HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2018 January 24.

Published in final edited form as:

Pac Symp Biocomput. 2018 ; 23: 280–291.

Best practices and lessons learned from reuse of 4 patient-derived metabolomics datasets in Alzheimer's disease*

Jessica D. Tenenbaum and

Department of Biostatistics & Bioinformatics, Duke University, Box 2721, Durham, NC 27710, USA, jessie.tenenbaum@duke.edu

Colette Blach

Duke Molecular Physiology Institute, Duke University, Box 104775, Durham, NC 27701, USA, colette.blach@duke.edu

Abstract

The importance of open data has been increasingly recognized in recent years. Although the sharing and reuse of clinical data for translational research lags behind best practices in biological science, a number of patient-derived datasets exist and have been published enabling translational research spanning multiple scales from molecular to organ level, and from patients to populations. In seeking to replicate metabolomic biomarker results in Alzheimer's disease our team identified three independent cohorts in which to compare findings. Accessing the datasets associated with these cohorts, understanding their content and provenance, and comparing variables between studies was a valuable exercise in exploring the principles of open data in practice. It also helped inform steps taken to make the original datasets available for use by other researchers. In this paper we describe best practices and lessons learned in attempting to identify, access, understand, and analyze these additional datasets to advance research reproducibility, as well as steps taken to facilitate sharing of our own data.

Keywords

FAIR; Open Data; Data Sharing; World Scientific Publishing

1. Background & Introduction

The importance of data sharing and reuse is increasingly recognized across the biomedical research landscape. Also receiving increased attention are the challenges of adhering to best practices in data sharing. In many cases, researchers and even data managers are not properly incentivized to put in the up-front time and effort required to make data discoverable, comprehensible, and interoperable. Even when projects do plan ahead for data sharing by incorporating the required effort into a budget and hiring experienced informatics

*This work is supported by 1RF1AG051550-01 and UL1TR001117

personnel, it is not always obvious how best to present data resources to facilitate discovery and uptake by others.

1.1. The FAIR guiding principles

Recognizing the urgent need to improve infrastructure for scholarly reuse of data, a group of stakeholders came together to develop what they referred to as "FAIR guiding principles", with FAIR as an acronym for Findable, Accessible, Interoperable, and Reusable.[1] These principles are meant to serve as guidelines and desiderata for good data stewardship. They are intended to enhance reusability of data, particularly from the machine perspective, enabling "computational agents" to identify, retrieve and analyze relevant datasets. A resource is 'F' (findable) if it has a globally unique and persistent identifier paired with rich metadata and is indexed in a searchable resource. 'A' (accessible) means that both data and metadata are retrievable using a standard, open protocol that allows for authentication as needed. The 'I' (interoperable) criteria relate to use of standards for knowledge representation. Finally, in order to be considered 'R' (reusable), a resource must have clearly defined and documented provenance and rules for usage.

The authors of the FAIR guiding principles make two important points that are relevant to the exercise described here: first, humans and machines face different challenges in the discovery and retrieval of relevant datasets. Humans have an intuitive sense of semantics and are able to interpret contextual clues such as icons, page structure, and narrative text. Machines lack these skills, but are far superior in scale and speed. In an ideal world, a resource enables discovery and reuse by both human and machine "stakeholders". Second, the FAIR authors assert that an optimal state in which computers are able to fully "understand" and operate on a digital object will likely rarely be achieved. Our intent in this work is not to fault any existing data resources, producers, or curators for in any way falling short of this theoretic optimal state. Rather, we seek to highlight ways in which existing datasets, all of which were made available before the FAIR guidelines were published, already adhere to these principles, and provide practical suggestions for how data producers going forward can make resources findable, accessible, interoperable, and reusable for both machines and humans.

1.2. The Alzheimer's Disease Metabolomics Consortium

The Alzheimer's Disease Metabolomics Consortium (ADMC- https://sites.duke.edu/adnimetab/) is a large, inter-institutional consortium that brings together centers of excellence of metabolomics, informatics and modeling to work collaboratively with Alzheimer's Disease experts to elucidate the molecular mechanisms of etiology and progression in AD. ADMC uses a systems approach in which metabolomics data are used to inform and complement genomics, proteomics, and neuroimaging data to provide novel insights about disease mechanisms.

The ADMC generated metabolomics data in collaboration with the Alzheimer's Disease Neuroimaging Initiative (ADNI) on the ADNI-1 cohort (see Section 2.1. below). These data were analyzed to identify peripheral metabolic changes in AD patients and correlate them with cerebrospinal fluid pathology markers, imaging features, and cognitive performance.

Desiring to validate findings in independent cohorts, we identified other extent sample collections and/or datasets for which similar clinical and molecular data had been collected, or could be generated prospectively (Figure 1). In this paper we assess the degree to which these datasets already adhere to FAIR criteria and identify additional desiderata for best practices in data sharing, especially for human users. Note that all of the datasets included here were discovered through distinctly human mechanisms: prior knowledge, networking, and past first-hand experience.

