Liberating Big Data:
Implications for a Traditional Consent Approach

Mark Repenshek, Ph.D.

Editor's Note: A version of this article was presented by Dr. Repenshek at CHA's Theology and Ethics Colloquium in March 2019.

The explosion of digitalization across industry has not escaped health care. One can see the explosion of health information in electronic medical records, health care information systems and handheld, wearable and smart devices. This has resulted in an exponential increase in the amount and variety of data, including sociodemographic data, data from social media, wearables and sensors, insurance claims and traditional clinical data. Current estimates by McKinsey Global Institute suggest, if used effectively, big data in U.S. health care could create a value of more than $300 billion every year, of which two-thirds would be in the form of reducing health care expenditures by about 8%.[1] This optimization will not be realized, however, if these data sit in unstructured or semi-structured form, without interoperability. The goal is to move to a completely fluid data system generating a "circulation of data" that has the potential to identify patterns that lead to improved health care quality, reduced costs and enable timely decision making.[2]

Industry also sees the potential opportunity toward these ends. By the middle of the decade, Health IT Analytics reports "health care organizations will join their peers from other industries spending $18 billion on deep learning technologies to analyze images, extract meaning from unstructured data, and support decision-making. Another $9 billion on business intelligence tools by 2023" and a staggering "$1.4 billion in blockchain vendors" by 2024."[3] The global health care blockchain market sat at a mere $34 million in 2017. Although some warn of a similar financial flood to that of the EMR with more hype than hope for value, the dollars are already flowing toward Big Data Analytics.

BACKGROUND

Big Data[4] was a term first coined in 2000 by Francis Diebold referencing an impending "explosion in the quantity of available and *potentially* relevant data."[5] Later in 2001, Doug Laney developed the key characteristics of Big Data, commonly known as the 3Vs—volume, velocity and variety.[6] Manyika et al., from the Global Institute regarded *value* as the fourth V and Feldman et al., added *veracity* as the fifth in the mantra when referring to health care.[7] The 5-Vs are often regarded today as the key characteristics of Big Data:

> *Volume*—refers to the quantity of Big Data in health care, which is estimated to increase dramatically to 35 zettabytes

by 2020[8] (that is 10^{21} or 1 and 22 zeros…I had to look it up).

Variety—refers to the different types of health care Big Data collected including heterogenous characteristics and the structured and unstructured nature of medical data.

Velocity—is the speed of data generation (i.e., real-time patient data) as well as data collection.

Veracity—refers to sources that influence accuracy such as inconsistencies, missing data, ambiguities, deception, fraud, duplication, span and latency (of acute concern in the use of Big Data in health care).

Value—represents cost-benefit to the decision maker through the ability to take meaningful action based on insights derived from data.[9]

In clinical practice, the 5-Vs translate to early detection of disease; accurate prediction of disease trajectory; identification of deviation from healthy state; changed disease trajectories; and detection of fraud. Clinicians would be equipped to personalize predictions, target treatment and tailor cost effectiveness algorithms to detect relatively low-frequency events that nonetheless may have significant clinical impact at an individual level. On an organizational level, the 5-Vs translate to "pinpointing patients who are the greatest consumers of health resources or at the greatest risk for adverse outcomes; identifying the treatments, programs and processes that do not deliver demonstrable benefits or cost too much;

The 5-Vs translate to early detection of disease; accurate prediction of disease trajectory; identification of deviation from healthy state; changed disease trajectories; and detection of fraud.

reduce readmissions by identifying environmental or lifestyle factors that increase risk or trigger adverse events and adjusting treatment algorithms accordingly; improving outcomes by examining vitals from at-home health monitors; managing population health by detecting vulnerability within patient populations during disease outbreaks or disasters; and bringing clinical, financial and operational data together to analyze resource utilization productively and in real time."[10]

