available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

Platinum Opinion

Harnessing Big Data for Health Care and Research: Are Urologists Ready?

Khurshid R. Ghani [a,b,*], *Kai Zheng* [c], *John T. Wei* [a], *Charles P. Friedman* [d]

[a] *Dow Division of Health Services Research, Department of Urology, University of Michigan, Ann Arbor, MI, USA;* [b] *Veterans Affairs Healthcare System, Ann Arbor, MI, USA;* [c] *Department of Health Management and Policy, School of Public Health, University of Michigan, Ann Arbor, MI, USA;* [d] *Department of Learning Health Sciences, Schools of Information and Public Health, University of Michigan, Ann Arbor, MI, USA*

"I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."—Alan Turing

Although we have not reached the heights of artificial intelligence predicted by Alan Turing [1], the British mathematician widely regarded as the father of theoretical computer science, there is no doubt that the last two decades have witnessed an unprecedented integration of information technology in our daily lives. Health care has not been immune to this change. The most obvious examples of this transformation have been the adoption of electronic health records (EHRs) and the use of picture archiving and communication systems (PACS) for radiologic images. Advances such as these have led to the creation of vast quantities of digital data—data sets so diverse and large that they have been given the term *big data*.

In this editorial, we introduce the field of big data in health care. The corporate sector leads in this arena, for which big data analytic techniques have helped identify consumer trends that guide product development and marketing. In particular, we discuss the implications of big data for urologic care and research, and we highlight recent examples of big data analytics using EHRs, registries, and social media.

Big data methods require advanced techniques and technologies to enable their capture, storage, distribution, management, and analysis. The three hallmarks of a big data set are volume, variety, and velocity (Fig. 1). *Volume* refers to the vast amounts of health-related data that are created and permanently stored. In health care, many factors contribute to the increase in data *variety*. The sources can reflect the totality of the patient experience, encompassing structured and unstructured data from EHRs; PACS; sensor-enriched medical devices; patient-reported outcomes; financial transactions; social media posts, including Twitter feeds and Facebook messages; and online patient forums. Finally, the data are often accumulated in real time and at a very high rate, or *velocity*.

Scrutinizing these data may uncover associations, patterns, and trends with the potential to advance patient care and lower costs [2]. It has been estimated that big data initiatives could save $300 billion per year in health care spending in the United States alone [3]. Furthermore, real-time analysis of these data sets may provide information that can guide management decisions and predict outcomes. However, the scale and complexity of the data mean that they cannot be sufficiently managed with existing tools and methods.

It is important to distinguish familiar methods of secondary data analysis from big data. Secondary data analysis pertains to any use of a data set for a purpose different from that which motivated its original collection. In health care, secondary analysis is often performed to examine care quality using data originally collected for billing purposes, but secondary analysis can also be performed on data originally collected in EHR systems to document care. Data sets used for secondary analysis may be of any size, and many are prospectively assembled for specific purposes. Very large secondary data sets, approaching the scale of big data, have been the source for many population-based studies in urology [4]. Marketscan (Truven Health Analytics, Ann Arbor, MI, USA) is one of the largest of these data sets, with an approximate size of

* Corresponding author. Department of Urology, University of Michigan, NCRC Building 16, 115W, 2800 Plymouth Road, Ann Arbor, MI 48109, USA. Tel. +1 734 615 4034; fax: +1 734 232 2400.
E-mail address: kghani@med.umich.edu (K.R. Ghani).

http://dx.doi.org/10.1016/j.eururo.2014.07.032
0302-2838/Published by Elsevier B.V. on behalf of European Association of Urology.