2. Methods

2.1. Datasets

Three cohorts were identified for use in validation of original findings (Table 1). In both the original analysis of the ADNI-1 cohort[2] and replication in the additional datasets, analysis required metabolomic data, demographics, and clinical data, e.g. cognitive tests, changes in AD status, and APOE genotype.

The ADNI-1 cohort on which the original analysis was performed is part of the Alzheimer's Disease Neuroimaging Initiative and comprises 200 normal controls, 400 individuals with MCI, and 200 subjects with mild AD. Metabolomics data were generated on baseline serum samples using the AbsoluteIDQ®-p180 kit (Biocrates AG).[3]

The Framingham heart study was initiated in 1948 to identify risk factors for heart disease, beginning with 5200 adult men and women from the town of Framingham, MA. In 1971 a second-generation "offspring" cohort was enrolled, consisting of 5,100 of the original participants' adult children and their spouses.[4] The offspring cohort had their second examination 8 years after enrollment, and subsequent visits approximately every 4 years after that, including imaging, cognitive assays, etc. On their fifth visit, blood was drawn and used to perform metabolomic profiling using a liquid chromatography / mass spectrometry (LC/MS) platform.[5] They did not the Biocrates p180 platform, however there was overlap in the specific metabolites measured including a number of amino acids, lysophosphatidylcholines, and sphingomyelins.

The Religious Orders Study (ROS) and the Memory and Aging Project (MAP) are both longitudinal cohort studies of aging and Alzheimer's disease (AD) run from Rush University. ROS enrolled individuals from more than 40 groups of religious orders (nuns, priests, brothers) across the United States for longitudinal clinical analysis and brain donation.[6] MAP was designed to complement the ROS study by using a similar structure and design as ROS, but enrolling participants with a wider range of life experiences and socioeconomic status.[7] The entire ROSMAP cohort consists of approximately 3000 participants. The ADMC has performed mass-spectrometry-based metabolomic profiling on both serum and post-mortem brain samples for a subset of the ROSMAP cohort using the AbsoluteIDQ®-p180 kit from Biocrates Life Sciences.

Finally, the MURDOCK Study is not an open dataset but rather a community-based longitudinal registry and biorepository based in Kannapolis, NC and run by Duke University with more than 12,000 participants enrolled.[8] A number of prospective disease-specific

"sub-studies" have been initiated from this registry, including a memory health study with approximately 800 participants. Blood and urine samples were collected at baseline enrollment along with demographic and clinical information. MURDOCK participants consent to give researchers access to their electronic health records for future study, and follow-up questionnaires are collected annually to ascertain longitudinal health status from the patient perspective. For the memory health study, participants were given assessments of cognitive status at enrollment and in a follow-up visit two years later. Metabolomic profiling was performed on baseline serum samples using the AbsoluteIDQ®-p180 kit.

2.2. Data governance

ADNI has a relatively straightforward process for applying for access. One must agree to an online Data Use Agreement and fill out a form that includes one's institutional affiliation and a description of the proposed use of the data. Annual status updates are requested via email, and failure to provide them results in access being rescinded.

Access to the Framingham data involves a more complex process. The Framingham data are stored in dbGaP. In order to request access, the applicant must have an approved IRB protocol for data analysis from their home institution. An application is then required that describes the proposed use of the data as well as a data management plan to keep data secure. Notably, the principal investigator's signature is not sufficient. Rather, an institutional signing authority is required to be involved, as well as an IT Director who has institutional (not just departmental) authority, e.g. the Chief Information Officer or Director of IT Security. A major hurdle for our inter-institutional consortium was the requirement that each institution obtain the data directly from dbGaP rather than access the data through our secure file share. Statistical collaborators at other institutions were thus required to obtain their own respective IRB protocol approval and apply for access through dbGaP including a named signing authority and IT contact. Even using Duke's protocol as a basis, this slowed things down considerably.

For the ROSMAP and MURDOCK studies, each has a process in place for a would-be collaborator to fill out a proposal for use of data and/or samples. A signed DUA is required between the source institution and each collaborating institution, as well as a material transfer agreement (MTA) where applicable. For both studies, the collaborator must then identify which specific variables are needed. MURDOCK additionally requires a data sharing document that specifies the mechanism of the data exchange.

3. Results

3.1. FAIR Assessment

We attempted to assess each dataset's adherence to the FAIR guiding principles. Note that we did not rely solely on machine-readable data and metadata particularly for the 'F', 'A', and 'R' criteria, but took into account resource owners' efforts to make datasets findable, accessible, and reusable for humans as well. The overall scores are provided in Table 2, with descriptions provided below.

