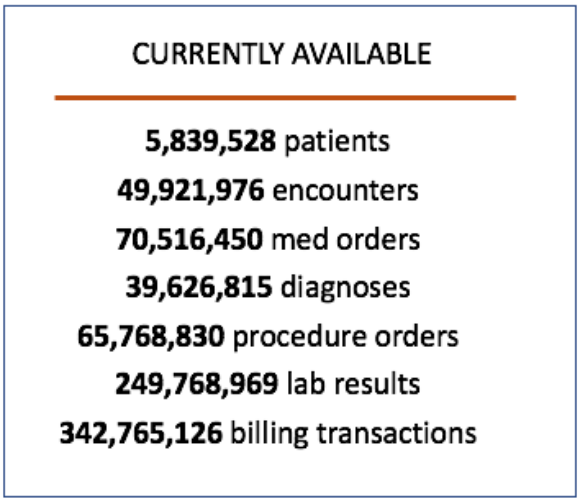
Clinical Data Lake and Data Warehouse

Cedars-Sinai Clinical Data Lake and Data Warehouse is deep, data rich, and one of the most research-ready data repositories of its kind. Containing more than a decade of medical data, it seamlessly brings together multiple internal and external data sources to provide researchers with access to approximately 50 million encounters for over 5 million patients. The associated diagnoses, labs, medications, and procedures number in the tens of millions each. The Enterprise Data Intelligence team at Cedars-Sinai works closely with researchers and clinicians to provide a wealth of knowledge about patients, their medical conditions, and their outcome. Since its launch, the data warehouse has contributed to numerus grant funded projects, clinical trials, published research articles, and commercial patent developments.



Data Types and Sources

The table below summarizes the avenues available to you to obtain the data of interest to your research

Demographic & other patient details: Basic patient information including date of birth, sex, race, ethnicity, recent height, weight and smoking history, along with as much identifying information as your IRB protocol permits.

Encounters and Admissions: Both inpatient and outpatient encounters, searchable by department, clinic name, or providers. The data includes detailed ADT (Admit-Discharge-Transfer) records for both inpatient and ambulatory visits, and inpatient treatment teams.

Billing Codes: Reimbursement and clinical codes for both diagnoses and procedures.

Lab Results: Results, both numeric and categorical / text, from diagnostic tests run on biospecimens.

Medication Orders & Administration: Drug orders and Medication Administration Records (MAR) for both in- and out-patients.

Clinical Documents: Clinical documents such as consultation notes, operative/procedures notes, and letters, including structured data captured via SmartForms.

Radiology Reports: Radiology reports, including imaging study accession numbers. Note that the associated DICOM images are also available for research in both de-identified and original (fully identified) form.

Pathology Reports: Pathology reports, including the biospecimen accession numbers.

Nursing Documentation: Flowsheet records, including point of care assessment results and vital sign readings, for both inpatient and outpatient.

Procedure Orders: All procedure orders, including referrals, and including information pertaining to ordersets aka. SmartSet orders.

DICOM images: Radiology images in DICOM file format for all imaging modalities. Images can be de-identified in bulk for machine learning projects.

Epic data other than Media Tab images: All data in Epic is available for research through the data warehouse. The only exception at this time is scanned documents on the Media Tab.

Clinical Ancillary Systems: Outside the EHR, we work closely with Our Enterprise Data Intelligence team to obtain clinical ancillary system data for research. We have Philips Bedside Monitor (vital signs) data, SciImage (Echocardiography), Muse (ECG), Syngo Via (Radiology), PowerPath (Pathology) and Aria (Radiation Oncology) datasets from the hospital, and can acquire others upon request.

Data Discovery at Scale

The Enterprise Data Intelligence Platform is consciously designed with speed, scalability, simplicity, and openness at its core and architected to handle the large analytical queries. The platform is available in the broadest range of storage, including on premise, in the cloud, or a hybrid model.

Using the latest data management architecture concepts, the data models support relational database management schemas to in-memory, column-oriented schemas. The flexibility in database schemas allows for data discovery scaled to the need of the project.

Machine Learning Platform

Our machine learning environment includes a number of analytic engines for various kinds of data mining and machine learning processing. The Automated Machine Learning Library includes over 100 algorithms made available to address common clinical data processing algorithms for supervised machine learning projects.

The machine learning platform is designed for flexibility and incorporates the open source statistical programming languages such as R, Python, Julia and the associated libraries. The data pipeline and platform are designed to assist developers to prepare, build, train, tune, and deploy high-quality machine learning (ML) models quickly.