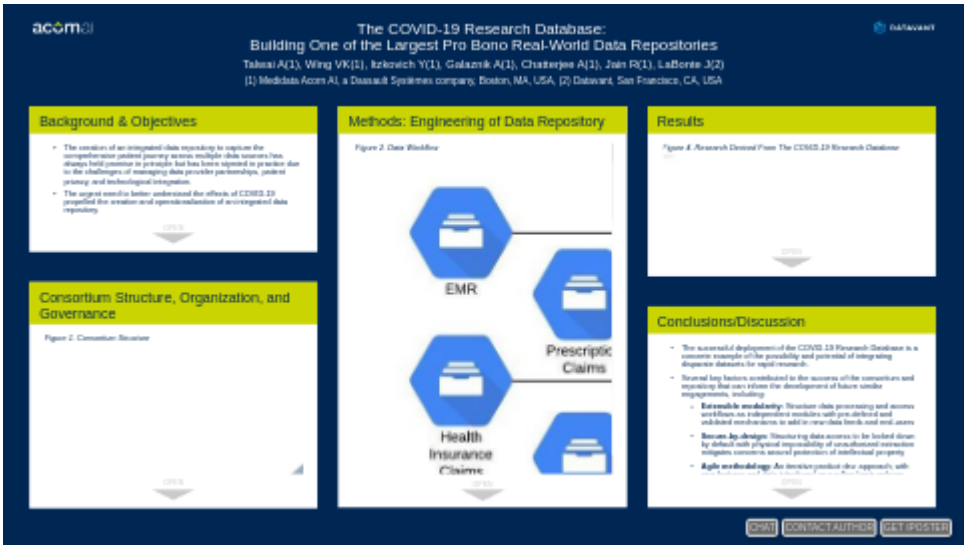
The COVID-19 Research Database: Building One of the Largest Pro Bono Real-World Data Repositories



Talwai A(1), Wing VK(1), Itzkovich Y(1), Galaznik A(1), Chatterjee A(1), Jain R(1), LaBonte J(2)

(1) Medidata Acorn AI, a Dassault Systèmes company, Boston, MA, USA, (2) Datavant, San Francisco, CA, USA

PRESENTED AT:

Virtual Poster Sponsor: PHAR

VIRTUAL ISPOR 2021

BACKGROUND & OBJECTIVES

- The creation of an integrated data repository to capture the comprehensive patient journey across multiple data sources has always held promise in principle but has been stymied in practice due to the challenges of managing data provider partnerships, patient privacy, and technological integration.

- The urgent need to better understand the effects of COVID-19 propelled the creation and operationalization of an integrated data repository.

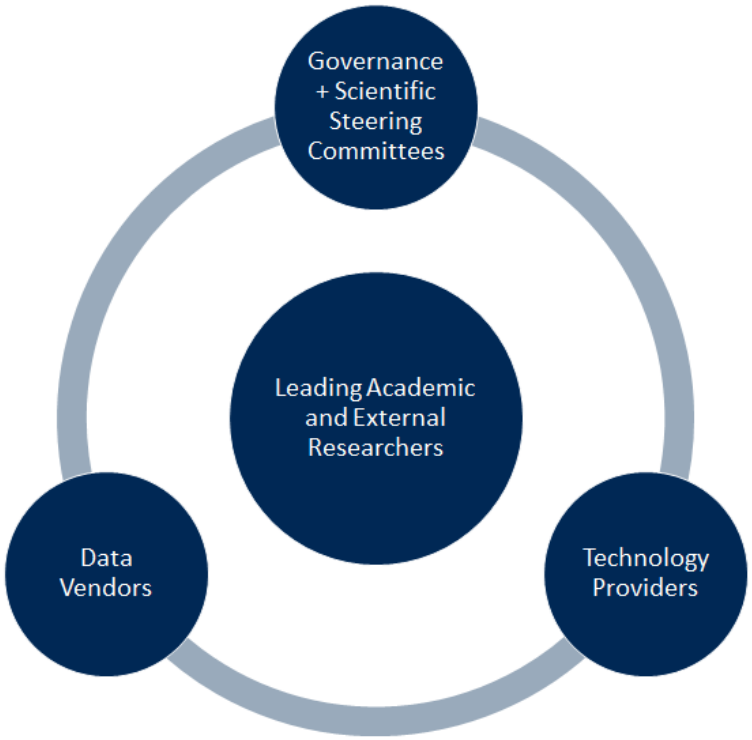
- Set up in the spring of 2020 and still in continued operation, the COVID-19 Research Database consortium is a pro-bono, cross-industry, cross-sector collaborative composed of institutions donating technology services, healthcare expertise, and de-identified data.

