BioMedical Informatics 258

Big Data in Healthcare-a two sided story

1. Introduction

Big data, a term coined to describe the massive amounts of digital data accumulated from a variety of sources, has emerged as the new buzzword in the field of healthcare. Indeed, acknowledging the great advances in data mining, computing power, large scale simulations and data storage capacity we could argue that we have entered the big data era. Big data comes with vast societal benefits in terms of driving biomedical innovation, ameliorating patient care and increasing the safety and efficiency of the healthcare system. Nevertheless, the big data culture raises several ethical concerns, the most prominent being the threat towards individual privacy. Thus, concerted efforts should be made to offset the salutatory uses of big data against the associated risks and relevant issues of concern.

In this essay, we first discuss the benefits of big data, complementing our analysis with anecdotal cases of medical breakthroughs that have emerged as results of big data analysis, while transforming scientific methods from hypothesis-driven to data- and evidence-driven, aspiring to better patient care. Thereafter, we turn our attention to the ethical concerns and risks associated with big data, placing emphasis on the changing landscape in terms of individual privacy and sensitive personal information.

2. Big data: big potential

Huge enthusiasm for the ample potential of big data in the healthcare sector has resulted in significant public policy initiatives (1–4). In this section, we present the great opportunities that emerge from an engineering perspective, in terms of big data analysis challenges, as well as the role of big data, both for driving biomedical innovation and ameliorating patient care.

2.1 Big challenges turning into big engineering opportunities

From an engineering standpoint, the advent of big data poses exciting new challenges associated with acquiring, processing and disseminating large biomedical datasets (5). The large volume of biomedical data, as well as the number and complexity of the variables recorded, require novel data analysis approaches. In an effort to make effective use of these large databases, the classical notion of data analysis falls short. In a similar vein, DNA sequencing, the process of determining the exact order of nucleotides within a DNA molecule, was originally

approached with laborious methodologies based on chromatography that took years to complete. However, recent scientific advancements in genomics have dramatically reduced both the time needed and the cost associated with DNA sequencing, promising exciting developments in the field of bioinformatics and personalized medicine (6,7). Similarly, novel dynamic data analysis and interpretation models that would effectively incorporate adaptive methodologies need to be devised.

The sheer volume of data, while an important challenge and a great opportunity for future research, is not the only one. Indeed, the heterogeneity of the different types and representations of data presents further challenges for engineers. Novel ways of coding and storing data, in ways that allow the unification of different datasets, are emerging as pertinent (5). Combining data from different health databases in order to generate new insights is critical. However, this may be impeded by data coded and stored in different manners. One may consider, for instance, the disparities arising from the different classification systems for mental disorders used in the U.S. and Europe. Along the same line, simple differences often arise on a more local level, in the same country or even city. Indeed, from one hospital to another, different units and/or different normal indications for hematological tests may be observed. Therefore, it is important to develop a strategy to transform different data and enable their unification and integration, at the stage of data interpretation. For this to be effective, it would have to happen in a completely automated manner. This further requires the differences in data semantics to be expressed in forms that can be digitally understood and later resolved by computers. Despite the promising progress in data integration, additional work is necessary to achieve an automated error-free difference resolution; especially, if one considers the vast variety of the structures of bioinformatics databases.

2.2 A new era of innovative research in healthcare

Developments in the use of medical records are taking place in several different areas. An area of significant interest is how big datasets can be utilized to identify positive and negative clinically relevant associations between a disease and/or a disorder and its genetic and environmental components, as well as health risk factors, within a large patient population such as observational databases (8,9). This approach may allow us to expand our understanding further in terms of the underlying mechanisms of disorders, as well as gain a new perspective regarding the interconnections between different contributing factors. A case in point is how research in South Africa found that the use of therapeutic vitamin B in HIV-positive patients delayed the progression of AIDS and death (10). This discovery was of particular importance at a time when HIV therapies were non-affordable in the region.

Another area of great potential for big data analysis is pharmacovigilance, defined by the World Health Organization as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem" (11). In 2012, a research group led by Stanford University Professor Dr. Altman reached a significant finding. Applying data mining techniques and extensive statistical analysis, they discovered that when taken together, "Paxil" (paroxetine), a highly popular antidepressant, and "Pravachol" (pravastatin), an equally popular cholesterol-reducing drug, have the synergetic side effect of increasing blood glucose to diabetic levels (12). However, since each medication in itself does not present such a side effect, Food and Drug Administration (FDA) granted its approval. The aforementioned discovery by Dr. Altman and his team prompted the FDA to issue a potentially life-saving warning to thousands of patients (13).

