

Bridge2AI
Face-to-Face Meeting
Science Talk

Open Data Sharing in the Era of AI: Challenges and Opportunities

Wednesday, May 21, 2025

Bhavesh Patel (AI-READI)
Research Professor
FAIR Data Innovations Hub
California Medical Innovations Institute

Bridge2AI is supported by NIH U54 HG012510, U54 HG012513, U54 HG012517,
OT2 OD032720, OT2 OD032742, OT2 OD032644, OT2 OD032701

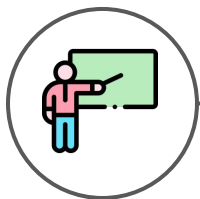


BRIDGE2AI

About This Presentation



10 min



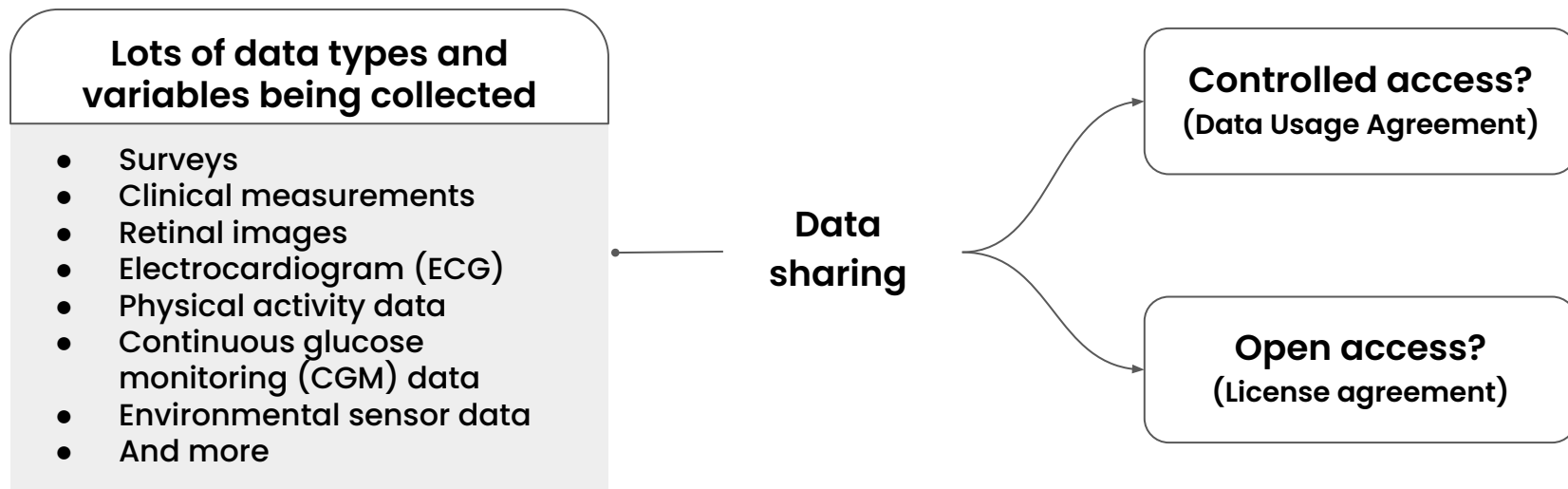
Learn about the AI-READI
data sharing journey



