses. There will likely be a number of potential uses, ranging from early detection of cognitive decline to monitoring disease progression and monitoring participants during treatment or clinical trials. Researchers accessing the final dataset will define their targeted context of use and bring their algorithms and analytical tools to the dataset, and as such will be expected to define their statistical analysis methods *a priori*. To prepare for success and enable the broadest possible analyses, we have carefully chosen cohorts that represent a heterogenous population that includes a sufficient sample size of each group for meaningful comparisons. The Study Centers have been selected based on several factors including cohort characteristics, size, ongoing study compatibility, language and others. Table 3 presents the Study Centers currently participating in this data collection project.

Table 2. The clinical centers/cohorts planned to participate in the study

Cohort	PI / Contact	Language	Rationale

The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) https://aibl.csiro.au/	Paul Maruff Larry Ward Colin Masters	English (Australian)	Large, well characterized, prospective longitudinal study of cognition and enriched for MCI and AD
Boston University (BU) ADRC https://www.bu.edu/alzresearch/	Rhoda Au	English	Leading digital biomarker efforts for ADRCs Synergies with other ongoing studies
Emory University Goizueta ADRC http://alzheimers.emory.edu/	Allan Levey Felicia Goldstein	English	Very large ongoing Healthy Aging Study (16,000 subjects)
Tulane Bogalusa Registry (BHS) https://bogalusaheartstudy.org/	Ileana De Anda-Duran	English	A 50-year epidemiological registry, focused on a biracial, semi-rural community in Bogalusa
Fundacio ACE, Institut Catala de Neurociencies Aplicades https://fundacioace.com	Sergi Valero	Spanish, Catalan	Includes mainly population with clinical diagnosis of MCI and AD
Barcelona Brain Health Initiative (BBHI) https://bbhi.cat/	Gabriele Cattaneo Javier Solana-Sánchez	Spanish, Catalan	Population-based cohort Includes deep phenotyping for 1,500 participants
Barcelona Beta Brain Research Center (BBRC) ALFA+beta-AARC Study https://www.barcelonabeta.org/en	Andreea Rădoi Gonzalo Sánchez-Benavides	Spanish, Catalan	Consists of asymptomatic, younger subjects Several proposed tests already being administered

4.3 Cohort Types

As AD presents across a large spectrum of symptoms, participants will range from Cognitively Normal (CN) to participants with subjective cognitive decline (SCD) but no detectable cognitive decline via traditional neuropsychological testing to prodromal/MCI, to mild AD, and eventually to AD. Participants are already enrolled in existing studies, and Table 4 summarizes the initial estimates of participants that will contribute to the study (numbers are prospective estimates and may vary as participants are enrolled into the study).

Table 3. A summary of the diagnoses across cohorts (projected as of Dec 2024)

SITE	TOTAL N	CN/SCD	MCI	AD
Ace	200	-	168	32
AIBL	300	176	97	27
Barcelona Beta Brain Research Center (BBRC)	200	198	2	-
BBHI Guttmann Institute	470	470	-	-

BU ADC	180	136	38	6
Emory ADC*	65	50	10	5
Tulane Bogalusa Registry (BHS)	400	380		20

*Site in the process of contracting, not contracting at the time this document was released.

4.4 Cohort Characterization

To enable the development of speech and language biomarkers that can be prognostic or used as a measure of disease progression, the selected cohorts will participate in annual assessments in clinic. Participants in the selected cohorts will be characterized using digital or traditional neuropsychological tests, genetic testing, MRI, and blood-based AD biomarkers, with a subset of participants also characterized by PET imaging or CSF biomarkers. Baseline participant characterization is critical to achieving a uniform dataset that will facilitate the generation of comparable ground truth / label a set of properties for further analytical processing. The addition of the proposed measurements has been designed to mitigate the burdens—of cost, time, and effort—for both study staff and participants. Table 5 presents an overview of participant characterization methods (initial assessment of data available as reported by each Site).

Table 4.Clinical information to be contributed by each Clinical Site (to be confirmed)

Cohort	Blood (pTau217)	CSF (Amyloid/Tau)	Amyloid PET	Tau PET	FDG PET	MRI	Neuropsych. tests	Speech Assessment
ACE								
AIBL								
BBHI								
BU ADRC								
BHS								
BBRC ALFA+								
BBRC AARC								
Emory ADRC*								
KI*								
CPLS*								

*Subject to contract execution

More specifically, based on the clinical characterization contributions from each site, harmonization tables have been created. Blood, MRI Imaging and neuropsychiatric testing methods have been captured for all the Sites in as much detail as possible.

4.5 Participating Sites Description

4.5.1 Boston University Alzheimer's Disease Research Center

The Boston University Alzheimer's Disease Research Center (BU ADRC) was established in 1996 as one of 31 centers in the US-funded by the National Institutes of Health to advance research on Alzheimer's disease and related conditions. The BU ADRC cohort (n>200) includes individuals diagnosed as cognitively normal, mild cognitively impaired or AD of mild severity. For those with cognitive impairment, a study partner is also recruited to assist with research participation. This cohort comes in for annual assessment that includes a blood draw from which all AD biomarkers are measured , a comprehensive cognitive assessment and annual diagnostic review.

4.5.2 Emory University Alzheimer's Disease Research Center

Founded in 2005, Emory's Goizueta Alzheimer's Disease Research Center (Emory ADRC) is one of 31 active centers in the nation supported by the National Institutes of Health. The goal of these centers is to bring scientists together to facilitate their research and help learn more about Alzheimer's and related diseases. The Goizueta ADRC is also committed to the education of health care professionals, persons with Alzheimer's disease, their families, and the community to aid in understanding, diagnosis and treatment of these illnesses. This active research site has multiple ongoing longitudinal studies, some of which are already obtaining voice recordings with the goal of identifying features that may indicate the early stages of cognitive decline.

4.5.3 Guttmann Institute (Barcelona Brain Health Initiative)

The Barcelona Brain Health Initiative (BBHI) is an ongoing prospective longitudinal study focused on identifying determinants of brain health. The main objectives are: (i) to characterize lifestyle, cognitive, behavioral and environmental markers related to a given individual's cognitive and mental functions in middle to old age, (ii) to assess the biological determinants predictive of maintenance of brain health, and (iii) to evaluate the impact of a controlled multi-dimensional lifestyle intervention on improving and maintaining brain health. In Phase II of the study, a sub-group of 1,000 individuals is undergoing detailed in-person evaluations repeated at eighteen-month intervals. These evaluations will provide deep phenotyping of brain function, including medical, neurological and psychiatric examinations, assessment of physical fitness, neuropsychological assessments, structural and functional brain magnetic resonance imaging, electroencephalography and perturbation-based non-invasive brain stimulation evaluations of brain activity, as well as collection of biological samples. Results of this initiative will critically contribute to a better understanding of the determinants to promote and maintain brain health over the lifespan.

4.5.4 Barcelona Beta Brain Research Center - ALFA+ Beta AARC

The Alfa Study is a research platform to identify early pathophysiological characteristics of Alzheimer's disease, as well as its early detection, and to be able to develop prevention strategies. It was launched in 2013, thanks to the impetus of the 'la Caixa'. It is one of the most complex and largest initiatives, with 2,700 cognitively unimpaired participants, dedicated to early detection and prevention of Alzheimer's. Participants are between 45 and 74 years old, and are mostly descendants of patients with Alzheimer's, so the cohort is enriched in genetic factors related to the disease.

4.5.5 Bogalusa Heart Study (Tulane University)

The Bogalusa Heart Study uniquely focused on a biracial, semi-rural community, consisting primarily of Black and White residents in Bogalusa, Louisiana to explore how cardiovascular risk factors vary across racial and socioeconomic lines, providing valuable insights into health disparities. By including children and young adults from diverse backgrounds, the study was able to examine how genetics, lifestyle, and environmental factors differently impacted Black and White participants, revealing important differences in the onset and progression of cardiovascular risk factors. The study's findings underscored the need for targeted public health interventions that consider demographic and socioeconomic diversity, setting a precedent for health research that addresses racial health disparities directly.

4.5.6 The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) Sites

The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) is a study to discover which biomarkers, cognitive characteristics, and health and lifestyle factors determine the subsequent development of symptomatic Alzheimer's Disease (AD). Launched on 14th November 2006, AIBL largest study of its kind in Australia, dedicated to the early detection of Alzheimer's disease and lifestyle interventions.

4.6 Project Plan

SpeechDx is a longitudinal study with speech data collection occurring every 3 months for 3 years. Allowing for 1 year to reach full enrollment, the calendar duration of the study will be 4 years. An initial pilot was conducted at the Boston University ADRC to test participant and staff burden, fine-tune the recruitment and data collection process, and prepare to scale the procedures across all sites. Findings from the pilot study were used to improve the data collection procedures and the data collection app design.

4.6.1 Pilot Phase

The pilot phase included a sample of 20 CN, community-based participants recruited from Boston University ADRC. Participants underwent data collection for 2 or 3 speech collection sessions as described below, allowing for more rapid trouble shooting of the app and scheduled activities. Patient and study staff burden was assessed along with any other required modifications to the procedures to ensure retention and feasibility. This process allowed the technical team to amend any operating procedures, data flows, and management before the main study. Unlike the main phase of the program where speech data is collected every 3 months and clinical assessments performed every 12-18 months, in this pilot phase, participants contributed clinical (synthetic) data once (in the beginning) and speech data weekly for 3 assessments. All data underwent the same processes that are planned for the full data collection phase as shown below:

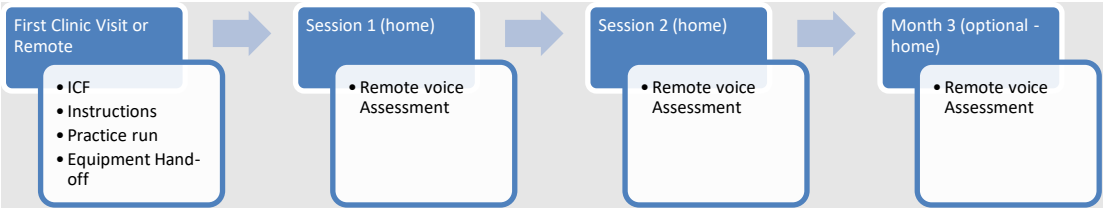


Figure 5. Pilot Activities Schedule

4.6.2 Full Data Collection Phase

Following the pilot data collection phase, the procedures and study app were refined before commencing full enrollment at Sites (Figure 3). Subjects are recruited into the study over a ~18 mo. period and will come into the clinic for their first visit, including all assessments collected by each site under each site's individual protocols (i.e., health history update, collection of demographics, blood draw etc.). At this visit, subjects will be provided with a study-provided device, and study staff will provide a demonstration and instructions for using the app. Subjects may then complete the practice speech assessment in the clinic during this first visit or at home. Voice assessments will be completed remotely every 3 months, with push notifications to prompt the subject to log into the app and complete their assessments. Follow-up phone

calls and/or home visits will be conducted as needed to assist with device/application use. For existing cohorts, enrollment can also be done remotely over the phone, with the tablets being pre-shipped to the participant's address.