ETHICAL ISSUES IN UTILIZING BIG DATA

These hoped for outcomes for Big Data analytics leads to questions concerning data management, privacy protection and oversight mechanisms. Prior to an examination of these questions, there are challenges that relate more directly to the composition of the data repositories themselves and the tools used to "liberate the data itself."[11] Big Data analytics utilize unstructured data sets, rarely clean data and models that often lack clearly defined underlying assumptions. As a result, the quality of evidence derived from digital health research employed to clinical ends is substantially impacted. Consider also the extent to which information on ethnicity, age, gender, socioeconomic status and geographical

distribution from unstructured datasets outside of the carefully specified criteria of current human subjects research protocols—say from user-oriented digital health applications—limits the generalizability and specificity of findings.[12] These simple examples illustrate the need for some sort of standards and possibly reference datasets to address reproducibility and veracity. Policy development also appears crucial relative to user-oriented digital health applications in order to enhance transparency of use and accountability of data integrity.

Although there is the acknowledged benefit of larger, more representative and diverse databases that are expected to address the issue of external validity that plagues randomized controlled trials,[13] there is a fundamental methodological shift at play with Big Data analytics. Clinical trials focus on hypothesis testing, whereas tools that mine large data repositories focus on hypothesis generating correlations between phenomena.[14] The implications of this shift remain to be examined, but it seems that even if the approach proves effective in establishing robust correlations for clinical intervention, there remains an important place for "controlled interventional, randomized trials on stratified patient cohorts to establish the safety and clinical utility of novel therapies or public health interventions."[15]

Finally, there is the matter of the traditional rubrics of privacy and security that are frequently cited as concerns in Big Data analytics. There is the obvious concern that as "more data sources become available and advanced analytics can be applied for various purposes,"[16] the matter of protecting privacy becomes increasingly complex. What creates such a complex environment for privacy in the area of Big Data analytics is, in part, the fact that traditional mechanisms for protecting the information of individuals are stretched to their limits, if not rendered unhelpful. Take, for example, the idea of anonymization as one such mechanism. Traditionally, the claim that data will be kept anonymous for potential users of a particular technology may no longer be a valid claim (or at least highly unlikely). Current anonymization technologies still leave open the possibility of re-identification given the sophistication of current analytic tools. Additionally, attempts to breach security systems in a variety of settings are becoming commonplace. According to the Breach Portal of HHS Office of Civil Rights on March 6, nearly 20,000 individuals' health care records were affected due to a hacking/IT incident. In February alone 2,112,618 individual health records were affected.[17] Yet a breach may not even be necessary.[18] The ability to utilize data from a variety of disparate repositories, namely through data circulation between individuals, devices and institutions where privacy has not been assured, can re-identify persons with startling accuracy.

An interesting example of data gathering absent breach is offered by Harvard Business School Professor Shoshana Zuboff in her book, *The Age of Surveillance Capitalism,* in which she references the Google-owned Nest thermostat. Two University of London scholars published a detailed analysis that examined if one entered the Nest ecosystem of connected devices and apps, each of which has its own terms of service for third-party data sharing, the purchase of a single device would entail the need to review nearly a thousand so-called "contracts." If a person did not accept the terms and conditions, "the thermostat may be deeply compromised, no longer supported by

the necessary updates meant to ensure its reliability and safety." Recently, Google informed users of the Nest Guard security technology that the device contained a microphone, "although never intended to be a secret," but had erroneously omitted it from the tech specs (the same was found true for Nest's smoke and carbon monoxide alarm).[19]

Although this may seem like a bleak backdrop against which a new approach may be formulated, it remains the case that robust security measures should be mandated and enforced, clearly articulated policy govern the use of data, security systems must be continuously monitored and evaluated for efficacy, and transparency and accountability need to be hallmarks of responses to breaches in privacy. What is essentially at stake is trust. In the case of health care, this means trust in health data access and utilization. Trust in Big Data and Big Data analytics will not come merely as a result of innovative models of consent, but perhaps, in part, through technology itself and perhaps a shift to a different understanding of privacy altogether.

