BMJ Open

Cross-sectional design and protocol for Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI)

Cynthia Owsley ,[1] Dawn S Matthies,[1] Gerald McGwin,[1,2] Jeffrey C Edberg,[3] Sally L Baxter ,[4] Linda M Zangwill,[4] Julia P Owen,[5] Cecilia S Lee,[5] AI-READI Consortium

To cite: Owsley C, Matthies DS, McGwin G, *et al*. Cross-sectional design and protocol for Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights (AI-READI). *BMJ Open* 2025;**15**:e097449. doi:10.1136/bmjopen-2024-097449

► Prepublication history for this paper is available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2024-097449).

Received 02 December 2024
Accepted 22 January 2025

[1]Ophthalmology and Visual Sciences, The University of Alabama at Birmingham, Birmingham, Alabama, USA
[2]Epidemiology, The University of Alabama at Birmingham School of Public Health, Birmingham, Alabama, USA
[3]Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA
[4]Ophthalmology, University of California San Diego, La Jolla, California, USA
[5]Ophthalmology, University of Washington, Seattle, Washington, USA

Correspondence to
Dr Cynthia Owsley;
cynthiaowsley@uabmc.edu

ABSTRACT

Introduction Artificial Intelligence Ready and Equitable for Diabetes Insights (AI-READI) is a data collection project on type 2 diabetes mellitus (T2DM) to facilitate the widespread use of artificial intelligence and machine learning (AI/ML) approaches to study salutogenesis (transitioning from T2DM to health resilience). The fundamental rationale for promoting health resilience in T2DM stems from its high prevalence of 10.5% of the world's adult population and its contribution to many adverse health events.

Methods AI-READI is a cross-sectional study whose target enrollment is 4000 people aged 40 and older, triple-balanced by self-reported race/ethnicity (Asian, black, Hispanic, white), T2DM (no diabetes, pre-diabetes and lifestyle-controlled diabetes, diabetes treated with oral medications or non-insulin injections and insulin-controlled diabetes) and biological sex (male, female) (Clinicaltrials.org approval number STUDY00016228). Data are collected in a multivariable protocol containing over 10 domains, including vitals, retinal imaging, electrocardiogram, cognitive function, continuous glucose monitoring, physical activity, home air quality, blood and urine collection for laboratory testing and psychosocial variables including social determinants of health. There are three study sites: Birmingham, Alabama; San Diego, California; and Seattle, Washington.

Ethics and dissemination AI-READI aims to establish standards, best practices and guidelines for collection, preparation and sharing of the data for the purposes of AI/ML, including guidance from bioethicists. Following Findable, Accessible, Interoperable, Reusable principles, AI-READI can be viewed as a model for future efforts to develop other medical/health data sets targeted for AI/ML. AI-READI opens the door for novel insights in understanding T2DM salutogenesis. The AI-READI Consortium are disseminating the principles and processes of designing and implementing the AI-READI data set through publications. Those who download and use AI-READI data are encouraged to publish their results in the scientific literature.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ The targeted sample size of 4000 persons is the largest publicly accessible data set currently available containing many multidomain variables relevant to type 2 diabetes mellitus (T2DM).
⇒ The sample is designed to be approximately balanced with respect to sex and race/ethnicities (Asians, blacks, Hispanics and whites), a demographic improvement over many previous epidemiological studies and clinical trials on T2DM.
⇒ The study population is designed to be balanced in T2DM severity groups of no diabetes, pre-diabetes, non-insulin controlled and insulin controlled, to facilitate artificial intelligence/machine learning model developments.
⇒ The study design process enacted ethical and equitable data collection and management practices as well as data sharing with adherence to the Findable, Accessible, Interoperable, Reusable principles.
⇒ A limitation of our biorepository is that there are a finite number of samples to share with scientists interested in using them in research. Procedures for reviewing and prioritising written requests will be developed before the biorepository is complete.

INTRODUCTION

Artificial Intelligence Ready and Equitable for Diabetes Insights (AI-READI)[1] is one of four National Institutes of Health-funded Bridge2AI (https://bridge2ai.org) projects that aim to generate flagship biomedical and behavioural data sets that are ethically sourced, trustworthy, well-defined and publicly available. AI-READI is a data generation project focused on type 2 diabetes mellitus (T2DM) to facilitate the widespread use of artificial intelligence and machine learning (AI/ML) approaches to study salutogenesis.[2] Salutogenesis in this context refers to the pathway from T2DM to health, which is the opposite of studying how healthy people proceed through pathogenesis resulting in T2DM. The data set generated by AI-READI, to be described below, is multimodal, given

the multifactorial and multisystemic nature of T2DM. The fundamental rationale for promoting health resilience in T2DM stems from its high prevalence of 10.5% of the world's adult population.[3] Persons with T2DM are at risk for many types of adverse health consequences such as stroke, kidney disease, heart disease, vision impairment, cognitive decline, peripheral neuropathy and physical inactivity.[4] Social determinants of health weigh heavily in threatening medical status in T2DM,[5] including reduced access to healthcare, decline in treatment adherence and challenges in seeking follow-up preventative care.[6] The prevalence of T2DM is higher in certain racial and ethnic populations,[7 8] and the deleterious consequences are exacerbated in these populations.[9 10]

There are certain unique features of AI-READI's design which make it ideally suited for the use of AI/ML analytic approaches, which opens the door for critical novel insights into understanding salutogenesis. First, the AI-READI team, a multidisciplinary group of investigators, including clinicians, data scientists, vision scientists, computer scientists, organisational scientists and ethicists, designed the protocol such that it is agnostic to hypotheses. Rather, the team assembled a data collection protocol composed of assessments from many health domains that are likely impacted by T2DM. Second, to address inherent limitations of many existing large data sets for AI/ML training, the cohort is designed to be balanced in race/ethnicity, T2DM severity and biological sex. Third, the data set is large, with a target sample size of 4000 persons from three geographical regions in the USA. Fourth, AI-READI places special emphasis on establishing standards, best practices and guidelines for collection, preparation and sharing of the data that can be used for future efforts to develop other data sets targeted for AI/ML. Fifth, AI-READI addresses challenges that have compromised AI implementation in clinical research by including ethical and equitable data collection and management, and adherence to Findable, Accessible, Interoperable, Reusable (FAIR) principles. These issues have been discussed previously[1] but are incorporated in administering the design and protocol described below.