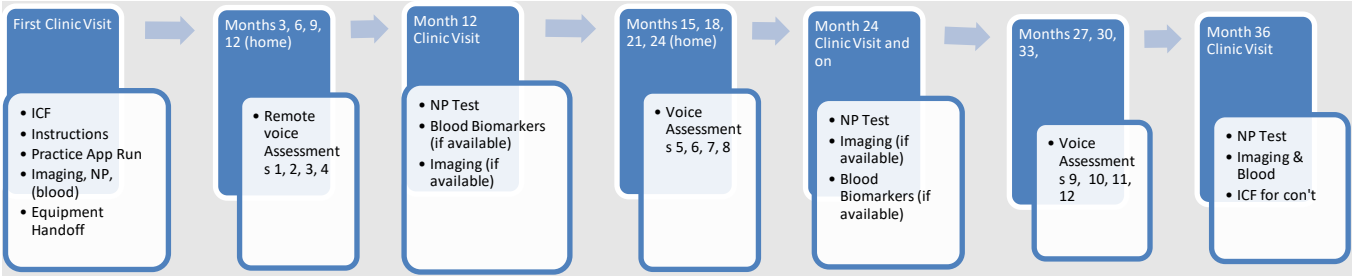


Figure 6. Full Data Collection Phase Plan

4.6.3 New IRB submissions and IRB amendments to the existing studies

Each site will be responsible for sponsoring and developing its own protocol based on these procedures, and for obtaining IRB approval and subject informed consent. The following elements are recommended to be incorporated into the ICF and IRB approvals/amendments that will be pursued as required by each study site.

- Permission to record voice in the process of recalling a story, telling a story, describing a picture, and answering a set of interview questions.
- To the extent required, US study sites will include a HIPAA authorization in the ICF unless a waiver of the authorization requirement has been approved by the applicable IRB.
- Speech Assessments will be completed using a study-provided device.
- Speech Assessments will be completed every 3 months at home, remotely.
- Each participant will be assigned a unique study ID, and no other specific identifiers will be collected
- All speech recordings will be treated as PII/personal data.
- Speech recordings will be transcribed and any personal identifiers inadvertently disclosed will be eliminated.
- Video or photos will not be collected.
- Data will be collected via a secure mobile app and uploaded to a cloud server that complies with applicable requirements and later securely transferred to a secure data repository provided by ADDI's AD Workbench.

4.7 Data Collection and Speech Eliciting Tasks

4.7.1 Recording Equipment

A study-provided Android device will be given to each subject. This approach has important advantages – notably, the ability to standardize device specifications (i.e. sensors, microphone etc.) and to increase inclusivity by facilitating the recruitment of low-income participants. Participants will be given the option of keeping the study-provided device. Study-provided devices will be pre-loaded with the study app. The tablet will also come with a list of preinstalled apps and will be managed to allow installation of several commonly used apps that have been pre-whitelisted.

The Samsung Tab A7 Lite tablet has been selected for this study (Figure 5). This device holds an 8.68-inch LCD screen with a resolution of 800 x 1340 pixels and has 3GB RAM. This product has been commercially available since mid-summer 2021 and is available for bulk purchases.

Devices will be provisioned from and managed by Mobility CG, who will be staging and kitting the devices along with a case and screen protector and will be providing a replacement/repair program in case of damage. Mobility CG will also install and QC the final version of the app on each device. All devices will be managed through Samsung Knox, which allows for concurrent management of the entire device fleet. Any updates or other settings will be managed for all devices through Knox. MCG will maintain that the tablets are up to date with the latest OS security updates.

Speech recordings will be collected for distinct speech tasks listed below during each session. These tasks were selected to represent a range of constrained, semi-constrained and unconstrained tasks, while leveraging robust and well-validated assessments. Further details for each assessment can be found in the Appendix.

4.8 Speech App

A custom app designed and implemented by Zühlke Engineering serves as the primary data collection vehicle for speech data collection in this study. The app is preloaded onto the Android tablets and was specifically built for the SpeechDx study.

The app sends push notifications to remind the participants several days before and during the opening of the test window. Additionally, the app will detect the device's posture to ensure the tablet is positioned upright for optimal sound acquisition.

To prevent recording corruption and session interruptions, the app checks the tablet's battery life at the start of each session and, if necessary, prompts the participant to connect to power. Before each session, participants are asked to go to a quiet place. The app will record 5 seconds of ambient noise to establish an audio baseline and determine if the background noise levels are too high. If noise levels are excessive, the participant will be prompted both verbally and in writing to resolve of noise source or repeat the session at another time.

The app is designed to accommodate users who may be hard of hearing by requiring them to set the volume to at least 80% before starting the assessment. Participants can subsequently lower the volume afterward if desired. Similarly, the app recommends a screen brightness of at least 60%, and it checks this setting, prompting participants to adjust brightness as needed.A visual how-to guide is presented in the case where one of these recommendations have not been met. Should the participant chooses not to comply with these recomendations, the app makes note of this while still proceeding to the main testing phase.

The app presents a series of stimuli, starting with a picture description task (using a new picture for every session), the PHQ-8 and Karolinska questionnaires, and three open-ended questions. The app then proceeds to the picture recall (asks the participant to recall what they saw in the original picture) and the two-story recall tests. Next, the PVT assessment is offered, followed by the storytelling assessment. Finally, a second picture description task (using the same picture for all sessions) is presented. Once the participant completes the session, the data is uploaded to the cloud service securely via an available internet connection.

Table 7. Data Collection App Components

User Check-in	Picture Description	Open Ended Questions	Mood/Condition/Sleep	Vigilance Assessment	Storytelling and Story Recall
The user ID has been auto filled since the last time entered. The first interaction of the user with the device will be in the clinical setup, hence the study staff will pre-code the user ID.	Loads one of the repository pictures according to a prespecified sequence	Load a repository question according to a prespecified sequence. App displays and reads out the topic/question.	User selects a sleepiness score from the displayed list.	PVT vigilance testing for intra-test condition characterization.	Recall: Load and read out a story from a recording to a prespecified sequence
App then records 5 seconds of ambient noise to create baseline and prompt participant to consolidate noise source.	App displays and reads out instructions for the test.	Audio rec is on.	Response is recorded along with date and participant ID.		Storytelling: show a repository picture according to a prespecified sequence. App displays instructions for storytelling task.
	Audio file along with metadata is encrypted (asymmetric) and stored in the local memory. Recording ends when user clicks “Next”.	Audio file along with metadata is encrypted (asymmetric) and stored in the local memory. Recording ends when user clicks “Next”.	User continues to next page.		Audio file along with metadata is encrypted (asymmetric) and stored in the local memory. Recording ends when user clicks “Next”.
	Recall: Without showing the picture again, the app asks participant to recall the second picture shown to him/her. Audio recording is on until the participant clicks “Next”				

4.8.1 Picture Description Task and Recall (Semi-constrained):

The battery consists of discrete pictures (examples are shown in the Appendix) with pictures developed in collaboration with the BU team. At each test administration, the user will be shown 1) an identical picture in all sessions and 2) a second, new picture from the repository. The images will be displayed, and the participant will be asked to verbally describe the image shown with the response being recorded. Audible and written instructions are provided. Before the end of the battery, the participant will be asked to recall the first picture and describe it (delayed recall), without the image being displayed on the screen.

4.8.2 Two Types of Story Recall Tests (Semi-Constrained):

In both versions of the Story Recall Tests, participants are asked to listen to a narrative story and then retell each story in their own words in as much detail as they can remember (immediate recall). The stimuli (stories) have been carefully designed to elicit naturalistic speech, but within a defined domain to increase reliability of analytic measures (semi-constrained speech production). The first Story Recall Test follows similar parameters to the Craft-21 story recall while the second Story Recall Task is meant to overload the participant with information to make the task more difficult. The task has an interaction with episodic memory, one of the established, early cognitive changes in Alzheimer's disease.

The task can be fully self-administered. For each story recall, the mobile application plays an audio recording with instructions to perform the task, including a retelling of a story, or a storytelling task. Once the instructions have been played, the application will start to record the participant. The stories are administered in pairs: Instructions for the first story are given, followed by immediate recall. This is then repeated for a second story. This administration is for test-retest reliability, and to elicit a longer duration of naturalistic speech. Different stimuli are used on different visit-days to mitigate learning effects.

Stories used in the story recall test are taken from a stimuli set of stories that have been balanced for important linguistic and discourse metrics, including: number of information units, number of words, number of sentences. Each recorded story is balanced for reading speed, using a text-to-speech service in the local accent. The stories are administered in the participant's native language, and names of people and places are localized with regards to the geography of the study. Within each assessment session, two stories are administered.

The resulting data from the Story Recall Test are audio recordings of the participant's recalls of each administered story. These recordings can be analyzed with or without the ground truth transcripts (which will be made available for researchers who will work with the resulting data). Transcripts can also be independently analyzed.

4.8.3 Storytelling (Unconstrained):

During the storytelling task, the participant will be shown one picture drawn from a library shown in the Appendix. This picture will be paired with an audible and written prompt that asks the participant to use their imagination to tell a story about a scenario involving the picture shown. The response is recorded.

4.8.4 Open ended questions (Unconstrained):

In this setup, participants will be asked to listen to and/or read an open-ended question and then reply to it. Their responses will be recorded. This process will be repeated three times with different questions. A list of topics for these

questions has been compiled based on literature and personal communication with Investigators. See Appendix for complete list and order of questions.

Karolinska Sleepiness Scale:

The Karolinska Sleepiness scale is a 9-point scale to assess sleepiness and alertness. It is a single screen with 9 options with only one option being selectable. The full content is attached in **Error! Reference source not found.**.

PHQ-8 Questionnaire:

The 8-item Patient Health Questionnaire depression scale (PHQ-8) is a well-established assessment that measures the severity of depressive disorders in clinical studies and clinical practice. The full content is attached in **Error! Reference source not found.**

Vigilance Assessment:

In the Vigilance Test (PVT), the test taker must only tap the screen when a red circle appears on the screen, and not when a white "+" sign appears. Participants metrics include response accuracy, number of correct hits, response time, and intertrial interval.

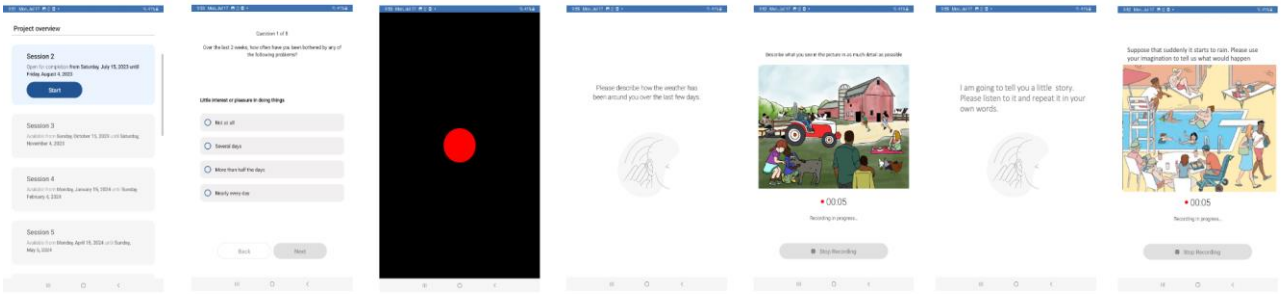


Figure 7. Screenshots from the SpeechDx app presenting the assessments

4.8.5 Order of stimuli:

The stimuli will be delivered in a predefined sequence shown in **Error! Reference source not found.**. After the c ompletion of a particular test, the app will advance to the next test by pressing the "Next" or arrow button.

