Big Data and Biomedical Informatics: A Challenging Opportunity

Riccardo Bellazzi
Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy

Summary

Big data are receiving an increasing attention in biomedicine and healthcare. It is therefore important to understand the reason why big data are assuming a crucial role for the biomedical informatics community. The capability of handling big data is becoming an enabler to carry out unprecedented research studies and to implement new models of healthcare delivery. Therefore, it is first necessary to deeply understand the four elements that constitute big data, namely Volume, Variety, Velocity, and Veracity, and their meaning in practice. Then, it is mandatory to understand where big data are present, and where they can be beneficially collected. There are research fields, such as translational bioinformatics, which need to rely on big data technologies to withstand the shock wave of data that is generated every day. Other areas, ranging from epidemiology to clinical care, can benefit from the exploitation of the large amounts of data that are nowadays available, from personal monitoring to primary care. However, building big data-enabled systems carries on relevant implications in terms of reproducibility of research studies and management of privacy and data access; proper actions should be taken to deal with these issues. An interesting consequence of the big data scenario is the availability of new software, methods, and tools, such as map-reduce, cloud computing, and concept drift machine learning algorithms, which will not only contribute to big data research, but may be beneficial in many biomedical informatics applications. The way forward with the big data opportunity will require properly applied engineering principles to design studies and applications, to avoid preconceptions or over-enthusiasms, to fully exploit the available technologies, and to improve data processing and data management regulations.

Keywords

Big data, data analytics, research reproducibility, cloud, NoSQL, map-reduce

Yearb Med Inform 2014:8-13
http://dx.doi.org/10.15265/IY-2014-0024
Published online May 22, 2014

Big Data: Why Bother?

Like other new terms that abruptly appeared on the scientific arena, the term "big data" has generated some doubts and concerns in both the research and business communities [1]. Several projects have dealt with large data collections, and several research labs have exploited computer clusters and multicore facilities for the last decade [2]. Thus, empowering the computational infrastructures by exploiting cloud-based solutions improving algorithms parallelization cannot be considered as a paradigmatic shift, but only as a technological step. So, why bother to invent a new word only to highlight the steady improvement of technology? Moreover, if "big data" refer to "social networks" data, their effective role in clinical or research studies seems to be distant enough to be of minor importance for the biomedical field.

The reality is that the term "big data" means much more than that, highlighting a challenge and an opportunity that the biomedical informatics domain must face in the next five years. Following the definition created by the IMIA working group on "Data Mining and Big Data Analytics", "Big Data are data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it" [3, 4]. It is now widely agreed that big data arise from the combination of four elements, whose joint occurrence represents an unprecedented combination [5, 6].

First, the data may be large (*Volume*) up to an extent that has been inconceivable until now. It is estimated that, in one minute, 640TB data are transferred, 100 thousands new tweets are created, and 204 million e-mails are sent over the Internet [5, 7]; Facebook has 10 million photos uploaded every hour, 3 billion "likes" or comments per day, YouTube increases by 1 hour of videos every second. Overall, it is estimated that 2.5 quintillions bytes are created each day. The total amount of healthcare data, estimated to be around 150 exabytes, is foreseen to explode in size when next generation sequencing will be used for diagnostic purposes and once the data from wireless health monitors - estimated to be 420 million in 2014 - will be integrated [5].

Volume alone, however, is not the only property of big data: the second one is diversity or *Variety*. The data stored are unstructured and structured texts, images, signals and streams, point-based numerical values, meta-data. Moreover, data may refer to different dimensional scales, from the molecular to the population one, and span over different time scales, from milliseconds to years. Properly managing and exploiting the variety to interpret and analyze the data is at least as complex as managing huge volumes [8].

The third property of big data is *Velocity,* i.e. the speed of analytical processing [9]. Sensor networks as well as the stock market need to process the huge amount of produced data streams timely, i.e. with the velocity that is required by the application (e.g. seconds or milliseconds for sensors, microseconds for the stock market). The implication from the data analytics viewpoint is that the analysis needs to be moved *close* to the data, shifting from the typical paradigm of batch processing to an online, distributed, analytical processing [10].

