Integrated Approach to Secondary Data Analysis at Ohio State: Access, Research Support, and Education

Guy Brock, PhD[1]; Amy Lehman, MAS[1]; Tanya Mathew, BDS, MS[2]; Lang Li, PhD[1];
Timothy Huerta, PhD, MS[1,3]; Henry Xiang, MD, MPH, PhD[4]; Rebecca Jackson, MD[5]

[1] Department of Biomedical Informatics, The Ohio State University, Columbus, OH; [2] Center for Clinical and Translational Science, The Ohio State University, Columbus, OH; [3] Department of Family Medicine, The Ohio State University, Columbus, OH; [4] Center for Pediatric Trauma Research & Center for Injury Research and Policy, Nationwide Children's Hospital, Columbus, OH; [5] Division of Endocrinology, Diabetes, and Metabolism, The Ohio State University, Columbus, OH

The Ohio State University CCTS is developing a comprehensive support system for researchers to work with existing datasets through *data curation and access*, *financial support* for innovative research ideas, and *education* about practical data issues and challenges.

DATA CURATION & ACCESS

CATALYST was formed by The Ohio State University College of Medicine to advance T3 research at Ohio State. As part of that mission, CATALYST has developed this DataCore for researchers across the College of Medicine to streamline and expand access to data that accelerates discovery and leads to funded research.

CATALYST seeks to curate selected data to serve as strategic resource for the college. The term "curation" is used to express the process that the DataCore takes to providing this resource to the community. The need for data curation is significant in healthcare-related research.

The CATALYST DataCore (CDC) is a shared resource available to researchers in The Ohio State University College of Medicine (COM) that brings large-scale clinical datasets into an analytic platform that is easy to access and simple to use to facilitate outcomes research.

The CDC will be a tool that streamlines research on secondary datasets. It is a shared resource available to researchers in COM that will reduce the costs associated with data licensing and the time associated with data acquisition and processing. The CDC contains large-scale clinical datasets such as Marketscan, the Healthcare Cost and Utilization Project (HCUP), and Centers for Medicare & Medicaid Services (CMS) Claims data. It’s an analytic platform that will empower researchers to ask their question in one datasource and get answers from many. The CDC will be easy to access by using automation to empower researchers to use an application to select the data they want and then automatically pull it into the statistical analysis software of their preference. The CDC is a source of truth for all the data held within it and includes clear instructions on how to use the data with a dataset that is already cleaned, merged, and harmonized.

Taken together, the CDC helps streamline, simplify, and automate scholarship and discovery.

The data commons will contain the following deidentified data:

- Healthcare Cost and Utilization Project (HCUP) from the Agency for Healthcare Research and Quality (AHRQ): admission-level data about hospital admissions, readmissions, and emergency department use
- Centers for Medicare & Medicaid Services (CMS) claims data: claims-level data about all claims filed through medicare
- American Hospital Association (AHA) Annual Survey (AHAAS) with Information Technology supplement (AHAIT): annualized survey about hospital demographics

- Health Information National Trends Survey (HINTS) from the National Cancer Institute (NCI): person-level survey that asks people general thoughts on cancer and cancer-related topics
- Patient-Centered Outcomes Research Institute (PCORI) PCORNET: patient EHR data along with PCORI studies in a unified data model
- Epic System's Cosmos: patient EHR data
- IBM's Truven: a set of clinical datasets of which one is a patient EHR dataset

The data commons is a single SQL database structured to allow for intra- and inter-dataset analysis and streamlined to empower ease of access. The data contained within the data commons has been structured such that it is merged with other years of the same data source and that the data has been corrected and validated. The data commons also contains metadata to explain what questions are being asked by each data source and what the responses to those questions mean.

For more information, please visit the CATALYST website:
https://wexnermedical.osu.edu/departments/catalyst-center/datacore

RESEARCH SUPPORT

In order to catalyze research and collaboration, we funded two pilot initiatives in 2019: Artificial Intelligence / Machine Learning and Secondary Analysis of Existing Datasets. Based on the success of these two initiatives, we plan to continue and expand these offerings for 2020.

Artificial Intelligence / Machine Learning Pilot Initiative

In September 2019, the CCTS announced a new funding opportunity to stimulate and support transformative, innovative, interdisciplinary pilot and early-stage studies that will leverage the power and impact of artificial intelligence and machine learning in new fields and applications and/or brings a new approach to the design or implementation of AI solutions.