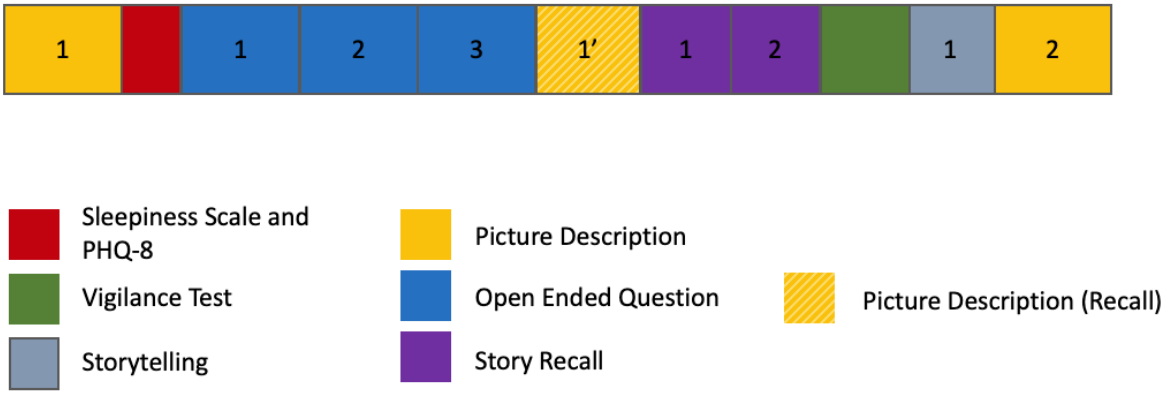


Figure 8. Order of stimuli presented in the speech app

4.8.6 Data Transfer/Storage

Data are encrypted with AES256 encryption and stored on the tablet using 256 AES encryption at rest. When internet connection is available, audio files will be uploaded using HTTPS service (ie. TLS 1.3 or above) from the tablet to a cloud server, where it will be stored encrypted at rest using 256 AES encryption. The audio files are erased from the device as soon as servers confirm they have received and securely stored the data. If the device cannot connect due to Internet connectivity issues, the encrypted data will stay on the device, and the app will continually retry until it successfully transfers the data. Data is temporarily stored on Azure blobs maintained by the Data Hosting Partner, Zühlke Engineering. A third-party service is utilized to remove potential Personal Identifiable Information that has inadvertently been exposed during the recordings (e.g., names, addresses, special achievements, or traits). Once this operation is completed, the audio clips are securely transferred (using secure FTP) to a secure data repository provided by AD Data Initiative's AD Curation Studio and supported by Aridhia's DRE platform, ("AD Data Initiative Curation Studio") where harmonization with de-identified clinical information is performed. Finally, the data is moved internally to AD Data Initiative's AD Workbench, a controlled-access space where it will live ever-after in encrypted format (FIPS 140-2 compliant 256 AES encryption). The AD Data Initiative AD Workbench data repository is protected by data privacy and security controls standard in biomedical and healthcare information technology, including the provision required to be GDPR compliant, including compliance to ISO27701 policies and processes and HITRUST Common Security Framework. All transmissions through AD Workbench interfaces are encrypted with HTTPS or TLS 1.2 and above protocols.

4.8.7 Data Types Captured

- Raw audio at 44100kHz of all assessments (9 per session: 3 questions, 2 pictures, 1 picture recall, 2 story recalls, 1 story telling) in FLAC lossless format
- Mood/Sleepiness review responses

- Vigilance Test results (correct answers, reaction time, etc.)
- Dictionary with the input question/picture/story list for each recording session

4.8.8 Pseudonymization of Voice Recordings and Transcription

Voice data can sometimes inadvertently capture Personal Identifying Information (PII)/personal data that could compromise participant identity if processed in a certain manner.

According to the National Institute of Standards and Technology (NIST) PII is defined as:

Direct Identifiers:

- Full name
- Social Security Number (SSN) or Tax Identification Number
- Passport number, Driver's license number, National ID number, Vehicle registration number, tax identifiers, employer identifiers
- Financial account numbers (e.g., credit card, bank account)
- Biometric data (e.g., fingerprints, facial recognition data, retinal scans)
- Email address (personal or work)
- Phone numbers (mobile, home, or work)
- Physical address (e.g., street, city, postal code)

Indirect Identifiers:

- Date of birth
- Place of birth
- Mother's maiden name
- IP address, Device identifiers (e.g., MAC address, IMEI)
- Login credentials (e.g., username and passwords)
- Geolocation data
- Behavioral data (e.g., purchasing history, browsing behavior)
- Employment details (e.g., job title, work location)
- Medical records or health-related data
- Awards and Recognitions
- Unique achievements that are publicly available
- Public Records of Accomplishments (book, patent, article, blog authorship)
- Personal Traits or Characteristics (birthmarks, scars, tattoos, rare talents that are publicized)
- Professional or Creative Work (art, sports, speeches etc)
- Cultural or Social Identifiers (tribal or cultural memberships if the group is small enough)

To avoid the risk of identification of a subject, and to maintain a database that can be shared with study partners as well as other researchers, all PII/personal data will be spliced out, to the extent possible, using a manual process as follows:

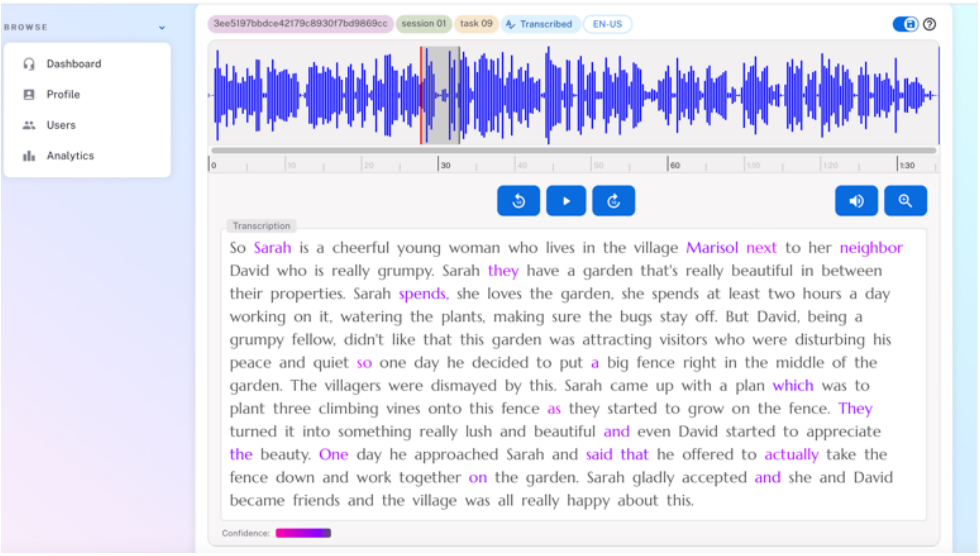
First, each audio clip contributed by each participant will be transferred to a secure server via internet connectivity and will be encrypted at rest. Audio clips are batched in quarterly buckets, and each clip in a quarterly bucket will be manually reviewed by ISO17100 certified translators/reviewers proficient in each language of the Speech Consortium dataset, contracted and trained by Circadic LLC. The reviewers receive special training on identification and removal of

personal data from an audio clip. This training includes a rule that provisions the removal of information that, either alone or in combination with other data, can identify a person within a pool of 1:100,000. For example, a reference to a small town with 2000 residents will be removed, whereas a reference to a large metropolitan area with over 1,000,000 inhabitants will not. An option to evaluate (escalate) a particular PII in question by another reviewer is provided. Audio clips are regularly quality-checked. If a segment of the recording is considered PII/personal data-exposing, that segment will be removed and digitally and permanently replaced by a tone of a certain frequency that will signify removal. This process will be done using specialized audio processing software designed and maintained by Circadic. Once recordings have been pseudonymized, they will be transferred to the ADDI's Curation Studio secure server, where they will live from then on. Upon confirmation of receipt at ADDI, the original files (that contain PII) will subsequently be deleted permanently and non-reversibly from the cloud servers.

In parallel, and during the same reviewer pass, each audio clip will be transcribed to provide a ground truth of the content of each clip. Other information about the clip will also be noted, such as a "silent" clip, in which a participant didn't speak at all, a "Noisy" clip, in which there can be heard other noises and voices, a "low volume" clip that flags clips that have very low levels of voice. This information will be provided on the Site dashboard so that Site managers can better provide advice to their participants.

To enable review of audio clips, a custom app will be developed to allow reviewers to manage the large number of audio files (the total number of clips exceeds 200,000), perform a first pass of the transcription using Automatic Speech Recognition (ASR), and allow for efficient splicing of the clip in case PII/personal data has been identified.

Once the recordings have been transcribed and free of personal data, they will be moved to ADDI's Curation Studio's secure server, where they will subsequently live.



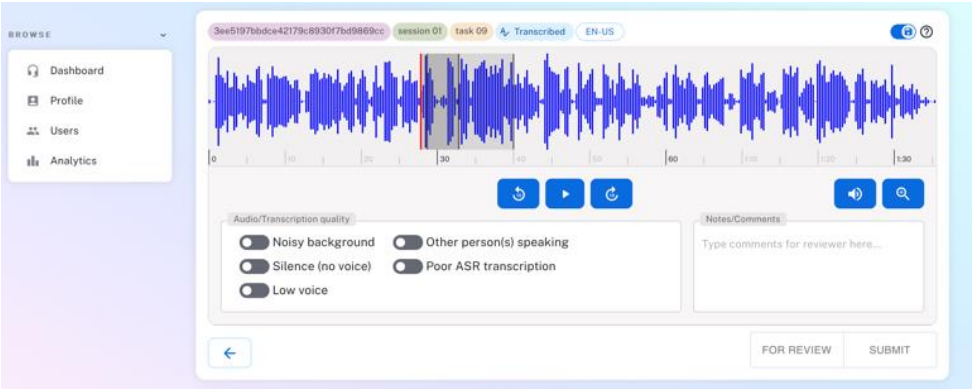


Figure 9. Interface design of the PII/personal data-splicing and transcription application

4.9 Clinical Characterization and Harmonization Plan

SpeechDx is a study that builds on existing and established AD research cohorts that have been carefully selected in terms of compatibility of operational aims. At all sites, participants will perform visits in accordance with each site's preexisting protocols, where each site will collect series of clinical data. This data may include blood and CSF biomarkers, NP testing, among other datapoints. Clinical data may include UDS formatted data and other (raw) data. This data will be contributed to the SpeechDx for use as ground truth data in machine learning model training operations. These procedures do not prescribe clinical data collection procedures to be followed, but rather suggests a process to facilitate incorporation into a single database that will harmonize all data fields under a common dictionary. Sites will individually specify what clinical data the site will contribute to the Speech Consortium and how such data will be collected.

Although every Site may have different procedures to collect data, the SpeechDx program requires the collection of the following datapoints, pseudoanonymized according to applicable standards:

-Full medical history, demographical information, preferably per UDS categories. (1 timepoint, and updates where applicable)

-Quarterly speech recordings via the study provided tablet for at least 3 years (at least 13 timepoints)

-Collection and storage of blood at least at baseline (+/- 6mo) and at the end of the study (+/- 6mo). Intermediate points are welcome to collect. Analysis according to the site's procedures, but should include ptau217. (at least 2 timepoints)

-MRI scanning at baseline (+/- 6mo) and at the end of the study (+/- 6mo). Scanning will occur according to the Site's procedures but should include T1 and T2 FLAIR sequences. (at least 2 timepoints)

-Neuropsychiatric testing yearly or every 18 months, according to the Site's procedures but must include Digit Span Forward and Backward, Category Fluency Animals/Vegetables, Trail Making A&B, Verbal Fluency, Symbol Digit Substitution, MoCA, MMSE. (at least 3 timepoints)

Table 5. Harmonization table for Blood biomarkers for each Site. All Sites are contributing ptau181, ApoE status. Efforts are being made to also include ptau217. All Sites store blood samples for future analysis. The exact assay and instrument used is provided.