    - The data repository constructed by the consortium contains integrated, linked datasets from multiple sources. This repository enables public health and policy researchers to use real-world data to better understand and combat the COVID-19 pandemic.

- The aim of this research is to chronicle the technical and collaborative efforts underpinning this success.

CONSORTIUM STRUCTURE, ORGANIZATION, AND GOVERNANCE

Figure 1. Consortium Structure



The consortium consists of the following:

- Governance Committee and Advisory Groups
  - Each consortium member represented on governance committee, which takes consortium-level decisions and ensures all stakeholder interests are protected
  - Expert advisory groups on privacy, patient advocacy, legal and strategy
- Scientific Steering Committee
  - Reviews research proposals for research for scientific rigor and peer-review standard; Comprised of leading academics and clinical researchers
- Targeted working groups
  - Discrete 2-5 member cross-organizational teams focusing on specific functions (e.g., outreach, legal, partnerships, etc.)
- Data governance
  - Reviews data extracts to be provided to each research group relevance and appropriateness, with sign-off / veto from data contributors
- Core operations team

- Creates and maintains data pipelines and tech workflows; on-boards and interfaces with researchers; oversees certification and export of results; handles project management efforts

METHODS: ENGINEERING OF DATA REPOSITORY

Figure 2. Data Workflow

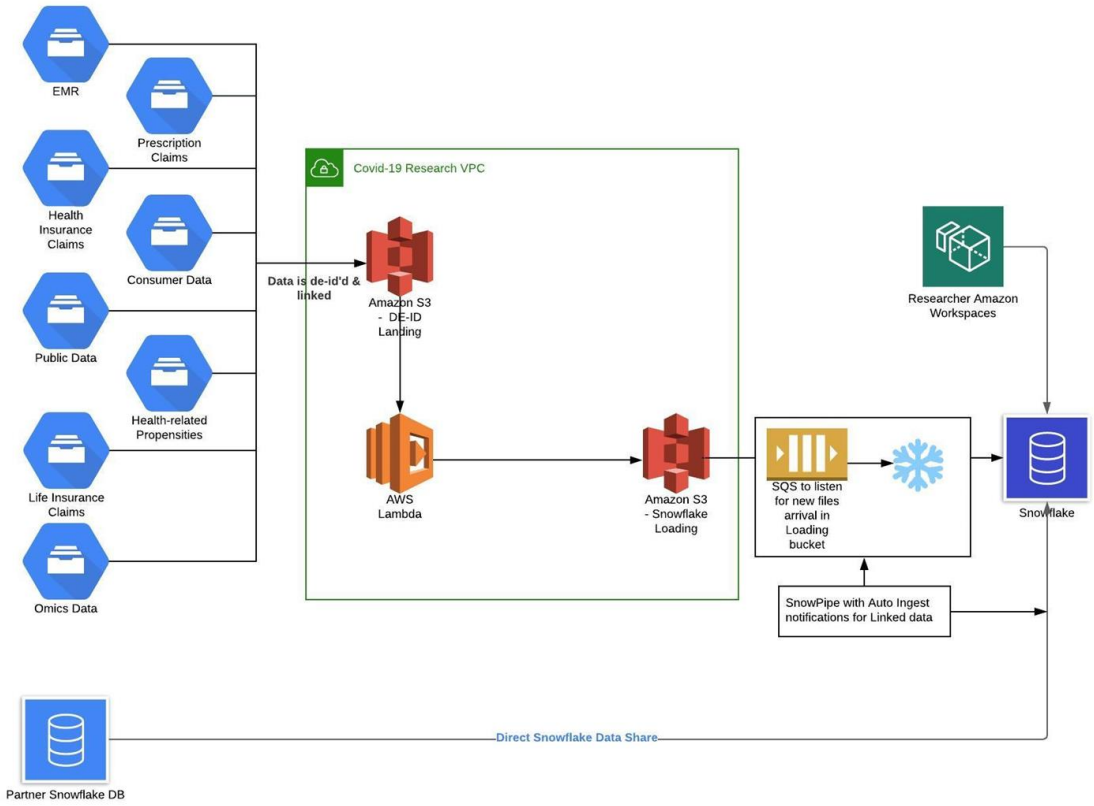
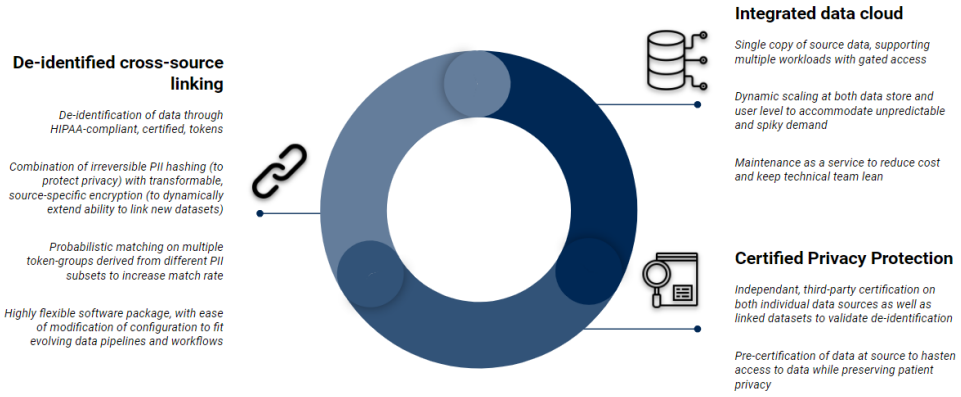