METHODS

AI-READI was approved by the Institutional Review Board (IRB) of the University of Washington (approval number STUDY00016228), including reliance agreements with the IRBs of University of Alabama at Birmingham and University of California, San Diego. Written informed consent is provided by all participants.

Patient and public involvement

A Community Advisory Board of 11 persons from the three sites including diversity in race and ethnicity as represented in the AI-READI sample contributes to the development of the protocol.

Design

AI-READI is a cross-sectional study whose target enrolment is 4000 people aged 40 and older, triple-balanced by self-reported race/ethnicity, T2DM presence and severity, and biological sex (figure 1). Building balanced data sets is critical for the development of unbiased machine learning models, so rather than targeting the demographic distribution of the US population, the study is triple-balanced by recruiting the following populations in equal proportions: four race/ethnic groups (Asian, black, white, Hispanic), four categories of T2DM (no diabetes, pre-diabetes/lifestyle-controlled diabetes, diabetes treated with oral medications or non-insulin-injectable medications, and insulin-controlled diabetes) and biological males and females. Participants are recruited across three data collection sites in different geographical locations: Birmingham, Alabama (University of Alabama at Birmingham (UAB)), San Diego, California (University of California, San Diego (UCSD)), and Seattle, Washington (University of Washington (UW)). All study groups are recruited from each site to capture diversity, but the proportion of each group will vary depending on the demographic prevalence of each group in the site's geographical area. Participants are required to speak, read and understand English. Pregnancy and type 1 diabetes are exclusionary for participation.

Source population

The study base is all patients aged ≥40 years of age who had a medical encounter within each health system site (UAB, UCSD, UW) between 2020 and 2025. Patients with T2DM and pre-diabetes are identified by screening electronic health records for ICD-10 diagnosis codes R73.09 and E11.X, respectively. Patients without diabetes will not have encounters with these ICD codes.

Recruitment

Enrolment began on 18 July 2023 and will continue until 30 November 2026. Participants are recruited in waves to facilitate the efficient sampling of the study base. The composition and size of each wave are influenced by the observed participation characteristics of race/ethnicity, gender, severity of T2DM and site, according to health records. As recruitment progresses, the composition of participants is monitored in terms of diversity and inclusion and will be adjusted by under- and oversampling groups as needed. Recruitment in waves also allows sufficient time for coordinators to respond to persons who are interested and follow-up on mailings with expediency. For each recruitment wave, a contact pool is identified by screening electronic health records at each site. Individuals in each pool are mailed a hardcopy invitation letter and sent an invitation email, both personalised to direct them into our online REDCap recruitment interface using links, access codes and QR codes. Once in the REDCap interface, individuals can read an overview of the research programme, expectations for participation and answers to Frequently Asked Questions. They are also given the opportunity to download the informed consent document, request a call back from study staff, complete a screening survey for qualification and enroll by signing

