COVID-19 RESEARCH DATABASE

A Case Study of Pragmatic Patient Privacy Protection

The current COVID-19 pandemic precipitated an urgent search for knowledge essential to our understanding, management, and control of the virus. Public health institutions responsible for having access to such knowledge were hard-pressed to meet the demand for patient-level and facility-based information required to support real-time operations. Furthermore, the institutions were under more than the normal operating stress as the pandemic was politicized. As a result, little to no information was available for medical researchers looking to explore and understand the public health, economic, and social effect of the pandemic in a timely manner using longitudinal patient health data. Much of the data desired by researchers, however, was readily available in real world health information contained in medical records and insurance claims. The COVID-19 Research Database, a collaboration of a host of public and private organizations in the healthcare data industry convened by Datavant, was created in April 2020 to make such data available to policy and public health researchers without charge.

A critical and uncompromising consideration of this database is the preservation of patient privacy. The database has been designed pragmatically to give researchers sufficient latitude to explore their ideas while minimizing re-identification risk to be consistent with HIPAA requirements. The implemented solution is a controlled analytics environment where the multitudinous datasets are accessed and explored; it has been made possible only through the partnership of an agile and responsive HIPAA certifier of health information (Mirador Analytics), a flexible technology host (Medidata Acorn AI), and patient and supportive data and technology partners.

All health datasets ingested into the database, from a plethora of philanthropic donors, are certified as being de-identified under HIPAA. Some of these datasets are linked, using an encrypted linking token derived from since-redacted PII (personally identifiable information) to allow a greater variety of research hypotheses to be addressed. These combined datasets are similarly certified as being de-identified, often first requiring implementation of modifications advised by the certifiers to sufficiently reduce the disclosure risk. Researcher access to the data is subject to strict governance rules, requiring all researchers to sign an end user agreement (EUA) and submit via a process designed and managed by the Health Care Cost Institute. Researchers are provided access only to those datasets that fits their research question. To ensure against unanticipated re-identification risk, datasets are encoded in a way that makes linking patient data from multiple datasets not possible without prior review and permission.

No information is permitted to be exported from the analytics environment without first being appraised for its re-identification risk. This is another task completed by the certifiers who—mindful of expediency—complete the task at least twice a week. This involves close scrutiny of what should be limited to insights, summaries and aggregations.

The utilized privacy approach has enabled timely access to real world information in a manner that is both privacy preserving and focused on extracting maximal utility from the information. While maximal utility and privacy preservation are often at odds, the controlled analytics environment has demonstrated this is possible for real world information.

The privacy protected research database has been adopted widely and supports over 1,600 researchers and over 100 research teams hailing from all of the top 30 US News and World Report ranked medical schools and several government and policy research organizations (including the CDC, NIH, VA, NBER, and the CBO). Over 130 projects have been submitted and many papers have been published. A sample of these are:

- The effects of state closure policy on the utilization of non-COVID-19 healthcare services (paper here)
- The disparity in infection rates between Black and Hispanic patients compared to white, non-Hispanic patients (findings here)
- How to balance economic and public health concerns in pursuing re-opening policies (findings here)
- The fluctuations in expected mortality driven by the pandemic (findings here)
- How mortality rates and infection rates have varied by socioeconomic group (findings here)
- The impact of COVID-19 on the use of preventative care (findings here)
- Charges of COVID-19 Diagnostic Testing and Antibody Testing Across Facility Types and States (findings here)
- Covid-19 Severe Outcome Risk Prediction (Private Machine Learning Model, Presented at the National Institute of Statistical Sciences' Data Science Conference on August 28

Some of the ongoing several ongoing studies include:

- Impact of telemedicine and impact by different specialties
- Therapeutic impact (impact of various drugs, from heparin to hydroxychloroquine)
- Looking at different claims as they break down—geography, socioeconomic factors
- Impact of socioeconomic factors and racial factors and cost

The database is expected to be maintained, supported, and expanded by the 30+ sponsors through mid-2021.