We assessed each resource on a scale from 1 to 5 with 1 signifying no adherence at all and 5 connoting perfect adherence to the principles. By definition, since we were able to re-use each dataset to some degree, none of them received a score of 1. Conversely, none of them received a perfect 5 in any of the four areas. A formal analysis enumerating each sub-criteria is beyond the scope of this review, but specific examples of how the different datasets demonstrated the guiding principles are described in the following sections, along with some areas for improvement.

3.1.1. Alzheimer's Disease Neuroimaging Initiative ADNI-1 Cohort—ADNI is indexed in the Neuroscience Information Framework (NIF) as a resource, though not as a dataset *per se* ('F'). Access to ADNI data generally requires log-in to the Laboratory of Neuro Imaging (LONI) Image and Data Archive (IDA)[9] and manual navigation through a web interface to identify the files of interest ('A'). Given that ADNI is a complex study in its second decade and involves a complicated protocol to collect clinical, genomic, demographic, imaging, and cognitive data on multiple sub-cohorts, the available data spans hundreds of files and thousands of variables. This can be challenging to navigate, particularly for researchers new to the study. ADNI mitigates these challenges through extensive documentation and data dictionaries ('R'). ADNI has data dictionaries for each data file and a single consolidated dictionary in .csv format that enables searching for terms and filtering by topic. ADNI also has a merged file containing the most important variables. A major strength of ADNI is that all data files are available not only as .csv but also as packages for R, SPSS, SAS, and Stata ('A', 'I'). LONI also has tools for visualization of the population by different parameters (Figure 2).

The extensive existing clinical dataset for the ADNI-1 cohort was collected under the NIA-funded Alzheimer's Disease Neuroimaging Initiative. All ADNI-1 data must follow ADNI data governance rules, which include the stipulation that data be distributed only through the Laboratory of Neuro Imaging (LONI) Image and Data Archive (IDA).[9] Metabolomic data were generated through the NIA-funded Accelerating Medicines Partnership - Alzheimer's Disease (AMP-AD), which has a different set of data governance rules, including the requirement that all data be accessible through Sage Bionetworks' Synapse platform.

In order to comply with AMP-AD rules that require access to datasets through Sage Bionetworks' Synapse platform[10], a project was created in Synapse with digital object identifiers (DOIs) that point to the relevant permanent URLs in LONI's IDA ('F'). Importantly, these DOIs can be versioned as the underlying data files are updated, e.g. when additional clinical data are collected. Though two separate logins are required, one for Synapse and one for LONI, the handoff is otherwise transparent and selecting "download" from the Synapse interface enables a user to download the appropriate file through the open http protocol directly after LONI authentication ('A').

3.1.2. Framingham—The study itself is indexed in both DataMed and NIF and has a permanent, versioned accession number in dbGaP ('F'). Data are downloadable from dbGaP through http once permission is obtained ('A'). Documentation is extensive, including annotated codebooks, procedures, variable statistics and publications ('R'). Again unsurprisingly for such a large, complex, and long-running study the documentation can be

overwhelming, particularly for someone new to the study. The complexity is partially mitigated by search tools on the Framingham web site for variables. File names include some amount of metadata, with documentation to help the user understand shorthand naming conventions. However, individual data dictionaries for respective files contain headers that describe file content clearly and in detail. Another helpful resource for understanding Framingham data is a spreadsheet listing all of the different data files along with what cohorts they apply to and what types of data they contain (‘R’) (Figure 3). Finally, although dbGaP does not make explicit use of the Data Use Ontology (DUO), which has been adopted by the Global Alliance for Genomics and Health (GA4GH, or the Global Alliance) to code consent information, it does reference concepts that appear in the DUO such as not-for-profit use only (‘F’).

3.1.3. ROSMAP—The Rush Alzheimer's Disease Center (RADC) has developed an elegant and user friendly “Research Resource Sharing Hub” designed to enable non-RADC investigators to navigate the complex set of data and biospecimens available for sharing (Figure 4) (‘F’). This website provides extensive documentation, the ability to generate reports on numbers of research participants matching specific criteria broken down by demographics (Figure 6), and the ability to submit a request for data and/or biospecimens (‘A’). Once our data request was approved, the Rush team extracted the required data and shared it via Dropbox.

3.1.4. MURDOCK—The MURDOCK Study is not an open dataset but rather a registry and biorepository intended to facilitate cohort identification and collaborative sub-studies. Thus, in contrast with the datasets described above, the MURDOCK Study currently has only five forms and hundreds of data elements compared to the many thousands found in Framingham or ADNI. The main MURDOCK Study website provides a link to an online data dictionary documenting the different data elements collected at the enrollment and follow-up stages of the study (‘R’). The website also gives a human readable overview of some demographic data and self-reported clinical history for a number of common diseases (‘R’) (Figure 5). It also provides information regarding the cognitive tests performed in the memory health study: attention and concentration, executive functions, memory, language, visual skills, conceptual thinking, calculations, and orientation. Once the data transfer is approved, the MURDOCK team extracts the specified field data and shares it using Box or ftp (‘A’).

