CURE ID
Researcher’s Guide
2023 Edition

A collaboration between:

- U.S. Food & Drug Administration (FDA)
- National Center for Advancing Translational Sciences of the National Institutes of Health (NCATS/NIH)
- Critical Path Institute’s CURE Drug Repurposing Collaboratory (CDRC)
- Society of Critical Care Medicine (SCCM)
- Johns Hopkins School of Medicine (JHU)
- Emory School of Medicine (Emory)
- The Infectious Diseases Data Observatory of Oxford University (IDDO)

Table of Contents

Introduction to CURE ID

CURE ID is a joint initiative between the US Food and Drug Administration (**FDA**) and the National Center for Advancing Translational Sciences (**NCATS**), a part of the National Institutes of Health (**NIH**). CURE ID is an online platform and mobile app that enables clinicians and patients to share their real-world experiences treating and being treated with repurposed drugs, respectively (see "What is drug repurposing" in FAQs for details). Users can report their repurposing experiences through the CURE ID case report form (CRF) from their computers, smartphones, or mobile devices. The platform promotes information sharing between healthcare professionals, researchers, pharmacists, and patients or care partners to better inform treatments, when there is a lack of adequate approved therapies.

The COVID-19 pandemic presented a challenge and an opportunity for CURE ID. With this platform, the aim was to build a repository of the repurposed drugs used as COVID-19 treatment(s) to better understand the health outcomes with these treatments. Healthcare providers directly submitted their case reports to CURE ID and cases from the published literature were also added manually. The repository was further enriched with cases from electronic health records (EHRs) of participating healthcare institutions and disease registries such as the Society of Critical Care Medicine (SCCM) Discovery Viral Infection and Respiratory Illness Universal Study (**VIRUS** – see "What is SCCM VIRUS" in FAQs for details). The automated extraction tool – the **EDGE tool** (See "What is the EDGE data automation tool" below in FAQs for details) – was developed to extract data from different EHRs and to convert them into the CURE ID format. This was made possible through partnerships with SCCM, Mayo Clinic, the Infectious Diseases Data Observatory (**IDDO** – see "What is IDDO" in FAQs for more details), Johns Hopkins University School of Medicine (**JHU**), and **Emory** School of Medicine.

The Sample Population (Cohort)

The project aimed to capture all cases of acute COVID-19 that required hospitalization from participating institutions. The sample population was defined as all inpatients at participating partner healthcare facilities since March 2020, with confirmed positive COVID-19 test within 14 days of admission. Outpatient encounters are not systematically captured. Patients with primary admission diagnoses related to trauma or surgery were excluded. Their COVID test was considered an ancillary finding rather than the cause of their hospitalization. Shift and truncate (SANT) de-identification was used to remove true calendar date information while still preserving temporal relationships. The displayed dates may be as early as 6 months before March 2020. All data collected in this project are completely de-identified and exclusively for the purpose of research.

