ADEPTIA

CUSTOMER SUCCESS STORY

Helping the NIH Fight the Pandemic

National Institutes of Health

Industry: U.S. biomedical research agency

Employees: 18,478

For more information: Visit nih.gov

Disclaimer: The National Institutes of Health is not endorsing Adeptia. By official policy, is it not permitted to endorse vendors.

Benefits

Centralized database: Final database is the largest-ever collection of row-level COVID-19 patient data in the United States.

Ability to track data: It's now possible to chart the spread of COVID-19 through the United States on a ZIP Code level.

Easy, quick access: Epidemiologists, disease experts, and other scientists can chart the course of disease, evaluate treatment options, and develop COVID-19 strategies.

Adeptia provided a self-service, AI-driven workflow that streamlined a manual, time-consuming process.

The Challenge

With the onset of the COVID-19 pandemic in early 2020, the NIH set a goal of promoting public policy at state and national levels to flatten the epidemic curve, which meant slowing the rate of new infections in order to ensure that people in need of health care resources, especially those who were critically ill, didn't need care at same time and thus overwhelm a system with limited resources.

To do this, the NIH needed to understand the pathophysiology and symptom progression of this new pandemic disease. It also had to address biological, environmental, and socioeconomic risk and protective factors, as well as identify treatments and rapidly build clinical decision support (CDS) and practice guidelines. Among the questions to answer:

- Which drugs were most likely to benefit a given patient?
- What treatments, risk factors, and social determinants of health had an impact on the disease course and outcome?
- How can we develop, adapt, and deploy CDS to keep up with a dynamic pandemic?

To address these questions and set public policy (with a goal of flattening the infection curve), it was necessary to collect and analyze a high volume of reliable patient-level, accurately attributed, nationally representative data related to:

- The presence and spread of COVID-19 infections
- The effectiveness of COVID-19 tests
- The effectiveness of drugs used to treat COVID-19 and its underlying symptoms
- Demographic statistics, including underlying health conditions to check correlation and how effectively drugs treated those underlying conditions
- Antibody tests on people who have had COVID-19 and have built up immunity to various strains
- Vaccine studies, testing the effectiveness through clinical trials

For all the COVID-related data to be of use to the NIH in its effort to flatten the infection curve, it needed to be **normalized, standardized, checked, assured for quality** and, finally, brought into a single database **in a format that could be easily accessed.**

Trouble is, the data needed to answer those questions was being collected all over the United States, in different labs, from different physicians groups and other health organizations. What's more, it was being collected in a variety of formats, using different terms and semantics, and focusing on different demographic meanings. For all the COVID-related data to be of use to the NIH in its effort to flatten the infection curve, it needed to be normalized, standardized, checked, assured for quality and, finally, brought into a single database in a format that could be easily accessed.

To accomplish this goal, the NIH formed the National COVID 19 Cohort Collaborative (N3C) to develop the means to rapidly collect and harmonize clinical, laboratory, and diagnostic data in order to address immediate needs and long-term consequences, and answer questions about COVID-19 treatments, risk factors, health outcomes, and other related questions.

(continued)

The challenge
(continued)

The N3C accepted data from any of the four common data models:

1. Observational Medical Outcomes Partnership (OMOP)
2. i2b2/ACT
3. the National Patient-Centered Clinical Research Network (PCORnet)
4. TriNetX

The phenotyping extraction process was performed with scripts, lightweight data quality checks, and an extraction code that transformed the data into zip files for transmission. (The four-month process started in April 2020.)

Once that data was received, it was reconstituted into target data models:

First Stage Ingestion

1. Expand the zip files in CSV (comma-separated value) format and check the data manifests
2. Reconstitute the data into native CDM (Common Data Model) formats
3. Hybrid data quality checks adapt the OHDSI (Observational Health Data Sciences and Informatics) Data Quality Dashboard

The end result of this first stage was to create a N3C-specific data quality dashboard that was then used to improve the process.

Second Stage Ingestion

1. Aberrant data was repaired or encoded using COVID LOINC (Logical Observation Identifier Names and Codes) Codes.
2. The CDM-formatted data was transformed into OMOP 5.3 format
3. The library of validated CDM was leveraged to OMOP maps

Partnering with Adeptia

Once the FIrst Stage Ingestion was completed, the received data was then moved through the Adeptia pipeline to perform the formal transformation from source data models into the target data model.

The National Center for Advancing Translational Science (NCATS) leveraged Adeptia's data integration solution to repair or encode aberrant data (in the COVID LOINC codes), then transform that source CDM into the OMOP 5.3 format. Adeptia's process allowed NCATS to complete this repetitive, meticulous process of combining data from thousands of research laboratories, hospitals, physicians groups, and other sources, data that was in different formats, using different semantics and having different demographic meanings into a single accessible data lake.

Adeptia automated and streamlined the NCATS effort, providing a self-service workflow so epidemiologists, infectious disease researchers, and others could access data more quickly from this single database. That accelerated the testing of various hypotheses and research, so the researchers could get answers sooner and inform public policy, making it possible to more quickly arrive at decisions on health care policy, testing, drug treatments, vaccines, etc.

The Final Merge

Once, thanks to Adeptia's process, the data had been transformed into the OMOP 5.3 format, NCATS was able to take the following steps:

1. The OMOP-versioned data from all sources was combined into an analytic database.
2. This database was migrated to the secure Palantir Analytic Platform (on AWS GovCloud).
3. To maintain security and confidentiality, no data on the Palantir environment can be downloaded.

Key statistics in this data include COVID positive patients, total patients, sites signed DTA, sites data ingested, rows of data, procedures, lab results, visits, observations, drug exposures.

The COVID-19 Database

This final database constitutes the largest-ever collection of row-level data in the United States focusing on COVID-level patients, and it allows unprecedented analytics, specifically machine learning, in a manner impossible with federated systems that are currently being used in many environments. The NIH believes the contributions of machine learning from the N3C data support will enable the discovery of information not easily found through other methods. It becomes possible to chart the spread of COVID-19 through areas on a ZIP Code level, with connections to underlying electronic medical records (EMRs), making it possible to predict a spread and epidemiological patterns associated with EMRs as predictors.

Adeptia's role in the NIH effort

Though Adeptia played only one step in the overall process, thanks to the company's AI-driven data integration solution, the NIH were able to combine disparate data from thousands of separate sources, in a variety of formats, and with a many demographic meanings. Adeptia provided a self-service, AI-driven workflow that streamlined what otherwise would have been a time-consuming, tedious process.

The end result? Epidemiologists, disease researchers, and other scientists and medical professionals could access data quickly and easily from a single database, allowing them to chart the course of the disease, evaluate treatment options, and develop a comprehensive strategy to address the COVID-19 pandemic.