Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives

Contribution of the IMIA Social Media Working Group

M. M. Hansen¹, T. Miron-Shatz², A. Y. S. Lau³, C. Paton⁴

¹ School of Nursing and Health Professions, University of San Francisco, San Francisco, California, USA
² Center for Medical Decision Making, Ono Academic College, Kiryat Ono, Israel
³ Centre for Health Informatics, Australian Institute of Health Innovation, University of New South Wales, Australia
⁴ George Institute for Global Health, University of Oxford, Oxford, UK

Introduction

Currently, multiple worldwide enterprises are asking key questions about “Big Data,” which has become a buzzword. For those who are willing to listen, Big Data is offering valuable patterns and predictions in the world today. It is not surprising this concept is recently receiving a lot of attention. According to Asigra, a Cloud Backup company since 1986, a staggering 90% of the data in the world today have been created only during the last two years [1]. And, it is predicted the worldwide number of Internet Protocol (IP) addresses will quadruple by 2015. Moreover, it is forecasted three billion people will be online creating close to eight zettabytes of data two years from now [1]. This amount of data may appear alarming while at the same time interesting when companies such as Google harness personal input data and forecast flu epidemics in collaboration with the Centers for Disease Control (CDC) [2].

Besides the legacy of electronic bulletin boards and listservs we now have large volumes of data being produced by multiple users of social media platforms [3]. While Electronic Medical Records (EMRs) contain a plethora of data, such as patient demographics, clinical and genomic data, and are known for assisting with the flow of health care, today they are seen as a way for performing large-scale and low-cost health care analysis and decision-making. EMR data sharing has its challenges, such as patient privacy and, privacy has to be a high priority in order to comply with the EU Directive 95/46/CE and the HIPAA privacy rule [4].

In regards to the increased use of Social Media tools, an example of Big Data is the fact that “32 billion searches” were performed via Twitter during the month of August 2012 [1]. Atule Butte (@atulebutte) tweets about wearable devices that assist [aspiring] fitness buffs to track their personal data [5]. As wearable devices become more popular and accepted, even for those with poor posture [5], personal quantifiable data will add to the exploding 2.5 quintillion data bytes per day [1]. The increased use of telehealth will further test the storage capacity for patient data and the innovative use of Google Glass by physicians will also add to the social and behavioral aspects of Big Data [6]. The healthcare industry has been slow to embrace Big Data due to the cost of adding analytic functions to existing EHRs, privacy issues, poor-quality data, and a lack of willingness to share data [7]. However, today more professionals are seeing the need to listen and act upon Big Data to benefit health outcomes through online communication and sharing of data. The aim of this paper is to provide the reader a glimpse of the literature centering on the challenges and opportunities in analytics of Big Data in science and health care. We begin by discussing the science of big data and the need to balance between quantity and quality, and then move on to small data and its challenges, which are a small scale reflection of the big data challenges.
The Science of Big Data

Over the past century, scientific advances in medicine have generally been made using a “frequentist” approach to statistical analysis: Samples of populations are studied and the results from the samples are extrapolated to estimate the effects of the intervention being studied. For most types of experiment, sampling data is sufficient to build an effective picture of the entire dataset and, statistically, we can give high levels of accuracy to predictions based on relatively small samples. Data collected in this way is often of very high quality. To ensure the sample is representative and accurate, the data is collected and ‘cleaned’ with great care. This extra care is often very expensive, however, and over the last few decades we have seen the costs of running large randomized control trials spiral upwards.

Big Data offers a potential solution to this issue. Although data produced from such sources as social networking communities, EHR systems, and wearable devices are generally of much lower quality than data carefully collected by researchers looking to answer specific questions, the sheer volume of the data may outweigh their messiness. In addition, there is also a trend to higher quality ‘big data’ collection such as the data produced in genomic analysis and structured data that can be generated from standard-compliant EHR systems. As the percentage of the population being sampled approaches 100%, messy data can have greater predictive power than highly cleaned and carefully collected data that might only be a sample of 1% of the target population for the researcher [8]. The quantity of data alters the way and approaches used to relate, utilize, and understand data.

