All of Us Research Program
Electronic Health Record
Data Management & Sharing Plan (DMSP), CDR v9

Summary Information	
Data type	Electronic Health Records (EHR)
Data source	Healthcare Provider Organizations (HPOs), Participant mediated EHR from CareEvolution (CE), Participant Technology Systems Center (PTSC), and Center for Linkage and Acquisition of Data (CLAD)
Dates reflected in data type	CDR v9 cut-off date: January 1, 2025
Cadence for updates	Twice a year ● Data submission dates ● 2025 Quarter 1 Submission: February 28, 2025
High-level summaries of the data type (materials/links)	*All of Us* EHR Collection and Data Transformation Methods CLAD Sourced EHR Data in *All of Us* (CDR v9)
Data quality (materials/links) ● **Summary** ● **Issue Tracker** ● **Change Logs**	● Summary ● Data Quality (DQ) Feedback ● Issue tracking: o DQ feedback is shared back with EHR Operations Team after review o Inquiries from researchers are tracked in this flow chart. o Issues with the data are listed in the CDR release notes ● Change logs: Any updates to the data are listed in the CDR release notes
***All of Us* Contact**	support@researchallofus.org

SUMMARY OF REVISIONS

Version	Date	Section	Revisions Made	Rationale
1	23 October 2025	All	N/A	1st version generated
1.1	23 July 2026	All	Modified the internal version for researchers	1st public version generated

All of Us Research Program
Electronic Health Record
Data Management & Sharing Plan (DMSP), CDR v9

Table of Contents

1. Types of Data Produced....3
2. Related Tools, Software, and/or Code....4
3. Standards for Data and Metadata....4
4. Data Preservation, Access, and Associated Timelines....5
5. Access, Distribution, or Reuse Considerations....7
6. Oversight of Data Management and Sharing....8

All of Us Research Program
Electronic Health Record
Data Management & Sharing Plan (DMSP), CDR v9

---

1. Types of Data Produced

1.1. Summarize the types and estimated amount of scientific data expected to be generated.

- Data type: EHR Data in Observational Medical Outcomes Partnership (OMOP) format from HPOs, PTSC, CE and CLAD.
    - The EHR data collected by the program is submitted by healthcare provider organizations, collected through health information networks (HINs), or directly shared by participants through their participant portals. The data are transformed into the OMOP common data model (CDM) structure. Recorded data elements are coded as source concepts (e.g., ICD 9 and ICD 10) and mapped to standard concepts (e.g., SNOMED) based on international health standards. OMOP categorizes clinical data concepts according to several major health domains.
- Amount/volume of data:
    - 481,918 total participants
        - 383,355 HPOs
        - Participant Mediated:
            - 24,386 CE (Directly Donated)
            - 58,806 PTSC
        - 15,371 CLAD (HIN)
- The level of aggregation: individual
- The degree of data processing or cleaning:
    - Data Curation Process – User Support
    - Participant Privacy Protections – User Support
    - Introduction to *All of Us* EHR Collection and Data Transformation Methods
    - CDR v9 Data Dictionary and Release Notes: Data Dictionaries – User Support

- Use case examples for the data type:
    - We expect researchers to use EHR data to create phenotype.

1.2. Describe which scientific data will be preserved and shared, and provide the rationale for this decision based on ethical, legal, and technical factors.

- Pre-processing/curation steps of the raw data: Data Curation Process – User Support
- EHR Operations Team Data Quality Process: Data Quality Feedback
- Introduction to *All of Us* EHR Collection and Data Transformation Methods
- Participant Privacy Protections – User Support
- CDR v9 Data Dictionary and Release Notes: Data Dictionaries – User Support
- CLAD Sourced EHR Data in *All of Us* (CDR v9)

1.3. List the metadata, other relevant data, or associated documentation that will be made accessible to facilitate interpretation of the scientific data.

- Introduction to *All of Us* EHR Collection and Data Transformation Methods
- Frequently Asked Questions About These Data Types – User Support
- CDR v9 Data Dictionary and Release Notes: Data Dictionaries – User Support
- CLAD Sourced EHR Data in *All of Us* (CDR v9)

2. Related Tools, Software, and/or Code

2.1. Indicate whether specialized/licensed software or tools are needed or recommended to access or manipulate data generated by the implementation partner.

- No specialized/licensed software or tools are needed.

2.2. Indicate whether custom-created code or in-house software or tools will be needed or created to access or manipulate data generated by the implementation partner. If so, how will these be made available to researchers?

Not applicable to this data. All standard Researcher Workbench functions are used for this data type.

3. Standards for Data and Metadata

3.1. Describe what standards, if any, will be applied to the scientific data and associated metadata (i.e., data formats, data dictionaries, data identifiers, definitions, unique identifiers) to enable interoperability of datasets and resources. If not applicable, indicate that no consensus standards exist.

