

INTRODUCTION

The a2 Collective represents the National Institute on Aging (NIA) Artificial Intelligence and Technology Collaboratories (AITC) for Aging Research program and is comprised of three AITCs centered at Johns Hopkins University (JH AITC), the University of Massachusetts Amherst (MassAITC), and the University of Pennsylvania (PennAITech), and a Coordinating Center (CC) managed by Rose Li and Associates, Inc. (RLA). The a2 Collective is committed to sharing resources, with broad availability of policies, practices, materials, protocols, data, and tools to facilitate collaboration across multiple centers as well as replication and reuse by a range of researchers and private entities. Developed by the a2 Collective, this Data and Resource Sharing Plan subscribes to the following overarching principles:

- Transparent sharing of non-proprietary resources and pre-competitive data, as permitted by consent and IRB guidelines, is essential for the expedited translation of research results into knowledge, products, and procedures to improve human health.
- a2 Collective members, including a2 pilot awardees, participating in data sharing, will be offered early access to these resources to ensure the non-duplication of foundational work for AI in aging and maximize the power of baseline analytics across centers.
- Resources produced, coordinated, and shared as part of the AITC program should be shared with the research community as broadly as possible while maintaining respect for data ownership and supporting timely commercialization with the fewest encumbrances possible.
- The utility of resources must be preserved for the future to the extent possible.
- The promotion and adoption of consensus open-access standards for data interoperability and harmonization of data reporting/coding into comparable and consistent renderings are important for sustaining coherent inferencing across domains.
- FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles – emphasizing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.

WHAT DATA/RESOURCES WILL BE GENERATED

Members of the a2 Collective are expected to generate a wide variety of data/resources, including:

- Pilot research data collected from human subjects (e.g., electronic health records, patient- or consumer-reported information, contactless sensor data, robotic parameters and outputs, video and image data, audio and acoustic data, aggregated focus group data)
- Software components to support the data collection process and/or processing of preexisting databases
- Databases of AI and technologies for healthy aging (e.g., metadata including descriptions of the study designs, interview guides, surveys, and data dictionaries of funded pilot projects)
- Protocols, tools, and templates concerning submission, review, award, and management of pilot competition applications
- Standard operating procedures (SOPs) for technology platform utilization
- Key program performance indicators and measures of program impact
- Background materials, webinars, and program recordings used for training and education

- Policies and other guidance assembled by special experts, such as that pertaining to AI ethics
- Proprietary data and software generated by private or corporate groups that is protected by copyright or patent or otherwise determined to be a trade secret

Any individual-level data generated by pilot awardees will be de-identified prior to sharing by the awardees, following IRB guidance, which could include HIPAA regulations for covered entities. Data deemed to be proprietary or forms of data that cannot be adequately de-identified (e.g., images, video and speech recordings, some forms of sensor data) might require restricted access (e.g., required Memorandum of Understanding, enclave use) to be shareable. Individual pilot awardees may negotiate what data are considered proprietary or subject to data sharing restrictions with the awarding AITC. Our expectation is that raw data will rarely be proprietary; rather, the inferences made from those data through AI or machine learning are likely to be proprietary. The rationale for classifying data as proprietary needs to be explicitly explained and can be included in the data sharing plan of the pilot awardee.

WHO WILL HAVE ACCESS TO STUDY DATA

External researchers seeking access to individual-level pilot data must satisfy the terms of a master Data Use Agreement (DUA) with the cognizant AITC, pilot awardee, or designated data repository. This master DUA requires:

1. A commitment to using the data only for research purposes and not to attempt to identify any individual participant
2. A commitment to securing the data using appropriate computer technology
3. A policy for retaining, destroying, or returning the data, as appropriate, after analyses are completed, and explanation of how compliance will be verified
4. Restrictions on redistribution of the data to third parties
5. Proper acknowledgement of the data resource and funding source

WHERE WILL THE DATA/RESOURCES BE AVAILABLE

Pilot awardees will be invited to store and share pre-competitive data of active interest in a common use data enclave hosted by JH AITC secure platforms, except when in conflict with IRB guidance. For proprietary data, efforts will be made to establish if pre-competitive components exist that can be shared without the release of sensitive information. Access to this data commons will be available to AITCs and pilot awardees participating in data sharing during the pilot project period of performance, with external researcher access permitted by IRB approval and according to AITC-specific protocols. Downloading of data outside of the enclave environment will be monitored and discouraged.