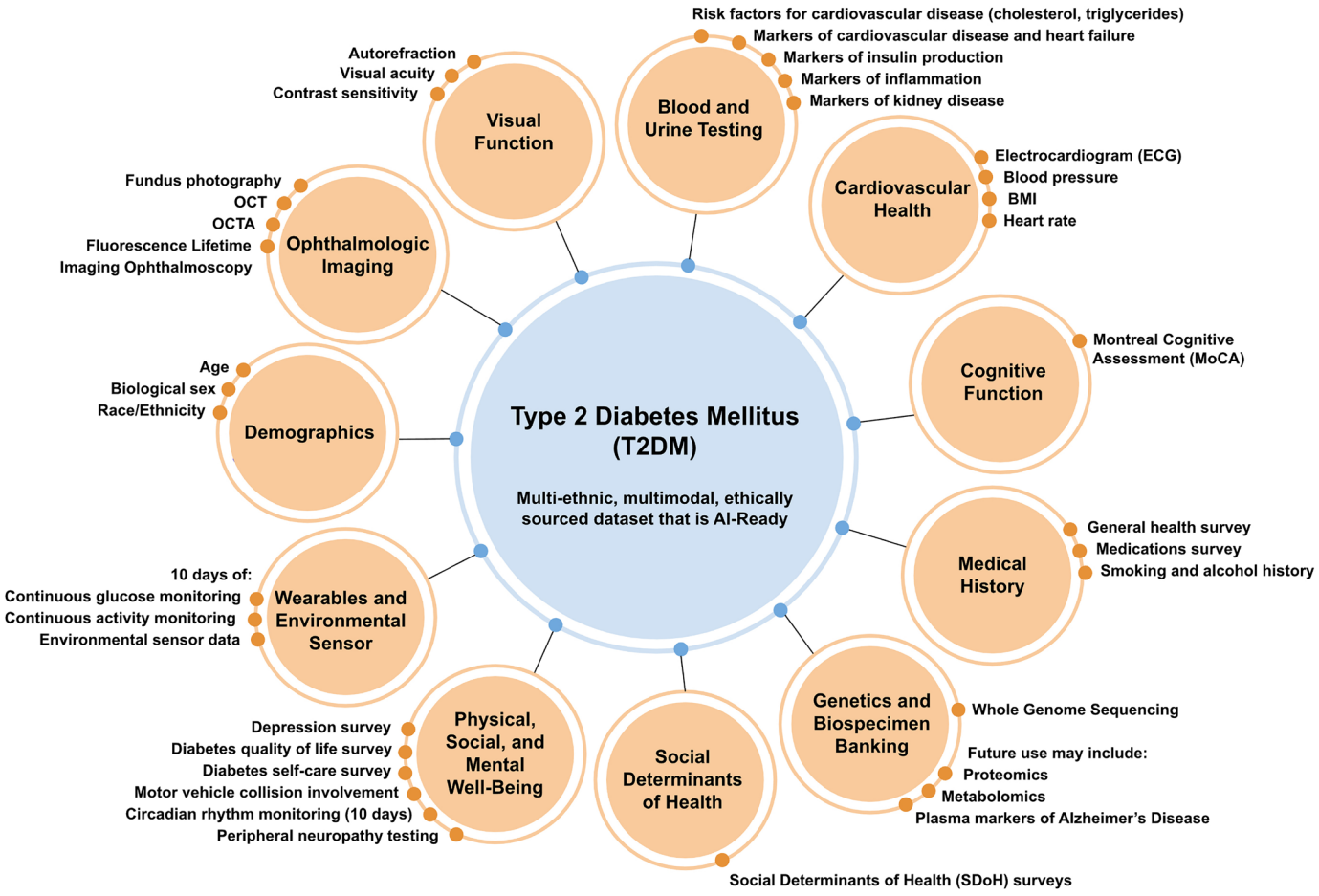

Figure 1 This project is generating an accessible, shared data set that includes a diverse set of health and behavioural domains and is harmonised across all variables to be artificial intelligence and machine learning ready. Descriptions of variables are available in tables 1–4. BMI, body mass index; OCTA, optical coherence tomography angiography.

our electronic consent. Once enrolled, they have access to study questionnaires (see below). Those who do not have access to the internet may call research staff for alternative methods of enrolment.

Protocol

The protocol is performed as a single visit lasting between 2.5 and 4 hours, depending on the participant. Participants are volunteers; therefore, there is selection bias known as volunteer bias which may limit the generalisability of the results among those who are not represented in the study population. For most participants, informed consent is performed remotely by computer, tablet or smart phone through a link in the hardcopy invitation letter or the invitation email which directed them into our REDCap database. Once enrolled, participants are presented with questionnaires for online completion. If participants did not select the option to consent electronically, it is performed in-person at the start of the visit, followed by the completion of all questionnaires. All research personnel at the three data sites who recruited, enrolled and collected data underwent training by the data manager at UAB and a detailed Manual of Procedures (MOP) for reference. Successful completion of a certification process is required for all coordinators, which involved mandatory institutional compliance training for human subject research. In addition, prior to enrolling and testing participants, all coordinators are required to successfully administer all protocol elements to at least three volunteer practice subjects, meeting the standards outlined in the MOP. Data managers at each site oversee the process. Preliminary pilot enrolment occurred between 18 July 2023 and 30 November 2023 in order to secure sufficient familiarity for all coordinators. The formal data collection process began on 1 December 2023. Preliminary data from the pilot enrollment period were included in the first released data set released in May 2024. A subsequent version of the data set that includes all data collected in the first year of the study (up to 31 July 2024) is planned for release in November 2024. Information on how data can be accessed is available at https://fairhub.io/

Figure 1 presents information on all the data domains collected in AI-READI. Questionnaires addressed several content domains as listed in table 1. Height, weight and waist and hip circumference are measured, followed by calculation of the waist-hip ratio and body mass index. Systolic blood pressure, diastolic blood pressure and heart rate were measured twice separated by 2 min with

Table 1 Questionnaires

Questionnaire	Domain addressed
Screening questionnaire	Eligibility (no type 1 diabetes mellitus or pregnancy), diabetes health status, diabetes treatments (lifestyle, medications), race and ethnicity, biological sex.
Demographic information	Date of birth, gender identification, marital status.
Center for Epidemiological Studies Depression Scale—10[24]	Screens for depressive symptoms.
Problem Areas in Diabetes5[25]	Diabetes-associated injuries, lifestyle changes and medical care.
Diabetes score (self-care)[26]	Focuses on diabetes self-management querying dietary habits, exercise, healthcare and foot care.
Dietary assessment[27]	Asks basic questions about food and drink habits over the past few months.
Ophthalmic survey	Asks about difficulties with vision and recent eye care.
Smoking, alcohol use, vaping and marijuana use*	Asks about the history of each of these behaviours.
General health	Asks about health history using this question: 'Has a doctor or other healthcare professional ever told you that you have/had?' followed by a list of chronic health conditions. Responses are yes/no; if yes, may be asked to specify a condition.
Social Determinants of Health (surveys[28]	Asks about social determinants of health and access to healthcare. Surveys on food and job insecurity, educational attainment, health insurance coverage, access to healthcare, housing and neighbourhood environment, perceptions of discrimination in medical settings and racial/ethnic discrimination (current and lifetime).
Current Medications with RxNorm codes[29]	Asks about all prescription and over-the-counter medications currently used. Includes pills, injections, creams, salves, sprays, eye drops,and dermal patches.