	ptau181		ptau217		Ab42		Ab40		NFL		GFAP		ApoE	Biobank?
	Assay	Instrument	Assay	Instrument	Assay	Instrument	Assay	Instrument	Assay	Instrument	Assay	Instrument	Method	
BU	Simoa v2.1 Advantage	Simoa HD-X	Alzpath	Simoa	SimoaNP4E	Simoa HD-X	SimoaNP4E	Simoa HD-X	SimoaNP4E	Simoa HD-X	SimoaNP4E	Simoa HD-X	96 SNPs assay using Standard BioTools' Biomark system	NCRAD for Clinical Core; locally for Registry
Emory	Quanterix	Simoa HD-X	Alzpath	Simoa									TBD	
BBHI	Neuro 3-Plex A	Simoa HD-X							SimoaNFLight	Simoa HD-X			LightBrain	Banked at a Guttman-affiliated institution
BBRC (ALFA+)	Simoa v2.1 Advantage	Simoa HD-X	Janssen and Lilly	Simoa and MSD	SimoaNP4E	Simoa HD-X	SimoaNP4E	Simoa HD-X	SimoaNP4E	Simoa HD-X	SimoaNP4E	Simoa HD-X	Collected	Locally with BBRC
BBRC (Beta)	Simoa v2.1 Advantage	Simoa HD-X	AlzPath	Simoa HD-X	SimoaNP4E	Simoa HD-X	SimoaNP4E	Simoa HD-X	SimoaNP4E	Simoa HD-X	SimoaNP4E	Simoa HD-X	Pending	Locally with BBRC
Ace	Lumipulse® G pTau 181 plasma IRC kit	Lumipulse G100	Fujirebio	Lumipulse G1200	Neuro 4-Plex E Lumipulse® G β-Amyloid 1-42	Simoa HD-1 Lumipulse G1200	Neuro 4-Plex E Lumipulse® G β-Amyloid 1-40	Simoa HD-1 Lumipulse G1200	Neuro 4-Plex E	Simoa HD-1	Neuro 4-Plex E	Simoa HD-1	Thermo Axiom Sp Array of Life Technologies Taqman	Private collection (ISCIII code C.0000299)
Bogalusa	Simoa	Simoa HD-1			Simoa	Simoa HD-1	Simoa	Simoa HD-1	Simoa	Simoa HD-1			TBD	Locally with BHS
AIBL	Simoa v2.1	Simoa HD-X	TBD	TBD									Collected but need info on method of collection	Locally with AIBL

Table 6. MRI Scanning Harmonization Table for all participating Sites. All Sites are contributing T1 and T2 FLAIR sequences and most of them contribute additional sequences. The magnet specifications as well as the sequence TE, TR, TI and resolution is presented.

	Magnet Strength	Magnet Model and coil type	T1	T2	T2 FLAIR	T2*	Difussion (DWI)	Perfusion (PWI)	fMRI	pCASL
BU	3T	Philips Ingenia Elition 32-ch coil	3D MPRAGE	Philips T2	ADNI-3 AP 3D FLAIR	3D multi-echo GRE, TR=56ms, TE 3-46ms (9x), Voxel=1x1x0.5mm	UCSF 3D DTI b=500 (30), b=1000 (48), b=2500 (96)		ADNI-3 AP EPI-BOLD	ADNI-3 3D PL Delay=2000ms
Emory	3T	Siemens MAGNETOM Prisma 32-ch coil	3D MPRAGE TR=2300ms TE=2.96ms TI=900ms Voxel=1mm ISO		3D FLAIR TR=4800ms TE=441ms TI=1650ms Voxel=1x1x1.2mm		EPSE TR=2600ms TE=80ms Voxel=2 mm ISO b=1000(64) b=2000(64) b=3000(64)		BOLD TR=1890ms TE=30ms Voxel= 1.5mm ISO 260 volume	
BBHI	3T	Siemens PRISMA, 32-ch coil	3D MPRAGE TR=2400ms TE=2.22ms, IT=1000ms, FOV=256mm, Voxel=0.8mm ISO	3D SPACE TR=3200ms, TE=563ms, FOV=256mm Voxel=0.8mm	FLAIR TR=9000ms TE=128ms IT=2500ms FOV=220mm Voxel=.9x.9x3.0mm		Hi Res TR=3230ms TE=89.2ms Voxel: 1.5mm ISO @100directions		T2-weighted EPI TR=800ms, TE=37ms, 750 volumes, FOV=208mm, Thickness =2mm	
BBRC	3T	Philips Ingenia CX 32-ch coil	3D TR=9.9ms, TE=4.6ms TI=900ms, Voxel=0.75mm ISO	3D TR=2500ms, TE=264ms, Voxel=1mm ISO	3D FLAIR TR=5500ms, TE=312, IR=1700ms, Voxel=1mm ISO					
Ace	3T	Siemens Magnetom Vida-XQ Numaris/X VA20A-04PW	T1 2D TR = 2200.0ms TE = 2.47ms IT = 900ms FOV = 250mm voxel 0.9mm ISO	T2 2D TR=8570ms TE=106ms Voxel=.2x.2x3 FOV=230mm	T2 3D FLAIR TR = 7000.0ms TE = 393.0ms IT = 2050ms voxel = 1.0mm ISO					
Bogalusa	3T	GE Discovery min 32 channel head coil	MPRAGE Voxel 1x1x1 FOV 256x256x176 TI=450		2D FLAIR Voxel .9x.9x3 FOV 256x256x50 TE=95 TR=9000 TI=2250		Axial 2D EPI Voxel 1.875x1.875x3.6 FOV 128x128x56 TE=min TR=8000	3D Spiral ASL Voxel XxXx5 FOV 512x12x72 NEX=3 time:9:45	Axial 2D gradient echo EPI Voxel 3x3x3 FOV 64x64x43 TE=35 TR=2500 NEX=1	
AIBL (Perth)	3T	MAGNETOM Vida	T1 MPRAGE FOV: 256x256 Voxel 1x1x1 TR = 2300 TE = 2.98 TI = 900 FA = 9		T2W FLAIR FOV: 256x256 Voxel 1x1x1 TR = 5000 TE = 391 TI = 1800 FA = 120	T2* ME-GRE FOV: 224x180 Voxel 1x1x1 TR = 30 TE = NA TI = NA FA = 15				
AIBL (Melbourne) Austin Hospital	3T	MAGNETOM Skyra	T1 MPRAGE FOV: 256x256 Voxel 1x1x1 TR = 2300 TE = 2.98 TI = 900 FA = 9		T2W FLAIR FOV: 256x256 Voxel 1x1x1 TR = 5000 TE = 391 TI = 1800 FA = 120	T2* ME-GRE FOV: 224x180 Voxel 1x1x1 TR = 40 TE = NA TI = NA FA = 15				
AIBL (Melbourne) Austin Hospital	3T	MAGNETOM Vida fit	T1 MPRAGE FOV: 256x256 Voxel 1x1x1 TR = 2300 TE = 2.98 TI = 900 FA = 9		T2W FLAIR FOV: 256x256 Voxel 1x1x1 TR = 5000 TE = 391 TI = 1800 FA = 120	T2* ME-GRE FOV: 224x180 Voxel 1x1x1 TR = 40 TE = NA TI = NA FA = 15				
AIBL (Melbourne) RMH	3T	MAGNETOM Prisma fit	T1 MPRAGE FOV: 320x320 Voxel .8x.8x.8 TR = 2400 TE = 2.31 TI = 1000 FA = 8		T2W FLAIR FOV: 320x320 Voxel .8x.8x.8 TR = 6700 TE = 463 TI = 2200 FA = 120	T2* ME-GRE FOV: 224x182 Voxel 1x1x1 TR = 50 TE = NA TI = NA FA = 15				
AIBL (Melbourne) RMH/BRI	3T	MAGNETOM Prisma fit	T1 MPRAGE FOV: 256x240 Voxel 1x1x1 TR = 2300 TE = 2.91 TI = 900		T2W FLAIR FOV: 256x240 Voxel 1x1x1 TR = 5000 TE = 388 TI = 1800	T2* ME-GRE FOV: 224x182 Voxel 1x1x1 TR = 30 TE = NA TI = NA				

Table 7. Harmonization table for neuropsychiatric testing. The first 5 tests have been matched across all sites. Each site has created a custom battery of tests presented below. (Y: Yes, N: No)

	BU Clinical Core (N=270)	BU Registry (N=180)	Emory	BBHI	BBRC (Alfa+)	BBRC (Beta)	Ace	Bogalusa	AIBL	AIBL (ADNeT co-enrolled)
Digit Span Forward and Backward	Y	Y	Y	Y	Y	TBD	Y	Y	Y	Y
Category Fluency	Y - Animals and Vegetables	Y - Animals and Vegetables	Y - Animals and Vegetables	Y - Animals, Vegetables,	Y - Animals and Vegetables	Y - Animals and Vegetables	Y - Animals	Y - Animals and Vegetables	Y - Animals, Fruits, and	Y - Animals, Fruits,
Trail Making Test A & B	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Verbal Fluency	Y - F and L	Y - F and L	Y - F and L	Y - P, F, and R	Y - P	Y - P	Y - P	Y - F, A, and S	Y - F, A, and S	Y - F, A, and S
Symbol Digit Modalities	Y	Y	Y	Y	Y - WAIS-IV	Y - WAIS-IV	Y	Y - via iPad	Y - WAIS-III	Y - WAIS-IV
MoCA	Y	Y	Y	Y	N	N	N	N	TBD	TBD
MMSE	N	TBD	N	Y	Y	Y	Y	Y	Y	Y
CDR	Y	Y	Y	N	Y	Y	Y	N	Y	Y
CDR Sum of Boxes	Y	Y	Y	N	Y	Y	N	N	Y	Y
Test of Memory Malingering	Y	Y	Y	N	N	N	N	N	N	N
Test of Memory Malingering (Retention)	Y	Y	Y	N	N	N	N	N	N	N
Judgment of Line Orientation	Y	Y	Y	N	N	N	N	N	N	N
Stroop Test	N	N	N	N	N	N	N	N	Y - Victoria version	Y - Victoria version
Controlled Oral Word Association Test	N	N	N	N	N	N	N	N	Y	Y
WAIS-IV Matrices	N	N	N	Y - WAIS-IV	Y - WAIS-IV	Y - WAIS-IV	N	N	N	N
Wide Range Achivement Tests 4th Edition	Y	Y	Y	N	N	N	N	Y	N	N
NAB LL Immediate and Delayed Recall	Y		Y	N	N	N	N	N	N	N
WMS-III Verbal Learning	N		N	N	N	N	Y - WMS-III	N	N	N
Multilingual Naming Test	Y		Y	N	N	N	N	N	N	N
Face Name Associate Memory Exam	N		N	Y	N	N	N	N	N	N
Boston Naming Test	N		N	N	Y - 30 item	N	Y - 15 item	Y - 60 item	Y - 30 item	Y - 15 item
WAIS-IV Cube Construction	N		N	Y - WAIS-IV	N	N	N	N	N	N
WAIS-III Block Design	N		N	N	N	N	Y - WAIS-III	N	N	N
RAVLT	N		N	Y	N	N	N	N	N	Y
CAVLT	N		N	N	N	N	N	N	Y	N
PVLRT	N		N	N	N	N	N	Y	N	N
Craft Story 21	Y		Y	N	N	N	N	N	N	N
WMS-III Logical Memory	N		N	N	N	N	N	N	Y - WMS-III	Y - WMS-III
WMS-IV Logical Memory	N		N	N	Y -WMS-IV	Y -WMS-IV	N	N	N	N
WMS-II Logical Memory	N		N	N	N	N	N	Y - WMS-II	N	N
Benson Complex Figure Copy	Y		Y	N	N	N	N	N	N	N
Rey Complex Figure Test	N		N	N	N	N	N	N	Y	Y
Craft Story 21 Delay	Y		Y	N	N	N	N	N	N	N
WMS-III Logical Memory Delayed	N		N	N	N	N	N	N	Y - WMS-III	Y - WMS-III
WMS-IV Logical Memory Delayed	N		N	N	Y -WMS-IV	Y -WMS-IV	N	N	N	N
WMS-II Logical Memory Delayed	N		N	N	N	N	N	Y - WMS-II	N	N
Benson Complex Figure Copy Delayed and	Y		Y	N	N	N	N	N	N	N
Rey Complex Figure Test Delayed and Recognition	N		N	N	N	N	N	N	Y	Y
NAB Mazes	Y		N	N	N	N	N	N	N	N
Finger Tapping	N		N	Y	N	N	N	N	N	N
Tap 30	N		N	Y	N	N	N	N	N	N
Face Name Associate Memory Exam - Delayed	N		N	Y	N	N	N	N	N	N
Letters and Numbers	N		N	Y	N	N	N	N	N	N
WAIS-IV Cancellation	N		N	Y	N	N	N	N	N	N
Corsi Cubes	N		N	Y	N	N	N	N	N	N
Poppelreuter's Test (responses)	N		N	N	N	N	Y	N	N	N
Global Orientation	N		N	N	N	N	Y	N	N	N
Imitation praxis	N		N	N	N	N	Y	N	N	N
Gobal praxis	N		N	N	N	N	Y	N	N	N
15-Objects test (responses)	N		N	N	N	N	Y	N	N	N
Luria's Clock test	N		N	N	N	N	Y	N	N	N
Automatic Inhibition SKT (seconds)	N		N	N	N	N	Y	N	N	N
Automatic Inhibition SKT (error)	N		N	N	N	N	Y	N	N	N
WAIS-III Similarities	N		N	N	N	N	Y	N	N	N
Weschler Test of Adult Reading	N		N	N	N	N	N	N	Y	N
Test of Premorbid Functioning	N		N	N	N	N	N	N	N	Y