Figure 3. Technology Features



Twelve de-identified data sets were donated by consortium data vendors and were integrated into one repository hosted and access-controlled by Medidata. The following data processing steps occurred between data delivery and accessibility for research **(Figure 2)** with key technology elements highlighted **(Figure 3)**:

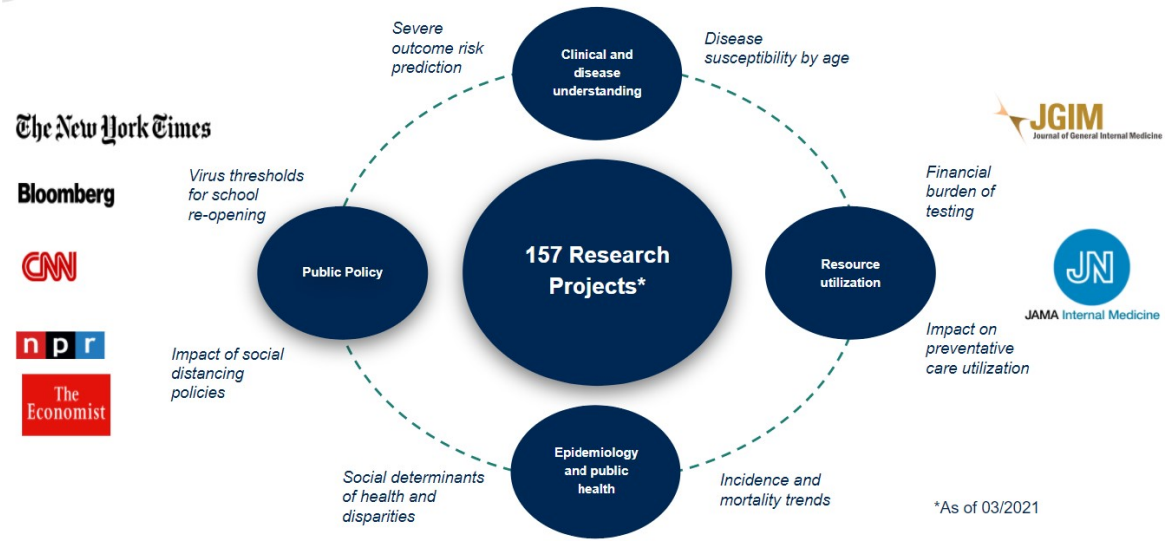
1. Data from each source was picked up on an automated cadence from a pre-specified location in the source environment and loaded into a dedicated virtual private cloud in the Medidata environment, isolated from both the public internet and other corporate networks.