There are numerous other cases where big data analysis has yielded groundbreaking results pertaining to pharmacovigilance. For instance, the discovery that "Vioxx" might instigate cardiac arrests leading to deaths was only enabled by analyzing clinical data and cost data aggregated by Kaiser Permanente, a managed care consortium based in California (14). The drug might not have been withdrawn had such an analysis not been undertaken, and thousands of patients would have been put at risk. Another example is that of "Megace" (megestrol acetate): originally formulated to treat breast and/or uterus cancer, it was reported to have significant weight gain side effects. Turning this potential lose-lose situation into a win-win scenario, the company re-introduced the drug as a means of treating cachexia and of preventing or reversing the unintended or significant weight loss of cancer or HIV-positive patients (15–17).

2.3 Improving patient care

Big data has the capacity to revolutionize not only biomedical engineering research, but also patient care (18). Continuous patient monitoring, assessing treatment outcomes and understanding the physiological effect that a disease and/or disorder has on a patient are just a few examples of the potential impact of big data analysis.

Biosensor modalities in portable devices, such as applications incorporated in smartphones (19), may be the next big data revolution in healthcare. Telemedicine solutions emerge as extremely promising in regards to improving patient care, quality and safety, as well as reducing healthcare costs (20,21). Remote biometrics monitoring using biological sensors may be facilitated by telemetric transmission of collected data to the supervising physician who could assess the patient without regular clinical visits. Medical devices that monitor blood sugar levels, cardiac beating and oxygen saturation, among others, are increasingly being promoted for home use, so that individuals can remotely monitor their own physiological and vital signs.

Whether deployed in rural settings to compensate for the shortage of healthcare specialists, or in urban settings to provide care in such cases as the elderly or infirm where patient transportation is difficult, telemedicine may prove to be live-saving, allowing patients to receive care from their homes. Such practices reduce the number of recurrent hospital admissions as healthcare professionals are remotely alerted to an impending medical issue, enabling them to act immediately towards restabilizing their patients. Moreover, the expenses and the risk of exposure to infections associated with hospital admissions diminish. Accurate and non-invasive means of observing the progression of diseases and disorders via big data analysis may further allow biomedical researchers to devise innovative ways of applying these insights as clinical trial endpoints. The potential is seemingly endless and appears to be attracting governmental interest. The government of the United Kingdom alone, in late 2011, announced its intention to fund the costs of telemedical devices and services for up to three million chronically ill patients (22).

Physicians may also use big data to generate individualized patient treatment solutions based on collective experience. In the past, physicians contextualized the available research evidence, to the best of their ability, by integrating it with their individual clinical expertise towards diagnosing and treating a patient (23). However, in the current “big data era”, this approach would be suboptimal since healthcare specialists find themselves unable to cope with the influx of big data. Nevertheless, recognizing that many aspects of healthcare depend on highly individual characteristics, such as variations in individual physiology and pathology, big data analytics has the potential to emerge as highly beneficial (18). Indeed, rigorous and systematic analysis of biomedical datasets may result in synthesizing high quality evidence on large population samples, thereby ameliorating clinical decision making in the diagnosis, investigation and treatment of individual patients. Using the experience of large populations of patients based on data evidence, healthcare professionals could be guided towards better informed decisions.

3. Big data: big privacy concerns

The growing availability of electronic health records and large medical databases afford research numerous opportunities. As presented earlier, patient medical information promises deeper insights into disease patterns, better understanding of treatment, and broader outlooks regarding increasing the effectiveness and efficiency of palliative patient care.

At the same time, big data analysis is associated with major ethical concerns. Indeed, the acquisition, management and interpretation of large sets of personal data of a sensitive nature, such as healthcare, raise growing concerns for privacy (24–26). Bearing in mind that managing patient privacy is both a technical and

sociological problem that must be addressed from both standpoints, if the promises of big data are to be realized, we go on to describe some of the most important challenges and risks to privacy presented by big data analytics and how technology might be used to mitigate them.

3.1 Privacy challenges

Health-related issues are highly personal and patients need to be confident that their medical records are confidential. Initially, de-identification of data was perceived as a means towards allowing different entities to reap benefits of large scale analytics while seemingly preserving personal privacy. Over the years, various methodologies have been deployed to distance data from personal identities, and de-identification has emerged as fundamental in the context of medical data and clinical trials (27).

However, it is very difficult to anonymize data in a way that cannot be reverse engineered if people are willing to put enough time and effort into it (25,28,29). Therefore, the question that emerges is whether healthcare data can be sufficiently anonymized in such a way that any effort to de-anonymize them would fall short of the benefits of doing so. In any event, de-identification strategies may reduce the information content too far to be useful in practice. As Paul Ohm puts it, former legal scholar from the University of Colorado, “Data can be either useful or perfectly anonymous, but never both” (25). Hence, efforts towards reducing disclosure risk should not be at the expense of research that may benefit society, especially to disadvantaged social strata.