3.2. Common challenges across datasets

3.2.1. Metadata summarization and complexity—Critical for every project was high level documentation to acquaint collaborators with study design and available data domains. Graphical overviews with linked details tend to be more informative and user-friendly than text-based summaries. In some cases, collection protocols were represented graphically; along with their corresponding naming conventions and file names. The best overviews included metrics for the data sets such as counts of different sample types, data types, etc. Metadata describing the processes, data files, fields and coded values were available from each of the projects and essential for data re-use. In almost all cases, metadata was largely human-readable and not computable or queryable (ROSMAP being the notable exception-see Figure 4).

Though none of the datasets described here were shared through metabolomics-specific repositories with computable metadata, progress has been made in establishing standards for metadata for metabolomic datasets. For example, EMBL-EBI (European Molecular Biology Laboratory- European Bioinformatics Institute)'s Metabolights data repository requires ISA-tab formatted metadata and provides a preconfigured downloadable ISACreator template. Use of ISA tools and the ISA standard does have some associated learning curve, but our team was able to make the ADNI1 p180 dataset ISA-compliant with significant help from knowledgeable curators for a recently accepted "data descriptor" (*Nat Sci Data, in press.*). According to a reviewer of this manuscript and documentation on GitHub (https://github.com/ISA-tools/isa-api), there exists a script, biocrates2isatab.py, that enables seamless conversion of Biocrates data to ISA-tab format, however we were unable to locate the script itself- perhaps it is not yet publicly available. Certainly the use of such tools and standards will help to ensure FAIR datasets moving forward.

Other important sources of study metadata are data dictionaries for each domain. Data dictionaries can take on different levels of rigor and utility. Ideally a data dictionary is provided in a tabular format so that it is searchable for specific terms, browsable to get a feel for the different data domains and variables included, and filterable by topic. If the data files themselves do not use standard identifiers for variables, the data dictionary may facilitate mapping variables to existing standards, e.g. mapping local identifiers for metabolites to standard identifiers such as INCHI Key or ChEBI ID. Although some variables may seem obvious enough not to need descriptive text, contextual information is often helpful, e.g. TimeStamp might be described as "Time stamp for blood draw" rather than "Time stamp."

An additional local use case was the ability to query and filter based on status of metabolomics assays, e.g. which biospecimens had been assayed on a specific platform, and connecting that information to clinical and demographic data. A tool with i2b2-like graphical querying functionality would enable a PI or researcher to assess how many participants had both metabolomic and imaging data, and a diagnosis of AD.

3.2.2. Data concept mapping across projects—Notable progress by the Metabolomics Standards Initiative and the Coordination of Standards in Metabolomics (COSMOS) initiative.[11, 12] But as with many biological domains other than genomics, adoption of metabolomic data standards has been slow. Metabolomic data itself adds a layer of complexity in that some observations of molecular species may be ambiguous, for example lacking the ability to differentiate between two molecules with the same atomic composition but with double bonds between different carbon atoms. It is therefore not possible in some cases to assign a specific identifier to a given experimental value, since the value actually represents *species A* OR *species B*. Since the same Biocrates kit was used for three of the four datasets, mapping of metabolites for those three sets is trivial. Mapping and some manual review are needed to map the overlapping species between the p180 kit and the LCMS platform results for the Framingham study. For example, lysophosphatidylcholine (carbon:double bond = 16:0), is referred to as "C16_0_LPC" and "lysoPC a C16:0" in Framingham and the Biocrates kit respectively. Analysis has not yet been performed to determine consistency among the Biocrates datasets, nor comparability between Biocrates

and the other LCMS platform, but this will be an important finding for future attempts to compare across metabolomic datasets.

In all observed cases, studies defined their own data elements rather than using existing concepts from existing terminologies such as SNOMED CT, LOINC, or PhenX. This resulted in some cases of significant semantic differences in variables of the same name, for example 'APOE' as a genotype vs. continuous variables representing RNA expression. Increased use of commonly accepted standards will increase interoperability of datasets moving forward.

Also related to interoperability, categorizing diagnoses was not consistent across studies and different protocols were used for consensus diagnosis. Although different assessments were used to evaluate cognitive impairment, they were each established, validated, standardized instruments. It was therefore possible to establish equivalent concepts across projects with input from clinical experts.

In all cases, no matter how detailed the codebooks or project descriptions, there was always some need to ask for assistance from the data owners and to document this additional information for the analysts. This included, for example, additional information conveyed within a variable name e.g. single letter codes within variable names identifying brain regions.

