Huge Data in Medicine Workshop 2020

V. K. Cody Bumgardner, PhD
cody@uky.edu
Caylin Hickey
caylin.hickey@uky.edu
Department of Pathology and Laboratory Medicine
University of Kentucky

Artificial intelligence (AI) has the potential to transform medical research and eventually the practice of healthcare. The recent coalescence of vast amounts of data, accelerated computing, deep learning, and other machine learning techniques has yielded remarkable results across research discovery and translational medicine. AI has emerged as a promising tool to improve patient outcomes and reduce the cost by providing insights from data across medical domains, from condition detection, diagnosis, and treatment response prediction. However, there remain challenges to overcome. The development of AI methods requires huge amounts of high-quality curated data. In fact, the volume of high-quality data used in model training can be directly correlated to model quality. The first major challenge to the fundamental use of AI in medicine is the lack of standard and accessible data readily for computational use. This challenge is comprised of not only storage needs, but also efforts in data transmission, metadata curation, and distributed processing, all adhering to privacy and policy preserving requirements.

While health care systems generate vast amounts of data to be exploited, most of the data is not accessible or usable for AI research. In many cases data, even for the same patient, data is generated and stored in disparate systems: personal medical records in Electronic Medical Record (EMR), radiology images in Picture Archiving and Communication Systems (PACS), laboratory results in Laboratory Information Systems (LIMS), pathology images in FDA-approved image platform, and genomic data in omic repositories. In our own laboratory, which is small in comparison to many reference laboratories, we generate huge amounts of data. Our next-generation sequencers (NGS) can generate over 300GB of genomic data in 24 hours and our whole-slide image (WSI) glass slide scanner in can generate over 600GB of digitized tissue image data in nearly half that time. Huge amounts of data also exist in PACS, LIMS, and other operational systems. As is often the case in clinical settings, huge amounts of data are distilled into human-readable interpretations, which becomes the accessible source of record, while raw data is maintained in the deriving systems. The problem is that many areas of research require access to both raw data and clinical reports. It is often the practice that research data must be extracted from operational systems, duplicating the data making huge storage requirements even larger. In addition, transdisciplinary efforts across omic, imaging, and medical record might require all data components to be accessible for training within the same platform, once again creating duplications and increasing data needs.

Curating data for AI training requires not only medical knowledge, but also understanding of computer models, access to multi-dimensional data sources, and experiencing dealing with huge volumes of data. These are barriers that few practitioners or scientists are able to overcome. Under current informatic practices, in order for researchers to develop AI methods for transdisciplinary research in medicine, they must have programmatic access to data, methods to correlate records, and expert knowledge to curate the data across domain-specific repositories, which is a serious burden for the researchers and a major barrier to current AI research.

To address the identified challenges of data curation across disparate data sources as a barrier to the advancement of AI in medicine, we have interest in three primary research activities: 1) Development of infrastructure for the ontology-based integration of data across heterogeneous data sources; 2) Development of methods for huge data collection, correlation, and curation in support of AI models consistent with the FAIR (findable, accessible, interoperable and reusable) data principles; 3) Development of high-level services allowing curated data to be processed in a distributed privacy-preserving way.

A paradigm shift has taken place in biomedical research that is characterized by the generation and use of massive patient-derived datasets requiring unprecedented storage and computational demands. Pathology departments generate a significant percentage of actionable patient data and by relation medical research data. Pathology data is diverse, ranging from large data sets generated by NGS and glass slide digitization, to millions of discrete data points generated through laboratory testing. Efforts to harness the massive volumes of data generated by pathology departments is a natural starting location for the broad collection, curation, modeling, and inference of transdisciplinary data across medical and scientific domains.