Challenges

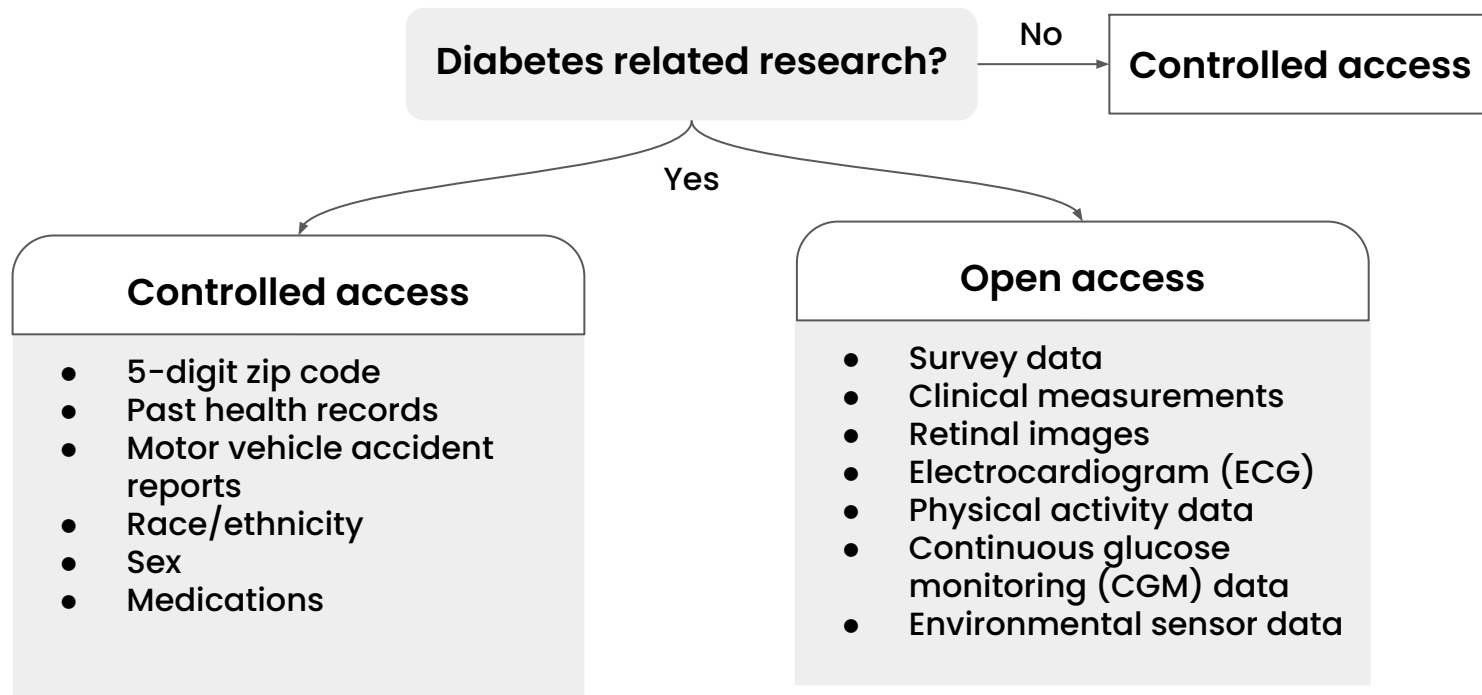
Challenges

How to share the AI-READI data?