*Marijuana use not assessed for participants living in Alabama because it is an illegal substance.

an automatic oscillometric, medically approved device. A 12-lead ECG is performed (Philips Pagewriter TC30 Cardiograph, Amsterdam, The Netherlands) while the participant sits in a reclining chair or lies supine; the position is recorded (0°, 30°, 60° or 90°) relative to the supine position. Peripheral neuropathy is performed using the monofilament test[11] assessing touch perception on both feet with shoes and socks removed. During testing, the participants' eyes are closed while responding yes/no whether they feel the 10 g filament in three locations on each foot (10 times per location). A general cognitive screener is carried out using the Montreal Cognitive Assessment[12] (MoCA), administered electronically on an iPad using the MoCA Duo Application (MoCA Cognition, Quebec, Canada). The total possible score is 30, with higher numbers representing better performance.

Visual acuity and contrast sensitivity are assessed under both photopic (daylight) conditions and mesopic (dim light) conditions. The right and left eyes are tested separately. Photopic letter visual acuity is measured with the Electronic Visual Acuity tester[13] (M&S Technology, Niles, Illinois) using the instrument's routine protocol. Mesopic acuity is measured by the participant viewing the display through a 2.0 neutral density (ND) filter which reduces the light level of the test to a mesopic level.[14] Visual acuity is measured while the participant views the display under best-corrected conditions, expressed as the logarithm of the minimum angle of resolution. Lower numbers are better resolution. Photopic letter contrast sensitivity is measured with the Mars chart (Mars Perceptrix, Chappaqua, New York)[15] using the routine procedure. Mesopic contrast sensitivity is assessed by viewing through the 2.0 ND filter. Contrast sensitivity is expressed as log sensitivity, with higher numbers meaning better sensitivity. Autorefraction data for each eye are obtained (spherical and cylindrical components expressed as diopter with cylindrical axis) using the autorefractor's standard protocol (Topcon KR 800, Topcon Healthcare, Oakland, New Jersey).

Blood (non-fasting) and urine samples are collected from participants during the visit. Whole blood, blood derivatives and urine are used for clinical lab testing, summarised in table 2. Clinical Laboratory Improvement Amendments-certified laboratories local to each data collection site perform complete blood count (CBC) testing on fresh whole blood samples; all other lab tests are performed by the University of Washington Nutrition and Obesity Research Center (NORC) testing facility using stored, frozen samples. Additionally, blood derivatives are sent to the biorepository at UAB's Center for Clinical and Translational Science (CCTS) and will be available to researchers for future ancillary studies (see table 3 for a summary and detailed discussion of the biorepository).

The retinal imaging protocol is designed to capture imaging from multiple devices. Both eyes are imaged on each participant. Table 4 details the scans and data formats used for each retinal imaging device. The imaging devices were chosen due to their different, yet partially overlapping, modalities. The specific scan types were chosen to provide widespread coverage of the retina with the overarching goal of detecting pixel-level differences in pathology. All retinal images are collected under dilated conditions, except for the Optomed (see table 4).

Table 2 Clinical laboratory tests and biorepository specimens

Test	Units	Reference range*	Rationale for inclusion
EDTA plasma tests			
N-terminal pro-B-type natriuretic peptide	pg/mL	Varies by age	Severity and outcome predictor of heart failure
Troponin-T	ng/L	Female<11; male<16	Marker of myocardial injury
C peptide	ng/mL	1.1–4.4	Indicator of insulin production
Insulin	ng/mL	0.0–24.9	Marker for diabetes
Serum tests			
C reactive protein, high sensitivity	mg/L	0.0–10.0	Inflammation marker; risk factor for T2DM
Total cholesterol	mg/dL	<200	Cardiovascular disease risk factor
Triglycerides	mg/dL	<150	Cardiovascular disease risk factor
Hign-density lipoprotein cholesterol	mg/dL	>39	Cardiovascular disease risk factor
Low-density lipoprotein cholesterol (calculated)	mg/dL	<130	Cardiovascular disease risk factor
Glucose	mg/dL	62–125	Marker for diabetes
Blood urea nitrogen	mg/dL	8.0–21.0	Marker for kidney and liver function
Creatinine	mg/dL	Female: 0.38–1.02; male: 0.51–1.18	Marker for kidney function
Blood urea nitrogen/creatinine ratio			Indicator of kidney function
Sodium	mEq/L	135–145	Marker of metabolic health
Potassium	mEq/L	3.6–5.2	Marker of metabolic health
Chloride	mEq/L	98–108	Marker of metabolic health
Carbon dioxide, total	mEq/L	22–32	Marker of metabolic health
Calcium	mg/dL	8.9–10.2	Marker of metabolic health
Protein, total	g/dL	6.0–8.2	Useful tool for overall health status
Albumin	g/dL	3.5–5.2	Useful tool for overall health status
Globulin, total (calculated)	g/dL		Useful tool for overall health status
A/G ratio (calculated)			Useful tool for overall health status
Bilirubin, total	mg/dL	0.2–1.3	Marker of liver health
Alkaline phosphatase	IU/L	Varies by age	Marker of liver health
Aspartate aminotransferase	IU/L	9–38	Marker of liver health
Alanine aminotransferase	IU/L	Female age 7–33 and male age 0–49: 10–64; male age 50+: 10–48	Marker of liver health
Whole blood tests			
HbA1c	%	4.0–6.0	Marker for diabetes
White blood cell	×10E3/µL	3.4–10.8	Useful tool for overall health status
Red blood cell	×10E6/µL	3.77–5.80	Useful tool for overall health status
Haemoglobin	g/dL	11.1–17.7	Useful tool for overall health status
Haematocrit	%	34.0–51.0	Useful tool for overall health status
MCV	fL	79–97	Useful tool for overall health status
MCH	pg	26.6–33.0	Useful tool for overall health status
MCHC	g/dL	31.5–35.7	Useful tool for overall health status
RDW	%	11.6–15.4	Useful tool for overall health status
Platelets	×10E3/µL	150–450	Useful tool for overall health status
Urine tests			

