

AI-READI: rethinking AI data collection, preparation and sharing in diabetes research and beyond

AI-READI Consortium



Here, we introduce Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI), a multidisciplinary data-generation project designed to create and share a multimodal dataset optimized for artificial intelligence research in type 2 diabetes mellitus.

The AI-READI project is one of four Data Generation Projects (DGPs) funded by Bridge2AI (<https://commonfund.nih.gov/bridge2ai>), a new NIH Common Fund Program aimed at setting the stage for the widespread adoption of AI in healthcare research. The primary goal of AI-READI is to collect and publicly share a multimodal, AI-ready dataset for studying the pathogenesis and salutogenesis (that is, the pathway from disease to health) of type 2 diabetes mellitus (T2DM).

T2DM is a growing public health threat, affecting >6% of the world's population, and at increasingly younger ages. Certain populations experience a greater burden of disease, exacerbating health disparities. Our data collection effort is centred around the principle of collecting a large multimodal dataset uniformly balanced for T2DM severity from a diverse group of participants so as to better understand this complex multifactorial disease using AI. Given the complexity of T2DM, AI-based approaches and multimodal model development may improve our understanding of T2DM.

However, a major barrier has been the lack of diverse datasets that are AI ready¹ for training AI models, which the AI-READI project is designed to address.

Here, we present the AI-READI dataset, give preliminary insights into our blueprint for making data AI-ready and provide our ongoing strategies for training a diverse workforce at the intersection of AI and biomedical research.

Introduction to the AI-READI dataset

AI-READI is enrolling 4,000 participants at three study sites over 4 years (2022–2026) (Supplementary Fig. 1). We plan to balance the racial and ethnic diversity of participants to address the under-representation in past research studies of groups that bear a higher burden of the disease. Having three data collection sites will increase geographic representation. Participants with a range of health states will be included to facilitate AI-based research into salutogenesis.

The recruitment process consists of a selection of individuals based on the available clinical diagnosis using ICD-10 codes and demographic data from the electronic health records of study sites. Each participant completes the study protocol, which includes questionnaires, on-site data collection and at-home data collection (Fig. 1). Blood is being collected to establish a large repository of participant plasma, serum and buffy coats at the University of Alabama at Birmingham. In addition, we are collecting PaxGene RNA tubes for future RNA isolation and cryopreserving peripheral blood mononuclear cells for future studies that could include functional immunologic studies or generation of induced pluripotent stem cells.

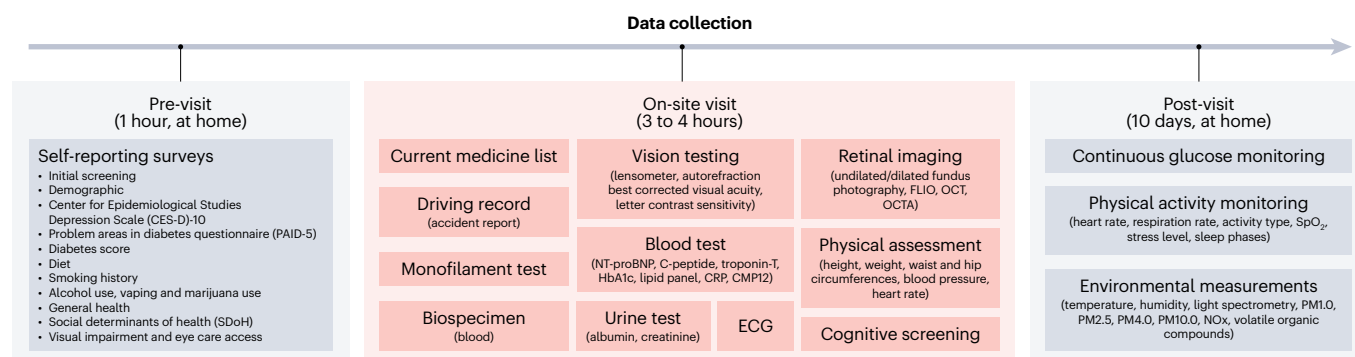


Fig. 1 | Data collection protocol for each participant in AI-READI.