Challenges

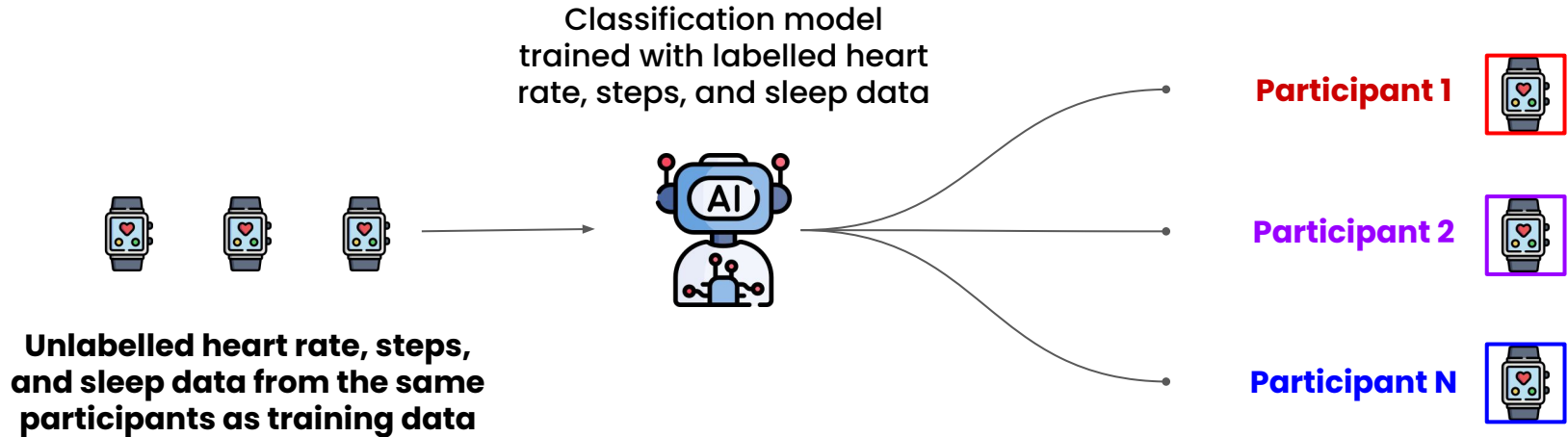
We looked at the informed consent



Challenges

Physical activity data

AI models can be trained to
identify physical activity data at an individual level



Challenges

Other data types

Literature suggests AI models can be trained to identify many data types at an individual level → **pseudo-reidentification**



Physical activity data (up to 99% accuracy)



Continuous glucose monitoring (CGM) data (86% accuracy)



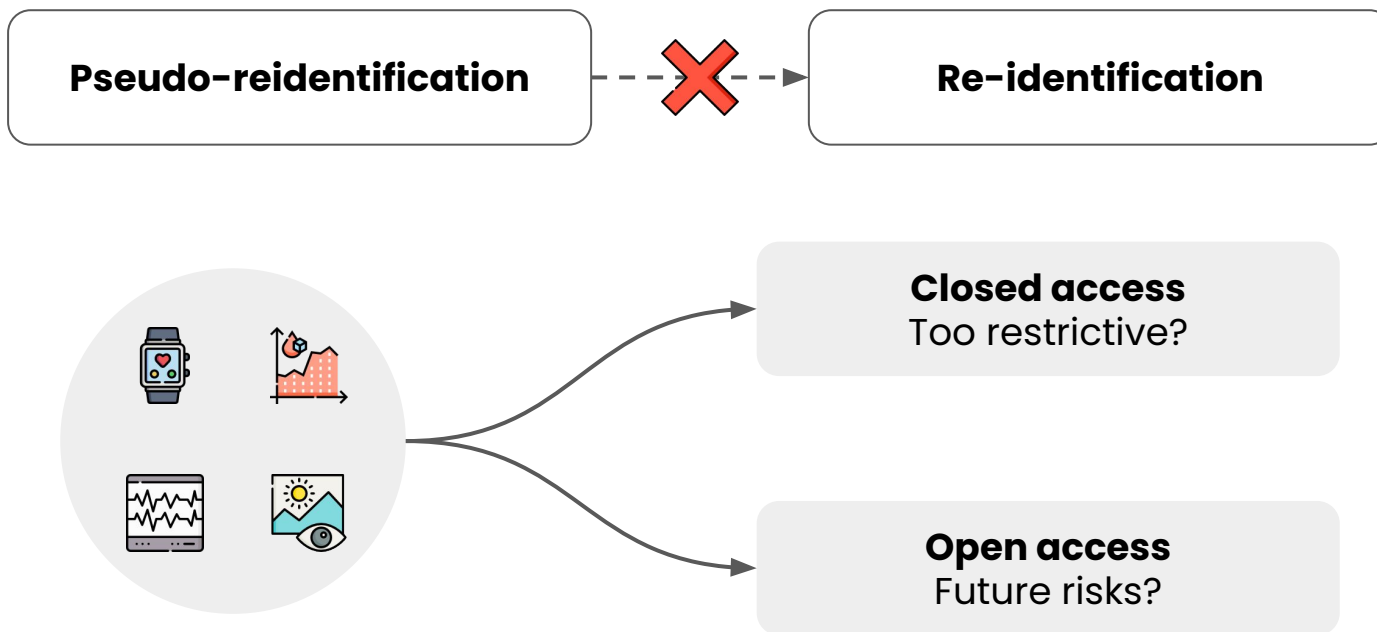
Electrocardiogram (ECG) data (up to 100% accuracy)



Retinal imaging data (up to 99% accuracy)

Challenges

How to share pseudo-reidentifiable data?