- Data and file formats: data from HPOs are in the CSV format. Additionally, there are FHIR payloads in the JSON format.
- The Data Dictionary lists all fields present within the program data model with relevant metadata, including description, data provenance, and whether the field was impacted by privacy methods.
- The EHR data collected by the program is submitted by HPOs, collected through HINs, or directly shared by participants through their participant portals. The data are transformed into the OMOP common data model structure. Recorded data elements are coded as source concepts (e.g., ICD 9 and ICD 10) and mapped to standard concepts (e.g., SNOMED) based on international health standards. OMOP categorizes clinical data concepts according to several major health domains.
- CDR v9 includes a new stream of EHR data sourced from Consolidated Clinical Document Architecture (C-CDA) documents. These records were acquired through the *All of Us* CLAD pilot, which used the eHealth Exchange health information network to retrieve participant-authorized records from Manifest MedEx, a California statewide health information exchange. This source returned C-CDA payloads that were mapped by CLAD to the OMOP Common Data Model and added by the *All of Us* Data and Research Center to the curated data repository alongside existing *All of Us* EHR data.
- Introduction to *All of Us* EHR Collection and Data Transformation Methods

3.2. Describe data formats created, used, and stored. If unusual data formats will be used, discuss the proposed solution for converting data into more accessible formats.

- EHR data in *All of Us* is organized into tables following the OMOP CDM version 5.3. Data related to visits, procedures, drugs, and conditions is stored in their respective tables, which are linked to the person table and the visit table to maintain relational integrity.
- CLAD HIE: C-CDA documents organize clinical data into XML sections such as problems, encounters, results, medications, and procedures. Each section was mapped into the OMOP Common Data Model. Mapping has two parts. Structural mapping decides which source fields populate which OMOP columns. Vocabulary

mapping decides which source codes correspond to which standard OMOP concepts
- Data Types and Organization – User Support
- Introduction to *All of Us* EHR Collection and Data Transformation Methods

4. Data Preservation, Access, and Associated Timelines

4.1. Describe where the data will be made publicly available.

Data is made available to researchers on the *All of Us* Researcher Workbench. Data will not be made available in other repositories.

4.2. Describe how the scientific data will be findable and identifiable for researchers.

- **Findability and Identifiability of EHR Data for Researchers:**
  - EHR data in *All of Us* is structured using the OMOP CDM and includes unique record identifiers (e.g., condition_occurrence_id and observation_id) and standardized concept IDs mapped to vocabularies such as SNOMED-CT, RxNorm, LOINC, and ICD-10. Researchers can locate and link data across domain tables (e.g., Condition Occurrence, Measurement, and Observation) using tools provided in the Researcher Workbench, such as Cohort Builder, Dataset Builder, and Concept Sets. Data dictionaries for the current CDR version provide detailed information on table structures, fields, and privacy protections.
- **Data Browser Accessibility:**
  - Researchers can explore aggregated and de-identified EHR data trends via the Data Browser, which provides insights into conditions, procedures, drugs, and measurements.
- **Accessing and Using Data via the Workbench:**
  - EHR data can be accessed through Researcher Workbench tools for cohort creation, dataset building, and querying standardized concepts. Researchers can use unique identifiers (e.g., person_id, visit_occurrence_id) to link data across tables, including Visit Occurrence for detailed analyses. Privacy protections and field suppression are applied to safeguard sensitive information, with updates provided in CDR data dictionaries for each release.
- Introduction to *All of Us* EHR Collection and Data Transformation Methods
- CLAD Sourced EHR Data in *All of Us* (CDR v9)

4.3. Describe what scientific data from 1.1. and 1.2 will be made available to other users and when data will become available.

- **Researchers may access the *All of Us* data through three separate tiers:**
  - **Public Tier** data includes only anonymized, aggregated data and is accessible to the public through the *All of Us* Research Hub's Data Browser.
  - **Registered Tier** data includes participant-level data with added transformations to protect participant privacy and is only available to registered researchers.
  - **Controlled Tier** data includes participant-level data with fewer transformations than those in the Registered Tier.
- Cadence for updates: Two times a year
  - 2025 EHR Data Submission Dates
    - 2025 Quarter 1 Submission was used for v9 CDR:
      - Data Cutoff Date: Wednesday, 1/1/25

- Initial Submission Date: Wednesday, 1/15/25
- Final Submission Date: Friday, 2/28/25

- Quality Checks are performed by the EHR Operations team when a data submission is made.
  - Data Quality Feedback
- During the data transformation process, a series of steps are completed to ensure data conforms to both OMOP and *All of Us* standards. Details can be referenced in the CDR data dictionary release notes. Steps include:
  - Validation checks that occur before and after data are transferred to the *All of Us* Data and Research Center (DRC) for CDR inclusion. The goal is to ensure that data conforms to OMOP naming conventions as well as data type and column structures.
  - Characterization checks that occur after data are transferred to the DRC and are completed before and after CDR inclusion. The goal is to use OMOP-supported tools like the Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (Achilles) to assess person-level data characterization and density.
  - Select tables, fields, and concepts are suppressed. The goal is to ensure that data conforms to the *All of Us* privacy protection principles and rules.
  - Select fields within the Measurement table undergo a small amount of cleaning. The goal is to help support data plausibility and utility for selecting data of high priority and research interest.
  - Introduction to *All of Us* EHR Collection and Data Transformation Methods.
- **CLAD HIE/HIN:** CDR v9 includes a new stream of EHR data sourced from Consolidated Clinical Document Architecture (C-CDA) documents. These records were acquired through the *All of Us* CLAD pilot, which used the eHealth Exchange health information network to retrieve participant-authorized records from Manifest MedEx, a California statewide health information exchange. This source returned C-CDA payloads that were mapped by CLAD to the OMOP Common Data Model and added by the *All of Us* Data and Research Center to the curated data repository alongside existing *All of Us* EHR data.
- Data was released as part of the CDR v9 in June 2026 in the Researcher Workbench.