3.2.3. Versioning and data provenance—Reproducible research requires the ability to track different versions of data as well as data provenance. Data sources can change for many different reasons, either because an error was discovered, or because additional data have become available. The Framingham study does a particularly good job versioning the data available in dbGaP, clearly identifying later releases of data for download after an embargo period, and dividing the data into two different groups based on participant consent. (One group consented to use for all research; the other consented for research use only by nonprofit entities.) For the ADNI-1 cohort, LONI has a policy that file names should not change so that researchers can always find the file they had previously downloaded. In addition, in order to adhere to DOI requirements for the AMP-AD project, LONI has enabled explicit versioning of data files within the IDA.

4. Conclusions

Based on our experience exploring publicly available datasets to validate translational findings we would add to the FAIR guiding principles the following best practices, particularly to enable data discovery and reuse by human beings: 1. Provide user-friendly metadata in the form of a graphical overview of data, sample types, instruments used at timepoints and counts; 2. Provide a data dictionary that is both browsable and searchable; and 3. Use common data elements wherever possible for data collection, whether from clinical terminologies or molecular databases.

It is easy for a group to become so familiar with their own data that they lose perspective on how it will be seen and interpreted by others. This exercise has helped inform our own work to make our data FAIR for other researchers, though we are not yet where we wish to be.

Understanding of data sharing use cases, a well-formed plan, and dedicated resources are needed to enable adherence to FAIR principles.

It is encouraging to see that real effort is being devoted to making scholarly data available for re-use. A decade ago, it would have been difficult to obtain even a single dataset for validation. Our experience with the three cohorts described above suggests that although we have a long way to go before data are FAIR for computational agents, significant progress is being made to make data resources findable, accessible, and reusable by human agents. Our experience also suggest that, as with clinical data, we have a long way to go before data are truly interoperable. The obstacles are largely not technical ones. Education in the issues described here, as well as the will and the resources through aligned incentives, will ensure that we continue to make progress toward a FAIRer research data landscape.

Acknowledgments

The authors would like to thank our contacts at the four described studies for their support in understanding the rich and complex datasets: Michael Donohue, Karen Crawford, and Arthur Toga from ADNI; Lauren Silva, Honghuang Lin, Chunyu Liu, and Rhoda Au from the Framingham Study, Debra Fleischman, Gregory Klein, and John Gibbons from ROSMAP, and Brenda Plassman, Lawrence Whitley and Heather MacDonald from the MURDOCK Memory Health Study. We also thank the reviewers for their helpful comments.

References

1. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016; 3
2. Toledo JB, et al. Metabolic network failures in Alzheimer's disease-A biochemical road map. Alzheimers Dement. 2017
3. St John-Williams L, et al. Targeted metabolomics and medication classification data from participants in the ADNI1 cohort. Nat Sci Data. 2017 (Accepted).
4. Kannel WB, et al. An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol. 1979; 110(3):281–90. [PubMed: 474565]
5. Ho JE, et al. Metabolomic Profiles of Body Mass Index in the Framingham Heart Study Reveal Distinct Cardiometabolic Phenotypes. PLoS One. 2016; 11(2):e0148361. [PubMed: 26863521]
6. Bennett DA, et al. Overview and findings from the religious orders study. Curr Alzheimer Res. 2012; 9(6):628–45. [PubMed: 22471860]
7. Bennett DA, et al. Overview and findings from the rush Memory and Aging Project. Curr Alzheimer Res. 2012; 9(6):646–63. [PubMed: 22471867]
8. Tenenbaum JD, et al. The MURDOCK Study: a long-term initiative for disease reclassification through advanced biomarker discovery and integration with electronic health records. Am J Transl Res. 2012; 4(3):291–301. [PubMed: 22937207]
9. Crawford KL, Neu SC, Toga AW. The image and data archive at the laboratory of neuro imaging. Neuroimage. 2016; 124:1080–1083. [PubMed: 25982516]
10. Omberg L, et al. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. Nature genetics. 2013; 45(10):1121–1126. [PubMed: 24071850]
11. Salek RM, et al. COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. Metabolomics. 2015; 11(6):1587–1597. [PubMed: 26491418]
12. Members MSIB, et al. The metabolomics standards initiative. Nat Biotechnol. 2007; 25(8):846–8. [PubMed: 17687353]