RETHINKING PRIVACY IN BIG DATA ANALYTICS

In rethinking privacy, we should consider blockchain technology. Blockchain is a peer-to-peer distributed, shared ledger technology for transactional applications. It utilizes transparency to build trust. In a health care context, all medical data would be stored off blockchain in a data repository called a data lake. Data lakes support interactive queries, text mining, text analytics and machine learning. All data would be encrypted and digitally signed to ensure privacy and authenticity of the information. The uniqueness of blockchain is that the user would have full access to all data (from those currently sequestered to siloed, unstructured data sets) and control over how data is used or shared. Through one's unique user identifier and dashboard application, the user could see who has permission to access the blockchain, view an audit log (including when and where data was accessed) and give and/or revoke access permissions to anyone. The environment of complete transparency and complete user control allows the individual user to make all decisions about what data is collected and how the data can be shared. Privacy is no longer ensured through multiple touchpoints of consent.

Blockchain technology may help to more positively establish the autonomy of groups and persons to specify the correct relationships that ought to exist between different organizations and associations within society.[20] Specific to health care, the technology may help to create "a single storage location for all health data, tracking personalized data in real-time and the security to set data access permissions at a granular level" [21] to the benefit of both research as well as the individual through personalized medicine. In more simple terms, blockchain technology creates a data environment in which the individual precedes Big Data and Big Data itself exists for the well-being of individuals. Individual rights, including that of ownership, are prior to the analytics and nothing is done by a higher or larger data analytics organization without the action of the individual. Access to information through blockchain technology is an example of the reprioritization of persons relative to their data. With blockchain, each participant is connected to the blockchain network with a secret private key and a public key that serves as an openly visible identifier. The pair is cryptographically linked such that

identification is possible in only one direction—through the individual. Therefore, the blockchain public/private key encryption creates identity permission layers that allow patients to share identity attributes on an as-needed-based only, thereby reducing vulnerabilities from storing PHI on all sides.

To be clear, I am not suggesting that blockchain is *the* next technology by which trust will be secured.[22] Rather, blockchain is an example of how technology can reframe the relationship between the individual and her data. It is also a technology that does not require layers of consent to ensure privacy. Rather, blockchain utilizes the concept of user control to determine where access should be granted in order to build trust related to utilization of the individual's data.

CONCLUSION

Liberating data to a completely fluid system requires trust. Public trust in health data use is of paramount importance. The discussion must evolve from traditional privacy protections brokered through consent intermediaries to privacy protections secured through transparent, consistent, rule-based participation and control. The correct relationship among these elements, I believe, is beyond mere innovations in consent. Use of innovative technologies, like blockchain, may be the very avenue through which trust is built.

Unfortunately, according to a poll of ISACA, a global non-profit offering IT governance leadership and resources, 47 percent of IT professionals say their executives are Big Data "Illiterate." The underlying uncertainty over fundamental analytics competencies in the same poll found that just 20 percent of respondents said that AI will be their top driver of transformation in the next few years…blockchain was at a mere 7 percent. The key takeaway from ISACA's 2017 Digital Transformation Barometer is that there is a direct correlation between digital literacy of an organization's leadership and that organization's overall appetite to examine, test and implement new emerging technologies."[23] We, as a field, need to be on the forefront of this discussion and not be a part of the 47 percent.

Mark Repenshek, Ph.D.
Vice President, Ethics and Church Relations
Ascension
St. Louis, Mo
Mark.repenshek@ascension.org

