JOHNS HOPKINS
UNIVERSITY & MEDICINE

How to Access National COVID-19 Data for Research

The National COVID Cohort Collaborative (N3C) is creating a central registry of patients who have been tested for COVID or have a clinical diagnosis of COVID. Contributing sites include CTSAs who chose to forward data on COVID cases from existing data collections such as Patient Centered Clinical Research Network (PCORNet), Accrual to Clinical Trials (ACT), Observational Health Data Sciences and Informatics (OHDSI), or TriNetX. The program is supported by HHS through many of its agencies including NIH, FDA, and CDC; the Department of Veterans Affairs is also contributing data. Data will include all demographics, medications, laboratories, diagnoses, procedures, and other structured data ultimately arising from EHRs for at least 1 year prior to testing. The registry will comprise a HIPAA limited dataset, as it will include real dates of service and geocodes (to census block resolution where available). The data will be updated from the contributing sites as frequently as practical, ideally twice a week.

The data will be harmonized by the National Center for Data to Health (NC2H) into a common data model, OMOP 5.3. Investigators can apply to analyze this data through a Data Access Committee, comprised of federal and stakeholder scientists. All analyses will be conducted on a secure instance of the Palantir data enclave, hosted by NCATS. Analyses tools will include conventional data science tools such as R, Python, and Jupyter. The data enclave will prohibit downloading the data, all analyses must be done on the platform. Investigators must attest to share all findings and code with the community. A separate synthetic derivative of the limited data set will be created, sampling from the statistical distributions of data in the real data set. Access to this dataset will require registration only, though it still must be analyzed on the data enclave platform.

Creating the N3C registry of row-level data as a limited (albeit protected) dataset of EHR data at a national level will be unprecedented in US clinical research. It will support novel machine learning analytics and discovery of important predictors associated with emergency visits, hospitalizations, ICU transfer, ventilator dependency, and death, amongst a myriad of related outcomes. It will have the scale, statistical power, and computing platform to address many questions the clinical and research communities seek to answer. Ideally it will be used by multidisciplinary teams engaging clinicians, statisticians, epidemiologists, biologists, health services researchers, and data scientists, who together can leverage this unique resource at this critical time in the COVID pandemic. Further, these activities can be conducted with minimal risk of data breach or inadvertent disclosure, as any effort to view or download row level data is generally disabled.

For more information, please contact Dr. Christopher Chute at chute@jhu.edu.