Clinical Data Transfer Operations

Clinical data management will be handled by MMS Holdings. Data will be assembled by MMS from each site and curated at the AD Data Initiative Curation Studio workspace. To do this, MMS, with the help of ADDF, will establish primary and secondary contact points with the data management officers at each site. MMS will obtain keys and transfer

protocols for data transfer and will also communicate with each site and obtain their employed data formats. Based on this information, MMS will determine the data format of the common database to which all Site databases will merge. MMS will then create sockets to transfer and store the data from each site onto the AD Data Initiative Curation Studio.

4.9.1 **Clinical Data Harmonization**

Uniformization is performed by merging or conforming incoming datastream dictionaries into a unified SpeechDx dataset. This process is be defined by a working group comprised of each site's data personnel, ADDF, MMS and ADDI. In the case where the site utilizes REDCap, the harmonization will be significantly facilitated since it works as an intermediate harmonization step. REDCap is a secure web application for building and managing online surveys and databases geared to support online and offline data capture for research studies and operations. Where needed, a script will perform the task of assigning the proper dictionary values to each entry that map each site's dictionary to a harmonized SpeechDx dictionary.

MMS, in coordination with the sites, will establish ways of recovering any missing data and ensure that the information stack is regularly up to date. Additionally, MMS will follow procedures that, at all instances, prevent identifying information being transferred or otherwise disclosed. In the case where PII/personal data may have been inadvertently disclosed by a site, all parties will take all measures to destroy such information, repair the channel that led to such disclosures, and promptly notify the site and ADDF of the incident.

Next, MMS will be responsible for matching SUID and SCUID for every data entry and will create an intuitive folder/database structure that can be queried longitudinally and cross sectionally to yield participant clinical data and audio clips. Informational material on database structure, including query examples, will be provided. MMS will work closely with AD Data Initiative to design, test, and maintain the database structure while the study is running.

To create a harmonized dataset that can be used in machine learning model training applications, ADDF will initially create a working group where staff site members and MMS will decide on dictionary rules that allow for proper channeling of the clinical data to their respective fields in a unified database structure. This group will meet as needed and will conclude when a protocol to harmonize information from all Sites has successfully and repeatably been agreed. *Data Harmonization* is the process of combining data of varying file formats, naming conventions, data fields, and transforming it into one cohesive dataset. It includes data centralization operations (bringing multiple site data into one database), data cleansing (removing duplicates or "orphan" entities), identifier matching, data normalization (e.g., establish one common measurement unit and coding), and, most importantly, data field matching.

Data entries from US-based Alzheimer's Disease Research Center sites are expected to comply with UDS standards. UDS is a NACC program to standardize evaluations and assessments conducted at the NIA-funded Alzheimer's Disease Research Centers (ADRCs). This will facilitate integration of the data from all US sites. US-based sites outside of the ADRC infrastructure and European sites do not necessarily comply with UDS, yet the latest consensus initiatives suggest a close match with this standardization [5]. Additionally, most sites perform additional assessments outside of the UDS standards; hence, custom harmonization rules will have to be created for those.

Once harmonization rules have been created and validated, MMS will create sockets that link each site's fields to the Speech Consortium database based on said rules. Automation scripts will be used as much as possible. Harmonization

of a quarterly batch will be completed before the end of the next quarter in order to temporally coincide with the PII/personal data splicing operation.

4.9.2 De-identification of Imaging Data

Imaging data that will be transferred from the sites to ADDI Curation Studio may contain MRI files. All identifying information will be scrapped from each MRI dataset by the Sites, prior to sharing with ADDI. MMS will check to see if inadvertent exposure of identity has occurred from the imaging dataset.

It has been shown that a person can be identified by means of 3D reconstruction of MRI slices. To make sure that no participant identification is possible, MRI data will be pseudonymized. This can be performed at the Sites using available software suites such as FreeSurfer, using the mri_deface module that removes face characteristics so that no face reconstruction is possible while maintaining the brain structures visualization intact.

4.9.3 Study Identifiers

For each new participant or for existing participants, each site has a **Site Unique Identifier (SUID)** assigned by the site**.** MMS will communicate with the sites to obtain such SUIDs at the beginning of the study (to capture existing SUIDs) and at regular timepoints (to capture new SUIDs). For each new or existing participant enrolled, MMS will generate a **Speech Consortium Study Unique Identifier (SCUID)** according to standard guidelines for generation of secure unique study identifiers. Once a participant is onboarded using the tablet app at a site, the SUID will be used to link the participant to that particular tablet. MMS will be responsible for maintaining the lookup information that allows SUIDs to be linked to SCUID. This information will be securely protected and not shared with other entities using standard data protection methods. Other study metrics such as annual visit data collection completed or pending can be provided.

4.9.4 Equipment Dissemination Management

The SpeechDx team will communicate with the sites to obtain the projected enrollment estimates on a quarterly basis and will communicate these needs to the hardware provider, Mobility CG (MCG), so that the estimated number of devices is shipped to each site on a regular basis.

4.10 Data Flows

SpeechDx joins two data streams: a speech and language data stream and a clinical characterization data stream. The two data streams will initially be independent and asynchronous, and matching of information using common identifiers will be done by MMS.

For the first 5 years of the program, data will live in the AD Data Initiative Curation Studio, and access will be limited to SpeechDx partners and participating sites. Following that period, data will be moved to the AD Data Initiative AD Workbench, where the Data Access Committee will review and grant access to applicants.

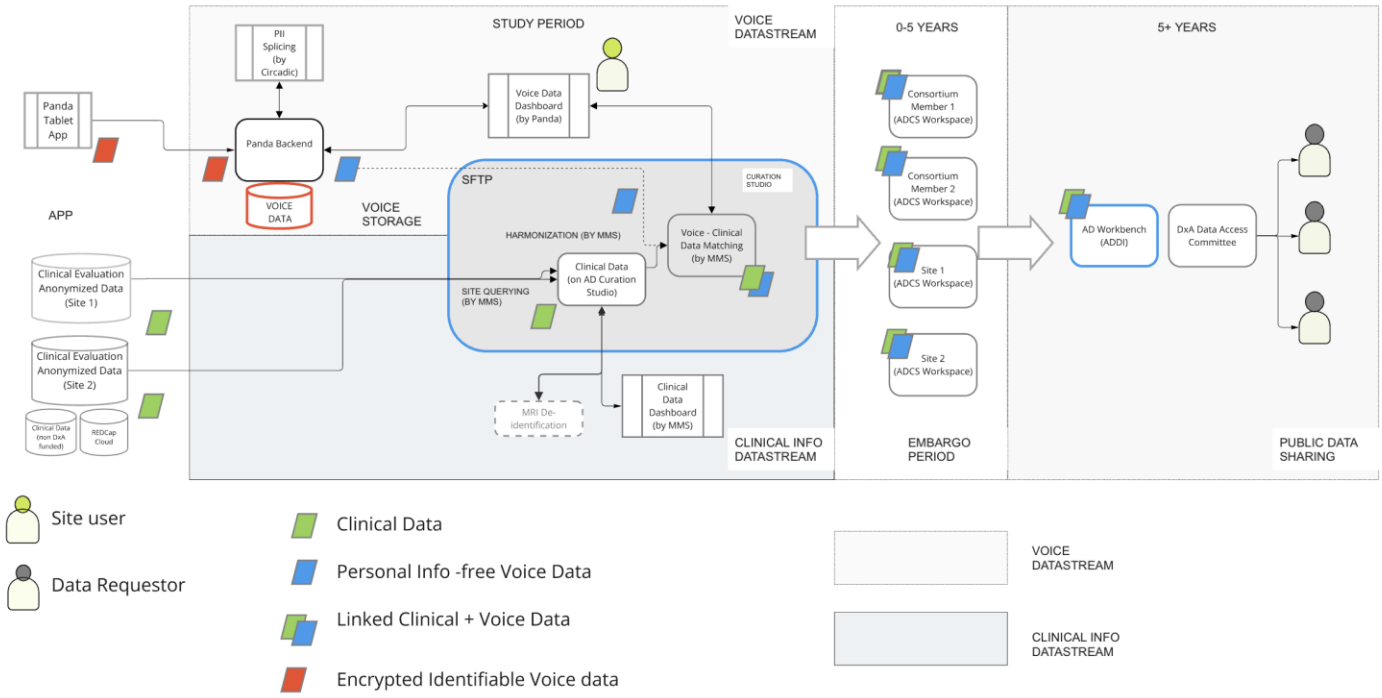


Figure 10 SpeechDx data flows

5.0 Study Population

5.1 Inclusion and Exclusion Criteria

Participants are eligible to be included in the trial only if they meet all the following criteria and they do not meet any of the exclusion criteria at screening. Individuals who do not meet the criteria for participation in this trial (screen failure) cannot be re-screened. Prospective approval of deviations to recruitment and enrollment criteria, also known as waivers or exemptions, are not permitted. Please refer to Table 8 for a list of inclusion and exclusion criteria.

Table 8. Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Signed informed consent form (ICF) Age ≥ 45 Access to at least 3G connectivity (or better) at home is required, though alternative options are available Native proficiency in the applicable language (currently English, Spanish, Catalan) or other supported languages	History of severe TBI, severe depression, schizophrenia, stroke, and/or brain tumors Significant and uncorrected vision impairment Significant and uncorrected hearing impairment Any neurological or psychiatric disorder that prohibits participants from taking the required tests

5.2 Recruitment Methods

The entire participant population will be recruited from partner sites' parent study patient pools. In the recruitment process, age, sex, and ethnicity may be assessed to ensure diversity in the data sample when applicable.

6.0 Study Procedures

A summary of app events is visualized in Figure 11.