Each participant who enrolls in the AI-READI study completes the study protocol, which includes pre-visit questionnaires, on-site data collection and post-visit data collection for 10 days using a Dexcom G6 for continuous glucose monitoring, Garmin Vivosmart 5 for physical activity monitoring and

a custom-built environmental sensor for indoor environmental parameter measurements. ECG, electrocardiogram; FLIO, fluorescence lifetime imaging ophthalmoscopy; OCT, optical coherence tomography; OCTA, optical coherence tomography angiography; PM, particulate matter.

Each data collection site uploads data at regular intervals to a cloud-based data management platform (FAIRhub) that is being developed for this project. Data are released annually, and each version is accessible through FAIRhub. At each release, two datasets are made available: a controlled-access set and a publicly accessible set with fewer requirements for access. The public set is stripped of Protected Health Information (PHI), as defined by the HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule via the “Safe Harbor” method, as well as information related to the sex and race/ethnicity of the participants to prevent stigmatization of findings. The first version of the public set was recently shared using a new licence developed as part of this project that allows reuse for any purpose but includes restrictions to protect participant privacy².

Blueprint for future data-generating projects

An additional aim of the AI-READI project is to share our blueprint for collecting and sharing AI-ready datasets so that future DGPs can follow them (Supplementary Fig. 2), including guidelines for project management, as well as data collection, management and sharing (<https://zenodo.org/communities/aireadi>; <https://github.com/AI-READI>; <https://aireadi.org>).

The integration of multiple disciplines has become a cornerstone for discoveries and innovation in biomedical research, so the successful preparation of an AI-ready dataset must be a multidisciplinary team effort^{3,4}. Therefore, an essential component of AI-READI is to implement team science strategies to understand the interaction patterns of our multi-team systems and then disseminate our findings (Supplementary Fig. 3).

We believe that, at a high level, making data AI ready involves two main sets of considerations: technical and ethical. The FAIR (findable, accessible, interoperable, reusable) principles provide high-level instructions for making data technically ready for reuse by AI systems². As part of the project, we are developing guidelines for making data types from the AI-READI project FAIR. We are using existing standards when available and working with their maintainers to extend them when necessary (Supplementary Table 1). We are also developing new standards, such as the Clinical Dataset Structure (CDS), a simple and intuitive way to organize clinical research datasets and include structured metadata (<https://cds-specification.readthedocs.io>).

Data used in developing AI systems are often a major source of downstream ethical issues⁵. To prevent such issues, ethical, legal and social implications are considered at every stage of the project cycle. Documenting practices regarding the creation, use and maintenance of clinical research datasets is critically important so that AI developers can easily understand the provenance and intended use of the data. We are reviewing existing documentation approaches such as datasheets and healthsheets to identify the most suitable one for biomedical data^{6,7}. AI algorithms can infer individual patient characteristics and could facilitate individual-level identification. To prevent reidentification attempts, we are implementing a robust data dissemination system, a strict licence agreement and data watermarking for both public and controlled sets.

Preparing and sharing AI-ready datasets can rapidly become time consuming and difficult for data-collecting researchers. To address this problem, we are developing a cloud-based data management, curation and sharing platform called FAIRhub (<https://fairhub.io/>). Inspired by existing user-friendly ‘FAIRification’ tools, the platform is being developed to include intuitive user interfaces and a suite of tools to simplify and automate the implementation of our guidelines for making data AI

ready⁸. After testing FAIRhub for the AI-READI dataset, we anticipate making the platform available for future DGPs desiring to manage and share AI-ready clinical research datasets.

American Indians and Alaska Natives (AI/AN) are unique and highly identifiable groups that could benefit from precision health AI studies. Through engagement with the Native Biodata Consortium, this project will provide a co-learning opportunity for understanding challenges specific to AI/AN communities and evaluating tools for addressing them, such as data usage agreements, contracts negotiations, transparent access agreements and some form(s) of assured return of benefit and sustainable input, control and monitoring of Tribal data.

Training future AI researchers

Concerns regarding bias and inequity in AI have been partially attributed to the lack of diversity in the AI workforce. In particular, women and racial and ethnic minority groups are under-represented.