In addition to just having more data, Big Data also generally refers to the application of machine learning for analyzing the data sets. Machine learning effectively turns the scientific method on its head. Instead of researchers creating a hypothesis and collecting data from a sample of the population, machine-learning algorithms plow through large data sets searching for hypotheses. They do this through a process of brute force classification (finding and matching clusters of correlations in the data) combined with a process of learning and feedback to make the process more efficient. Machine learning algorithms are generally quite simple and are really just looking for associations between different elements of the data.

Because of this, we need to take the results of Big Data machine learning algorithms for what they are: new hypotheses rather than firm predictions. Researchers can test the hypotheses to a limited extent by dividing the datasets or re-running the algorithms on new data collected. But to gather the best evidence on a particular question, it may still be necessary to run a prospective “frequentist”-style trial to test any strong hypothesis that come out of the machine learning process, particularly when trying to answer questions about human health.

Healthcare Sector

While researchers are still debating the definitions and boundaries of Big Data in health, benefits of health-related Big Data have been demonstrated in three areas so far, namely to 1) prevent disease, 2) identify modifiable risk factors for disease, and 3) design interventions for health behavior change [9]. Organizations worldwide are recognizing the Big Data movement and introducing new initiatives for knowledge discovery and data-driven decision-making. For example, the National Institute of Health (NIH) is establishing the Big Data to Knowledge (BD2K) and Infrastructure Plus Program, which provides a shared computational environment (e.g. data standards, ontologies, data catalogues, virtualized cloud computing) to facilitate large-scale biomedical data analysis for the NIH community [10]. Specifically, the NIH US Library of Medicine hosts an impressive set of data sharing repositories [11], which primarily accept submissions of biomedical data and other information sharing systems from NIH-funded investigators. In addition, the United Nations (UN) is launching the Global Pulse project, which advocates for the ‘data philanthropy’ movement by asking organizations and individuals to contribute data, resources, and skills to help understand the impact of UN development programs and ways to improve their outreach on affected populations and regions [12].

Big Data streams in health can be broadly summarized into three categories [13]. Traditional medical data is primarily originated from the health system (e.g. EMRs, personal and family health history, medication history, lab reports, pathology results), where the objective of these analyses is to derive a better understanding of disease outcomes and their risk factors, reduce health system costs, and improve its efficiency [13]. “Omics” data refer to large-scale datasets in the biological and molecular fields (e.g. genomics, microbiomics, proteomics, and metabolomics), where the aim of these analyses is to understand the mechanisms of diseases and accelerate the individualization of medical treatments (e.g. “precision medicine”) [3,6]. As pointed out by Alice Whitmore, in the Stanford Big Data in Biomedicine Conference (2013), genomic testing and mapping could, for example, point to women in high risk of developing breast cancer, which would allow allocating them preventive care, and reduce the need for large scale, potentially hazardous interventions, for other low-risk women [14]. Last but not least, data from social media and the quantified-self movement essentially consist, of signs and behaviors on how individuals (or groups of individuals) use the Internet, social media, mobile applications (apps), sensor devices, wearable computing devices, or other technological and non-technological tools to better inform and enhance their health.