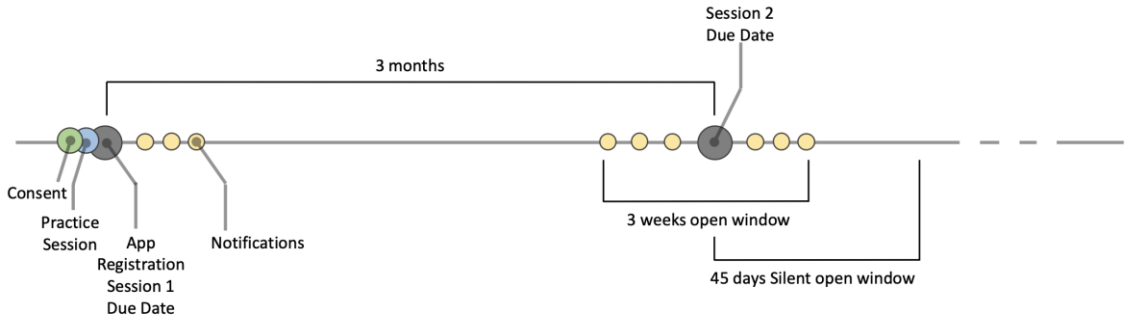


Figure 11. Schematic of app events with the Consent, app Registration, Practice Session, Sessions Dates and Notifications

Staff Training (App)

Study staff will be trained to provide remote support to participants during the duration of the study. The following procedure will be carried out to train and certify study staff:

- Complete an overview of the study & typical data collection procedures for remote assessments with the site PI
- Complete a short training session with the ADDF team or vendors regarding the study app. During training, the sites will learn how to use the administrative portal and apps and will be provided with a detailed site instruction manual or site training.
- Obtain application credentials and log into the study app and change their password
- Complete a practice assessment on the study app

Participant Training (App)

Participants will complete a brief training on app usage. Said training can be done at the clinic or remotely via teleconference as follows:

- Describe the basic functions of the tablet, how to turn it on and off, adjust volume, brightness
- Locate the SpeechDx study app icon and help them familiarize with the Android interface
- Log into the study app using credentials provided by the study team. This includes a 5-digit code that is automatically generated by the study dashboard. This one-time use only code can be shared with the participant using a text message or email or spelled out on the phone
- Select the practice session and run through it, with the rater explaining each step.

Once this has been completed, the participant will be ready to complete their first assessment remotely.

App Procedures and Flow

The participant receives mobile notifications prior to the due date of the test. They are advised to acknowledge them and mark the dates. On the day of the test, once the participant logs in, the recording app will experience the following steps:

- First, a picture will be displayed on the screen. The pool of pictures to be displayed will be updated when the app updates or connects to the internet (i.e. it will not be a live display from the pool, but rather a periodically updated file set). A visual and oral prompt will ask the participant to describe the picture. When the participant is done describing the picture, he/she can click "next."
- The PHQ-8 questionnaire is then presented in a multiple-choice format, where each question is presented on a screen. Once a response is provided, the "Next" button takes the participant to the next question. The user can also hit the back button to review/correct an answer.
- In the open-ended questions test, a topic for discussion will be presented both visually and verbally. The pool of topics to be presented will be updated when the app updates or connects to the internet (i.e. it will not be a live presentation from the pool, but rather a periodically updated file set). Once the topic is presented, recording is set to ON and will record as long as the background noise level is sufficiently low. When the participant is done discussing the topic, he/she can click "next."
- The participant is then instructed to recall the picture that was shown in the beginning, recording starts as soon as they hit "Next" button.
- Next, the app will then play a recording of a story that was previously recorded. The pool of stories and storytelling tasks with instructions will be presented to the participant. When the participant is done telling the story, he/she can click "next.". This process will be repeated twice for two different stories.
- A vigilance (game) test will be displayed. The user is asked to tap anywhere on the screen when a red circle appears on the screen as fast as she can, but not when a white "+" sign appears. The correct answers as well as the response time will be collected.
- The app will then show a picture with an audio and written prompt for the participant to use their imagination to tell a story about one or more characters in the picture.
- Finally, the app will present the participant with another picture (always the same in all sessions) and instruct them to describe it in their own words. This is the last test.
- Lastly, the participant is advised to wait for a few minutes before closing the app or turning off the tablet, to allow the app to connect to the cloud and upload the data.

6.1 Participant Visits

Participants will come into the clinic for in-person assessments once annually or every 18 months depending on the Site visit schedules, and all other assessments will be conducted remotely.

Baseline Visit: This visit will serve as a combined screening and baseline visit. During this visit, participants will be screened against the inclusion and exclusion criteria and, should they meet these criteria, proceed to a consent interview. During the first (baseline) visit, study staff will explain the study procedures to the participant, obtain a signed ICF, provide the device, and assist the subject with initial set-up. Informed consent will be acquired from all participants. All participants recruited from existing studies at each site will have completed ICFs. ICFs or supplemental/amended ICFs detailing the specifics of this study, including the key components listed above, will be acquired during the initial visit with each subject. Note that participants can choose to withdraw consent or restrict how their samples are used.

Participants will sign their own consent. In cases when the participants functioning is too impaired to sign independently, consent will be obtained from a Legally Authorized Representative (LAR). The LAR must sign a separate form indicating that they have appropriate legal rights to act as the LAR for the participants. All consent procedures and

content will be required to abide by applicable laws and regulations. Once consent is obtained, study staff will onboard the participant and speech data will be collected.

Quarterly Assessments: Once every 3 months, the study application will send the subject push notification(s) via the Android device to prompt the subject to complete their assessment. Email reminders will also be used to facilitate remote compliance, and participants will be given a 29-day window to complete the battery of tasks. The recruiting centers will have insight into participant compliance adherence through a trial schematic available in the Portal. The schematic will flag assessments that are overdue or missed. In addition, participants will receive notifications on their device to alert them of upcoming scheduled assessments to help facilitate remote compliance.

Annual Visits: Subjects will come into the clinic for an in-person assessment annually or every 18 months, in accordance with the protocol for each Site's parent study. In this visit, study staff will collect any needed health information, biospecimens, and/or neuropsychiatric data as defined by the parent study protocol at each site.

Table 9. Schedule of Activities

ASSESSMENT	Baseline Visit (0)	Months 0, 3,6,9	Annual Visit	Months 12, 15,18,21	Annual Visit	Months 24, 27,30,33,36	Last Visit
Informed Consent	X						
Application Training	X						
Blood Draw, Storage and Analysis (ptau217)	X		X		X		X
MRI Imaging (T1 and T2 FLAIR)	X						X
Neuropsychological Testing *	X		X		X		X
SpeechDx Battery		XXXX		XXXX		XXXX	

*At least Digit Span Forward and Backward, Category Fluency Animals/Vegetables, Trail Making A&B, Verbal Fluency, Symbol Digit Substitution, MoCA, MMSE

6.2 Withdrawal or Discontinuation of Participant

Participants will be discontinued under the following circumstances:

- Enrollment in any other clinical study judged not to be scientifically or medically compatible with this study
- Participation in the study needs to be stopped for medical, safety, regulatory, or other reasons consistent with applicable laws and regulations

- Participant decision
- PI decision
- Withdrawal from the parent study

Any data collected up to the point of participant withdrawal or discontinuation may be used in the study, unless the participant requests their data to be eliminated according to GDPR policies.

Any participant who does not contribute at least two consecutive recording sessions, without any medical reason, will be deemed a dropout and discontinued from the study. In the case where a participant is considered a drop out, the tablet device will be reclaimed and repurposed for another participant. Staff will contact participant/care giver, conduct an exit survey, cease e-mail notifications, obtain the device, reset it to factory settings, and follow procedures for donation or recycling or participant retention of the device.

Throughout the duration of the study, participants will be allowed to use the Android study device provided for their personal use. At the end of the study, study staff will perform a remote factory reset on the device. Participants have the option to retain or return the device. In the case where participants opt to return the device, returned devices will be collected and either recycled or donated through philanthropic channels.

6.3 Discontinuation of the Study

The study will be discontinued at the applicable individual Site if the PI judges it necessary for medical, safety, regulatory, or other reasons consistent with applicable laws, regulations, and good clinical practice (GCP). In case the parent study is discontinued for any reason in one or more Sites, the speech data collection will also be discontinued for that/those sites.

7.0 Data Ingestion and Dissemination Plan

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. The AD Initiative AD Workspace addresses aspects of a) uniformization/harmonization of the clinical data, and b) access (open, limited, nested) and enabling virtual processing of datasets within the repository to maintain patient privacy.

During the first 5 years of the study, voice data and transcripts (from the tablets) as well as clinical data (from the Sites) will be accessible to SpeechDx Partners as well as the participating Sites (5-year exclusivity period). Terms and conditions of becoming a SpeechDx Partner will be negotiated with each requesting party. During this period, data will be disseminated in batches every six months. ADDF will follow a venture philanthropy model, investing directly in companies and/or seeking royalty agreements in order to reinvest any returns in future Alzheimer's research.

After the exclusivity period, the dataset will be made accessible to the broader public under the control of a Data Access Committee.

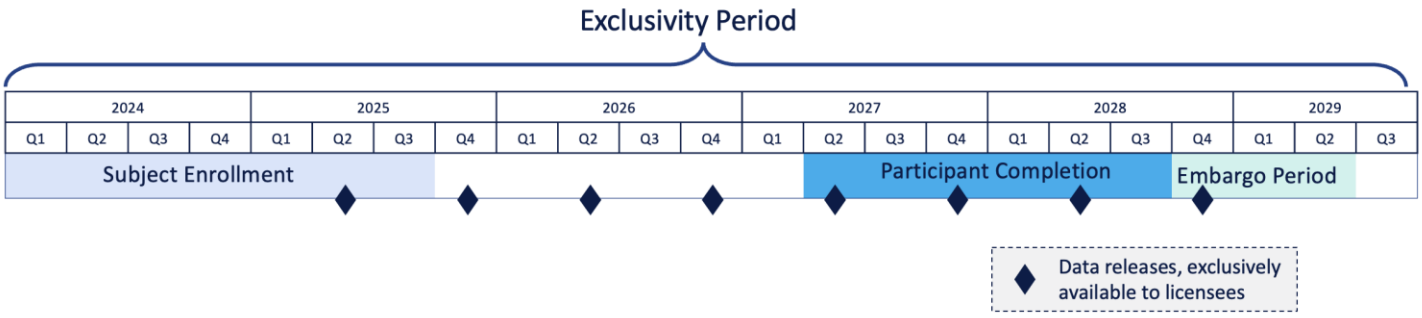


Figure 12. Timeline of data releases during the exclusivity period

A Data Access Committee (DAC) will be created to grant access to the SpeechDx dataset after the completion of the data acquisition (plus optionally an embargo period that will be determined by ADDF). For clarity, **the Data Access Committee will govern sharing of the data past the 5-year exclusivity period; the DAC will not govern sharing of the data during the 5-year exclusivity period.**

The DAC will be comprised of representatives from the ADDF and representatives from Gates Ventures, as well as representatives from Sites and will be authorized (or a delegate of either party's choosing) to grant access to requestors of the SpeechDx dataset based on the evaluation of their proposed use of data. The DAC will be tasked with evaluating the proper and ethical use of the dataset following the exclusivity period of the project. For the Committee's evaluation, a written submission of the applicant's proposed use of the dataset will be required. A unanimous positive vote from all parties will be required for access to be granted, and any negative votes will require a clear rationale and may be challenged by the other party. The DAC is expected to make decisions according to principles of open data access, and

rejections can be made in instances where data security and/or ethical data use are under question or in cases where the intended use lies outside of the scope of this program (e.g., not involving speech analysis). If a Site participating in the DAC does not agree with a particular data access request for reasonable cause, the Site has the option to request their data not to be shared with that particular requestor.

Researchers wishing to access the data will:

- Undergo identity verification by submitting an affidavit signed by an authorized institutional official or public notary or by providing proof of a professional license.
- Submit an intended use statement, which will be made public.
- Submit proof of review and approval from an accredited ethics board or IRB for a secondary analysis plan.
- Agree to ethical data use and study-specific data terms of use.

Database access will be provided through AD Data Initiative-provided Virtual Machines. The dataset will not be accessible from the internet by any other means.

As approaches to maintaining privacy evolve, these rules may be updated as needed to align with current best practices. The process of releasing the dataset to a requesting entity may be modified accordingly. All data sharing activities follow the permissions laid out in the consent form.