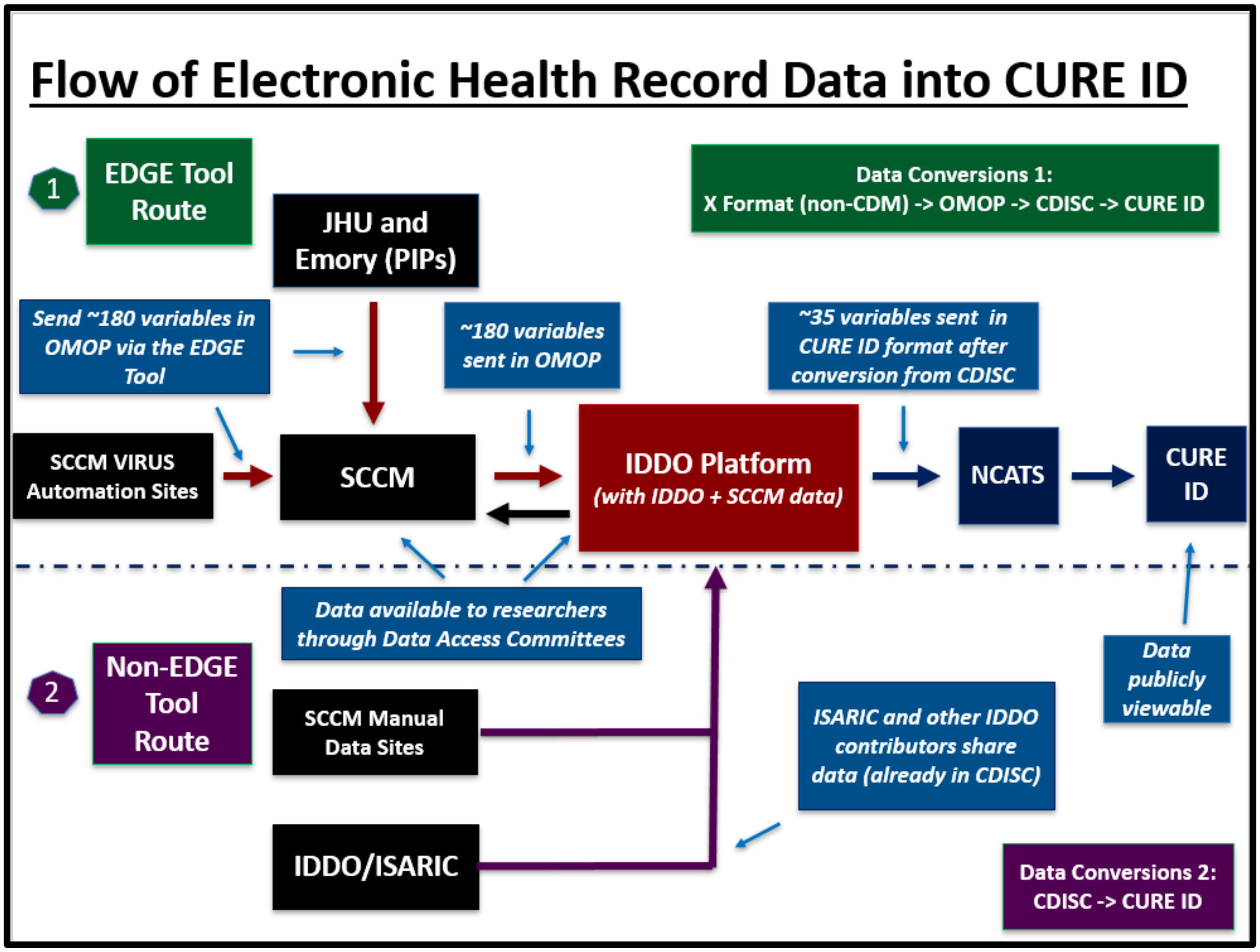
The expanded CURE ID repository houses 117,173 COVID-19 case reports. This large collection of cases from different sources of real-world data (registries, EHRs, clinician-generated data, etc.) may help to identify signals of potentially safe and effective off-label use of approved drugs. By adopting this approach, CURE ID aims to forge an efficient pathway to narrow the potential drug candidates for repurposing and inform the study of these candidates in randomized controlled trials.

This document describes the nature of the data in CURE ID (source, data flow, quality, etc.) and how investigators can access this de-identified patient-level data.

Summary of Data Processing and Pathway

For participating institutions (sites) that do not use a common data model, the EDGE tool maps the data from their institutional-specific EHRs to the Observational Medical Outcomes Partnership Common Data Model (**OMOP** CDM – see “What is OMOP” in FAQs for more details), and automates the extraction of the required data fields into the CURE ID case report form. Most sites in this category are existing SCCM VIRUS Registry sites. Once data has gone through the de-identification process, OMOP transformation, and data quality assessment, it is submitted to the Mayo Clinic or SCCM. This data, consisting of ~180 variables, then flows to SCCM VIRUS for site de-identification variables and additional quality check, and then on to IDDO. IDDO converts the data from OMOP to Clinical Data Interchange Standards Consortium’s Study Data Tabulation Model (**CDISC SDTM** – see “What is CDISC SDTM” in FAQs for details), so that it can be aggregated into the COVID-19 data platform at IDDO, which includes data from the International Severe Acute Respiratory and Emerging Infection Consortium (**ISARIC** – see “What is ISARIC” in FAQs for more details). IDDO makes a subset of this data (~40 variables) available to NCATS after having converted it to the CURE ID format. NCATS then incorporates the data into the CURE ID database. NCATS is also responsible for continued collection of cases manually extracted from the published literature, and those submitted by healthcare providers. See Figure 1.