The last piece of the puzzle is represented by a crucial component in biomedical informatics, the *Veracity* or the uncertainty of the data. Very large data collections may often combine various sources of variable reliability and trust. Moreover, data may often not be collected for research purposes, but they rather are "process" data that show snapshots

of a working system. How to derive real "signals" in such a noisy environment is a crucial component of the analytical process [11, 12].

In the following, I will address some issues that I feel relevant for big data in biomedicine, starting from the identification of contexts where the four Vs' are essential, moving to the potential side effects of big data and finishing with a quick look at recent "big data technologies".

Big Data: Must-have or Nice-to-have?

A crucial question is to determine where big data are essential components of biomedical research and healthcare management.

An area that is clearly facing a "going-big or perish" scenario is represented by molecular biology and molecular medicine research [13, 14]. High throughput analysis, such as next generation sequencing (NGS), generates terabytes of data, and proper management and organization of information is mandatory to achieve the goal of scientific discovery while showing good value for money [15, 16]. This need becomes more and more essential when molecular information is integrated with imaging, like in neuroimaging genetics [17], and with phenotype data, like in projects such as the Emerge network [18]. Interestingly, exploiting big data technologies, such as large scale databases and parallel computing, is not only necessary for those online data repositories usually maintained by large research centers, but is also a need for any research lab that has NGS facilities. Building research management software nowadays is a big-data exercise, since these tools must manage large volumes of data of diverse nature, including raw signals, annotations, images, phenotypes, and textual reports [19, 20, 21]. Moreover, in particular for molecular diagnostic purposes, data are continuously generated and never erased, thus leading to a "Volume and Velocity" challenge to ensure accessibility of information.

Biomedical research has also been boosted by the application of tools developed for automated literature analysis and by the analysis of the biomedical knowledge bases available on the Internet. The overall area known as "integrative bioinformatics" is, actually, a "big data" analysis area, in which Volume and Variety play an essential role [22].

Public health services have started the transition towards big data centers. For example, in Italy, local governments (Regional and Local Health Care Agencies) have the duty of handling citizens' health leveraging on detailed and granular data collection. All hospital admissions, discharges, drug prescriptions, and specialists' visits are stored and analyzed [23, 24]. Thanks to the easy geo-localization obtained by the citizens' address, as well as by the address of hospitals, pharmacies, GPs, and ambulatories, it is possible to derive geographic health maps over years for the entire population of a region [25]. The start of national programs for health informatics allows obtaining an integrated large source of information that holds the premise of providing the complete picture on the health of a certain area. Such projects are already a reality in other European countries, like Denmark [26, 27]. A potential future dimension for a "big-data"-enabled public health is the integration of other sources of information, in particular concerning the environment, such as pollutants, car traffic, heating, grocery stores and markets, food consumptions. Some of those data are "open" and their inclusion could be beneficial for the risk stratification of the population [28].

Moving from a "must-be-big-data-enabled" to "should-be-big-data-enabled", we certainly need to analyze the hospital case. The Medical Informatics community is well aware that hospitals are sources of data characterized by an extremely variable format, by the need of being processed fast, by measurement and data collection errors and, finally, thanks to the improved measurement capability, by their increasing volume [29,30]. Only the "information silo" syndrome that often affects hospital information systems has prevented hospitals from being a running example of application of big data technologies [31]. We are currently witnessing examples of "technology push" given by the combination of pervasive computing, and big data storage and retrieval solutions. Let us suppose that a hospital has "full" monitoring capabilities of surgical interventions. Basic clinical data are collected in an EMR, while vital signs and signals are monitored by operating room instruments. Moreover, sensors provide temperature, humidity, and pressure in the room. Finally, the entire hospital is equipped with sensors and monitored, including the functioning of elevators, heating, and air conditioning. All process data are stored in a high performance data warehouse for flexible future retrieval. Such technological infrastructure may allow a full auditing of hospital performance and of surgical interventions, thus leading to quality control and potential corrective actions. Handling of exceptions, such as delays in the surgery time due to malfunctioning of devices in the operating room, or of other equipment, such as elevators, may be possible [32].