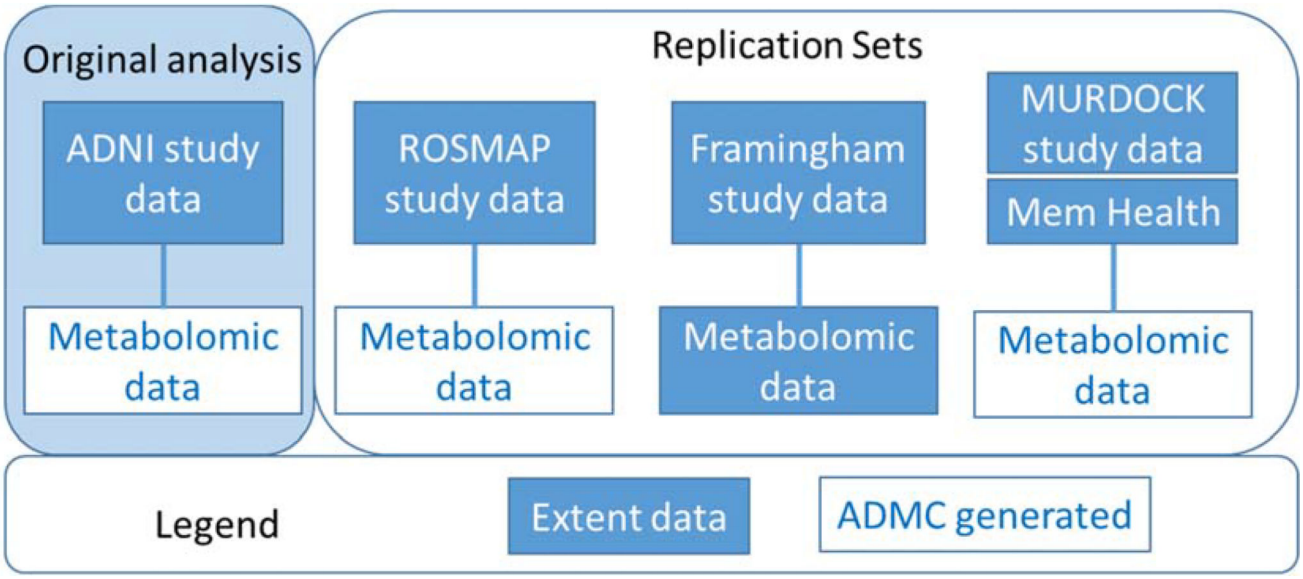


Figure 1.

Metabolomic profiling was performed on the ADNI-1 cohort. The resulting metabolomic dataset was combined with clinical data collected on the ADNI-1 cohort, including AD-related markers and cognitive tests, to identify biomarkers in AD. Three additional cohorts were identified for which either metabolomic data had been collected (Framingham) or biospecimens were available (MURDOCK and ROSMAP). The ADMC performed metabolomic profiling on serum samples from ROSMAP and MURDOCK. Analysis of these datasets is ongoing.