Figure 1. Key Architecture Components Data Sources and Formats



Flow of Electronic Health Record Data into CURE ID

1 EDGE Tool Route

JHU and Emory (PIPs)

Data Conversions 1:
X Format (non-CDM) -> OMOP -> CDISC -> CURE ID

Send ~180 variables in OMOP via the EDGE Tool

~180 variables sent in OMOP

~35 variables sent in CURE ID format after conversion from CDISC

SCCM VIRUS Automation Sites

SCCM

IDDO Platform
(with IDDO + SCCM data)

NCATS

CURE ID

Data available to researchers through Data Access Committees

2 Non-EDGE Tool Route

SCCM Manual Data Sites

ISARIC and other IDDO contributors share data (already in CDISC)

Data publicly viewable

IDDO/ISARIC

Data Conversions 2:
CDISC -> CURE ID

Characterization of Data Sources

1. **SCCM**
   SCCM in collaboration with the CURE Drug Repurposing Collaboratory (CDRC) and CURE ID built an infrastructure to support development of a COVID-19 observational dataset, specifically focusing on repurposed and novel medication use. CURE ID data adopted from the SCCM VIRUS Registry was derived through two key methods:

   1.1 Manual VIRUS Registry Sites Data from partnering institutions that are participating in the VIRUS registry is manually extracted**.** *(Parent of Legacy VIRUS data)*

   DATA MODEL: VIRUS (REDCAP) Case Report Form

   1.2 EDGE Tool Data: Data from partnering institutions (VIRUS Sites and EHR data) are processed by the EDGE tool.

   DATA MODEL: OMOP

2. **ISARIC**
   The ISARIC COVID-19 clinical data platform comprises the individual patient-level datasets collected as part of clinical care and follow-up, clinical trials, and observational research either retrospectively or prospectively. At present, data from ISARIC is not available on CURE ID itself, due to data sharing restrictions. ISARIC data is currently only available on IDDO, as part of the IDDO COVID-19 data platform, where it can be analyzed in combination with the SCCM VIRUS data.

   DATA MODEL: CDISC SDTM

3. **CURE ID CRF**
   Data in CURE ID is captured in the form of internally developed Case Report Forms (CRFs) that are specific to CURE ID. NCATS and FDA are responsible for continued collection of cases manually extracted from the published literature, and those submitted by healthcare providers and patients/caregivers.

   The CURE ID CRF form data is only available for exploration in the CURE ID Website and Mobile App, and it is not available for download.

   DATA MODEL: CURE ID Specific Data Model

Accessing the Data

Note: IDDO and NCATS jointly share the responsibilities of the Data Coordinating Center (DCC). IDDO is the DCC for the large dataset and NCATS for the small dataset.

The patient-level data can be accessed at three points:

1. **Small Dataset:** 40 variables in the CURE ID CRF format
The 40 variable "small dataset" carved out by IDDO from the large dataset forms the basis for the openly available CURE ID case report form. Researchers can access this subset of data with the 40 row-level data in the CURE ID case report form. It is available with an online tool for exploration, but users are not able to download the data. Users can explore the data openly on the CURE ID website or by downloading the "CURE ID" App from the Apple App store or Google Play store.

2. **SCCM**
SCCM will be hosting de-identified data (the full 180 variables in the "large dataset"), derived out of the CURE ID project, in addition to the full data derived out of the SCCM VIRUS COVID-19 Registry.

   Investigators must submit an ancillary study proposal via an online submission portal to access this dataset: Society of Critical Care Medicine - Discovery Ancillary Proposal Submission Portal (secure-platform.com).

   The SCCM Discovery Publication Workgroup will assign two reviewers per proposal. The timeline for approval is up to 30 days. Once approved, investigators receive access to the de-identified, requested subset of data. Feedback on proposals is shared with investigators. Investigators also have access to resources and collaborations through the Discovery Critical Care Research Network.