A rather new interesting area, which we can classify as "could-be-big-data-enabled" (or handle with care) involves disease surveillance and pharmacovigilance [11]. The scientific community is currently investigating whether the analysis of Internet searches and social media is a valuable path to extract useful information from data in this context, as witnessed by the "Google Flu" system [33, 34]. However, signals in search engines and social networks are very weak and the noise, on the contrary, is quite high (so, even if Volume is high, the Veracity may be very low). Such signals must be analyzed in the light of a combination of information coming from a Variety of sources, thus integrating EHRs, social media, knowledge bases, and literature analysis [35, 36].

Finally, there is another domain that may have a great input from big data management and analytics: medical decision-making. The Watson system (without any personal evaluation on the specific validity of the solution in clinical practice, which still needs to be assessed by clinical studies) is a demonstration that new-generation electronic decision support systems must be "big-data-enabled" to allow a better implementation of the "precision medicine" agenda [37]. Precision will not only be related to the molecular characterization of the disease, but also to the proper tailoring of the enormous amount of information available, including guidelines, to the specific patient's data [38]. As a natural consequence, in the near future,

together with "big-data-enabled" decision support systems for physicians, we will live in a world with "big-data-enabled" patients, i.e. patients that are well informed[1] and who may exploit automated decision support services on the web [39]. Medical informatics should take the opportunity of the convergence of increased patient involvement and improved capabilities of analyzing large collections of data, including social networks, to develop new tools that will be able to give a better understanding not only of patients' data but also of patients' preferences and experiences [40].

Big Data Side Effects

Entering the big data era will potentially have serious side effects, which need to be anticipated, and hopefully prevented. I will analyze two of the side effects: reproducibility of scientific results and the policies to deal with privacy and data reuse.

The reproducibility of results is an important issue that too often has remained concealed in the scientific literature. The very nature of big data makes this issue extremely hard to handle. As reported in a brilliant analysis by Furlanello [41], a interesting example is the publications in major biomedical journals of a set of biomarkers extracted from high throughput molecular data in oncology, which were then retracted because the data analysis performed was shown not to be reproducible due to a mislabeling of the data [42]. Such cases have highlighted that the combination of high data volume and of the complexity of the data analysis process, which includes data cleaning, data integration, preprocessing, modeling, and validation, makes full reproducibility difficult to achieve. A group of researchers and statisticians later performed a study on the published RNA microarray analyses, showing that more than 50% of the studies were not reproducible [43]. In order to eliminate the risk of other similar cases and to provide a roadmap, they launched a reproducibility initiative that published a set of guidelines promoting open source communities to share code to be used in the data analysis process [44, 45]. At the same time, the Gigascience initiative, a collaboration between Beijing Genomic Institute Shenzhen and BioMed Central, started to create an open-access data platform able to provide software workflows and databases on a cloud computing platform to enable researchers to implement their data analysis pipelines and make them available to others for reproducibility purposes [46].

Nevertheless, we expect these challenges to remain when dealing with data that are so big to require distributed analysis in order to be processed, or, even worse, when deriving knowledge from data with high volatility, for example acquired during online processes from wearable sensors. In this case, it will be necessary to define properly the type of scientific studies and assure the quality of scientific evidence supposed to be derived from the analysis of big data. There still will be cases in which it will be necessary to have snapshots of data and transfer them into cloud-based services for assessing reproducibility. However, in the majority of situations, this will not be possible due to the very nature and volume of data. For this reason, it will be important to require the sharing of methods and tools, and have a clear definition of the data analysis process and pipelines. Efforts towards the formalization of modeling activities seem the only way to ensure a "process-based" reproducibility rather than a complete reuse of the data themselves [47]. Moreover, it would be of great interest to assess the stochastic properties of processes generating big data to describe under which conditions the results obtained can be considered valid. Adopting known analytics standards, following the example provided by Good Research for Comparative Effectiveness (GRACE) principles in the area of comparative effectiveness, is a viable way forward [48].