8.0 Use of Collected Data

The current study is limited in scope to data collection only, and data analysis is not included. We hypothesize that extensive interrogation of this dataset by the biomedical research community will yield important new findings to advance speech- and language-based biomarkers for Alzheimer's disease and related dementias. Therefore, the intent of this study is to make the well-characterized, de-identified, harmonized dataset collected herein accessible to the research community, including industry, allowing researchers to bring their own bioinformatics/data science approaches to the dataset and fuel future research and commercial opportunities. The dataset may be made available for and on both commercial and noncommercial bases. In all cases, use of the dataset must comply with national and local regulations, including the AI Act for EU member countries.

9.0 Data Security and Confidentiality

9.1.1 **Data Security**

The AD Workbench (AD Data Initiative) data repository is protected by data privacy and security controls standard in biomedical and healthcare information technology, including the provisions required to be GDPR compliant, including the implementation of ISO 27701 policies and processes. Stored data are encrypted in transit. Transmissions through AD Workbench interfaces are encrypted with HTTPS or TLS 1.2 protocols or above. By default, Microsoft Azure encrypts data at rest using FIPS 140-2 compliant 256 AES encryption for storage accounts and virtual machine disks. All user access to the AD Workbench is via HTTPS URL protected by a rooted certificate issues by DigCert SHA2 Secure Server CA, utilising sha256RSA signature algorithm with sha256 signature hashing algorithm and requires Two-factor authentication. All uploads go through a malware scanning process. Data access will be approved by the SpeechDx Data Access Committee (DAC) and facilitated by the AD Data Initiative.

Data will only be identified by the person's unique study (SCUID)

Recording(s) and associated data will be kept for at least 10 years but are subject to deletion or restriction of use at participants' requests. Participants can request to no longer share their data with new requesting entities at any time by notifying study personnel. This does not hold for data that has already been shared up to that point in time. Foregoing data may continue to be processed and shared provided that there are other legal bases for doing so. Participating Sites will have to inform their subjects on legal bases as applicable to their national and local regulations.

Recording(s) will be used for the purposes of advancing speech- and language-based biomarkers, including educational or commercial purposes, and/or analysis by the research team, or for future unspecified use.Compensation will be clearly stated – including the option to keep the study-provided device following completion of the study.

Risks associated with speech and voice assessments are minimal and there is a small chance testing will produce some fatigue. Investigators' experiences with these tasks, however, suggest that fatigue is unlikely and has little impact on the participant. There is also a risk of loss of privacy, and measures are being taken to pseudonymize and secure data as stated above.

Data Security Plan: Speech recordings and cognitive data will be collected for distinct tasks via a single custom app on Android tablets, which will serve as the study data collection vehicle for both speech data collection and selected cognitive assessments. Data are encrypted and stored on the tablet only until it has been securely transferred to cloud servers, when it is erased. This usually takes several few seconds, unless there is an Internet connectivity issue. As each section of every test is completed, it is encrypted with RSA public key encryption on top of AES256 encryption and stored. Once the data are encrypted, they cannot be decrypted without the private key; even if the device is stolen and breached, the data on the device cannot be decrypted. If the device cannot connect due to Internet connectivity issues, the encrypted data will stay on the device, and the app will continually retry until it successfully transfers the data.

The app collected data are encrypted in transit and at rest on web servers that satisfy applicable privacy and security law requirements. All data are hosted on Azure EU-based servers behind secure API gateways and are access-controlled,

so that only trained and approved employees can access them. Metadata only includes non-identifying information, such as time & date for the session and information about the participant's device.

Production encryption keys are accessible only on the servers that require those keys in order to function, and access to those servers is protected by multiple layers of security (VPN access with multi-factor authentication and SSH keys). Access to those servers and encryption keys is held only by 2 senior employees of the AMS team (Zuhlke Engineering), and those keys are kept securely in encrypted storage. Key transfer is only performed when necessary, over encrypted secure connections.

As new data protection technologies emerge, the above-mentioned data protection methods may be adjusted accordingly to provide the maximum level of security.

The recordings can sometimes inadvertently capture additional PII/personal data that could compromise participant identity if processed in a certain manner. To avoid this and to maintain a database that can be shared with study partners as well as other researchers, all PII/personal descriptors will spliced out using a manual process as follows: First, all data will be transferred to cloud servers maintained and operated by the Data Hosting Vendor. Then, all voice recordings will be transcribed and manually examined by trained ISO certified language professionals. If a segment of the recording is considered PII/personal descriptors-exposing, that segment will be removed and replaced by a tone of a certain frequency that will signify that PII/personal descriptors has been removed and redacted from the transcription. This process will be done via a custom-built audio processing and transcription software. Once any personal descriptors have been removed, they will be transferred to ADDI Curation Studio, supported by Aridhia DRE, in quarterly batches.

Data will be centralized on a data-sharing platform or repository managed by ADDI. The AD Workbench is built on the Aridhia Digital Research Environment (DRE), a set of web services and tools that make it easier for researchers to aggregate, organize, analyze, and share scientific data, code, and insights. The AD Workbench is designed according to the Findable, Accessible, Interoperable, Reusable (FAIR) principles and provides data governance tools to ensure data is protected and accessible to scientists while respecting patients' rights. AD Workbench technology is REST-based, modular, and built from industry-standard components, intended to leverage public platforms and tools, including secure cloud storage and compute, data repositories, data visualization tools, and analytical tools such as RStudio and Jupyter Notebooks. The ADDI technical team supports integrations with external systems.

The AD Workbench is protected by data privacy and security controls standard in biomedical and healthcare information technology, including the provisions required to be GDPR compliant. A validated HIPAA/HITRUST assessment was completed in June 2021 followed by HITRUST approval and certification. Stored data are encrypted with FIPS 140-2 compliant 256 AES encryption and is protected by data privacy and security controls standard in biomedical and healthcare information technology, including the provision required to be GDPR compliant through implementation of ISO27701 policies and HITRUST CSF. Transmissions through AD Workbench interfaces are encrypted with HTTPS or TLS 1.2 or above protocols. The AD Workbench operates under comprehensive governance policies designed to enable cross-disciplinary research while safeguarding data. These policies include well-documented Terms and Conditions of Use, a standardized data use procedure, and adaptive privacy policies. Data access will be governed according to criteria stated in Section 8.0.

All identifiable clinical information about the participants, their medical conditions, and other study data will not be transmitted to the cloud. Such data will be secured by the PIs and site staff in accordance with all local, state, national,

and supranational laws, regulations, and Ethics Board policies regarding collection and distribution of patient information. Participants' identifying information will not be disseminated in any case.

10.0 Risks and Benefits

10.1 Potential Risks Associated with the Study

This study presents minimal risk (does not involve any significant risk beyond what a person would experience in their daily activities). It is a prospective exploratory study; no drug intervention will be provided.

There is a potential risk of loss of privacy associated with this study. To ensure confidentiality of study subjects, study sites will assign a random unique identifier to each individual and remove direct identifiers from the data. The risk of identifiability from voice recordings remains. To mitigate this risk, data will be password-protected and encrypted, as described above. In addition, to prevent researchers from listening to the recordings and to limit exposure to their scripts running on the ADDI server, no option will be provided to install a list of audio reproduction programs (e.g. VLC, Audacity, etc.).

10.2 Adverse Events

Adverse events (AE) will not be actively solicited from study participants since no medical products are being provided or evaluated in this study. Although this is a minimal risk study, AEs and serious adverse events (SAEs) may still occur.

If a participant spontaneously reports an AE to the study staff or study researchers, it will be evaluated and will be reported to the sponsor and necessary parties/authorities, as appropriate. Adverse events or device-related complaints will be recorded and submitted, as appropriate.

10.3 Potential Benefits Associated with the Study

Participants of this study will not receive any direct benefit from the proposed research. However, evidence and algorithms generated by the data collected in this study will be used to discover digital biomarkers in speech that can help detect dementia earlier, and once diagnosed, monitor its progression.

11.0 Intellectual property (IP)

The final unified, harmonized and de-identified speech, clinical, and related data resulting from the participation of multiple Sites and up to 3,000 study participants across Sites as set out above (the Dataset), and the collected speech recordings will be solely owned by the Alzheimer's Drug Discovery Foundation. The design and presentation of the test battery organized and maintained by or for ADDF will be owned by ADDF and/or its licensor(s). The Dataset will be made available to applicable researchers pursuant to the terms of any applicable grant, consortium membership, or other agreements. It is anticipated that new IP will be generated by both academic and for-profit researchers through their analyses of the Dataset. Subject to the terms of any applicable grant, consortium membership, or other agreements, such new IP generated will be owned by the inventor(s) or their employers. Subject to applicable law, relevant informed consents, and any rights of the applicable data subjects, Sites retain ownership of clinical (i.e., medical) data of participants from their Study Center.

Study Governance Considerations

Clinical data collection protocols are already in place at each Site for the current studies/cohorts with local IRBs. Each Site, as the sponsor of the individual collection efforts at their location(s), will determine if they will need to create an amendment to an existing protocol or create a new protocol.

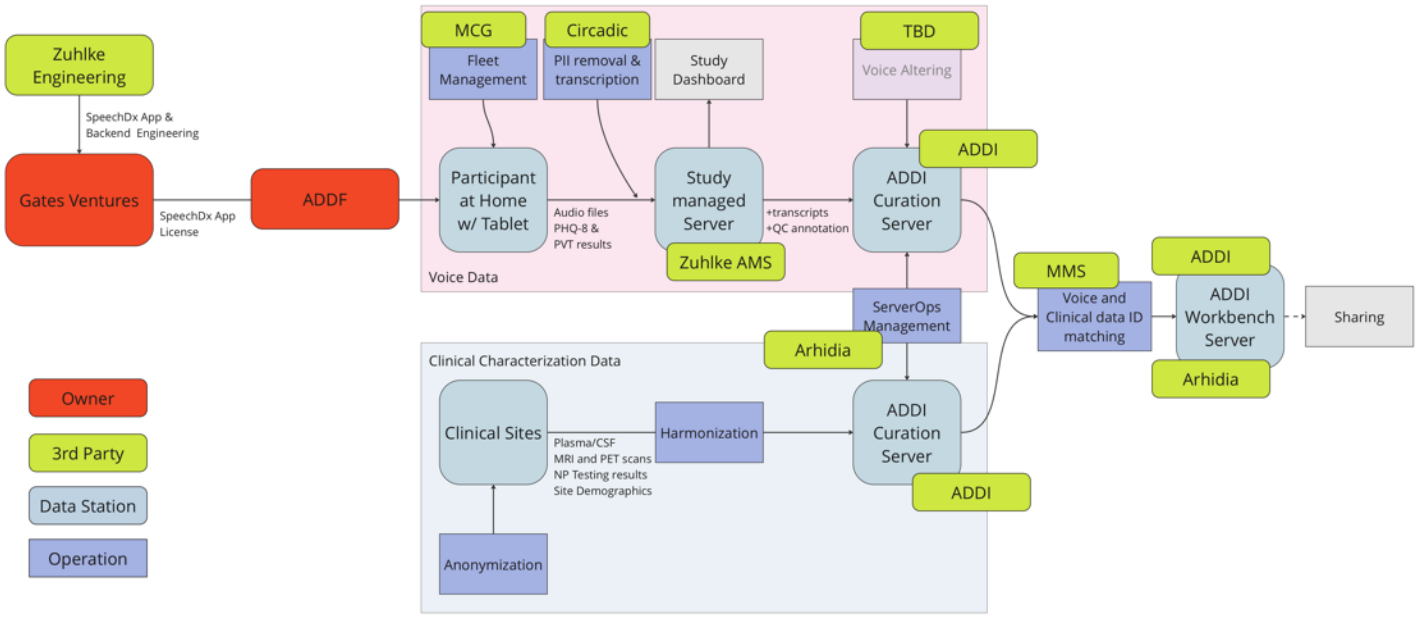


Figure 13. A summary of the entities involved and their function in the data collection, management and storage for the SpeechDx program

