The Rise of Big Data in Oncology

Kristen L. Fessele

Objectives: *To describe big data and data science in the context of oncology nursing care.*

Data Sources: *Peer-reviewed and lay publications.*

Conclusion: *The rapid expansion of real-world evidence from sources such as the electronic health record, genomic sequencing, administrative claims and other data sources has outstripped the ability of clinicians and researchers to manually review and analyze it. To promote high-quality, high-value cancer care, big data platforms must be constructed from standardized data sources to support extraction of meaningful, comparable insights.*

Implications for Nursing Practice: *Nurses must advocate for the use of standardized vocabularies and common data elements that represent terms and concepts that are meaningful to patient care.*

Key Words: *electronic health records, meaningful use, artificial intelligence, neoplasms.*

The term "big data" first appeared in the literature in 1997 by researchers at NASA as they described the challenges to store the volume of information generated as a result of a new, data-intensive type of computational work.[1] In 2008, a white paper entitled "Big-Data Computing: Creating revolutionary breakthroughs in commerce, science and society," highlighted the rapid integration of data-driven strategies across settings ranging from Wal-Mart's (then) 4 petabyte (4000 trillion bytes) data warehouse to the 15 petabytes of data projected to be generated annually by the Large Hadron Collider particle accelerator project,[2] and is credited with widespread adoption of the term.[3]

As the pace of data generation has increased, challenges beyond the ability to simply store huge volumes of data have become apparent, including the ability to effectively manage and successfully analyze the information collected.[4] The ability to gain insight from within these vast depots of data has become the new challenge across disciplines, including health care.

Knowledge generation in health care has traditionally arisen from time-intensive activities such as prospective clinical trials, with a frequently cited lag of 17 years from the initial dissemination of

Kristen L. Fessele PhD, RN, ANP-BC, AOCN®: *Nurse Scientist, Memorial Sloan Kettering Cancer Center, Office of Nursing Research, New York, NY.*

Address correspondence to Kristen L. Fessele, PhD, RN, ANP-BC, AOCN®, Memorial Sloan Kettering Cancer Center, Office of Nursing Research, 205 East 64th St. Room 251, New York, NY 10065. e-mail: fesselek@mskcc.org

research results to implementation in clinical practice.[5] To counter this, in 2007, the Institute of Medicine published The Learning Healthcare System,[6] stressing the need for a new paradigm to more rapidly and continuously integrate the best evidence from both rigorous clinical research with knowledge gained as a "natural outgrowth of patient care. . . to ensure innovation, quality, safety and value in health care."[6(p. 37)]

Key to the support of this shift was the increasing adoption of electronic health records (EHRs), in part spurred as a result of the HITECH Act,[7] which promoted their uptake and outlined a 5-year, multi-stage plan to implement standards for "meaningful use" intended to increase data capture and sharing, improve clinical workflow, and eventual demonstration of improved patient outcomes.[8] While EHRs represent tremendous opportunity to capture and share longitudinal patient information across multiple primary care and specialized settings, these benefits have been difficult to realize. Even within a single health care system utilizing the same EHR, HIPAA (Health Insurance Portability and Accountability Act) requirements and lack of standardized documentation practices continue to limit real-time access to patient information.[9] At a population level, these tools on their own are not suitable to organize or allow easy retrieval of data for quality reporting or research purposes; currently, individual EHR products predominantly function as a data capture tool and repository for patient information, without a deep analytic component.

TYPES OF AND CHALLENGES RELATED TO DATA SOURCES

Day to day patient care activities generate enormous amounts of data as a natural byproduct of interactions within the health care system. Clinical and demographic data can encompass not just provider documentation, but also information from computerized provider order entry (CPOE) and clinical decision-making software, laboratory and radiology results (both written reports and the digital imaging files themselves), automated output from patient monitoring devices, and registration and financial data. This information may be classified into two major categories; structured and unstructured data (Table 1).

Structured data are typically those that can be organized into searchable tables, usually generated from a pre-planned set of answer choices. The values may be selected from a drop-down menu by a clinician as they interact with an EHR, for example, or may include a list of diagnostic or billing codes. Unstructured data, which may comprise up to 80% of all health care information, are those that cannot be as easily searched or organized, such as responses entered into free-text fields by patients or clinicians, narrative notes, hand-written or scanned documents, images, etc.[11]

As the concept of big data has been adopted across various fields, some underlying principles have evolved, such as a growing list of characteristics that may describe very large data sets. For example, what began as a classic list of three "V"s

TABLE 1.
Examples of structured versus unstructured data

Type of data	Examples	Comments
Structured	Age Treatment codes (ie, HCPCS) Disease codes (ie, ICD) Lab results	Data that are in coded format can be retrieved, aggregated, and compared more easily, but may lack context about the clinical situation
Unstructured	Scanned documents Handwritten notes Narrative text (eg, visit notes, procedure and diagnostic imaging reports, e-mail correspondence, etc.) Images (eg, film or digital CT scan files, etc.)	Narrative text generated by clinicians or patients may be a more rich source of information about treatment decision-making rationale, etc. Challenges remain to extract pertinent information from unstructured sources

Data from Raghupathi et al.[10]