ENDNOTES

[1] L. Haar, "Big Data Expected to Have Big Impact on Diagnostic Imaging." 2014 accessed at: http://www.diagnositicimaging.com/siim-2014/big-data-expected-have-big-impact-diagnositc-imaging
[2] TB Murdoch and AS Detsky, "The Inevitable Application of Big Data to Health Care." 309 (2016): 5-6.
[3] J Bresnick, "Deep Learning, Blockchain, Big Data to See Huge Growth in Health care." *Health IT Analytics,* December 11, 2018. Accessed at: https://healthitanalytics.com/news/deep-learning-blockchain-big-data-to-see-huge-growth-in-health care
[4] ID Dinov, "Volume and Value in Big Health care Data." *J. Med Stat Inf* 4 (2016), accessed at: http://dx.doi.org/10.7243/2053-7662-4-3
[5] FX Diebold, *Big data dynamic factor models for macroeconomic measuring and forecasting.* (Adv. Eco. Econmo. Eigth World Congr. Econmo. Soc., 2003): 115-122. Accessed at: https://www.sas.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF
[6] D. Laney, "META delta." Appl. Deliv. Strateg. 949 (2001) accessed at: http://dx.doi.org/10.1016/j.infsof.2008.09.005.
[7] B. Feldman, EM Martin, T. Skotnes, "Big Data in health care—hype and hope, Dr. Bonnie 360 degree." *Bus. Dev. Dignit. Heal.* 2012(2013): 122-125. http://www.riss.kr/link?id=A99883549.
[8] Hao Gui, Rong Zheng, C Ma, "An architecture for health care big data management and analysis." International Conf. Heal. Inf Sci (2016): 154-160.
[9] N Mehta and A Pandit, "Concurrence of big data analytics and health care: A systematic review." *Int. J. Med. Info.* 114 (2018): 57-65, 59.
[10] W Raghupathi and V Raghupathi, "Big data analytics in health care: promise and potential." *Heal. Inf. Sci. Syst.* 2(2014): 3. Accessed at: http://dx.doi.org/10.1186/2047-2501-2-3.
[11] V Effy, H Tobias, A Afua, B Alessandro, "Digital Health: Meeting the Ethical and Policy Challenges." *Swiss Medical Weekly* 148 (2018): 1-9, 3; MJ Khoury and JP Evans. "A public health perspective on a national precision medicine cohort: balancing long-term knowledge generation with early health benefit." *JAMA* 313, 21 (2015): 2117-8.
[12] Effy, et al., 3.
[13] PM Rothwell. "External validity of randomized controlled trials: "To whom do the results of this trial apply?" *Lancet* 365, 9453 (2005): 82-93.
[14] Choong Ho Lee and Hyung-Jin Yoon, "Medical big data: promise and challenges." *Kidney Res Clin Pract* 36 (2017): 3-11, 5
[15] Effy, et al., 3.
[16] Effy, et al., 3.
[17] U.S. Department of Health and Human Services, Office for Civil Rights. Breach Portal, found at: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf;jsessionid=EDDA197408BB24351B594EE9538DC377. Accessed on March 6, 2019.
[18] D Gayle, A Topping, I Sample, S Marsh, and V Dodd, "NHS seeks to recover from global cyber-attack as security concerns resurface." *The Guardian* May 13, 2017. Accessed at: https://www.theguardian.com/society/2017/may/12/hospitals-across-england-hit-by-large-scale-cyber-attack
[19] N Bastone, "After a big privacy backlash, Google's Nest explains which of its products have microphones and why." *Business Insider,* February 24, 2019 accessed at: https://www.businessinsider.com/google-nest-products-with-microphones-2019-2.
[20] ME Allsopp, "Principle of Subsidiarity." In, *Catholic Social Thought* QA 49
[21] LA Linn and MB Koo, "Blockchain for Health Data and its Potential Use in Health IT and Health Care Related Research." Accessed on March 3, 2019 found at: https://www.healthit.gov/sites/default/files/11-74-ablockchainforhealth care.pdf
[22] M Orcutt, "Once hailed as unhackable, blockchains are now getting hacked." February 19, 2019, access on March 27, 2019 found at: https://www.technologyreview.com/s/612974/once-hailed-as-unhackable-blockchains-are-now-getting-hacked/.
[23] J Bresnick, "Executives Are Big Data 'Illiterate'" *Health IT Analytics* found at: https://healthitanalytics.com/news/47-of-it-pros-say-their-executives-are-big-data-illiterate. Accessed on February 28, 2019.