Northwestern University Feinberg School of Medicine
Galter Health Sciences Library & Learning Center

Big Data: Fresh Data for Fresh Approaches

By: Sara Gonzales, Data Librarian

A big data revolution has been taking place in the biomedical field over the past decade. Advances in bioinformatics have allowed ever more sophisticated analyses of large datasets in genomics and the basic sciences, leading to breakthroughs in medical knowledge and diagnostics. Data-driven approaches have also had a transformative effect for clinicians, including those involved in research and patient care. In the Feinberg and Northwestern communities, there are many resources available for big data analyses for clinical studies, some of which we'll cover here.

There are several routes to gathering the data for secondary data analyses, including direct contact with mentors and colleagues, leveraging nationally available datasets such as NHANES, and harvesting datasets from catalogs of research outputs such as Zenodo, Harvard's Dataverse, and Feinberg's institutional repository DigitalHub. Depositing datasets to such catalogs is increasingly required by funders and journals for the purpose of sharing datasets with the wider research community, but how can the community make best use of this deposited data?

Northwestern's Enterprise Data Warehouse

Computational analyses of batches of datasets allow patterns to be found and discoveries to be made that may not be possible using smaller datasets. Well-described, quality datasets are vital, as Raghu Chakravarthi, Senior Vice President of a data management, integration, and analytics company attests: "Collecting metadata about data, mining real-time data using anomaly detection techniques for figuring out the outliers, and applying machine learning to cleanse data is the way to improve data quality."[1] In addition, fresh, up-to-date data can be vital for clinical studies. At Northwestern a prime source for recent clinical data is the Enterprise Data Warehouse (EDW).

Each day, Northwestern's EDW "loads 2.8 billion new data elements from 142 separate sources, including electronic health records, pathology data from the hospital and research laboratories, biomarker data from research databases and research transactional data from our eIRB and other institutional systems."[2] Bioinformatics and data science professionals can be consulted through the university to perform analyses on data gathered from the EDW, as well as to support informatics groups and research projects. NUIT Research Computing can offer assistance with everything from high-performance computational analyses to secure data storage. Most recently Galter Health Sciences Library & Learning Center launched the DataLab, a one-stop resource where researchers collecting, storing, and analyzing data can schedule consultations, learn about data wrangling classes and clinical research support, and find out more about innovations in data storage and management currently being implemented through Galter's involvement in the National Center for Data to Health grant (Grant Number U24TR002306), sponsored by NCATS.

As data continues to be produced exponentially in all fields, the same will hold true for healthcare and biomedical data. By keeping attuned to the data resources available through the Feinberg School of Medicine and Northwestern, researchers can be ready to apply the latest analytical techniques to the freshest data, and make discoveries that will fuel both improved patient care and the research of the future.

1. Harris, Richard. "Data science and the currency of the future." https://appdevelopermagazine.com/data-science-and-the-currency-of-the-future/
2. Northwestern University Clinical and Translational Sciences Institute. "Enterprise Data Warehouse." https://www.nucats.northwestern.edu/resources/data-science-and-informatics/nmedw/index.html

Printed: Tuesday, June 25, 2024 6:53 AM
Source: https://galter.northwestern.edu/news/big-data-fresh-data-for-fresh-approaches.pdf