The Bridge2AI Program has therefore defined workforce development as one of its key pillars. AI-READI hosts a year-long research internship programme (https://shileyeye.ucsd.edu/research/ai_readi) to provide immersive training for individuals interested in working at the nexus of biomedicine and AI. The programme includes a 2-week data science/programming bootcamp followed by a yearlong curriculum of didactic lectures and mentored research. Interns are mentored by one or more AI-READI investigators and are involved in various aspects of the project, including developing healthsheets to describe the AI-READI data, mapping project data into common data models, and analysing equity and ethical issues. The inaugural cohort in academic year 2023–2024 consisted of 10 interns, among whom 70% were women, 30% were Black and, overall, 50% were from under-represented backgrounds based on NIH criteria. The programme is continuing broad outreach and recruitment efforts and is disseminating its strategies to serve other similar programmes.

Discussion

The AI-READI project is ambitious in scope and may be viewed as a milestone in the field of big data and AI health research. We expect that the flagship dataset will lead to novel discoveries into the T2DM pathogenesis and salutogenesis while benefiting a broader population, given the diversity of the study participants. In addition, our blueprint has the potential to remove one of the major bottlenecks to the widespread use of AI in healthcare: the availability of data that is off-the-shelf ready for AI-based analysis.

AI-READI Consortium*

*A list of authors and their affiliations appears at the end of the paper.

Published online: 8 November 2024

References

1. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
2. Contreras, J. et al. License terms for reusing the AI-READI dataset. *Zenodo* <https://doi.org/10.5281/zenodo.10642459> (2024).
3. Hackman, J. R. & Katz, N. *Handb. Soc. Psychol.* **2**, 1208–1251 (2010).
4. Lemieux-Charles, L. & McGuire, W. L. *Med. Care Res. Rev.* **63**, 263–300 (2006).
5. Abramoff, M. D. et al. *NPJ Digit. Med.* **6**, 170 (2023).
6. Gebru, T. et al. *Commun. ACM* **64**, 86–92 (2021).
7. Rostamzadeh, N. et al. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* 1943–1961 (Association for Computing Machinery, 2022).
8. Patel, B., Soundarajan, S., Ménager, H. & Hu, Z. *Sci. Data* **10**, 557 (2023).

Acknowledgements

This work was supported by the US National Institutes of Health (NIH) through grants OT2OD032644 and P30 DK035816. We thank the Microsoft AI for Good Lab for supporting the

Comment

cloud services needed for the project. We thank Topcon Corporation (Tokyo, Japan), Optomed (Oulu, Finland), iCare World (Raleigh, NC) and Carl Zeiss (Oberkochen, Germany) for loaning their devices for research purposes at no cost. We thank Heidelberg Engineering (Heidelberg, Germany), Dexcom (San Diego, CA) and Garmin (Olathe, KS) for research discounts on study devices. We also thank the study participants and the AI-READI Advisory Council.

Author contributions

The Writing Committee members created the first draft, which was reviewed, edited and approved by all the authors.

Competing interests

S.B.: funding — NIH, University of California Office of the President, Research to Prevent Blindness; consultant — Topcon; equipment — Optomed. V.R.d.S.: funding — NSF, UCSD Social Sciences, Sanford Institute for Empathy and Compassion (Center for Empathy and Technology), Intel, Mathworks, UCSD instructional improvement grant, equipment funding from Adobe and NVIDIA, Kavli Institute for Brain and Mind, IBM, past funding from Sony; member — Cognitive Science Society Governing Board. K.F.: member — Institutional Review Board for the All of Us Research Program, Digital Ethics Advisory Panel for Merck KGaA (Merck Germany). C.S.L.: funding — NIH, Alzheimer's Disease Drug Discovery Foundation,