An additional angle to consider here, hitherto unknown, is how to share healthcare data in real time while preserving both patient privacy and data integrity (23). Access control and tracking to maintain data confidentiality and encryption to preserve data integrity may impede the healthcare community from gaining real-time access to accurate patient data, particularly since patients may change hospitals and general practitioners throughout the course of their lives. However, fast access to the medical records of a patient is quintessential in order for a hospital and/or general practitioner to provide the best possible care. Consequently, it is an unconditional requirement that security for big data information sharing is re-examined so as to strike a balance between access and privacy.

On the other hand, the risks to personal privacy and what it might mean to gain or lose ownership over personal medical records are emerging as topics of discussion, naturally generating major ethical concerns.

3.2 Ethical concerns

As information and data accumulate, they give rise to more uncertainties regarding data ownership: who is to say where these records may or may not be used and who will be granted access to them? Such questions are pressing for answers. Do patients own their medical records or is it a social resource and our ethical responsibility to share it? The National Health Service (NHS) in England has implemented an opt-out policy, meaning that by default everyone is considered to allow their de-identified medical records to be used for research purposes, unless they actively choose to opt out (4).

As has always been the case, technology is not inherently good or bad, it is merely a question of what use we collectively decide to put it to. The balance between risk and innovation is a true challenge. Let us consider an example from the field of biomedical engineering.

Over the past years, the cost of genome sequencing has rapidly declined, and the $1000 genome sequencing technique is seemingly just around the corner. The $1000 genome sequencing methodology may allow researchers to commence sequencing tumors, thus enabling the customization of cancer therapies based on the unique genetic profile and tumor characteristics of their patients. In the United Kingdom, funding was provided for a project involving the genome sequence of 100,000 cancer patients and patients with rare diseases with the ultimate intention of later using the results both for the patients and the development of novel diagnostics and treatments (30).

Nevertheless, one has to wonder, once nearly everyone will be able to have their genome sequenced, what effect will that have on privacy, health insurance and healthcare decision making? Will people be advised according to their genetic profile to modify their behavior, or be requested to do so? Taking this further, given their access to healthcare data, would insurance companies demand that their clients adopt a set dietary regime as a condition of their policy? Would lung cancer patients be covered by insurance companies or even be deemed eligible for transplantation if their records show that they have been regular smokers since their teens?

The possibility of insurance companies or other entities using their access to healthcare data to stigmatize patients, limit treatment access and devise new policymaking strategies looms overhead. Strict laws should govern healthcare electronic records to assure that access is only granted to healthcare professionals and researchers on the premise that they aim towards alleviating human pain and suffering and improving patient care.

4. Summary and conclusions

In this essay, we first attempted to describe how big data can revolutionize both biomedical engineering research and patient care. The advent of big data in medical practice poses exciting challenges concerning data acquisition, storage and management. These challenges emerge as areas of particularly promising future opportunities from an engineering perspective. Moreover, big data analytics marks the dawn of a new era in healthcare innovation. This becomes evident when reviewing biomedical breakthroughs based on big data analytics, such as genome sequencing and pharmacovigilance. Indeed, such efforts may help to form deeper insights into disease patterns, allowing us to better assess treatment outcomes while improving our understanding of the physiological impact that disease can have on a patient. Thereafter, we reflected on how the use of medical data records is believed to have the potential to promote preventive care and tailor patient care. Among others, we addressed efforts towards incorporating continuous home-based patient monitoring and discussed how telemedical devices can ameliorate the performance, efficiency and quality of healthcare provided while also lowering medical costs.

On the other hand, acknowledging that big data technology raises several challenging questions pertaining to personal privacy, we ventured to address the associated ethical concerns from both a technological and sociological perspective. While ethics may be an abstract concept, the real-world ramifications cannot be ignored. Big data itself may be ethically neutral, but their use is not; it has moral implications. Hence, collective efforts should be made to inform and align big data actions with ethical values. Acknowledging their impact, attention should concentrate on framing them in a way that is understandable to a common group of individuals working towards a single end, improving patient care. Only if the ethical concerns and risks arising from big data are identified, scrutinized and addressed will the public trust the endeavors made and take advantage of the significantly important benefits big data innovations have to offer.

References

1. Coiera E. Building a National Health IT System from the Middle Out. J Am Med Inform Assoc JAMIA. 2009;16(3):271–3.

2. Morrison Z, Robertson A, Cresswell K, Crowe S, Sheikh A. Understanding Contrasting Approaches to Nationwide Implementations of Electronic Health Record Systems: England, the USA and Australia. J Healthc Eng. 2011 Mar 1;2(1):25–42.