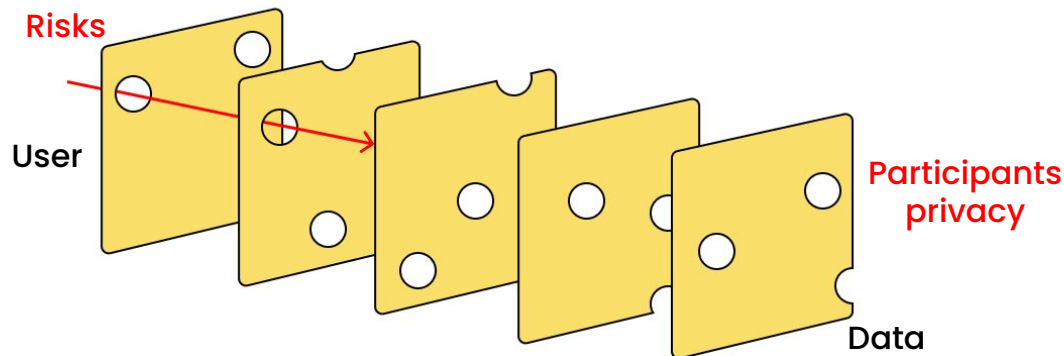


Opportunities

Opportunities

We developed a new open data access model

Swiss-cheese model of open data sharing
implemented in a new data sharing platform called FAIRhub