Continued

Table 2 Continued

Test	Units	Reference range*	Rationale for inclusion
Urine creatinine	mg/dL	N/A	Urinary biomarker for early diabetic nephropathy
Urine albumin	mg/dL	N/A	Urinary biomarker for early diabetic nephropathy
Biorepository specimens			
Buffy coats (from EDTA anticoagulated vacutainers)	–	–	DNA isolations for future studies such as whole genome sequencing
Genomic DNA	–	–	Genomic applications
Plasma (from EDTA anticoagulated vacutainers)	–	–	Future proteomics and metabolomics studies
Serum	–	–	Future proteomics and metabolomics studies
PAXgene vacutainers	–	–	Future RNA isolations
Peripheral blood mononuclear cells	–	–	Peripheral blood mononuclear cells for future immunological studies; generation of iPSCs

*Reference range is a set of values that represent low and high ends of results that are considered normal. The above reference ranges were provided by the testing laboratories.
HbA1c, hemoglobin A1c; iPSCs, induced pluripotent stem cells; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemobglobin concentration; MCV, mean corpuscular volume; RDW, red blood cell distribution width.

All participant images are exported from imaging devices in their raw format and some images required conversion to a DICOM standard format prior to upload to an external AI-READI data storage site (discussed below).

At the end of their in-person visit, participants are sent home with three data monitoring devices as listed in table 4: (1) home environmental sensor, (2) Garmin fitness tracker and (3) Dexcom Continuous Glucose Monitor (CGM). Coordinators provide instructions on how to use these monitoring devices continuously for 10 days. Participants are instructed to return the devices to study staff by overnight mail (mailing materials and fees provided by the site) or in-person after 10 days. The environmental sensor and the Garmin fitness tracker were specifically chosen with participant privacy as a primary concern. These devices do not capture video or audio, are not equipped with GPS location monitoring and are not synced with participant-owned devices. Participants are masked to the results from the Dexcom CGM during the monitoring period. Participants return the monitoring devices with a data sheet indicating which wrist the fitness tracker was worn on, their dominant hand and where the environmental sensor was placed in the home.

The following results are returned to participants either as they leave the visit or later by HIPAA-compliant, encrypted email. Results from heart rate, blood pressure and visual acuity assessments are provided to the participant on an exam card before they leave the visit, along with instructions in lay language for interpreting the data and any follow-up recommendations. Following the 10-day at-home monitoring period, a Dexcom report providing average blood glucose levels (hourly, daily and overall) and an overall Glucose Management Indicator is provided by encrypted email. At the time of the yearly data release (see below), laboratory test results are sent with information on normative values by Health Insurance Portabillity and Accountability-compliant, encrypted email.

Coordinators alert participants to several incidental findings before they leave a visit. Participants are recommended to go for emergency care for the following reasons: individuals with systolic blood pressure readings of over 180 mm Hg or less than 100 mm Hg (with any symptoms of haemodynamic instability), diastolic blood pressure readings of greater than 120 mm Hg or less than 60 mm Hg (along with any symptoms of haemodynamic instability) or heart rate readings of greater than 100 bpm or less than 60 bpm (if not known to be usual for them and with any symptoms of haemodynamic instability). Retinal imaging technicians are trained to detect certain conditions that are potentially life or vision threatening (retinal detachment, tumour and optic disc oedema). If one of these conditions is suspected, immediate follow-up with the participant is performed with a referral to the emergency department (disc oedema) or referral to an ophthalmologist for immediate care (retinal detachment, tumour).

Data management

REDCap is used for collecting patient-reported data from questionnaires and the following clinical data from the visit: medications assessment, vitals, visual acuity, contrast sensitivity, monofilament test results and CBC test results. Data from the following devices are exported in their raw

Table 3 Biospecimen collection including processing and purpose

Specimen collection	Sample type	Location	Processing details	Purpose
EDTA vacutainers	Whole blood	Local clinical lab	None	Complete blood count analysis
	Whole blood	UW NORC*	None	Testing for HbA1c
	Plasma	UW NORC*	Processed locally at data site on the day of collection	Testing for NT-proBNP, troponin-T, C-peptide and insulin
	Plasma	Biobanking at UAB CCTS	Processed locally at data site on the day of collection	Future proteomics and metabolomics studies
	Buffy coats	Biobanking at UAB CCTS	Processed locally at data site on the day of collection	DNA extractions for future genomics
Serum separator vacutainers	Serum	Biobanking at UAB CCTS	Processed locally at data site on the day of collection	Future proteomics and metabolomics studies
	Serum	UW NORC*	Processed locally at data site on the day of collection	Testing for CRP-HS, glucose, BUN, creatinine, carbon dioxide, total protein, albumin, globulin, bilirubin, alkaline phosphatase, AST, ALT, electrolytes and lipids
Urine collection kit	Urine	UW NORC*	Processed locally at data site on the day of collection	Testing for creatinine and albumin
CPT mononuclear cell preparation tubes	Peripheral blood mononuclear cells	Biobanking at UAB CCTS	UAB—processed locally on day of collection UCSD/UW—tubes sent to UAB for next-day processing	Future immunological studies; generation of iPSCs
PAXgene RNA vacutainers	Stabilised, unisolated RNA	Biobanking at UAB CCTS	UAB—processed locally on day of collection UCSD/UW—tubes sent to UAB for next-day processing	Future gene expression studies