5. Access, Distribution, or Reuse Considerations

5.1. Describe and justify any factors affecting subsequent access, distribution, or reuse of scientific data.

- Overview of *All of Us* Research Program Policies for Researchers
  - **Data User Code of Conduct (DUCC):** Researchers must read and sign the DUCC, adhere to ethical guidelines, comply with laws, respect participant privacy, and use data solely for biomedical or health research. It prohibits discriminatory research and sets restrictions on data sharing without explicit permission.
  - **Data and Statistics Dissemination Policy:** Restrictions on publishing or distributing participant counts of 1 to 20 to protect privacy, with strategies provided for obscuring such data. Exceptions are rare but can be requested.
  - **Publication and Presentation Policy:** Requires submission of peer-reviewed manuscripts to PubMed Central for immediate public access, notification of upcoming publications or presentations, and adherence to Data and Statistics

Dissemination Policy regarding participant counts.
    - **Stigmatizing Research and Ethical Conduct Policies:** Prohibits research that could lead to social detriment or stigma. Emphasizes inclusivity, ethical research practices, and adherence to the Federal Policy for the Protection of Human Subjects (The Common Rule).
    - **Data Access Framework and Egress Alert Policy:** Inclusive access framework with three data tiers (public, registered, controlled). Users must obtain a "data passport" for higher access levels. Egress alerts are triggered by security policy violations, primarily related to data downloads.
- *All of Us* Consent Process: https://allofus.nih.gov/article/all-us-consent-process
- Participant Privacy Protections – User Support

5.2. Describe how access to scientific data will be controlled (i.e., made available by a data repository only after approval).

Access to individual-level participant data is controlled via *All of Us* Data Use and Registration Agreements, which must be completed and approved before researchers gain access to the *All of Us* Researcher Workbench. Additionally, researchers must complete identity verification, responsible conduct of research training, and the Data User Code of Conduct. Additional training is required for researchers with access to the Controlled Tier of the Workbench. More information on data access can be found here.

5.3. Describe protections for privacy, rights, and confidentiality of human research participants.

- Participant Privacy Protections
- **De-identification and Data Transformation:** The *All of Us* Research Program employs rigorous methods to remove personally identifying information (PII) from participant records. Data are also subject to transformations such as date shifting and generalization to protect privacy and minimize re-identification risks.
- **Tiered Data Access:** Data are categorized into three tiers—Public, Registered, and Controlled—based on privacy risks and access requirements. Public Tier includes only anonymized, aggregated data; Registered Tier includes participant-level data with significant privacy transformations; Controlled Tier includes more detailed participant-level data with fewer transformations.
- **Suppression of Sensitive Information:** Explicit identifiers, free-text fields, and sensitive information (e.g., geolocation data, race/ethnicity subcategories, and certain diagnosis codes) are suppressed or generalized to protect participant confidentiality. For example, geographic data smaller than the US state level and specific race/ethnicity subcategories are suppressed in the Registered Tier.
- **Additional Protective Measures:** The program includes privacy experts to ensure data curation methods effectively protect participant privacy. The Controlled Tier data undergoes fewer transformations but continues to exclude direct identifiers and apply necessary suppressions to protect participant confidentiality.

- *All of Us* Data and Statistics Dissemination Policy

5.4. Describe security measures that will be in place to protect the data from unauthorized access.

- Access to individual-level participant data is controlled via *All of Us* Data Use and Registration Agreements, which must be completed and approved before researchers gain access to the *All of Us* Researcher Workbench. Additionally, researchers must complete identity verification, responsible conduct of research training, and the Data User Code of Conduct. Additional training is required for researchers with access to

the Controlled Tier of the Workbench. More information on data access can be found here.
- https://www.researchallofus.org/privacy-security-protocols/

5.5. Describe if there are factors that limit the ability to share data (e.g., proprietary nature or commercialization of the data).

All of Us Research Program Policies for Researchers

6. Oversight of Data Management and Sharing

6.1. Provide details/roles of personnel responsible for implementing the DMSP, in line with allowable cost guidance for data management and sharing (e.g., data curation, quality checks, developing supporting documentation).

- DRC Roles: Curation team, Curation project management, EHR Operations project management
- NIH Roles: review, finalize, and approve the DMSP

6.2. Provide details of data management and sharing oversight practices throughout the research project workflow. State how often compliance with the DMSP will be verified (e.g., every X months, on the first of each month).

- The DMSP will be updated when changes are made. It will also be verified annually until the next CDR release, when the DMSP is updated.