A second potential side effect of the big data era is the threat for privacy and the resulting need for policies determining data sharing and data reuse. In this case, we must first define the nature of the problem. Exploiting the big data opportunity enables new kind of studies and knowledge discovery. Big data allow population-based analyses to unveil correlations between basic human health behaviors and common diseases, to enable individual–level studies involving phenotype, genotype and exposure data, and finally to build personal health records enriched by quantified-self data [28]. All those scenarios have implications on privacy management, since data may be used for a purpose different from the reason why they have been collected. In the case of population-based analyses, a serious issue is the implementation of a secure and reliable system for anonymizing data: the potential re-identification of patients is a risk that increases with the dimensionality of the data collection [49]. The need of managing data at the individual level is crucial for supporting biomedical research: in this case, an entire new model for governing research studies seems to be the proper solution. To this end, IMIA has started initiatives with all stakeholders to support trustworthy data use [50]. Such initiatives target the level of regulation, and amendment to laws, such as the EU's General Data Protection Regulation that will substitute the current EU Data Protection Directive 95/46/EC [51, 52]. Finally, data protection regulations would require that available big data-enabled health information systems allow building and maintaining a personal health record that contains all personal data, i.e. clinical, genetic and environmental (exposome) data [28]. Although this record should be under the citizens' control, the technological infrastructure and the corresponding regulations should be properly designed and planned to allow the implementation of this repository on a virtual platform, which may contain information physically stored in internationally distributed locations.

Big Data Technologies: Software, Algorithms, and Architectures

Big data have become relevant because they are increasingly present in many sectors of human activity and need specific methods, algorithms, and tools to be stored, managed,

[1] See, for example, the "e-patient Dave", advocate of patients' engagement (http://www.epatientdave.com/)

and processed. The first areas that had to deal with big data were big science projects, like particle physics experiments: the data center of CERN, for example, stores more than 100 Petabyte, which is only a portion of the data generated[2]. Subsequently, the paradigmatic applications handling big data have become web search engines and social networks. In all cases, the increasing growth of data characterized by at least two of the four Vs' has motivated the need for developing technological solutions as a crucial enabling factor of the core activities (business or research). In other words, without "big-data" technologies neither Google, nor Facebook, nor the Higgs boson experimental discovery would have ever succeeded.

A noteworthy technological result is represented by the changes in parallel programming driven by big data. A very successful paradigm is now represented by Map-Reduce [53], a programming model developed and implemented by Google, aiming at simplifying parallelization. This is done by organizing the computational steps in the code using two main functions: i) Map, which accepts input data as key-value pairs, performs computations, and outputs other key/value pairs, and ii) Reduce, which processes key-value pairs showing the same key to derive the final result. An application/algorithm is thus implemented as a sequence of tasks, each with a Map and a Reduce phase. The Map-Reduce paradigm is simpler to implement than other techniques for parallel programming, which require a fine tuning of low level programming languages; it enables developers to use high level programming languages (i.e. Java, Ruby, Python) more efficiently, but requires a complex architecture, whose core element is a distributed file system. The Apache software foundation has developed one of the most widely used Map-Reduce implementations: Hadoop [14, 54, 55].

New database technologies are also offering answers to the scalability problem. A plethora of new solutions have been recently presented under the umbrella of NOSQL (Not Only SQL) data management systems [56]. Such systems are designed to provide easy horizontal scaling (i.e. the data may grow horizontally involving more nodes in a computer cluster) to represent data without the burden of relational modeling (i.e. some NOSQL databases are "document oriented" and are able to store and query collections of documents). The price to pay is the lack of a standard query language and the high variability among different solutions; the consequence is the need of pre-programming views and queries [57]. The shift in technology holds the promise of building new databases and data warehouses oriented towards the collection of not only high volumes, but also of highly variable data naturally gathered in distributed environments [58].

