# NIH's *Big Data to Knowledge* initiative and the advancement of biomedical informatics

doi:10.1136/amiajnl-2014-002666

Lucila Ohno-Machado, *Editor-in-Chief*

Two influential reports on data and computation from advisory committees to the NIH leadership resulted in important initiatives: (1) the report from the Working Group on Biomedical Computing for the Advisory Committee to the Director (ACD) of the National Institutes of Health (NIH) in 1999[i] led to the *Biomedical Information Science and Technology Initiative* (BISTI), and (2) the more recent report from the Working Group on Data and Informatics for the ACD in 2012[ii] led to the *Big Data to Knowledge* (BD2K) initiative.[iii] Several AMIA members participated in this working group. Both reports recommended strong support for data science and computation in biomedical sciences and healthcare.

The BD2K initiative was launched at NIH in 2013 through the development of several focused workshops, calls for proposals for centers of excellence, for a data discovery index, for training programs, and through the creation of the new position of Associate Director of Data Sciences, reporting directly to the NIH director. According to Francis Collins, the charge is to *"lead an NIH-wide priority initiative to take better advantage of the exponential growth of biomedical research datasets, which is an area of critical importance to biomedical research. The era of 'Big Data' has arrived, and it is vital that the NIH play a major role in coordinating access to and analysis of many different data types that make up this revolution in biological information."* The first NIH Director of Data Sciences is a member of AMIA and an elected fellow of the American College of Medical Informatics. Phillip Bourne, former Editor-in-Chief of PLoS Computational Biology, developer of the Protein Data Bank, and Professor of Pharmacology at the University California San Diego, introduces this special *Big Data* focus issue of *JAMIA* with an editorial describing his vision of a 'Digital Enterprise'.

Bourne's vision reflects a new reality in which informatics has moved from the periphery of the healthcare and biomedical research enterprise to the center of action. *JAMIA* has been documenting solutions to data acquisition, management, and knowledge generation, and will be a premier venue for reporting on BD2K and related activities. In this first special issue of *JAMIA* on *Big Data*, we present creative solutions to challenges in data acquisition, organization, and analysis, with a particular emphasis on electronic health record data.

1. *Technical and policy infrastructure for data acquisition, efficient storage, and management.* Articles by LeDuc *et al* (***See page 195***), White *et al* (***See page 379***),[iv] and Sahoo *et al* (***See page 263***) focus on technical infrastructure for research data, and articles by Bloomrosen *et al* (***See page 204***) and Akagu *et al* (***See page 374***) focus on secondary use of healthcare data, from a sociotechnical perspective, which includes privacy concerns. Goldwater *et al* (***See page 280***) describes how the acquisition of electronic health data can be feasible in institutions that compose the U.S. federal safety net. EHRs are often distributed, so techniques for record linkage are important to integrate the data – Kum *et al* (***See page 212***) and Rajasekaran *et al* (***See page 252***) describe algorithms for this task. Additionally, new data modalities are increasingly augmenting the EHR, and some of these data can challenge current organizational structures and storage capabilities. Tenenbaum *et al* (***See page 200***)[iv] discusses standards for 'omics' data, and Li *et al* (***See page 363***) describes a novel algorithm for genomic data compression.
2. *Data processing and organization.* EHR phenotyping, which was the focus of *JAMIA's* December 2013 issue, refers to data processing for accurate characterization of disease status and health conditions using data collected in the process of care. A review by Shivade *et al* (***See page 221***) provides the context for EHR phenotyping, and articles by Tate *et al* (***See page 292***) and Melton *et al* (***See page 299***) describe algorithms for EHR phenotyping based on structured data and resources for structured data derivation from clinical notes. Rosenman *et al* (***See page 345***) focuses on database queries on hospitalizations for acute congestive hearth failure, while Dentler *et al* (***See page 285***) and Perotte *et al* (***See page 231***) focus on quality measures and diagnosis code assignment on EHRs.
3. *Knowledge generation.* The articles by Iyer *et al* (***See page 353***), Friedman *et al* (***See page 308***), and Liu *et al* (***See page 245***) focus on detecting drug-related adverse events based on EHRs and related sources. Huston *et al* (***See page 238***) uses data mining techniques to suggest drug repurposing. Data mining techniques for predicting hospital readmissions, early detection of neonatal sepsis, outcome of septic patients, and changes in hypertension control are presented in articles by He *et al* (***See page 272***), Mani *et al* (***See page 326***), Tagkopouls *et al* (***See page 315***), and Sun *et al* (***See page 337***), respectively.

*Big Data* is a big deal in biomedical research and healthcare. I hope our readers enjoy this special issue and continue to submit the products of *Big Data* initiatives such as BD2K for dissemination through *JAMIA*. In the highly diverse biomedical informatics community, professionals with expertise in library science, statistics, management, computer science, software engineering, natural language processing, and implementation science focus on biomedical and healthcare data. Our community is uniquely positioned to translate these data into actionable knowledge to promote health and advance science.

[i]http://www.bisti.nih.gov/library/june_1999_Rpt.asp (accessed 19 Jan 2014).
[ii]http://acd.od.nih.gov/Data%20and%20Informatics%20Working%20Group%20Report.pdf (accessed 19 Jan 2014).
[iii]http://http://bd2k.nih.gov/ (accessed 24 Jan 2014).

[iv]Watch journal club webinars by White and Italia, and by Tenenbaum and Haendel at http://dbmi.ucsd.edu/display/DBMI/UCSD-iDASH+Journal+Club