This section presents examples of health-related Big Data projects, with an emphasis on data from social media and the quantified-self movement (Table 1). For big data research related to EMRs, digital enterprise, genetic data and omics sources, readers can refer to the following reviews and perspectives conducted recently [15, 16, 17, 18, 19].
<table>
<thead>
<tr>
<th>Data type</th>
<th>How has it been used in health?</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantified-self data (via devices, self-reporting, or sensors)</td>
<td>Engaged in the self-tracking of signs and behaviors as n=1 individual or in groups, where there is often a proactive stance toward acting on the information [13]</td>
<td>Food consumption [20]</td>
</tr>
<tr>
<td></td>
<td>Provides richer and more detailed data on potential risk factors (biological, physical, behavioral or environmental) [13]</td>
<td>Information diet [21]</td>
</tr>
<tr>
<td></td>
<td>Allows data collection over potentially longer follow-up periods than is currently possible using standard questionnaires [13]</td>
<td>Smile triggered electromyogrph (EMG) muscle to create unexpected moments of joy in human interaction [22]</td>
</tr>
<tr>
<td>Location-based information</td>
<td>Information derived from Global Positioning Systems (GPS), Geographic Information Systems (GIS), and other open source mapping and visualization projects</td>
<td>Weather patterns, pollution levels, allergens, traffic patterns, water quality, walkability of neighborhood, and access to fresh fruit and vegetables (such as supermarkets) [34, 35, 36]</td>
</tr>
<tr>
<td>Twitter (Note: a 2011 study has suggested that 8.5% of English-language tweets relate to illness, and 1.6% relate to health [46])</td>
<td>Assesses disease spread in real-time</td>
<td>Weather patterns, pollution levels, allergens, traffic patterns, water quality, walkability of neighborhood, and access to fresh fruit and vegetables (such as supermarkets) [34, 35, 36]</td>
</tr>
<tr>
<td></td>
<td>Assesses sentiments and moods</td>
<td>Weather patterns, pollution levels, allergens, traffic patterns, water quality, walkability of neighborhood, and access to fresh fruit and vegetables (such as supermarkets) [34, 35, 36]</td>
</tr>
<tr>
<td></td>
<td>Facilitates emergency services by allowing far the wide-scale broadcast of available resources, enabling people in need of medical assistance to locate help</td>
<td>Weather patterns, pollution levels, allergens, traffic patterns, water quality, walkability of neighborhood, and access to fresh fruit and vegetables (such as supermarkets) [34, 35, 36]</td>
</tr>
<tr>
<td></td>
<td>Facilitates crisis mapping (e.g. where eyewitness reports are plotted on interactive maps. These data can help target areas for emergency services and additional resources)</td>
<td>Weather patterns, pollution levels, allergens, traffic patterns, water quality, walkability of neighborhood, and access to fresh fruit and vegetables (such as supermarkets) [34, 35, 36]</td>
</tr>
<tr>
<td></td>
<td>Facilitates discourse on non-emergency healthcare (e.g. broadcasts of public health messages, quantity medical misconception)</td>
<td>Weather patterns, pollution levels, allergens, traffic patterns, water quality, walkability of neighborhood, and access to fresh fruit and vegetables (such as supermarkets) [34, 35, 36]</td>
</tr>
<tr>
<td>Health-related social networking sites</td>
<td>Facilitates sharing of personal health data and advice amongst patients and consumers</td>
<td>Quantify medical misconceptions (e.g. concussion) [38]</td>
</tr>
<tr>
<td></td>
<td>Monitors spread of infectious diseases via crowd surveillance</td>
<td>The spread of poor medical compliance (e.g. antibiotic use) [39]</td>
</tr>
<tr>
<td></td>
<td>Facilitates emergency services</td>
<td>Trends of cardiac arrest and resuscitation communication [40]</td>
</tr>
<tr>
<td></td>
<td>Monitors public health data and social determinants of health</td>
<td>Cervical and breast cancer screening [41]</td>
</tr>
<tr>
<td>HealthMap</td>
<td>Facilitates emergency services</td>
<td>Postpartum depression [42]</td>
</tr>
<tr>
<td></td>
<td>Facilitates crisis mapping (e.g. where eyewitness reports are plotted on interactive maps. These data can help target areas for emergency services and additional resources)</td>
<td>Inflenza A H1N1 outbreak (disease activity and public concern) [43]</td>
</tr>
<tr>
<td></td>
<td>Facilitates discourse on non-emergency healthcare (e.g. broadcasts of public health messages, quantity medical misconception)</td>
<td>2010 Haitian cholera outbreak [44]</td>
</tr>
<tr>
<td>Other social networking sites (e.g. online discussion boards, Facebook)</td>
<td>Monitors how patients use social media to discuss their concerns and issues</td>
<td>Emergency situations from Boston marathon explosion [45]</td>
</tr>
<tr>
<td></td>
<td>Provides awareness of what the “person in the street” is saying [56]</td>
<td>PatientsLikeMe [47]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease surveillance sites which collect participant-reported symptoms and utilizes informal online data sources to analyze, map, and disseminate information about infectious disease outbreaks (e.g. Flu Near You, HealthMap, GermTracker, Sickweather) [37, 48, 49, 50]</td>
</tr>
<tr>
<td>Search queries and Web logs</td>
<td>Found to be highly predictive for a wide range of population-level health behaviors</td>
<td>Side effects and associated medication adherence behaviors (e.g. drug switching and discontinuation) [51]</td>
</tr>
<tr>
<td></td>
<td>Search keyword selection has been found to be critical for aiming at reliable curated health content</td>
<td>Google and Yahoo search queries have been used to predict epidemics of illnesses, such as:</td>
</tr>
<tr>
<td></td>
<td>“Click” stream navigational data from web logs are found to be informative of individual characteristics such as mental health and dietary preferences [57]</td>
<td>- Influenza (Google 2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Dengue fever [52]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Seasonality of mental health, depression and suicide [53]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Prevalence of Lyme disease [54]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Prevalence of smoking and electronic cigarette use [55]</td>
</tr>
</tbody>
</table>

