Big Data

Big data is the term commonly applied to massive data sets that are generated through a variety of sources, including environmental and body sensors, mobile devices, electronic medical records, imaging and laboratory studies, and administrative claims data. The National Institutes of Health (2015) described big data as referring to:

the complexity, challenges, and new opportunities presented by the combined analysis of data. In biomedical research, these data sources include the diverse, complex, disorganized, massive, and multimodal data being generated by researchers, hospitals, and mobile devices around the world. (para. 1)

The promise of big data is that it may provide an opportunity to create strong linkages between the causes and outcomes of disease states, allow for the development of more precise drug therapies to achieve the goals of personalized medicine, and improve the understanding of health behaviors (Khoury & Ioannidis, 2014).

However, with promise comes the possibility of peril. The existence of large data sets from a variety of sources has the potential to create false alarms by identifying presumed associations with disease outcomes. Often, data elements within specific databases are not well understood. When data elements from several sources are combined, clarity of meaning can be further eroded and an assumption exists that equivalence can be imputed (Wigan & Clarke, 2013). Further, there may be an increase in spurious correlations and associations that occur statistically but have no clinical relevance. A key concern is that “data are observational in nature and are fraught with many biases such as selection, confounding variables, and lack of generalizability” (Khoury & Ioannidis, 2014, p. 1054).

Within the discipline of nursing, it is important to note that of the amount of data nurses collect and record in electronic health records, very little is available for big data analysis (Keenan, 2014). Further, most federal administrative claims databases do not contain nursing data. Thus, one priority for nurse leaders is to “understand structured data and processes that support analytic methods to control costs and improve patient quality and safety” (Westra et al., 2015, p. 305). This has important implications for nursing education in course design and in clinical application exercises. The importance of using standardized terms to describe nursing phenomena of interest; using evidence-based, standardized assessments; and discouraging the use of free-text documentation should be emphasized from the beginning of undergraduate programs and throughout the educational process (Healthcare Information and Management Systems Society, 2015). The American Nurses Association has recognized terminologies that meet data standards (Westra, Delaney, Konicek, & Keenan, 2008). These must be used to accelerate nursing knowledge development in the big data era.

At the graduate level, the challenge of big data requires learning new methods of inquiry and analysis known as data science. According to Brennan and Bakken (2015):

Data science is both a philosophy and a set of techniques to address data that are openly accessible and distributed across multiple locations, explored and analyzed with sharable routines, and fraught with uncertainty, such that data provenance (the trace of the data source and all subsequent modifications) is as important in big data explorations as precise variable definitions are in traditional inquiry. (p. 478)

Given the variety of data sources and new techniques, data science is most often conducted by interprofessional teams. Thus, graduate programs must prepare researchers not only in new analytic methods but also in team-based research processes.

Brennan and Bakken (2015) also stated:

The key starting point to take advantage of the opportunities afforded by big data and data science methods, as well as to advance the methods, is to begin with asking good, well-focused questions. (p. 481)

Questions that are guided by, but not constrained by, theory will facilitate discovery of meaningful relationships within a data set, instead of spurious associations. When standardized nursing data are included in these analyses, insights into patient response to symptom management modalities will emerge.

The privacy and security of health information will remain a challenge as big data moves forward. Health professionals are accustomed to securing data collected for care and research purposes. However, big data comes from sources that have not been routinely subjected to regulation. With individuals both knowingly (through online portals such
as PatientsLikeMe®) and unknowingly sharing health and behavioral data (especially through social media and a variety of mobile fitness applications), further discussions and delineation of best practices will be required.

Ownership of big data is complicated. Those who organize the data typically describe the rights or access to data versus ownership (Wigan & Clarke, 2013). Ownership is complicated when health records are considered, given that those records tend to be personal in nature. Even when data are de-identified, it takes few data points to attribute a record to an individual, especially given the array of sensors and data points that may contribute to the database. Rudimentary secondary data analysis typically involves one or a small number of databases where ownership is defined. Multiorganizational federated databases tend to ascribe ownership to contributors instead of to the data aggregators. Thus, data rights remain with owners, and data usage agreements are defined to facilitate analysis and generation of new knowledge (Wigan & Clarke, 2013).

Big data holds great promise for the generation of nursing knowledge. It is imperative to prepare nursing students to take advantage of emerging practices and techniques to advance the profession.

References

Amy J. Barton, PhD, RN, FAAN
Assistant Editor
The author has disclosed no potential conflicts of interest, financial or otherwise.
doi:10.3928/01484834-20160216-01