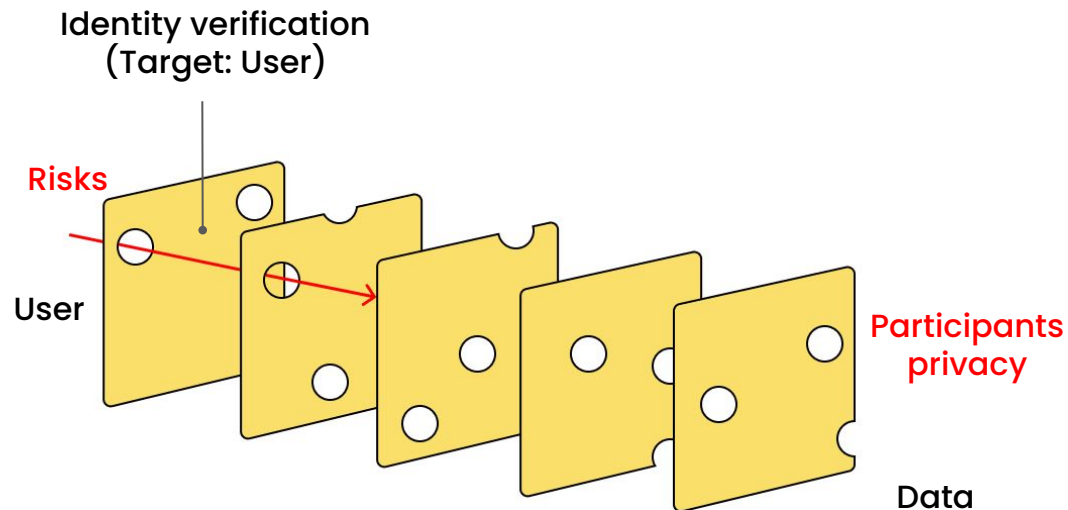
Multiple steps to access the data

Each designed to help minimize risks to participants privacy and prevent misuse

Not foolproof individually but stronger when used together

Opportunities

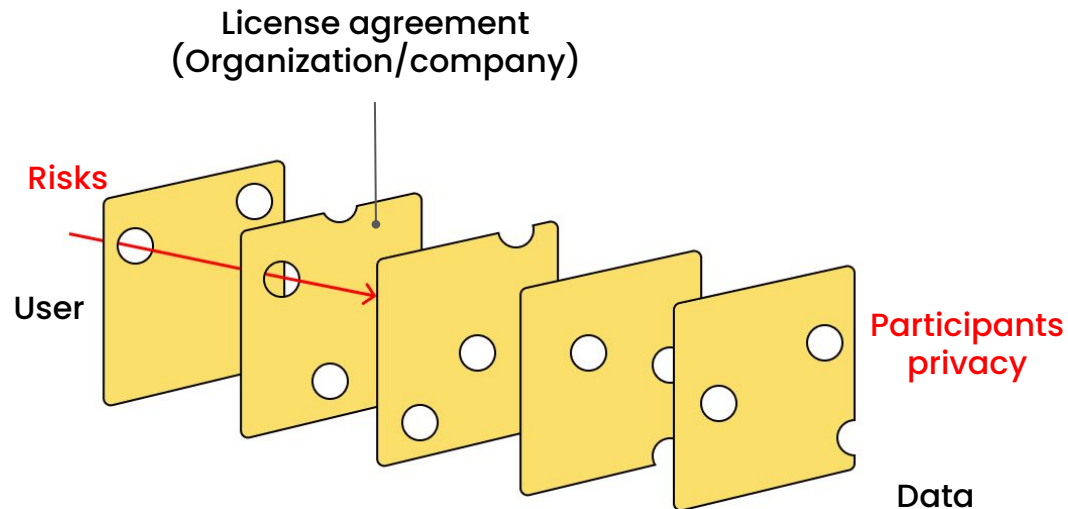
Layer 1: Identity verified data access







Anyone accessing the data needs to login through an identity verification system using their institutional email

Opportunities

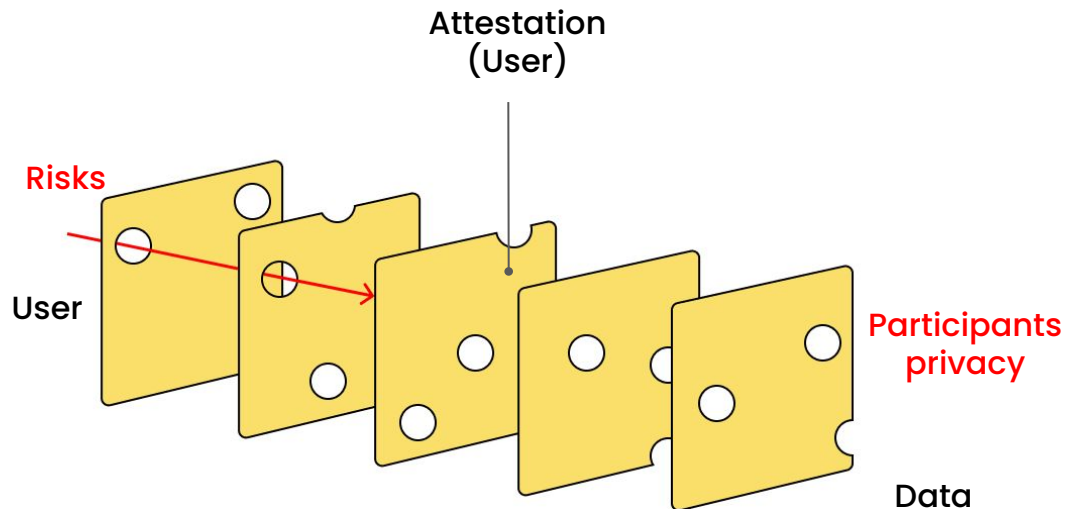
Layer 2: New license agreement



-  Re-identification of study participants or harm
-  Distribution of the data
-  Distribution of models that “memorize” the data
-  Commercial use

