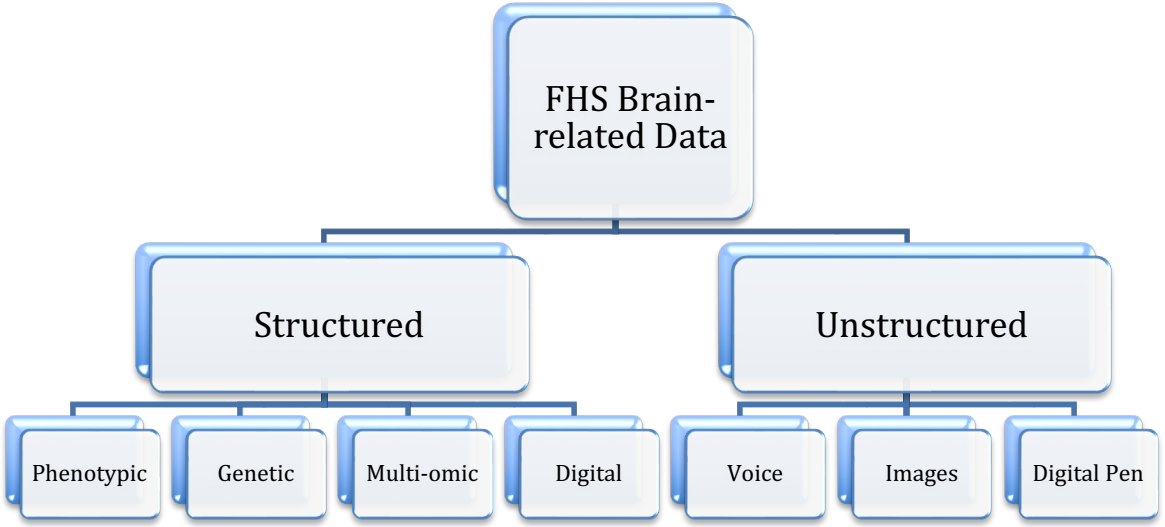
Data Repository

Currently, FHS has more than 600 conventional datasets that contain data collected from both the FHS core study and various ancillary studies. These datasets range from simple demographic and self-reporting data points to more complex multi-omic and digital data. In addition to these structured data, FHS has also been gathering unstructured data, especially in the recent years.

As depicted in the hierarchical flowchart below, FHS brain-related data can be broadly classified into structured and unstructured data, which can further divided into various categories. For structured data, they are available either at FHS or selected data-sharing sites (e.g., dbGAP, BioLINCC). Unstructured data are only at FHS.



Phenotype Data

The majority of FHS data falls under the structured phenotypic data. Under this category, one can find commonly used data such as demographic, self-reported responses to questionnaires, clinical outcomes, lab test results, etc

	Description	Data Format	File Type
Demographic Anthropometric	• Demographic: gender, age at each exam cycle • Anthropometric: Height, weight, blood pressure at each exam cycle	• Row-and-column • Consolidated longitudinal time series format	csv xlsx sas7bdat
Questionnaires	• Self-reported responses during clinic exam interviews • Self-reported responses for mailed questionnaires	• Row-and-column • Largely individual datasets • Few consolidated longitudinal time series format	csv xlsx sas7bdat
Investigational	• Non-invasive tests: Urine dipstick, physical stress test, lung function test • Invasive tests: Blood serum	• Row-and-column • Largely individual datasets • Few consolidated longitudinal time series format	csv xlsx sas7bdat
Clinical Outcomes	• Self-reported outcomes • Adjudicated outcomes • Survival status	• Row-and-column • Panel time series format	csv xlsx sas7bdat

Genetic Data

Over the past two decades, DNA has been collected from blood samples and from immortalized cell lines obtained from Original Cohort participants, members of the Offspring Cohort and the Third Generation Cohort (over 9,300 participants). Researchers are encouraged to request genetic data via dbGAP.

	Description	Data Format	File Type
Whole Genome Sequencing	Whole genome sequencing, mean 30X coverage, (~4100 participants)	General feature format	BAM VCF
Array-based genotypes and imputation to whole genome sequencing	~8000 participants		VCF
Whole Exome Sequencing	Whole exome sequencing (~2000 participants)	General feature format	BAM VCF

Multi-omic Data

Many of the multi-omic datasets comes under the SABRe projects – to identify the biomarker signatures of metabolic risk factors.

	Description	File Type
Gene Expression	Gene expression profiling, 18,000K (~5600 participants)	csv
DNA Methylation	DNA methylation, 45,000K (~4200 participants)	csv
	microRNA profiling (~7500 participants)	
Metabolomics	High throughput metabolite profiling (~2500 participants)	csv sas7bdat
Proteomics	Discovery proteomics in case-control studies of subclinical atherosclerosis, metabolic syndrome, general population	csv sas7bdat
Immunoassay	80 circulating protein biomarkers of atherosclerosis and metabolic syndrome (~ 7400 participants)	csv sas7bdat

Digital Data (Structured)

Many of the multi-omic datasets comes under the SABRe projects – to identify the biomarker signatures of metabolic risk factors.

	Description	Data Format	File Type
Radiological Scans	Derived measurements from X-rays Derived measurements from CT scans Derived measurements from MRI scans	• Row-and-column • Individual datasets	csv xlsx sas7bdat
Ultrasound	Derived measurements from ultrasound scans	• Row-and-column • Individual datasets	csv xlsx sas7bdat
EKG, Physical	Derived measurements from EKG Derived measurements from physical activities devices	• Row-and-column • Individual datasets	csv xlsx sas7bdat
Novel data	Derived measurements from digital pens Derived features from voice files (TBC)	• Row-and-column • Individual datasets • Panel time series format	csv xlsx

Digital Data (Unstructured)

		Description	File Type
Voice	Raw voice Recordings	Unedited version of a digital voice recording (DVR) that may contain personally identifiable information (PII)	dss , dvf, m3u m4a, mp3, wav, wma
	Censored Voice Recordings	A DVR that has been censored for PII using information from an MTI transcription	dss, dvf, m3u m4a, mp3, wav, wma
	Transcriptions	A human made transcription of a DVR. These transcriptions are time stamped, diarized, and generally do not contain PII.	txt
Image	MRI Brain	MRI brain scans since 2000 Defaced images	DICOM
	PET/Tau Scans	PiB Aβ and FTP tau PET imaging	NIFTI
Pen	Digital Clock Drawing (dCDT)	Real-time pen motion recording during the digital clock drawing test	csk
	Digital Pen (dPen)	Real-time pen motion recording during other neuropsychological tests	txt