Small Data — Do patients Make Sense of Their Data and Use It to Improve Health?

While the paper focuses on Big Data, this section focuses on how patients (and people in general) use the small, personal data that is generated on their personal apps and tracking devices. Indeed, “the quantified self is a natural progression from the current practice of the patient being monitored by health professionals to individuals monitoring themselves” [59]. Some have identified a trend of “citizen science,” in which non-professionally trained individuals conduct science-related activities [60]. This begs the question of whether self-monitoring, and informed use of tracking information by patients - not to mention the ability to become a mini-expert, identifying trends, and acquiring specialized, quasi-scientific knowledge of one’s disease or condition - are prevalent and easily obtainable. Several issues, known from psychological research, suggest obtaining this goal is far from trivial.

Primarily, to use data, one first has to make sense of it. Yet comprehension cannot be taken for granted. Studies examining how people understand simple probabilistic information pertaining to prostate or breast cancer have found mistake rates to hover around 50% [61, 62]. Furthermore, miscomprehension also occurred when students were presented with information on prenatal testing [63]. This suggests whatever data or trends we expect patients to benefit from, need to be tested for clarity and understandability with low health literacy taken into account [64]. Comprehension is further hindered when people, physicians included, are presented with more than 3 pieces of information at a time [65]. In addition, once one has made sense of data, one also needs to be motivated to change the behavior. An interesting case comes from the Federal Drug Administration (FDA) warning the administration of cough and cold medication to children under the age of two. A comparison of experienced parents, who had raised children over the critical age of two, and inexperienced parents found more than half (53.3%) of inexperienced parents adhered to the FDA warning,
compared with just over a quarter (28.4%) of experienced parents [66]. The researchers concluded that experience, such as having given a child cough and cold medication numerous times in the past, with no adverse effects, was more influential than information delivered through a warning. In the context of using one’s own data to improve one’s health, it might be tracking health indicators, especially if routinely performed, will serve as actual experience and will motivate human action.

However, comprehension of information, and motivation for change, are not always enough. Patients required to detect a change, for example, those following repeated measurements of their blood sugar levels, may not always know what to do in order to reduce it: Should they change their medication? Eat differently? Exercise more? This is where a healthcare professional’s involvement is called for. And Big Data provides just the opportunity. As Kim [67] suggests: people involved in the quantified self movement will still want to share information with their physicians and healthcare providers. That way they can receive better, more personalized care that is based on their health conditions, diet, and level of physical activity [67]. Just like Feinberg [68] reminds us, patients may wish to have varying degrees of involvement in the treatment process, and, we can extrapolate – patients may have varying degrees of ability to determine the required course of action based on self-tracked information. Yet, apps, devices, and wearables are for the most part sold to consumers, regardless of the physician’s awareness or input. Not only are the physicians unaware the tracking device was purchased, interoperability, legal, and privacy issues may prevent healthcare professionals from approaching this data or making use of it.

A recent attempt to help patients integrate input from various self-tracking sources, to make sense of it, and even to connect it to medical records, comes from a US insurer, Aetna, which developed an access-free platform for such integration. While everyone can use the platform, only Aetna members have access to their medical information [69]. Reservations aside, tracking devices, apps, and other means of collecting patient and consumer input, have the potential to empower and inform patients, as well as to advance science. In some cases, this happens through patient participation in online and other data collection endeavors, such as the ones on PatientsLikeMe [47], a website inviting patients to monitor their disease and share data so knowledge is accumulated regarding their condition. For example, amyotrophic lateral sclerosis (ALS) patients reported their use of limbs and associated it with disease onset, which allowed for the identification of trends in onset. This detailed information may not have been available otherwise [70]. Patient partnership in entering the data in a personal health record (PHR), and the ownership they feel of the process that may happen in their home and is controlled by them, rather than by a health professional, may assist in introducing greater trust.