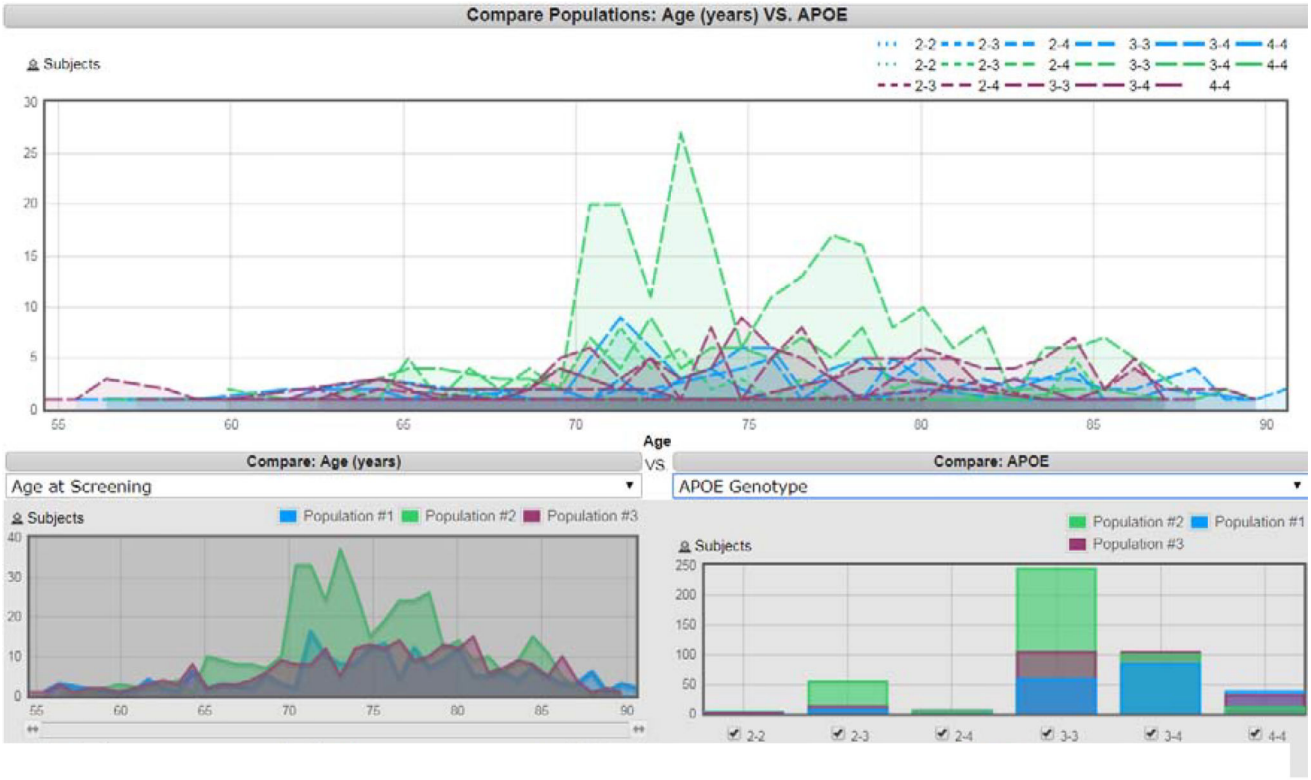


Figure 2.
Visualization tools on the ADNI's data archive website.

Nature of Data	Data Set Name	Cohort: Gen 1 Original	Gen 2 Offspring	Gen 2 New Offspring	Gen 3 Generation 3	Omni 1	Omni 2	Time Interval	Interview / Physical Exam / Examiner's Opinion: Original data collectio: Clinic Exam	Neuropsychology	Health Status Update	Validated (though medical records) and/or Reviewed and/or Derived: Menopause	Foot Study	MMSE, CERAD/Stroop	Rheumatic Heart	ICD Codes	Dementia	Atrial Fibrillation	Exam Dates, Age, Sex	Food Frequency with Der	Cancer	Diabetes	Cardiovascular	Survival	Endpoints	Social Network	Stroke Related	Medications	Neuropsychology Brain MRI S	Bone Related
Year of Birth in 5 Year Intervals	birthyr_alls	●	●		●			Exam 1											●											
Initial Neuropsych Battery Assoc with Brain MRI	neuropsych_2005s	●	●					thru 2005																					●	
SOE (Sequence of Events)	soe_2007s	●	●	●	●			thru 2007																	●					
Survival File & Follow-up CVD Events	srv_2007s	●	●					thru 2007																●						

Figure 3.
Framingham spreadsheet of data files and related metadata.

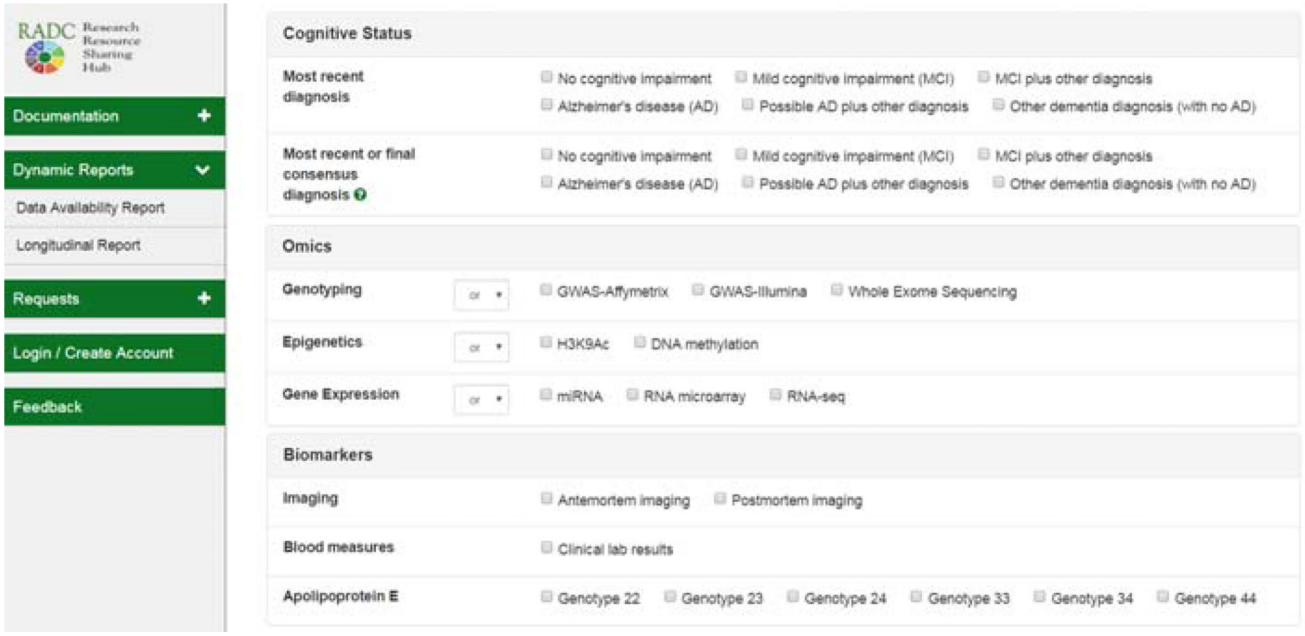


Figure 4.

