Research Techniques Made Simple: An Introduction to Use and Analysis of Big Data in Dermatology

Mackenzie R. Wehner[1], Katherine A. Levandoski[2], Martin Kulldorff[3] and Maryam M. Asgari[2]

Big data is a term used for any collection of datasets whose size and complexity exceeds the capabilities of traditional data processing applications. Big data repositories, including those for molecular, clinical, and epidemiology data, offer unprecedented research opportunities to help guide scientific advancement. Advantages of big data can include ease and low cost of collection, ability to approach prospectively and retrospectively, utility for hypothesis generation in addition to hypothesis testing, and the promise of precision medicine. Limitations include cost and difficulty of storing and processing data; need for advanced techniques for formatting and analysis; and concerns about accuracy, reliability, and security. We discuss sources of big data and tools for its analysis to help inform the treatment and management of dermatologic diseases.

Journal of Investigative Dermatology (2017) **137,** e153–e158; doi:10.1016/j.jid.2017.04.019

CME Activity Dates: 20 July 2017
Expiration Date: 19 July 2018
Estimated Time to Complete: 1 hour

Planning Committee/Speaker Disclosure: All authors, planning committee members, CME committee members and staff involved with this activity as content validation reviewers have no financial relationship(s) with commercial interests to disclose relative to the content of this CME activity.

Commercial Support Acknowledgment: This CME activity is supported by an educational grant from Lilly USA, LLC.

Description: This article, designed for dermatologists, residents, fellows, and related healthcare providers, seeks to reduce the growing divide between dermatology clinical practice and the basic science/current research methodologies on which many diagnostic and therapeutic advances are built.

Objectives: At the conclusion of this activity, learners should be better able to:

- Recognize the newest techniques in biomedical research.
- Describe how these techniques can be utilized and their limitations.
- Describe the potential impact of these techniques.

CME Accreditation and Credit Designation: This activity has been planned and implemented in accordance with the accreditation requirements and policies of the Accreditation Council for Continuing Medical Education through the joint providership of William Beaumont Hospital and the Society for Investigative Dermatology. William Beaumont Hospital is accredited by the ACCME to provide continuing medical education for physicians.
William Beaumont Hospital designates this enduring material for a maximum of 1.0 *AMA PRA Category 1 Credit(s)*™. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Method of Physician Participation in Learning Process: The content can be read from the Journal of Investigative Dermatology website: http://www.jidonline.org/current. Tests for CME credits may only be submitted online at https://beaumont.cloud-cme.com/RTMS-August17 – click 'CME on Demand' and locate the article to complete the test. Fax or other copies will not be accepted. To receive credits, learners must review the CME accreditation information; view the entire article, complete the post-test with a minimum performance level of 60%; and complete the online evaluation form in order to claim CME credit. The CME credit code for this activity is: 21310. For questions about CME credit email cme@beaumont.edu.

WHAT ARE BIG DATA?

Big data are commonly defined as data so large or complex that traditional data processing and analytic approaches are inadequate. The *3 Vs* that characterize big data are volume (amount of data), velocity (speed at which data are generated and processed), and variety (types of data) (Laney, 2001), all of which have been growing rapidly (Figure 1). Although there is no predefined threshold for volume, in general, anything 1 petabyte (10^{15} bytes, or the approximate size of 1 million human genomes) or greater is considered big data (Figure 2). The ability to monitor, record, and store information from large populations from

[1]*Department of Dermatology, University of Pennsylvania, Philadelphia, Pennsylvania, USA;* [2]*Department of Dermatology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA; and* [3]*Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA*

Correspondence: Maryam M. Asgari, Department of Dermatology, Massachusetts General Hospital, 50 Staniford Street, Suite 230A, Boston, Massachusetts 02114, USA. E-mail: harvardskinstudies@partners.org

SUMMARY POINTS

- Big data describes any collection of datasets whose size and complexity exceeds the capabilities of traditional data processing applications.
- Big data has the potential to help inform the treatment and management of dermatologic diseases through improved risk assessment, surveillance, diagnosis, and treatment methods.
- While big data presents spectacular research opportunities, there are important limitations to consider, including storage costs, processing challenges, and concerns about accuracy, reliably, and security.

sources including electronic medical records, insurance claims, surveys, disease registries, biospecimens, apps and social media, the internet, and personal monitoring devices has shepherded the era of big data into use in health care. The volume of health care data in the United States in 2017 is rapidly approaching zettabyte levels (iHT2, 2013). This wealth of structured and unstructured data has the potential to substantially affect health care delivery through improved risk assessment, surveillance, diagnosis, and treatment methods.

