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

Summary Information	
Data type	Clinical Notes
Data source	Healthcare Provider Organizations (HPOs) and Participant mediated Electronic Health Record (EHR) from CareEvolution (CE) and Participant Technology Systems Center (PTSC)
Dates reflected in data type	CDR v9 cut-off date: January 1, 2025. Notes data was included in the Feb 28,2025 submission.
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)	● Clinical notes—narrative entries within EHRs—offer rich, unstructured data that can enhance insights beyond structured fields. To harness this information, the CLAMP (Clinical Language Annotation, Modeling, and Processing) tool is used to extract medical concepts such as conditions, procedures, medications, and lab results with high accuracy. These extracted entities are mapped to OMOP concept IDs and stored in the "Note_NLP" table as "NLP-derived" concepts, forming the clinical notes dataset within the CDR. ● Clinical Notes Data QC Report
Data quality (materials/links) ● **Summary** ● **Issue Tracker** ● **Change Logs**	● Summary: Clinical Notes Data QC Report ● Issue tracking: o Data Quality (DQ) feedback is shared back with the 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	3 March 2026	All	N/A	1st version generated
1.1	23 July 2026	All	Modified the internal version for researchers	1st public version generated

Table of Contents

All of Us Research Program
Electronic Health Record Clinical Notes
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.

- Clinical notes—narrative entries within EHRs—offer rich, unstructured data that can enhance insights beyond structured fields. To harness this information, the CLAMP (Clinical Language Annotation, Modeling, and Processing) tool is used to extract medical concepts such as conditions, procedures, medications, and lab results with high accuracy. These extracted entities are mapped to OMOP concept IDs and stored in the "Note_NLP" table as "NLP-derived" concepts, forming the clinical notes dataset within the CDR.
- Participants: ~98k
- Note_NLP: ~95M
- Aggregated to the individual level

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.

- Process outlined here: Clinical Notes Data QC Report
- During the data transformation process, a series of steps are completed to ensure data conforms to both OMOP and *All of Us* standards. Details are 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 (e.g., removed). The goal is to ensure that data conforms to the *All of Us* privacy protection principles and rules.
- Introduction to All of Us Electronic Health Record (EHR) Collection and Data Transformation Methods
- Data Curation Process – User Support

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

- Clinical Notes Data QC Report
- Data provenance including sources is identified in the Data Dictionary

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?

No custom-created code or software is needed.

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.

- **Common data model standard:** Clinical notes are represented in the OMOP Common Data Model (CDM) structure. Notes reside in an OMOP-aligned NOTE table, and natural language processing (NLP) results reside in an OMOP-aligned NOTE_NLP table. This standard table/field structure supports consistent interpretation and cross-site interoperability.
- **Standardized concepts for NLP-derived data:** Concepts extracted from free-text notes through NLP are normalized to standardized clinical concepts consistent with OMOP vocabularies. NLP-derived concepts loaded into OMOP clinical tables are labeled as "NLP-derived" to distinguish derived content from source-recorded data.
- **Identifiers and harmonization across sites:** OMOP conventions for record identifiers and required fields support merging note data from multiple HPO sites into a single unioned dataset without changing meaning across sites.
- **Metadata and documentation:** Data dictionaries and OMOP table definitions describe table purpose, field meaning, provenance (source note vs. NLP-derived), and any relevant derived-data indicators to support reuse after publication.

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.

- **Submitted data format:** HPO sites submit OMOP-formatted clinical note records as JSONL (JSON Lines) files to their designated Google Cloud Storage buckets.
- **Primary storage and processing format:** Submitted JSONL files are ingested into BigQuery and stored as structured BigQuery tables aligned to OMOP CDM. Tables from each HPO are merged into a single unioned NOTE table containing notes across all HPOs.
- **Annotation and evaluation sample format:** Sample rows exported for manual annotation and baseline evaluation are provided as plain text files within a secure FISMA VM environment.
- **NLP output format:** CLAMP produces NLP outputs in **O**MOP NOTE_NLP table format (structured to match NOTE_NLP) for loading back into BigQuery through a custom pipeline.
- **Validation for published research data:** NLP outputs are validated against a manually annotated gold standard for benchmarking/baseline evaluation. Published NLP-derived concept outputs meet a minimum performance threshold of ≥80% agreement with the gold standard for concept extraction.

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.

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

Data will not be made available in the data browser or cohort builder. The researcher needs to query the dataset and join it to 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.

- Quality checks outlined here: Clinical Notes Data QC Report
- Standard Curation Quality Checks were also performed.
- Data was released into the Researcher Workbench for researchers in June 2026.

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, which. 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.