3. Office of Science and Technology Policy, Executive Office of the President. Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments [Internet]. Washington (DC): Executive Office of the President; 2012. Available from: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

4. Everyone to be "research patient."BBC [Internet]. 2011 Dec 5; Available from: http://www.bbc.co.uk/news/uk-16026827

5. Keen J, Calinescu R, Paige R, Rooksby J. Big data + politics = open data: The case of health care data in England. Policy Internet. 2013;5(2):228–43.

6. Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. Genomics. 2009 Feb;93(2):105–11.

7. Cordero P, Ashley EA. Whole-genome sequencing in personalized therapeutics. Clin Pharmacol Ther. 2012 Jun;91(6):1001–9.

8. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol. 2011 Aug;7(8):e1002141.

9. Holmes AB, Hawson A, Liu F, Friedman C, Khiabanian H, Rabadan R. Discovering disease associations by integrating electronic clinical data and medical literature. PloS One. 2011;6(6):e21132.

10. Kanter AS, Spencer DC, Steinberg MH, Soltysik R, Yarnold PR, Graham NM. Supplemental vitamin B and progression to AIDS and death in black South African patients infected with HIV. J Acquir Immune Defic Syndr 1999. 1999 Jul 1;21(3):252–3.

11. WHO | Pharmacovigilance [Internet]. WHO. Available from: http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/

12. Tatonetti NP, Denny JC, Murphy SN, Fernald GH, Krishnan G, Castro V, et al. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clin Pharmacol Ther. 2011 Jul;90(1):133–42.

13. Tatonetti NP, Fernald GH, Altman RB. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. J Am Med Inform Assoc JAMIA. 2012 Feb;19(1):79–85.

14. Rubin R. How did Vioxx debacle happen? USA TODAY [Internet]. 2004 Oct 12; Available from: http://usatoday30.usatoday.com/news/health/2004-10-12-vioxx-cover_x.htm

15. Loprinzi CL, Ellison NM, Schaid DJ, Krook JE, Athmann LM, Dose AM, et al. Controlled Trial of Megestrol Acetate for the Treatment of Cancer Anorexia and Cachexia. J Natl Cancer Inst. 1990 Jul 4;82(13):1127–32.

16. Yeh S, Schuster MW. Megestrol acetate in cachexia and anorexia. Int J Nanomedicine. 2006 Dec;1(4):411–6.

17. Ruiz Garcia V, López-Briz E, Carbonell Sanchis R, Gonzalvez Perales JL, Bort-Marti S. Megestrol acetate for treatment of anorexia-cachexia syndrome. Cochrane Database Syst Rev. 2013;3:CD004310.

18. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 Jun;13(6):395–405.

19. Klasnja P, Pratt W. Healthcare in the pocket: Mapping the space of mobile-phone health interventions. J Biomed Inform. 2012 Feb;45(1):184–98.

20. Black AD, Car J, Pagliari C, Anandan C, Cresswell K, Bokun T, et al. The Impact of eHealth on the Quality and Safety of Health Care: A Systematic Overview. PLoS Med. 2011 Jan 18;8(1):e1000387.

21. McLean S, Sheikh A, Cresswell K, Nurmatov U, Mukherjee M, Hemmi A, et al. The impact of telehealthcare on the quality and safety of care: a systematic overview. PloS One. 2013;8(8):e71238.

22. 3ML | 3 million lives [Internet]. Available from: http://3millionlives.co.uk/

23. Jee K, Kim G-H. Potentiality of Big Data in the Medical Sector: Focus on How to Reshape the Healthcare System. Healthc Inform Res. 2013;19(2):79.

24. McGuire AL, Gibbs RA. No Longer De-Identified. Science. 2006 Apr 21;312(5772):370–1.

25. Ohm P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization [Internet]. Rochester, NY: Social Science Research Network; 2009 Aug. Report No.: ID 1450006. Available from: http://papers.ssrn.com/abstract=1450006

26. Rothstein MA. Is Deidentification Sufficient to Protect Health Privacy in Research? Am J Bioeth AJOB. 2010 Sep;10(9):3–11.

27. Hon WK, Millard C, Walden I. The problem of "personal data" in cloud computing: what information is regulated?—the cloud of unknowing. Int Data Priv Law. 2011 Nov 1;1(4):211–28.

28. Sweeney L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon Univ Data Priv Work Pap 3. 2000;

29. Narayanan A, Shmatikov V. Robust De-anonymization of Large Sparse Datasets. Proc 2008 IEEE Symp Secur Priv. 2008;111–25.

30. Hawkes N. Cameron announces 100m for "unlocking the power of DNA data."BMJ. 2012 Dec 11;345(dec11 1):e8413–e8413.