WHAT ARE SOME BIG DATA SOURCES IN HEALTH CARE?

There are many big data sources in health care. OptumLabs (https://www.optumlabs.com), an open collaborative research center, provides de-identified clinical data from electronic health records and claims data for over 100 million insured members (Borah, 2016). Sentinel (https://www.sentinelinitiative.org), a US Food and Drug Administration initiative, uses data from electronic health records, insurance claims, and registries to monitor postmarketing, real-world safety of medicines. Sentinel data were used to estimate the validity of *International Classification of Diseases—Ninth Revision* codes (Centers for Disease Control, 1998) for ascertaining Stevens-Johnson syndrome and toxic epidermal necrolysis in 12 collaborating research units, covering almost 60 million people (Davis et al., 2015). UK Biobank and Kaiser Permanente Biobank are examples of medical data and tissue samples collected for research purposes. UK Biobank (www.ukbiobank.ac.uk) is a cohort of 500,000 participants in the UK who have provided baseline information and blood, urine, and saliva samples and who are being followed prospectively through their regular care. The Kaiser Permanente Research Biobank (https://www.dor.kaiser.org/external/DORExternal/rpgeh) is composed of 220,000 health plan members who have contributed genetic and electronic health record data. This was recently used in a large genome-wide association study of cutaneous squamous cell carcinoma, which identified 10 single-nucleotide polymorphisms associated with cutaneous squamous cell carcinoma at genome-wide significance and provided new insights into the genetics of heritable cutaneous squamous cell carcinoma risks (Asgari et al., 2016). For genomic data, such as those found in biobanks, the National Center for Biotechnology Information has developed the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo), which acts as a public archive and repository of microarray, next-generation sequencing, and high-throughput functional genomic data. Geographic information systems, such as the National Cancer Institute Geographic Information Systems and Science for Cancer Control (https://gis.cancer.gov), capture geographic data that allow for mapping of disease trends. Solar UV radiation data are available through this system, and the association between cutaneous melanoma incidence rates and county-level UV exposure has been examined (Richards et al., 2011).

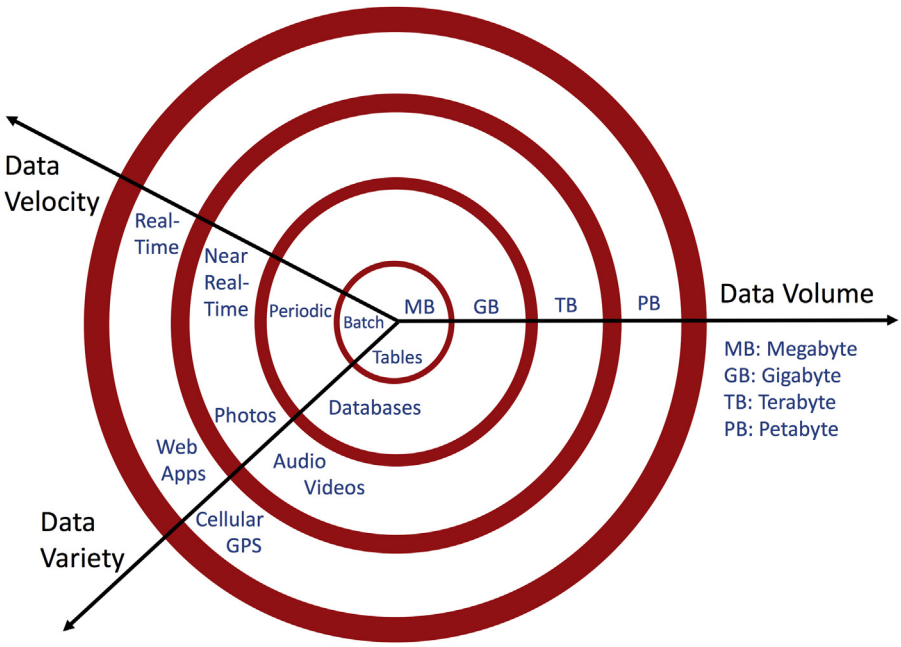

Figure 1. The *3* Vs of big data. The *3* Vs of big data are volume (amount of data), velocity (speed at which data is generated), and variety (number of types of data), all of which have been growing rapidly. After "The 3Vs That Define Big Data," Diya Soubra, Data Science Central, http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data. GPS, global positioning system.