A screenshot (edited) of Rush's Research Resource Sharing Hub, enabling users to query for available data for research participants who meet specific criteria.

Self-Reported Diseases	Percent of Total Cohort
Coronary Artery Disease	7.3%
Cancer	23.6%
Diabetes	18.4%
High Cholesterol	44.5%
Osteoarthritis	23.9%
Depression	28.3%
Other Mental Illness	5.2%

Calculated BMI	Percent of Total Cohort
Underweight (<18.5)	1.1%
Normal (18.5-24.9)	26.9%
Overweight (25.0-29.9)	33.0%
Obese (>29.9)	37.1%
No BMI Recorded	1.9%

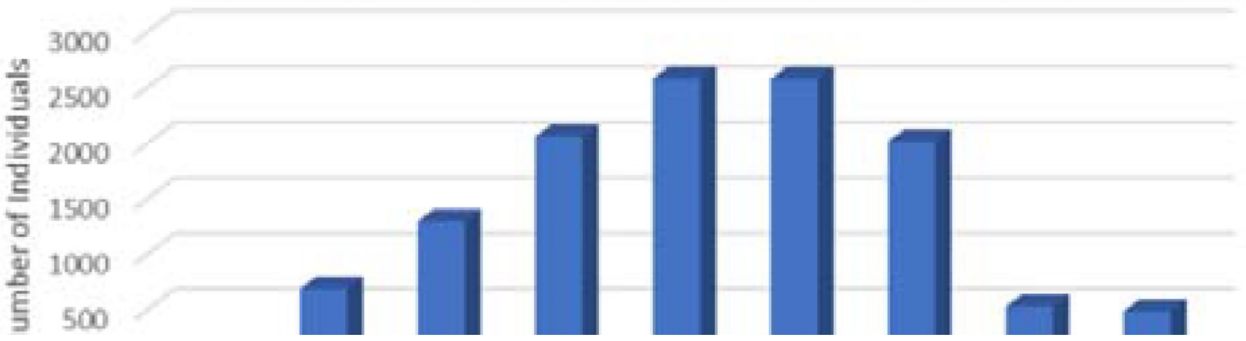
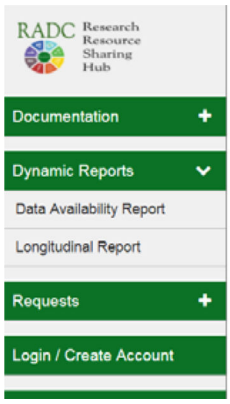


Figure 5.
Self-reported clinical history, BMI, and age in the MURDOCK Registry found on the public facing MURDOCK Study website.



Ethnicity by sex	Female	Male	Total
Latino	136	41	177
Non-Latino	2263	846	3109
Unknown	0	0	0
Total	2399	887	3286

Race by sex	Female	Male	Total
White	2195	849	3044
Black/African-American	177	32	209
Asian/Pacific-Islander	14	4	18
Other	13	2	15
Total	2399	887	3286

Education	Total participants	Mean education (years)	Sample standard deviation (years)	Participants without college	Percent without college
All	3286	16.2	3.8	667	20.3%
Female	2399	15.9	3.6	511	21.3%
Male	887	16.9	4.2	156	17.6%
White	3044	16.3	3.7	600	19.7%
Black/African-American	209	15.3	3.8	55	26.3%
Asian/Pacific-Islander	18	14.1	4.6	7	38.9%
Other	15	13.8	6.3	5	33.3%

Figure 6.
Tabular results of a query of a Rush Research Resource Sharing Hub query for frequency data of ROSMAP participants.

Table 1

Overview of datasets included in evaluation.

Dataset	Full name	Study URL	Data URL
ADNI	Alzheimer's disease neuroimaging initiative	http://adni.loni.usc.edu/	http://adni.loni.usc.edu/data-samples/access-data/
Framingham	Framingham Heart Study	https://www.framinghamheartstudy.org/	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?id=phs000007
ROSMAP	Religious Orders Study and Memory and Aging Project	https://www.synapse.org/#!Synapse:syn3219045	https://www.radc.rush.edu/
MURDOCK	Measurement to Understand Reclassification of Disease of Cabarrus/ Kannapolis	https://www.murdock-study.com/	https://www.murdock-study.com/services/data-dictionary/

Table 2

Scoring of compliance with FAIR principles for each dataset. Legend: 1- no adherence; 2- minimal evidence of adherence; 3- some adherence; 4- good adherence; 5- follows principles to the letter. The MURDOCK Study is not included here because it is not an open data set but rather a registry and biorepository for collaborative research.

Dataset	Findable	Accessible	Interoperable	Reusable
ADNI	3	3	2	4
Framingham	4	3	2	4
ROSMAP	4	2	2	4