Opportunities

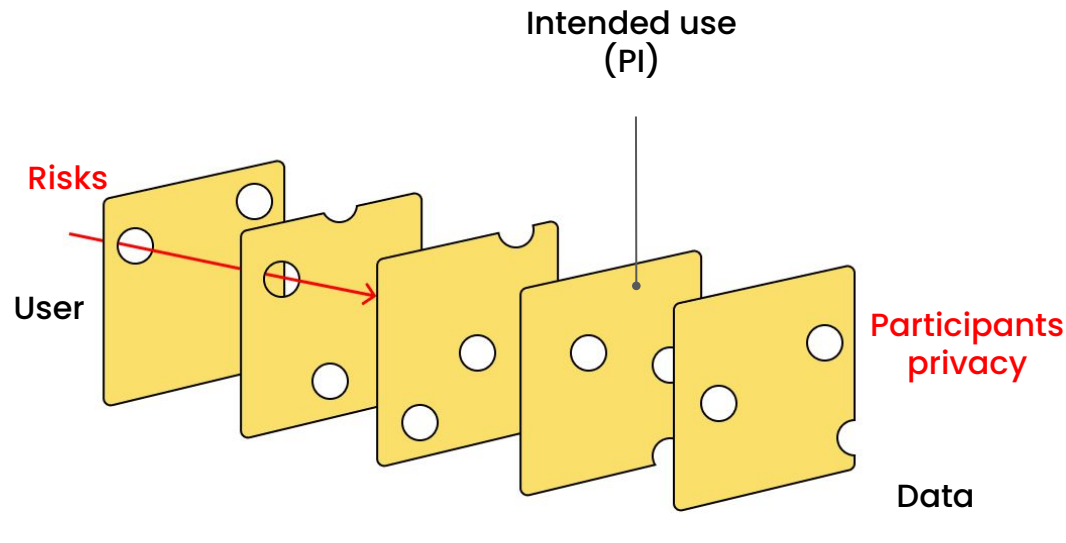
Layer 3: Attestation



Users must type out character by character and attest to the main conditions of the license

Opportunities

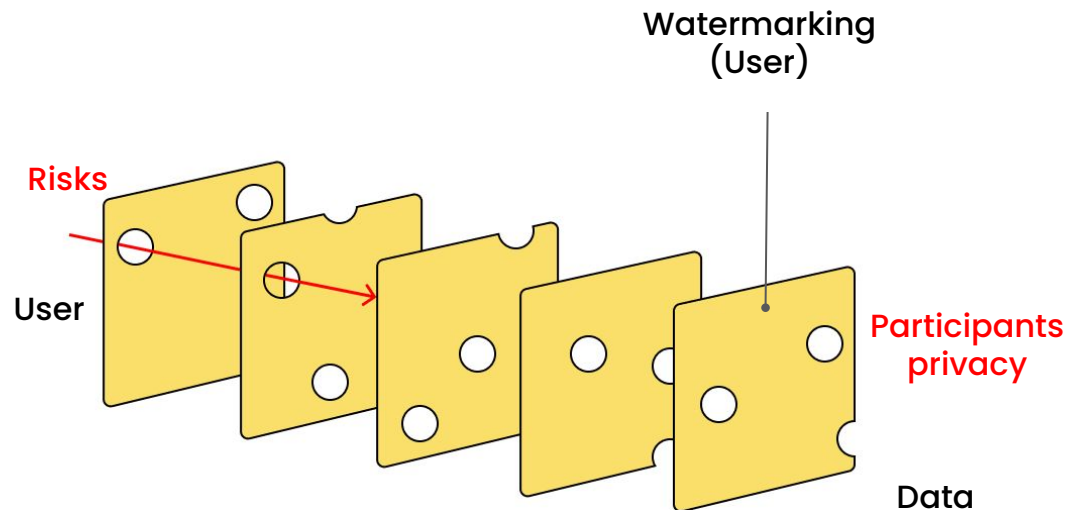
Layer 4: Intended use of the data (publicly shared)



Users must describe their intended use of the data which is shared publicly with their name

Opportunities

Layer 5: Data watermarking for each user



Each user gets a uniquely watermarked dataset tied to their identity

If the dataset appears online we can track who leaked it

Opportunities

Implemented in FAIRhub (fairhub.io/datasets/2)