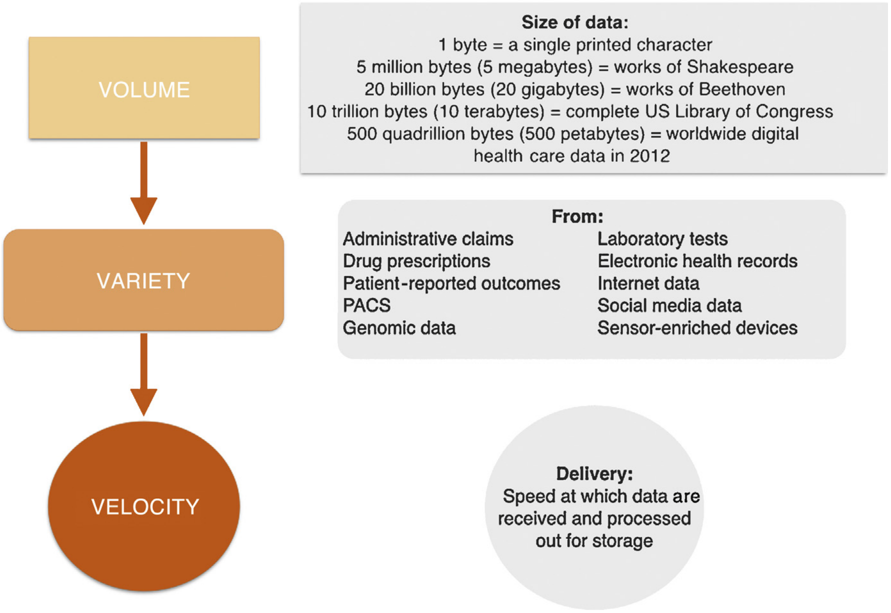
VOLUME
Size of data:
1 byte = a single printed character
5 million bytes (5 megabytes) = works of Shakespeare
20 billion bytes (20 gigabytes) = works of Beethoven
10 trillion bytes (10 terabytes) = complete US Library of Congress
500 quadrillion bytes (500 petabytes) = worldwide digital health care data in 2012
VARIETY
From:
Administrative claims
Drug prescriptions
Patient-reported outcomes
PACS
Genomic data
Laboratory tests
Electronic health records
Internet data
Social media data
Sensor-enriched devices
VELOCITY
Delivery:
Speed at which data are received and processed out for storage

Fig. 1 – The 3Vs of big data: high volume, high variety, and high velocity. PACS = picture archiving and communication systems.

3.5 terabytes (10^{12} bytes), and it has been used to explore provider-level variation in urology [5].

The term *big data* refers to ultralarge bodies of data that have not been prospectively limited in size or scope by the intent to address specific research questions or disease conditions, and that grow continuously and rapidly. There is no defined size of a data set that qualifies it as big data, but these data sets often exceed 1 petabyte (10^{15} bytes). Data stores approaching this size demand unique methods for management and advanced analytics to process them for useful insights. The pooled raw data have to be made ready for storage and processing using one of many technologies, including cloud-based applications, data warehousing, or massive parallel processing by way of multiple computational nodes [6]. Analytic techniques include natural language processing, machine learning, and data mining, which are all inherently different from currently used statistical methods for health services research. Machine learning is a branch of artificial intelligence concerned with the construction and study of systems that can learn from data. Machine learning focuses on prediction, based on known properties learned from sets of training data. In contrast, data mining focuses on the discovery of unknown regularities embedded in the data. Data mining and machine learning are key elements in the business models of Google and Facebook. Google has even applied these algorithms to detect influenza epidemics using search queries [7].

The EHR enables big data in health. As patient care is routinely documented in EHRs, the data are fed to large repositories that, as they grow in size and scope, become big data resources. One of the largest EHR-derived repositories is that managed by the US Department of Veterans Affairs (VA) for its integrated health care system. This veritable arsenal contains >30 million patient records, including 3.2 billion clinical orders, 1.8 billion prescriptions, and 2 billion clinical text notes. This hybrid body of structured and unstructured data can be analyzed using automated clinical reasoning, natural language processing, and predictive analytics to identify risks and improve patient care. In one study using the VA data warehouse, sociodemographic characteristics, medical conditions, vital signs, medications, and laboratory tests of 4.6 million patients were analyzed to accurately identify patients at increased risk of hospitalization or death [8]. The potential to study urologic diseases in this population through access to specific laboratory values (eg, chronic kidney disease outcomes in patients undergoing radical or partial nephrectomy) is enormous.

EHRs can also be mined to guide clinical decisions in almost real time. When presented with a patient who had a rare combination of conditions, generating extreme uncertainty over the best treatment strategy, physicians at the Lucile Packard Children's Hospital (Palo Alto, CA, USA) used their institutional EHR-based data warehouse to perform an automated cohort review of patients with similar presentation [9]. The analysis, which took <4 h, provided very helpful suggestions for the best clinical pathway for that patient. It is conceivable that similar real-time big data analytics may allow urologists to identify patients at risk of readmission after cystectomy or to determine whether a 75-yr-old patient with intermediate-risk prostate cancer is better served with radiation therapy or radical prostatectomy, based on outcomes of patients with similar profiles in the same institution.