12.0 References

1. Kourtis, L.C., Regele, O.B., Wright, J.M. *et al.* Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *npj Digital Med* 2019, 2:9.
2. Veronica Boschi , Eleonora Catricalà, Monica Consonni Boschi *et al*. Connected Speech in Neurodegenerative Language Disorders: A Review. *Front Psychol*. 2017, 8:269.
3. Mueller, K.D., Hermann, B., Mecollari, J, *et al*. Connected Speech and Language in Mild Cognitive Impairment and Alzheimer's Disease: A review of picture description tasks. *J Clin Exper Neuropsychol* 2018,40, 9: 917–939.
4. Bjorklund, N.L., Fillit, H., Malzbender K, *et al.* The Need for a Harmonized Speech Dataset for Alzheimer's Disease Biomarker Development *Explor Med* 2020, 1:359-363.
5. Boccardi M. et al. Harmonizing neuropsychological assessment for mild neurocognitive disorders in Europe. *Alzheimers Dement* 2022 18(1):29-42
6. Amini, S. *et al*. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. *Alzheimers Dement.* **20**, 15525279 (2022).
7. García-Gutiérrez, F. *et al*. Assessing the progression of Alzheimer's disease through spontaneous speech analysis: a machine-learning approach based on acoustic features. *Front. Neurosci.* **17**, 1221401 (2023).
8. Eyigoz, E. *et al*. Linguistic markers predict onset of alzheimer's disease. *EClinicalMed.* **28**, 100583 (2020)
9. König, A. *et al*. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **1**, 112–124 (2015).
10. Hajjar, I. *et al*. Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early Alzheimer's disease. *Alzheimers Dement.* **15**, e12393 (2022).
11. Ezzati, A. *et al*. Predicting MCI to AD conversion using baseline language performance and brain structure volumes. *J. Alzheimers Dis.* **73**, 501–513 (2020).
12. Chou, C.-J. *et al*. Screening for early Alzheimer's disease: enhancing diagnosis with linguistic features and biomarkers. *Front. Aging Neurosci.* **16**, 1451326 (2024).
13. Robin, J. *et al*. Automated detection of progressive speech changes in early Alzheimer's disease. *Alzheimers Dement.* **15**, e12445 (2023).
14. Roberts RO et al. Prevalence and outcomes of amyloid positivity among persons without dementia in a longitudinal, population-based setting. *JAMA Neurol. (2018);**75(8):**970-979.*

13.0 Appendix

13.1 Picture description task Stimuli

ADDF has obtained licensing rights to the following picture stimuli from Boston University

Pictures designed by Boston University.

13.2 Storytelling Assessment Stimuli

ADDF has obtained licensing rights to the following picture stimuli from Boston University. For each story telling task, the mobile application displays a picture (see below) and a prompt that introduces an outside element to the picture and ask the participant to use their imagination to tell a story about what would happen next.

Pictures designed by Boston University.

13.3 Story Recall Stimuli

ADDF, in partnership with Boston University has created stories for the story recall test. Stories have been adapted in Spanish and Catalan by native speakers.

STOR 11: James loved to cook and dreamed of opening his own restaurant but was afraid it would fail. After his friend Sofia encouraged him to take a chance, he decided to open a small cafe. He worked hard to make every dish delicious. People in town soon began to notice the light-filled space, his freshly made pastas, and the vegetables he grew in his own garden. Over time, his dining room became popular in the community. People even began to come from other places to taste his cooking and enjoy the atmosphere. He was surprised and happy to see how much other people loved his cooking, and he was grateful to his friend for supporting him. To thank her, he named his favorite chocolate dessert on the menu after her and told her she would always be welcome to eat for free.

STOR 12: Alice loved reading books and wanted to share her passion with others. She decided to ask her classmates at school to help her donate books to the community. The other students were excited about the idea and brought her many textbooks and novels. In just one month, she collected over two hundred. She organized them and took them to the local library. The librarian, Mr. Brown, was impressed when he saw her bringing him all the books. He thanked her and told her that he was sure many people would want to borrow them. She felt proud that she was able to make a difference in the lives of others. Determined to keep making her town a better place, she began to volunteer there every week. She decided that one day she would become a writer to share her love with even more people.

STOR 13: Max loved playing soccer. He trained every day after school and dreamed of becoming a professional player. One afternoon, while practicing in a park, he kicked the ball too hard, and it landed in a nearby pond. He was upset because he thought he had lost his favorite ball forever. The next day, his neighbor, Mrs. Johnson, called him over and told him she had found his ball in the pond and taken it home to dry it off. He was very grateful, thanked her for her help, and invited her to watch his soccer game that weekend. She agreed and showed up wearing a team jersey and a big smile. He played his best game yet and even scored the winning goal. They became good friends after that, and he always remembered the lesson she taught him about the kindness of others.

STOR 14: Daniel loved to explore the woods behind his grandmother’s house. One day, he found a pond hidden behind some trees. He could see round pebbles at the bottom and many colorful fish swimming around through the crystal-clear water. He watched them for hours and enjoyed the peacefulness. As he was getting ready to leave, he noticed a trail leading away from the pond. He followed it to a large hill he had never seen before. He climbed to the top and was amazed by the view. He started visiting these places often, bringing a notebook to write about what he saw. After a while, he realized that he wanted to share the beauty with others, so he invited his friend Ava to join him. She was awed by the experience, and it inspired her to spend more time outdoors. From then on, they often explored nature together.

STOR 15: Andrew lived in a small town near the coast. He loved going to the beach and collecting seashells. One day, he found a beautiful shell the size of his palm that was unlike any he had ever seen before. He took it home to show to his neighbor, Ms. Harris, who worked as a scientist. When she saw it, she became very excited and told him that it was from an ancient species of sea creature that was thought to be extinct. She said that it was very rare and worth thousands of dollars. He was proud to have made such an important discovery and wanted to share what he found, so he decided to donate it to the natural history museum where others could see and learn from it. He continued to collect seashells, hoping to find more treasures that could teach him about the world around him.

STOR 21: Maria and George were two neighbors who lived in San Pedro and had always been competitive with each other. They enjoyed participating in various neighborhood activities and trying to outdo one another. When the annual San Pedro Charity Run was announced, they both decided to compete and prove who was the faster runner. For four weeks leading up to the race, Maria and George trained diligently. They each bought a new pair of running shoes and dedicated their mornings and evenings to jogging and doing exercises to improve their stamina and speed. Both of them also sought advice from experienced runners. On the day of the race, the entire neighborhood gathered to watch and cheer for the participants. As the starting horn sounded, Maria and George raced alongside each other, matching each other's pace step for step. Throughout the 5-mile course, neither of them managed to gain a significant lead. As they approached the finish line, George stumbled but quickly recovered. Maria, noticing this, decided to slow down just enough to make sure that they would cross the finish line together. As they finished the race side by side, they realized that their friendly rivalry had actually brought them closer. From that day on, Maria and George became great friends, learning that sometimes, it's not about winning but about supporting one another and enjoying the journey together.

STOR 22: Sarah was a cheerful and friendly woman who lived in a small village called Mirasol. She lived next to a grumpy young man named David, and their houses were separated by a beautiful garden. Sarah spent at least two hours every day watering the plants, pulling up weeds, and keeping bugs off of the flowers, creating a colorful oasis that brought joy to the entire village. One sunny morning, David decided to build a tall fence around his property, right in the middle of the garden. Sarah was upset and asked him to reconsider, but David refused to yield. He claimed that the garden attracted too many visitors, disturbing his peace and quiet. The villagers were dismayed by the fence, but Sarah came up with a plan. She planted three climbing vines along her side of the fence. As the vines grew taller, they transformed the cold barrier into a living, breathing wall of lush greenery and vibrant red roses. Over time, even David began to appreciate the beauty of the thriving vines. One day, he approached Sarah with an offer to remove the fence and to work together to maintain the garden. Sarah gladly accepted, and they began a partnership that lasted for many years. The once divided garden grew more beautiful than ever, and Sarah and David even became good friends, bringing the village closer together.

STOR 23: In a peaceful town called Solara, a woman named Joan was preparing for her first-ever cooking competition at the community center. She decided to make a special dish using five key ingredients: potatoes, carrots, onions, bell peppers, and tomatoes. Joan only had three hours to complete her culinary masterpiece. As she began chopping the vegetables, she realized she was missing the tomatoes. Panicking, she quickly put on their shoes, grabbed her wallet, and ran to the nearby market, Harvest Grocers. Upon reaching the market, Joan searched for the tomatoes. By the time she found them, there were only two left, and another shopper named Alex had already put them into his shopping cart. Sensing Joan's urgency, Alex kindly offered one tomato to Joan, deciding to share the scarce resource. Grateful, Joan thanked Alex and hurried back home to resume cooking. She incorporated the tomato into their dish and managed to finish just in time. She rushed to the community center, where she presented her dish to the judges. To her delight, Joan’s dish was a hit, and she was declared the winner of the cooking competition. As a token of appreciation, Joan shared her victory with Alex, inviting him to a celebratory dinner. The two formed a lasting friendship, bonding over their love for cooking.

STOR 24: Paula lived in a small town called Arbora, where she loved creating paintings and sculptures. Her family was very proud of her, and they encouraged her to pursue her passion for the arts. One day, Paula’s school announced an art competition. The winner would have their artwork displayed in the city council building. Paula was excited and immediately began working on her entry. She spent weeks crafting a beautiful watercolor painting of two scarlet red birds flying over the town's lake during the sunset. When the day of the competition arrived, Paula was nervous but excited to showcase her work. She was proud of her art, but when she walked into the museum where all of the other entries were on display, she noticed that many of the other paintings were much more detailed and intricate. The judges announced the winner, and it wasn't Paula. She was disappointed but tried to remain positive. As she was leaving the museum, a woman approached her, introduced herself as Mrs. Garcia, the owner of Gallery 34, an art gallery in the center of town. She told Paula how much she loved her painting and offered to display Paula’s artwork in her gallery. Paula was overjoyed and grateful for the opportunity. With her work in the gallery, Paula began to gain recognition as a talented young artist, and she continued to create beautiful works of art for many years to come. She learned that winning isn't everything and that sometimes, disappointing moments can lead to even greater opportunities.

STOR 25: Julia, a librarian in a small rural town called Rio Verde, found an old journal belonging to Laura, a woman who lived in the town 100 years ago. Laura had been the first woman to graduate from the medical school in their state and had become a respected doctor. She had lived in a modest house that had now been vacant for many years. Julia was so inspired by Laura's story that she decided to raise money to buy Laura's old house and turn it into a museum to preserve her legacy. With the help of her friends, Julia spent 6 months restoring the home and collecting old furniture, clothing, photographs and objects belonging to Laura and her family. She carefully arranged the objects, creating an immersive experience for visitors to step back in time and experience Laura's life. Julia created a study room in the museum with Laura's old wooden study desk, rocking chair, bookcase with scientific books, doctor's bag, and stethoscope. The museum became a beloved spot in the town and Julia felt proud to have made a difference by preserving Laura's legacy. Even after she retired, Julia would visit the museum and think about Mary's remarkable life and how it had influenced her own.

13.4 Open-Ended Questions Stimuli

Repeatable Open-Ended Questions

- Please describe how the weather has been around you over the last few days. (Q1)
- Please describe some of the foods you have eaten over the past few days. (Q2)

Rotating Open-Ended Questions

- Please describe what you usually do during the day. (Q3)
- Please describe how would you get yourself ready for a special event (like a wedding, sporting event, or family celebration). (Q4)
- Do you like using the smartphone? What do you most like using it for? What do you not like using it for? (Q5)
- What is your favorite food to eat? Why? (Q6)

	Baseline	3 month	6 month	9 month	12 month	15 month	Repeat
Speech Consortium Open-Ended Questions	Q1 Q2 Q3	Q1 Q2 Q4	Q1 Q2 Q5	Q1 Q2 Q6	Q1 Q2 Q3	Q1 Q2 Q4	