*Samples analysed by the UW NORC laboratory (serum, plasma, whole blood and urine) are first stored locally at UAB and UCSD at −70°C, then batch-shipped to UW on dry ice approximately once per quarter.
ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; CRP-HS, C reactive protein, high sensitivity; Electrolytes, sodium, potassium, chloride, calcium; Lipid analysis, total cholesterol, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, triglycerides; NT-proBNP, N-terminal pro-B-type natriuretic peptide; UAB, University of Alabama at Birmingham; UAB CCTS, University of Alabama at Birmingham Centre for Clinical and Translational Science; UCSD, University of California, San Diego; UW, University of Washington; UW NORC, University of Washington Nutrition Obesity Research Center.

format to local storage: ECG (.xml), MoCA Duo application (total score, section subscores, Memory Index Score and task completion times; .csv), retinal imaging (table 4) and at-home monitoring devices (environmental sensors, CGMs, fitness trackers; Table 3). Blood and urine testing by the NORC lab is provided in .csv format. All data are mapped to applicable data standard formats, such as the Observational Medical Outcomes Partnership Common Data Model for clinical data and the DICOM format for retinal imaging. All data are uploaded at regular intervals to the AI-READI-specific data management platform called FAIRhub (https://fairhub.io/), a Microsoft Azure cloud-based platform developed for this project. For some devices, the data are transformed from its proprietary state into a standard model format prior to upload to FAIRhub (tables 3 and 4). The Data Manager at each data site is responsible for the quality control of all data before uploading to FAIRhub. All data are stored and shared using FAIRhub as 'AI-ready' which enables immediate AI/ML research without reformatting and preprocessing. More details about the AI-readiness of the data set are provided in the data set documentation (https://docs.aireadi.org/).

Data availability

A more detailed explanation of FAIRhub was provided previously[1] and at https://aireadi.org. Deidentified data are released approximately yearly in two data sets: a controlled access set and publicly accessible set with fewer requirements for access. The controlled access data set contains all data available in the publicly accessible set, plus more sensitive data that are only available for use by scientists whose institutions have completed legal and privacy use agreements with the AI-READI research programme. The publicly accessible data set from pilot data collection was made available in May 2024. All data collected through 31 July 2024 were made available in November 2024.[16] Information regarding data set

Table 4 Retinal imaging and at-home monitoring devices

Imaging device	Scans collected	Data format	Manufacturer	Number of images per participant
Aurora IQ	Fundus: macula and disc centred (CFP) colour fundus photo (undilated)	DICOM	Optomed (Oulu, Finland)	4
EIDON widefield truecolor confocal fundus system	Ultra-wide Field Central CFP Ultra-wide Field Nasal CFP Ultra-wide Field Temporal CFP Ultra-wide Field Central IR Ultra-ide Field Central FAF Mosaic of Ultra-wide Field images	DICOM	iCare USA, Inc (Raleigh, North Carolina)	12
Spectralis HRA optical coherence tomography (OCT) and optical coherence tomography angiography (OCTA)	Optic Nerve Head-Radial Circle-High Resolution (OCT) Posterior Pole Macula-HR-61 lines (OCT) Macula-20×20-HS-512 lines (OCTA)	DICOM	Heidelberg Engineering GmbH (Heidelberg, Germany)	6
Maestro2 3D OCT-1	3D Wide (H) 12×9–512×128 (OCT) 3D Macula 6×6–512×128 (OCT) Macula 6×6–360×360 (OCTA)	Exported in .fda file format; converted to DICOM for the data set	Topcon Healthcare (Itabashi-ku, Tokyo, Japan)	6
Triton DRI OCT	3D(H)+Rad 12×9–512×256 (OCT) Macula 6×6–320×320 (OCTA) Macula 12×12–512×512 (OCTA)	Exported in .fda file format; converted to DICOM for the data set	Topcon Healthcare (Itabashi-ku, Tokyo, Japan)	6
Cirrus 5000	Disc Cube, –200×200 (OCT) Macula Cube, –512×128 (OCT) Macula, –6×6 (OCTA) Disc, –6×6 (OCTA)	DICOM	Carl Zeiss Meditec AG (Jena, Germany)	8
Fluorescence Lifetime Imaging Ophthalmoscopy	Macula-centred	Exported in .sdt file format; converted to DICOM for the data set	Heidelberg Engineering GmbH (Heidelberg, Germany)	2
Monitoring device	**Data variables collected**	**Data format**	**Manufacturer**	
Dexcom G6 Continuous Glucose Monitor	Blood glucose measurements (mg/dL) every 5 min	.csv	Dexcom, Inc. (San Diego, California)	
Garmin VivoSmart 5 Physical Activity Monitor	Number of steps Heart rate Sleep duration (circadian and diurnal rhythm) Oxygen saturation	Exported in .FIT file format; converted to mHealth standard for the data set	Garmin Ltd (Schaffhausen, Switzerland)	
Environmental Sensor	Ambient temperature Relative humidity Nitrogen oxides (NO and NO_2) Volatile organic compounds Particulate matter (PM1.0, PM2.5, PM4, PM10) Multi-spectral light intensity measurements (11)	.csv	Custom designed, Karalis Johnson Retina Center, University of Washington (Seattle, Washington)[30]	

availability and access may be found at https://aireadi.org/goals/data-sharing. The requirements for controlled access data sets are being currently developed by the Data Access Committee.