Registries such as the National Institutes of Health (NIH) Surveillance Epidemiology and End Results program [10] have been used for these types of studies, but these registries have a significant limitation. Registries are constructed to obtain data on defined outcomes in advance, making it difficult to address unforeseen questions, conduct studies across morbidities, or collect new information at a later date. Data collection may rely on manual extraction, with the potential for missing information [11]. Using the power of big data analytics to interface registries with EHRs may address some of these limitations. The soon-to-be-launched American Urological Association Quality Registry aims to provide a comprehensive clinical registry by eliminating the burden of data entry by way of natural language extraction from the patient's electronic records [12]. Starting with prostate cancer, the quality of care within and between institutions can be compared and benchmarked against national data.

Similar big data approaches have been applied to cancer genomics and health-related social media. The American Society of Clinical Oncology is developing CancerLinQ, a platform that can use information from EHRs, eventually including genomic data, to provide a targeted approach for cancer therapy [13]. Using social media, attitudes and sentiments toward public health matters may be gauged by analyzing millions of tweets using supervised machine learning methods [14]. More important, EHRs and big data analytics may provide the opportunity to create "learning health systems," in which physicians learn from each patient at every visit and close the feedback loop for clinical decision making in real time [15]. To adapt to these new approaches, urologists will have to ask the right questions. Familiar analytic methods may have to be supplemented with advanced analytics of machine learning or data mining. In the United States, initiatives such as the NIH Big Data to Knowledge program aim to train a new generation of physicians and researchers in these techniques [16].

It is clear that big data is not a panacea for all the unanswered questions regarding the delivery, access, and quality of health care. Increases in data size do not reduce possible bias in the data set. Great care must be exercised in drawing inferences of causation from statistical association. The true causal factor may be a variable highly correlated with a variable that is included in the data set. This problem, known as *unmeasured confounding*, is being addressed by researchers, but much work in this area remains [17]. The challenge will be in determining the essential elements and interactions that matter from data that may be unstructured and may have significant background noise. It can also be argued that big data can never address the required detail when dealing with inherently complex systems such as the human body and its disease processes. Yet proponents of big data argue that this very issue of complexity has been addressed successfully with computational science in other fields. The reality is likely to lie somewhere in between [18]. Finally, if big data analytics proves to be a source of new knowledge, new methods will be required for curating and archiving that knowledge, as well as for making it accessible at the point of care for decision making. Big data must not become an end in itself.

Conflicts of interest: The authors have nothing to disclose.

References

[1] Turing AM. Computing machinery and intelligence. Mind 1950;59:433–60.

[2] Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309:1351–2.

[3] Groves P, Kayyali B, Knott D, Van Kuiken S. The "big data" revolution in US healthcare. McKinsey& Company Web site. http://healthcare.mckinsey.com/big-data-revolution-us-healthcare. Published April 2013.

[4] Schlomer BJ, Copp HL. Secondary data analysis of large data sets in urology: successes and errors to avoid. J Urol 2014;191:587–96.

[5] Hollingsworth JM, Wolf Jr JS, Faerber GJ, Roberts WW, Dunn RL, Hollenbeck BK. Understanding the barriers to the dissemination of medical expulsive therapy. J Urol 2010;184:2368–72.

[6] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inform Sci Systems 2014;2:3.

[7] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012–4.

[8] Wang L, Porter B, Maynard C, et al. Predicting risk of hospitalization or death among patients receiving primary care in the Veterans Health Administration. Med Care 2013;51:368–73.

[9] Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med 2011;365:1758–9.

[10] Hu JC, Gandaglia G, Karakiewicz PI, et al. Comparative effectiveness of robot-assisted versus open radical prostatectomy cancer control. Eur Urol 2014;66:666–72.

[11] In H, Bilimoria KY, Stewart AK, et al. Cancer recurrence: an important but missing variable in national cancer registries. Ann Surg Oncol 2014;21:1520–9.

[12] AUA Quality (AQUA) Registry. American Urological Association Web site. http://www.auanet.org/resources/quality-registry.cfm.

[13] Savage N. Bioinformatics: big data versus the big C. Nature 2014;509:S66–7.

[14] Salathe M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Computat Biol 2011;7:e1002199.

[15] Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010;2(57), cm29.

[16] Vieweg J. Big data in biomedical research. AUA News 2014;19:21.

[17] Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. BMJ 2009;338:b81.

[18] Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science 2014;343:1203–5.