If a recommendation is generated based on a patient’s personal data, it might be perceived as better suited to them, trustworthy, and the patient will be more likely to act accordingly. This may help circumvent the issue of relatively low trust in government health agencies such as the FDA [71] as opposed to far greater trust in, for example, one’s pediatrician, who of course you know in person [66]. Patients may derive additional benefits from reporting and tracking their medical data, benefits that may be different from developing an expertise in their disease. For example, patients benefit from the ability to know how well they are doing in comparison with others [72]. And patients who reported their symptoms and other personal health information on PatientsLikeMe reported an increased comfort in sharing such information [73]. Notably, this does not require comprehending the meaning and trends in one’s information. It comes from the mere opportunity to share one’s data, and to have it accepted without judgment. It may translate to these patients feeling more secure and being more open when discussing their condition outside the realm of the health social network. This suggests small data is beneficial to patients on many levels, which may be quite different than the Big Data angle.

Connectivism

The connectivist approach takes ideas from brain models and neural networks in learning from technologies [74]. Therefore, a few of the principles related to connectivism are that learning may reside in machines, maintaining connections is necessary to create constant learning and, up-to-date knowledge is the core of connectivist learning moments. Connectivism as an analogy to health is evident. Health requires not only knowledge but also a connected relationship between the provider and the patient, and personalization such that interventions are tailored to the patient’s unique preferences and form of conduct, such as drug adherence. Different people have different reasons for non-adherence to medications [75]. Furthermore, connectivism may serve as an underlying theory for how massive amounts of data collected through various technologies, connect humans and afford interactions in science, healthcare, and education. Hussain [76] explored the underlying principles associated with Siemens’ [74] connectivist theory of learning that is historically considered the go-to theory supporting learning in the digital age. Hussain posits connectivism may need to be reconsidered in the advent of “ambient mobile pervasive communication (p.14)” consisting of filtering mechanisms and smart agents. And, this query has been investigated with an overarching suggestion connectivism still remains a strong theory for understanding Big Data and its initial links to human interactions with technology.

Concluding Thoughts

Recognizing, understanding, and using Big Data in terms of scientific research and healthcare are necessary at this time in order to arrive at best evidence in a world of ever increasing data. Further investigation into the limitations of Big Data, such as inconsistencies regarding standards, policy, ethics, gaps in structured databases and finding a way to contain and deliver Big Data in a meaningful way to health care practitioners is interesting and necessary. This review presents just a glimpse of current and cogent literature
illustrating and supporting the use of Big Data in two areas. Another area to consider is education because of online education and today’s classroom milieu-ubiquitous powerful mobile learning devices becoming more mainstream. The fascinating concepts of Big Data and analytics are not to be ignored in this unprecedented era of innovative technologies that create colossal volumes of both structured and unstructured data. Future papers directed at issues surrounding the open problem of “Quo vadis” (data privacy), confidentiality, and learning analytics are needed. The confluence of Big Data interpretations will continue given the proliferation of data from scientific led endeavors, accelerating healthcare innovations, and the rise of Big Data in higher education as a result of embedding technologies and the proliferation of e-Learning in higher education.

Acknowledgments

The authors declare no conflicts of interest.

Conflict of Interest

References


38. Sullivan SJ, Schneiders AG, Cheang CW, Kitto
47. PatientsLikeMe. [Internet]. [Cited November 30, 2013]. Available from: http://www.patientslike.me/
65. Miron-Shatz T, Barron G, Hanoch Y, Gummerman M, Doniger GM. To give or not to give: A word to the wise, two to the wiser: Parental experience and adherence to the food and drug administration warning about over-the-counter cough and cold medicine usage. Judgment and Decision Making 2010;5(6):428-36.

Correspondence to
Margaret Hansen
School of Nursing and Health Professions
University of San Francisco
San Francisco, California
USA
E-mail: maggie2hansen@gmail.com