2. Each source file was run through an automated ETL process to obtain an updated data extract in a pre-set source-specific schema and loaded into Snowflake's cloud-based data store.

3. Based on researcher needs and data rights, a subset of these data sources were cross-linked using Datavant's tokenization technology to knit together patient data from disparate sources in a privacy preserving manner:
    - Patient data at the individual sources were run through Datavant's patent-pending de-identification technology which replaces identifiable patient information with an encrypted token.
    - Each subset of sources was run through Datavant's Link software, which replaced their source-specific encryption with a common (but still de-identified) encryption scheme.
    - Patients records data across multiple data sources were thus able to be identified and joined on the encrypted patient key, while still preserving personally identifiable information (PII).
    - De-identified datasets were certified by an independent patient privacy certifier.
4. Each researcher approved by the scientific and data governance committees were provisioned with a separate, secure workspace pre-loaded with analytical tools (ex: SAS, R, Python) and able to access, query and perform analysis on the data store.
    - Researchers were granted role-based access to specific data feeds relevant to their approved research proposal as opposed to the whole data store, in order to better protect contributor intellectual property (IP).
    - Researchers could analyze and process data in their workspaces, and share code and results within their research group, but were restricted from exporting data.
    - The sole route for data leaving the workspaces and the dedicated environment was through a manually gated channel for publishing (aggregated) results, which were reviewed by external patient privacy certifiers before being approved for release.
    - All other ingress and egress routes from their workspaces and data store were black-holed to ensure data privacy and security.
5. Data in the repository are updated regularly with the latest patient data. An automated process was engineered to streamline new data ingestion, de-identification and linkage, and availability to researchers.

RESULTS

Figure 4. Research Derived From The COVID-19 Research Database



- Approximately 17 terabytes of data, consisting of 87 billion records from over 250 million unique patients and over 2 million patients with COVID-19, from 12 different data providers were loaded onto the platform.
- Data types ingested include medical, pharmacy, and life insurance claims, electronic health records, mortality, consumer, -omics, and health propensities.
- Over 350 academic, scientific, and medical researchers accessed the repository and conducted more than 150 research studies, which have resulted in a number of publications as well as mentions in national media **(Figure 4)**.
- Major news and journalism outlets have reported on these research findings.
- Topics of research included:
  - Public policy
  - Clinical and disease understanding
  - Epidemiology and public health
  - Resource utilization

CONCLUSIONS/DISCUSSION

- The successful deployment of the COVID-19 Research Database is a concrete example of the possibility and potential of integrating disparate datasets for rapid research.

- Several key factors contributed to the success of the consortium and repository that can inform the development of future similar engagements, including:

  - **Extensible modularity:** Structure data processing and access workflows as independent modules with pre-defined and validated mechanisms to add in new data feeds and end-users

  - **Secure-by-design:** Structuring data access to be locked down by default with physical impossibility of unauthorized extraction mitigates concerns around protection of intellectual property

  - **Agile methodology:** An iterative product dev. approach, with new features and data introduced on a rolling basis reduces time-to-access and prevents roadblocks from derailing broader effort

  - **Automate to scale:** Investing in upfront automation of the data ingestion, linking and provisioning process allows for rapid growth in both users and data

  - **Structured onboarding:** Proactively anticipate challenges in end-user experience and mitigate through comprehensive onboarding process and collateral

  - **Stakeholder rights:** Inclusive governance committee and veto rights on core interests builds trust and willingness to participate

  - **Credible convenor:** Sponsorship and involvement from widely respected and ‘neutral’ figures in setting up the consortium ensures openness to engage

  - **High-bar wide-scope for applications:** Allow end-users latitude to explore a wide range of use-cases and research topics, but enforce bar on quality of output

- Future efforts that can be employed using this technology include:

  - Prospective tokenization during enrollment in source (e.g., registries, trial, etc.)

  - Linkage between clinical trials and observational datasets to allow monitoring of adverse events and effectiveness after completion of the trial

  - Integrated social determinants of health data through linkage of census profiles, social datasets, and consumer profiles, to patient healthcare data

  - Linkage to non-traditional healthcare data, such as wearables and social media

  - Population-level data from decentralized health systems

  - Patient-centric longitudinal value-based contracts

  - Burden sharing across payer entities and life cycles