Biospecimen processing and repository

Blood (53 mL) is collected for clinical lab assays and biobanking purposes, and a urine specimen is also collected during the study encounter (table 2). The collection materials, sample type, specimen handling and location details are summarised in table 3. Local processing for plasma, serum and buffy coats is performed using standardised operating procedures, ensuring consistent handling of these biospecimens for subsequent biobanking and clinical lab analyses. One EDTA vacutainer is dedicated for delivery to a local clinical testing lab for CBC analysis. Whole blood, plasma, serum and urine collected for central clinical lab assays are batch-shipped to the UW NORC on a regular basis

(table 3). Likewise, plasma, serum and buffy coats collected at UW and UCSD are batch-shipped to the UAB CCTS for integration into a central study-wide biobank. Genomic DNA extractions from stored buffy coats are performed at UAB CCTS. Because of the complexity in isolating and cryopreserving peripheral blood mononuclear cells (PBMCs), centralised processing of the CPT tubes occurs at the UAB CCTS. While this strategy does introduce an additional time variable for isolation and cryopreservation of PBMC from participants recruited at UW and UCSD, the ability to pre-centrifuge these tubes locally followed by overnight delivery to UAB allows for consistency in initial vacutainer handling. Finally, the PAXgene RNA vacutainers are stable for up to 72 hours after collection, allowing for overnight shipment of vacutainers from UW and UCSD to the UAB CCTS. Biobanked samples will eventually be available to scientists according to procedures and policies that are in development.

AI-READI has several strengths which make it highly suitable for AI/ML studies on T2DM. The targeted sample size of 4000 persons is the largest publicly accessible data set currently available containing many multidomain variables relevant to T2DM. Furthermore, after 1 year of enrolment, the sites are currently on schedule for reaching a final sample of 4000 persons within 3 years of data collection. The sample will be approximately balanced with respect to Asians, blacks, Hispanics and whites, a demographic improvement over many previous epidemiological and clinical trials on T2DM. A meta-epidemiological review of T2DM studies published from 2000 to 2020 indicated that samples lacked racial and ethnic diversity,[17] thus making AI-READI an important step forward in addressing these demographic inequities in research. The study design process enacted ethical and equitable data collection and management practices, as well as data sharing with adherence to the FAIR principles.

Challenges will also be encountered. It is a well-documented observation in health research that recruiting black men into studies is more difficult than recruiting other populations,[18] yet this population represents a disproportionate percentage of burden in T2DM.[19] Factors such as medical mistrust and fear of safety stem from historical events such as the Tuskegee syphilis study.[20] Researchers have identified strategies to encourage interest among black men to participate in research such as tailoring printed materials and using a personalised, participatory approach to recruitment.[21] These trends for under-recruitment and for disease burden have also been reported for Hispanic[22] and Asian persons.[23] If sample balancing on race/ethnicity emerges as a challenge in AI-READI, we will implement new recruitment strategies to overcome the imbalance. Requiring that eligible participants speak, read and understand English may have created challenges in recruiting Hispanic and/or Asian persons. A limitation of our biorepository is that there are a finite number of samples to share with scientists interested in using them in research. Procedures for reviewing and prioritising written requests will be developed before the biorepository is complete.

In summary, AI-READI is an NIH Common Fund Bridge2AI-supported data collection effort to generate a 4000-person data set with T2DM that is suitable for AI/ML. It is triple-balanced with respect to race/ethnicity, biological sex and T2DM severity (including those who do not have T2DM). It is multimodal with over 10 variable domains that can be associated with adverse conditions in T2DM, yet it is hypothesis agnostic. AI-READI emphasises making data 'AI-ready' and establishing standards, best practices and guidelines for collection, preparation and sharing of the data. This approach can be viewed as an example for future efforts to develop other health data sets targeted for AI/ML. AI-READI opens the door for novel insights into understanding T2DM salutogenesis.

ETHICS AND DISSEMINATION

AI-READI aims to establish standards, best practices and guidelines for collection, preparation and sharing of the data for the purposes of AI/ML, including guidance from bioethicists. Following FAIR principles, AI-READI can be viewed as a model for future efforts to develop other medical/health data sets targeted for AI/ML. AI-READI opens the door for novel insights into understanding T2DM salutogenesis. The AI-READI Consortium are disseminating the principles and processes of designing and implementing the AI-READI data set through publications. Those who download and use AI-READI data are encouraged to publish their results in the scientific literature.

X Cynthia Owsley @cynthiaowsley

Collaborators AI-READI Consortium. Amir Bahmani, Sally Baxter, Edward Boyko, Christopher Chute, Aaron Cohen, Jorge Contreras, Garrison Cottrell, Virginia de Sa, Jeffrey Edberg, Nicole Ehrhardt, Nicholas Evans, Irl Hirsch, Michelle Hribar, Samantha Hurst, Aaron Lee, Cecilia Lee, T.Y. Alvin Liu, Bonnie Maldenado, Gerald McGwin Jr., Shannon McWeeney, Cynthia Owsley, Bhavesh Patel, Sara Singer, Michael Snyder, Bradley Voytek, Joseph Yracheta, Linda Zangwill.

Contributors CO is the guarantor. All authors are submitting authors. They meet the four criteria of the IJMJE Recommendations 2019 as follows: substantial contributions to the conception or design of the work; drafting the work or revising it critically for important intellectual content; final approval of the version to be published; and agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The AI-READI Consortium are collaborators (group authorship). CO is the corresponding author.

Funding This research is supported by National Institutes of Health grants OT2OD032644, P30DK035816, and UL1TR003096 and Research to Prevent Blindness.