The screenshot shows the FAIRhub website interface. At the top, the FAIRhub logo is on the left, and navigation links 'Find Datasets', 'Share datasets', 'About', and 'Contact' are on the right. The main heading is 'Flagship Dataset of Type 2 Diabetes from the AI-READI Project', with 'AI-READI Consortium' below it. A red rectangle highlights two buttons: 'Access this dataset' (with a download icon) and 'View the dataset documentation' (with a document icon). To the right is the AI-READI Consortium logo. Below the buttons is a horizontal menu with links: 'About' (highlighted), 'Healthsheet', 'Study Dashboard', 'Study Metadata', 'Dataset Metadata', 'Dataset Structure Preview', 'Dataset Quality Dashboard', and 'Dataset Uses'. The 'About' section contains an 'Info' box stating 'This page provides an overview of the dataset and associated study.' and an 'Overview of the study' section with a detailed paragraph about the AI-READI project. On the right side, there are three panels: 'Usage statistics' showing 11914 Views, 3 Cited by, and 510 Access approved; a file size panel showing 2.01 TB and 165,051 Files; a 'License' section with a link to 'Health Data License'; and a 'Keywords' section with tags for 'Diabetes mellitus', 'Machine Learning', 'Artificial Intelligence', and 'Electrocardiography'.

FAIRhub

Find Datasets Share datasets About Contact

Flagship Dataset of Type 2 Diabetes from the AI-READI Project

AI-READI Consortium

[Access this dataset](#) [View the dataset documentation](#)

[About](#) [Healthsheet](#) [Study Dashboard](#) [Study Metadata](#) [Dataset Metadata](#) [Dataset Structure Preview](#) [Dataset Quality Dashboard](#) [Dataset Uses](#)

Info
This page provides an overview of the dataset and associated study.

Overview of the study

The Artificial Intelligence Ready and Exploratory Atlas for Diabetes Insights (AI-READI) project seeks to create a flagship ethically-sourced dataset to enable future generations of artificial intelligence/machine learning (AI/ML) research to provide critical insights into type 2 diabetes mellitus (T2DM), including salutogenic pathways to return to health. The ability to understand and affect the course of complex, multi-organ diseases such as T2DM has been limited by a lack of well-designed, high quality, large, and inclusive multimodal datasets. The AI-READI team of investigators will aim to collect a cross-sectional dataset of 4,000 people and longitudinal data from 10% of the study cohort across the US. The study cohort will be balanced for diabetes disease stage. Data collection will be specifically designed to permit downstream pseudo-time manifold analysis, an approach used to predict disease trajectories by collecting and learning from complex, multimodal data from participants with differing disease severity (normal to insulin-dependent T2DM). The long-term objective for this project is to develop a foundational dataset in T2DM, agnostic to existing classification criteria or biases, which can be used to reconstruct a temporal atlas of T2DM development and reversal towards health (i.e., salutogenesis). Data will be optimized for downstream AI/ML research and made publicly available.

Usage statistics

11914 Views 3 Cited by 510 Access approved

All versions Current version

[More info on how stats are collected...](#)

2.01 TB 165,051 Files

License

[Health Data License](#)

Keywords

Diabetes mellitus Machine Learning Artificial Intelligence Electrocardiography



Closing remarks

Summary

Challenges

Some data types considered safe for open sharing are identifiable at an individual level (**pseudo-reidentification**)

Opportunities

We developed a new **swiss cheese model of open data sharing** to minimize risks to participants privacy and misuse

Check it out!

You can check out this data sharing model for the AI-READI dataset at **fairhub.io**

AI-READI Team



Joseph Yracheta
Native BioData



Bhavesh Patel
CALMI



Linda Zangwill
UCSD



Sally Baxter
UCSD



Nick Evans
University of
Massachusetts



Sara Singer
Stanford



Amir Bahmani
Stanford



Samm Hurst
UCSD



Jorge Contreras
University of Utah



Christopher Chute
JHU



Michelle Hribar
OHSU



Cecilia Lee
UW



Aaron Lee
UW



Cynthia Owsley
UAB



Gerald McGwin
UAB



Shannon McWeeney
OHSU

Developers

Interns

Clinical Research Coordinators



Program and Project Managers

Thank You!



bpatel@calmi2.org



fairdataihub.org