Budgets up to $50,000 in direct costs were considered for projects that can be completed within one year.

- **37** initial abstracts were received and scored by 12 AI committee members, ranked and discussed.
  - **19** projects were selected and invited to submit full applications
  - **17** full proposals were received and reviewed by academic (n=32) and external peer (n=23) reviewers.

The review process involved the development and creation of Data Science standards for review and Peer Review Criteria geared to Patient and Community reviewers, spearheaded by Dr. Tanya Matthew, Research Specialist for the CCTS and Adjust Assistant Professor, College of Dentistry.

The following proposals were funded after peer-review:

1. Deep learning model of histopathologic images to serve as a proxy to predict recurrence risk and chemotherapy benefits in ER-positive/HER2-negative breast cancers
2. Computer-Aided Detection of Advanced Neoplasia in Intraductal Papillary Mucinous Neoplasms Using Confocal Laser Endomicroscopy
3. DeepCross: a visualized deep learning for signaling pathway crosstalk of prostate cancer disease progression
4. Deep Transfer Learning of Drug Sensitivity by Integrating Bulk and Single-cell RNA-Seq data
5. Development of a Deep-Learning Model Prototype to Identify Children Exposed to Parental Incarceration

Secondary Analysis of Existing Datasets Pilot Initiative

In July 2019, the CCTS announced a new funding opportunity to foster innovative studies focused on secondary analyses of existing clinical and translational datasets (publicly available and accessible national databases or locally available databases generally accessible to faculty at The Ohio State University or Nationwide Children's Hospitals).

Budgets up to $15,000 in direct costs were considered for projects that can be completed within 6 months.

- **29** proposals were received, selected **7** for funding after peer review by 61 reviewers

  Data Source of Proposals:
  - Claims: 8
  - EHR: 7
  - Molecular / Omics: 6
  - Cross-Sectional / Cohort: 5
  - Imaging: 2
  - Others: 3

  Human Health Focus of Proposals:
  - Cardiometabolic / Endocrine: 10
  - Cancer: 7
  - Pediatrics: 6
  - Neurological: 3
  - Public Health: 2
  - Digestive Disease: 1

- Funded applications in three major areas of focus:

  Basic Science:
  1. Systems biology analysis of cancer stem cells in ovarian cancer evolution from tumor initiation to drug resistance
  2. Deriving a whole-brain functional connectivity neuromarker in Alzheimer's Disease
  3. Exogenous sequences in unmapped transcriptome data from the ORIEN consortium

  Clinical:
  1. Computational drug repurposing for coronary artery diseases (CAD) using MarketScan data
  2. Improving outcomes for infants with severe bronchopulmonary dysplasia through deep phenotyping

  Population:
  1. Clarifying the effects of media exposure on ADHD vulnerability
  2. Racial disparities in receipt of pain treatment among cancer survivors

EDUCATION

The Biostatistics, Epidemiology, and Research Design (BERD) Core of the CCTS at OSU has developed a new educational seminar series called "**Rethink, Reuse, Research: Secondary Analysis of Existing Data**". The series consists of one hour seminars geared towards researchers and clinicians – each seminar will focus on a particular data source and be led by CCTS BERD members (collaborating biostatisticians or faculty) who are familiar with the data.

Educational goals:

1. For each dataset discussed in the series:
   a. Understand the background and overview of the original study design.
   b. Identify the types of data contained in dataset.
   c. Describe examples of previous research that have been accomplished using the data, highlighting researchers at OSU for possible collaborations and networking.
   d. Identify major limitations or practical issues in working with the data that must be taken into consideration for research.
2. Understand the process to access data and identify BERD member(s) to collaborate with for research.

Our current schedule is listed below. Survey responses from OSU faculty and staff were factored into selecting data sources.

Date	Topic
February 2021	Women's Health Initiative (WHI) Investigator Data
March 2021	Medicare / Medicaid
April 2021	National Cancer Database (NCDB)
May 2021	PCORnet®, The National Patient-Centered Clinical Research Network
June 2021	National Clinical Trials Network / NCI's Community Oncology Research Program (NCTN / NCORP) Data Archive
July 2021	National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) biorepository
August 2021	Pediatric data from Partners for Kids (PFK)
September 2021	Surveillance, Epidemiology, and End Results (SEER) cancer incidence database
October 2021	WHI Life and Longevity After Cancer (LILAC) + ancillary studies