The AITCs intend to model Bridge to Artificial Intelligence (Bridge2AI) initiatives and guidelines to harmonize, consolidate, and protect pilot project data, including the adoption of open-access, international, consensus clinical and biomedical data standards. In addition, the consortium data enclave will embrace the use of firewalls and encryption, and adherence to data

safety and security requirements. Data generated by AITCs (e.g., AITC-initiated clinical translation and validation studies) may be stored on individual AITCs' secure platforms.

Members of the a2 Collective agree to archive meaningful pilot project data and underlying raw, pre-competitive data (if any) at the NIA-funded National Archive of Computerized Data on Aging (NACDA), located within the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan in accordance with timelines and accommodations made for sensitive and proprietary data that may be subject to restrictions.

Focus group data will be aggregated into tables and the aggregated data will be shared through publications or white papers.

In addition to posting on a2 member websites, relevant papers related to a2 Collective research and development activity and accompanying supplemental files will be uploaded to archives that are available to key audiences (e.g., PennAITech uses Scholarly Commons to store and share metadata and provide details on requesting access to the data for secondary analysis purposes, which is free for Penn investigators). Digital Object Identifiers (DOIs) will be generated and provided through services available to all (e.g., FigShare, Zenodo).

Non-proprietary software components developed to support the data collection process will use open source software platforms where possible to facilitate sharing and will be made available through GitHub.

WHEN WILL DATA/RESOURCES BE SHARED

Resources produced, coordinated, and shared as part of the AITC program will be shared with the research community at large as broadly as possible while maintaining respect for data ownership and supporting timely commercialization with the fewest encumbrances possible if applicable. Pilot awardees may negotiate allowances on a case-by-case basis with the awarding AITC regarding proprietary resources, such as algorithms that are generated by private companies for the purpose of commercialization or are otherwise protected by patents or copyright laws.

To the extent possible, pilot project data will be readied for researcher use or for deposit into NACDA by the end date of the pilot grant period, upon acceptance of the data for publication or public disclosure of a submitted patent application, when deemed appropriate by patent or copyright laws, or upon release of proprietary resources, whichever is earlier. Pilot awardees may negotiate alternative embargo periods on a case-by-case basis with the awarding AITC but data cannot be embargoed beyond the life of the AITC grant.

HOW WILL RESEARCHERS LOCATE AND ACCESS THE DATA/RESOURCES

Information about availability and access to data and resources will be included in any publications and presentations authored or coauthored by a2 Collective members, as well as acknowledgement of the AITC program (a2 Collective) and funding source in any publications and presentations. NACDA, which is an NIA-funded repository, has policies and procedures in place that will provide data access to qualified researchers, fully consistent with NIH data sharing policies and applicable laws and regulations.

A catalog of all de-identified data sets associated with an AITC's pilot awardees will be available on the respective AITC's website, and these datasets and associated documentation will be shared through DUAs.

INITIATIVE TO HARMONIZE STUDY DATA PRIOR TO ARCHIVING

Pilot awardees are expected to communicate quarterly with the awarding AITC mentoring team to review progress, data capture, existing data standards, and data catalogues in order to ensure that no issues have arisen regarding data integration and quality. Reporting on data collection and sharing will be part of the AITC's integrated reporting and mentoring process and will be shared with the AITC's data and resource sharing lead(s).

Contingent on supplemental funding, the AITCs will harmonize data governance approaches where advisable (e.g., data transfer agreements, data use agreements, data use requests, codes of conduct, and attribution, as well as publication policies and use of accepted standard nomenclatures). For the domains of data relevant to this project, the AITCs intend to adopt known best practices for standards-based representation of subject-level raw data, and give careful consideration to the appropriate anticipated standard, maturity and sufficiency, ease of extension, likelihood of new standards, and AI/ML readiness of data, with the goal of unifying data attributes across multiple data sources and across data types and establishing sensible interoperability targets.

The AITCs are committed to bringing their IRB protocols into alignment with this Data and Resource Sharing Policy.

REFERENCES

Cole CL, Sengupta S, et al. 2021. Ten principles for data sharing and commercialization. *Journal of the American Medical Informatics Association*, 28 (3) March: 646-659.

<https://doi.org/10.1093/jamia/ocaa260>

OECD. 2019. Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-Use Across Societies. <https://doi.org/10.1787/276aaca8-en>

<https://www.oecd-ilibrary.org/sites/b4d546a9-en/index.html?itemId=/content/component/b4d546a9-en>