Competing interests CO: Johnson & Johnson Vision, Sanofi (consultant). SLB: Topcon (consultant, travel). LZ: AbbVie, Topcon Medical Systems (consultant); AISight Health Inc. (stock or stock options). DM, GMcG, JE, JO, CL: none.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Cynthia Owsley http://orcid.org/0000-0003-3424-011X
Sally L Baxter http://orcid.org/0000-0002-5271-7690

REFERENCES

1 AI-READI Consortium. AI-READI: rethinking AI data collection, preparation and sharing in diabetes research and beyond. *Nat Metab* 2024;6:2210–2.
2 Kalra S, Baruah MP, Sahay R. Salutogenesis in Type 2 Diabetes Care: A Biopsychosocial Perspective. *Indian J Endocrinol Metab* 2018;22:169–72.
3 International Diabetes Federation. IDF diabetes atlas. 2021.
4 Bonora E, DeFronzo RA. *Diabetes complications, comorbidities and related disorders*. Springer, 2020.
5 Hill-Briggs F, Fitzpatrick SL. Overview of Social Determinants of Health in the Development of Diabetes. *Diabetes Care* 2023;46:1590–8.
6 Khunti N, Khunti N, Khunti K. Adherence to type 2 diabetes management. *Br J Diabetes* 2019;19:99–104.
7 Cheng YJ, Kanaya AM, Araneta MRG, *et al*. Prevalence of Diabetes by Race and Ethnicity in the United States, 2011-2016. *JAMA* 2019;322:2389–98.
8 Unnikrishnan R, Pradeepa R, Joshi SR, *et al*. Type 2 Diabetes: Demystifying the Global Epidemic. *Diabetes* 2017;66:1432–42.
9 Wilson C, Alam R, Latif S, *et al*. Patient access to healthcare services and optimisation of self-management for ethnic minority populations living with diabetes: a systematic review. *Health Soc Care Community* 2012;20:1–19.
10 Office of Minority Health. Diabetes and american Indians/Alaskan natives (US department of health and human services). 2021. Available: https://minorityhealth.hhs.gov/diabetes-and-american-indiansalaska-natives
11 Boulton AJM, Armstrong DG, Albert SF, *et al*. Comprehensive foot examination and risk assessment: a report of the task force of the foot care interest group of the American Diabetes Association, with endorsement by the American Association of Clinical Endocrinologists. *Diabetes Care* 2008;31:1679–85.
12 Nasreddine ZS, Phillips NA, Bédirian V, *et al*. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* 2005;53:695–9.
13 Beck RW, Moke PS, Turpin AH, *et al*. A computerized method of visual acuity testing: adaptation of the early treatment of diabetic retinopathy study testing protocol. *Am J Ophthalmol* 2003;135:194–205.
14 Owsley C, Swain TA, McGwin G Jr, *et al*. How Vision Is Impaired From Aging to Early and Intermediate Age-Related Macular Degeneration: Insights From ALSTAR2 Baseline. *Transl Vis Sci Technol* 2022;11:17.
15 Arditi A. Improving the design of the letter contrast sensitivity test. *Invest Ophthalmol Vis Sci* 2005;46:2225–9.
16 AI-READI Consortium. Flagship dataset of type 2 diabetes from the AI-READI project (1.0.0) FAIRhub. Available: https://docs.aireadi.org/
17 Ahmed R, de Souza RJ, Li V, *et al*. Twenty years of participation of racialised groups in type 2 diabetes randomised clinical trials: a meta-epidemiological review. *Diabetologia* 2024;67:443–58.
18 Randolph S, Coakley T, Shears J. Recruiting and engaging African-American men in health research. *Nurse Res* 2018;26:8–12.
19 Liburd LC, Namageyo-Funa A, Jack L Jr. Understanding 'masculinity' and the challenges of managing type-2 diabetes among African-American men. *J Natl Med Assoc* 2007;99:550–2.
20 Scharff DP, Mathews KJ, Jackson P, *et al*. More than Tuskegee: understanding mistrust about research participation. *J Health Care Poor Underserved* 2010;21:879–97.
21 Woods VD, Montgomery SB, Herring RP. Recruiting Black/African American men for research on prostate cancer prevention. *Cancer* 2004;100:1017–25.
22 O Rojo M, Jing J, Wells C, *et al*. Hispanics' Perceptions of Participation in Research Studies and Solutions for Improvement in Participation. *J Family Med Community Health* 2024;11:1–9.
23 Lim J-W, Paek M-S. Recruiting Chinese- and Korean-Americans in Cancer Survivorship Research: Challenges and Lessons Learned. *J Cancer Educ* 2016;31:108–14.
24 Andresen EM, Malmgren JA, Carter WB, *et al*. Screening for depression in well older adults: evaluation of a short form of the CES-D (Center for Epidemiologic Studies Depression Scale). *Am J Prev Med* 1994;10:77–84.
25 McGuire BE, Morrison TG, Hermanns N, *et al*. Short-form measures of diabetes-related emotional distress: the Problem Areas in Diabetes Scale (PAID)-5 and PAID-1. *Diabetologia* 2010;53:66–9.
26 Hashim MJ, Nurulain SM, Riaz M, *et al*. Diabetes Score questionnaire for lifestyle change in patients with type 2 diabetes. *Clin Diabetes* 2020;9:379–86.
27 Paxton AE, Strycker LA, Toobert DJ, *et al*. Starting the conversation performance of a brief dietary assessment and intervention tool for health professionals. *Am J Prev Med* 2011;40:67–71.
28 Hamilton CM, Strader LC, Pratt JG, *et al*. The PhenX Toolkit: get the most from your measures. *Am J Epidemiol* 2011;174:253–60.
29 Unified Medical Language System (UMLS). RxNorm. Available: https://www.nlm.nih.gov/research/umls/rxnorm/index.html [Accessed 7 Aug 2024].
30 Shaffer J, Gim N, Wei R, *et al*. Portable environmental sensor enabling studies of exposome on ocular health. *Invest Ophthalmol Vis Sci* 2024;65:6370.