3. **Combined Dataset (SCCM VIRUS Registry + IDDO COVID-19 Data Platform)**
IDDO is hosting the CURE ID "combined large dataset" of approximately 180 variables that contains detailed laboratory values and longitudinal physiologic measures derived from the VIRUS Registry and aggregated with the IDDO COVID-19 data platform, which includes the ISARIC COVID-19 database. Investigators wishing to access the full dataset of 180 variables (including IDDO/ISARIC data), must submit a request to IDDO. The request is reviewed by an independent Data Access Committee. To date, all requests have been ultimately approved, however, some requests are sent back to investigators for clarification or strengthening of ethical considerations. To submit a request, researchers must download and complete a Data Access Application Form which can be found in the IDDO COVID-19 Data Access Guidelines. Applications should include details of the variables needed for the analysis, as listed on the Data Inventory.

   Note: The large dataset data can be requested through VIRUS or IDDO. However, if requested through IDDO, data from both IDDO and VIRUS could be made available.

Data Security

Multiple layers of protections have been implemented to ensure patient privacy. Outside parties are not able to access electronic health records (EHR), gain access to institutional networks, or ask for any permissions to get around any firewalls.

Participating institutions remove patient IDs using the EDGE tool developed by JHU.

There are three layers of data security protections which are as follows:

1. **Institution Based De-identification and Access**
   Only trusted team members at a participating institution can implement the tools to harmonize and de-identify the data. De-identified data includes unique patient identifiers and relative dates of care, but these data bear no systematic relationship with the original or "true" data. Our de-identification process minimizes the likelihood that anyone can re-identify patients captured in the dataset.

2. **Removal of Identifying Information**
   The provided algorithm generates random person identifiers to replace EHR's medical record numbers, rather than rely on a cipher or encryption algorithm which can be reverse engineered. The process also shifts all dates in a patients record. This shift is randomly generated for each patient, so the dates associated with their care are still relevant to each other, but no longer reflect the patient's actual dates of hospitalization. This preservation of temporal relations in data while maintaining privacy is called Shift and Truncate (SANT).

3. **Site De-identification**
   The SCCM VIRUS coordinating center, Mayo Clinic, is removing sites' IDs and curating data from all participating hospitals into one large dataset and assigning new person identification numbers, so that no researchers accessing the data can know the hospital at which a particular patient was treated.

Quality Assurance and Control

Data Quality and De-identification Contributors:

Mayo Clinic, SCCM VIRUS, IDDO, JHU, NCATS

IDDO and NCATS perform quality assurance activities (e.g., data quality checks on plausibility, conformance, and completeness of data values) and ongoing troubleshooting.

Frequently Asked Questions

What is Drug Repurposing?

A: Drug repurposing is when clinicians use existing drugs in new ways such as, for new diseases, new aspects of diseases, new populations, in new doses, or in new combinations. This off-label use of existing approved drugs is known as drug repurposing.

What is the EDGE Data Automation Tool?

A: The EDGE Tool is a series of resources to expedite the implementation of the OMOP common data model's extraction, transformation, and loading (ETL) process, including de-identification procedures, a data quality dashboard, and a platform for building exportable cohort definitions. This includes graphic user interfaces (GUI) to aid in mapping concepts (see "What is concept mapping" below) from a hospital's proprietary data model to OMOP. The EDGE Tool can assist in extracting data from discrete fields or those with defined ontologies (e.g., drop-down menus, type-ahead fields, or fields restricted to integers). Specific examples of discrete data include flowsheet rows in Epic (e.g., vital signs, nursing assessments), measurements (e.g., laboratory results) and certain past medical history fields and assessment forms. The EDGE Tool is not currently capable of extracting data from unstructured fields such as notes or reports (e.g., imaging results, history, and physical).

The EDGE tool was developed by JHU to automate the extraction of data from different electronic health records (EHRs) and convert it into the OMOP format. JHU deployed the EDGE tool within a cohort of recruited sites through the SCCM VIRUS Registry and Emory and supported the institutional partners in this process.