Gates Ventures, Research to Prevent Blindness. T.Y.A.L.: funding — Research to Prevent Blindness, Dr. H. James and Carole Free Career Development Award. B.P.: funding — NIH. L.M.Z.: funding — NEI, NIH, The Glaucoma Foundation, Heidelberg Engineering, DRCR Retina Network/JAEB Center for Health Research, The Krupp Foundation; receipt of equipment, materials, software — Optomed, iCare, Topcon, Heidelberg Engineering, Carl Zeiss Meditec, Optovue/Visionix; consultant — Abbvie, Topcon Medical Systems; co-founder, inventor, board member, equity holder — AlSight Health Inc. S.H.: funding — NIH, NIH/NARCH, RWJF, UCSD Herbert Wertheim School of Public Health. H.I.: funding — NIH; founder, stock holder — Gobiquity, Inc. A.Y.L.: funding — Santen, Topcon, Carl Zeiss Meditec, Regeneron, Amazon, Meta, Research to Prevent Blindness; personal fees — Genentech, Sanofi, US FDA, Johnson and Johnson, Boehringer Ingelheim, Gyroscope; non-financial support — iCareWorld, Optomed, Heidelberg, Microsoft. S.M.: Funding — NIH, Edward P. Evans Foundation, OHSU Knight Cancer Institute; receipt of in-kind contribution — Nike. C.N.: funding — NIH, NSF, PCORI. C.O.: consultant — Johnson and Johnson. L.H.: funding — NIH Grant UL1TR001442; consultant — Bristol Myers Squibb. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42255-024-01165-x>.

AI-READI Consortium

Writing Committee

Sally L. Baxter¹, Virginia R. de Sa¹, Kadija Ferryman², Prachee Jain³, Cecilia S. Lee⁴, Jennifer Li-Pook-Than³, T. Y. Alvin Liu², Julia P. Owen⁴, Bhavesh Patel⁵, Qilu Yu⁶ & Linda M. Zangwill¹

Principal Investigators

Amir Bahmani³, Sally L. Baxter¹, Christopher G. Chute², Jeffrey C. Edberg⁷, Kadija Ferryman², Samantha Hurst¹, Hiroshi Ishikawa⁸, Cecilia S. Lee⁴, Aaron Y. Lee⁴✉, T. Y. Alvin Liu², Gerald McGwin⁷, Shannon McWeeney⁸, Camille Nebeker¹, Cynthia Owsley⁷, Bhavesh Patel⁵, Sara J. Singer³ & Linda M. Zangwill¹

Research, Technical and Clinical Staff

Riddhiman Adib⁸, Mohammad Adibuzzaman⁸, Arash Alavi³, Catherine Ashley⁴, Adrienne Baer³, Erik Benton⁸, Marian Blazes⁴, Aaron Cohen⁸, Benjamin Cordier⁸, Katie Crist¹, Colleen Cuddy³, Virginia R. de Sa¹, Aydan Gasimova⁵, Nayoon Gim⁴, Stephanie Hong², Prachee Jain³, Trina Kim⁴, Jennifer Li-Pook-Than³, Wei-Chun Lin⁸, Jessica Mitchell², Caitlyn Ngadisastira⁴, Victoria Patronilo¹, Jamie Shaffer⁴, Sanjay Soundarajan⁵ & Kevin Zhao⁴

Project Managers

Caroline Drolet⁴, Abigail Lucero⁸, Dawn Matthies⁷, Julia P. Owen⁴, Hanna Pittock³, Kate Watkins³ & Brittany York¹

Interns

Charles E. Amankwa¹, Monique Bangudi¹, Nada Haboudal¹, Shahin Hallaj¹, Anna Heinke¹, Lingling Huang¹, Fritz Gerald P. Kalaw¹, Apoorva Karsolia¹, Hadi Khazaei^{8,9}, Muna Mohammed¹⁰ & Kyongmi Simpkins¹

NIH Program Scientists

Xujing Wang¹¹ & Qilu Yu⁶

¹University of California San Diego, La Jolla, CA, USA. ²Johns Hopkins University, Baltimore, MD, USA. ³Stanford University, Stanford, CA, USA.

⁴University of Washington, Seattle, WA, USA. ⁵FAIR Data Innovations Hub, California Medical Innovations Institute, San Diego, CA, USA. ⁶National Center for Complementary and Integrative Health, NIH, Bethesda, MD, USA. ⁷University of Alabama at Birmingham, Birmingham, AL, USA. ⁸Oregon Health & Science University, Portland, OR, USA. ⁹Portland State University, Portland, OR, USA. ¹⁰Meharry Medical College, Nashville, TN, USA. ¹¹National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), NIH, Bethesda, MD, USA. ✉e-mail: leeay@uw.edu