What is OMOP?

A: The OMOP (Observational Medical Outcomes Partnership) Common Data Model (CDM), developed by the OHDSI (The Observational Health Data Sciences and Informatics) community, allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format. The OMOP CDM organizes EHR data into a standard set of tables using standard vocabularies such as LOINC, RxNorm, SNOMED, and ICD-10. Once data from an EHR has undergone the extract, translate and load (ETL) process into the OMOP CDM, it may be consolidated and analyzed with data from other EHRs that have undergone the ETL process.

What is Concept Mapping?

A: A concept is a specific type of data captured in the EHR. In a common data model, concept IDs help standardize the process of how data is obtained, captured, and stored. It is a clear definition outlining acceptable methods or devices for capturing the measurements. For example, in OMOP, these different ways of measuring, capturing, and recording the patient's pulse are stored in concept IDs.

What is SCCM VIRUS?

A: The Society of Critical Care Medicine's **(SCCM)** Discovery Viral Infection and Respiratory Illness Universal Study **(VIRUS)** is an international pandemic registry that provides information regarding intensive care treatments and outcomes for patients with coronavirus disease 2019**.** The VIRUS Registry is a global COVID-19 registry that tracks ICU and hospital care patterns. SCCM is encouraging enrollment of VIRUS participating sites, as well as national and international collaboration on ancillary studies.

The de-identified, HIPAA compliant database was developed to capture both core data collection fields containing clinical information collected for all patients, and an enhanced data set of daily physiologic, laboratory, and treatment information. The case report forms were adapted from the ISARIC/World Health Organization template data collection forms with focus on an ICU-specific context. Participating institutions retain all rights to the data contributed and have full access to their own data through the ancillary study submission process. All participating sites for whom publications are derived from their data being a part of the analysis, are given collaborative co-authorships on these publications. Participating institutions who have a full instance of OMOP can use the data to support other research, including quality assurance, or participation in other registries.

What is IDDO?

A: The Infectious Disease Data Observatory (IDDO) at Oxford University is a platform for housing, harmonizing, and sharing individual-patient level data from registries, observational studies, and clinical trials. IDDO assembles clinical, laboratory and epidemiological data on a collaborative platform to be shared with research and humanitarian communities for the purpose of generating new evidence and insights.

The CURE ID data, once in OMOP, is aggregated and transferred to IDDO. IDDO converts the data from OMOP to the Clinical Data Interchange Standards Consortium's (CDISC) Study Data Tabulation Model (SDTM). The IDDO platform further aggregates this data with the existing IDDO COVID-19 platform to form a unified SDTM database, which includes over 800,000 individual patient's data from the ISARIC COVID-19 dataset.

What is CDISC SDTM?

A: The Clinical Data Interchange Standards Consortium's (CDISC) Study Data Tabulation Model (**CDISC SDTM**) is the data model used for organizing data, standardizing structure for human clinical trial data tabulations, and for non-clinical study data tabulations that are to be submitted as part of a product application to a regulatory authority such as the United States Food and Drug Administration (FDA). FDA has put considerable effort to develop a repository for all submitted trial data as well as a suite of standard review tools to access, manipulate, and view the tabulations. SDTM includes domain classes such as interventions, events, or findings.

What is ISARIC?

A: International Severe Acute Respiratory and emerging Infection Consortium (**ISARIC**) is a global federation of clinical research networks that aims to generate and disseminate clinical research evidence whenever and wherever outbreak-prone infectious diseases occur. ISARIC provides tools suitable for different researchers` needs, including:

- Data collection only
- Contributing clinical data
- Research protocols
- Guidance to start your study
- ISARIC's affiliated studies

The ISARIC COVID-19 clinical database is one of the world's largest and richest standardized collections of comprehensive COVID-19 clinical data for hospitalized patients. The ISARIC COVID-19 clinical database is comprised of individual patient-level datasets collected as a part of clinical care and follow-up, as well as observational research conducted either retrospectively or prospectively.

IDDO and ISARIC are based at the University of Oxford where the platform is hosted. COVID-19 clinical research resources are fully adaptable and completely free to use.