From the point of view of data analysis, machine learning, and decision support, there are various algorithms and tools that seem particularly important to analyze big data [59]. In particular, while the Map-Reduce paradigm may allow rewriting existing algorithms in a distributed architecture, distributed intelligence strategies may allow performing decentralized computations, reducing the burden of data transfer and data integration. Moreover, looking at aspects other than volume, a group of algorithms may effectively deal with big data velocity. Time processing constraints are well approached by "anytime algorithms". These algorithms return a valid solution to a problem even if they are stopped at any time before they are completed, being designed to progressively find better solutions as they proceed [60]. If data are generated by fast, non-stationary, processes, their analysis can be approached by methods able to deal with the so-called concept drift. Concept drift learners are able to monitor input data and adapt, when needed, the learners to new acquired data. Such approaches are particularly suitable to deal with data streams [61, 62, 63].

Together with software and data management tools, new IT architectures are also necessary to support big data management and analysis; in particular, cloud-computing seems a crucial solution to enable high performances while containing building and operational costs [64, 65, 66]. Currently, there are several cloud types, which may be suited for almost all needs in healthcare and biomedicine. Private clouds provide services to one organization only, and are located inside the organization itself or at a third party provider, which allows controlling the infrastructure without the need for hardware management. Community clouds seem promising for research purposes since the infrastructure is shared by several organizations that have common needs. Public clouds provide high-level services to generic users, a solution that allows taking full advantage of both the elasticity and heterogeneity characteristics of the services offered, but allows less control infrastructure. Finally, hybrid clouds are the composition of two or more types of clouds, bound together by standard or proprietary technologies [67, 68]. Privacy concerns are currently being discussed, with the aim of enabling cloud-based services for the biomedical and healthcare sectors [69].

Final Suggestions and Remarks

The "big data revolution" is only at its beginnings, but it looks like to be inevitable, as recently reported by Murdoch and Detsky [8] and witnessed by the NIH initiative "Big data to knowledge" [70]. High throughput data gathering, in particular in the "–omics" sciences, has allowed researchers to generate hypotheses on the basis of a data driven approach, thus enabling the possibility of finding the "needle in the haystack" by resorting to powerful machine learning and data analysis methods. With big data, our confidence that the needle IS in the haystack has increased but this confidence must be supported by methods and tools smart enough to find it [71]. The problem, of course, beside the development of suitable technologies, is data quality and results evaluation, which, as mentioned in this paper, is probably the most difficult challenge to deal with.

Even if it is too early to propose guidelines to deal with this peculiar problem, it is certainly possible to focus on four issues that, in my opinion, need to be taken into account by the biomedical informatics community as recommendations.

i) Apply engineering principles. When planning a research study that takes place

[2] http://home.web.cern.ch/about/updates/2013/02/cern-data-centre-passes-100-petabytes

in a big data context, it is mandatory to analyze big data sources and characterize them in terms of:

a. Volume, to optimize the storage, access, and scalability of the application.
b. Variety, to carefully plan pre- and post-processing and analytics.
c. Velocity, to define the IT architecture, in particular in terms of distributing computation, and to select appropriate algorithms.
d. Veracity, to understand the quality of evidence that can be derived from the study.

For example, let's suppose that we need to plan a research study that integrates NGS and clinical data. In this case, it is first crucial to define the IT infrastructure components, taking into account volume and velocity aspects. Moreover, given the complex nature of the data, i.e. its variety, it will be important to select the algorithms and tools to run the analysis of data, which are able, for example, to integrate data and free text, and, if needed, find implementations based on the map-reduce paradigm. Finally, the assessment of the quality of evidence to be derived is crucial; it will depend on the data sources to be integrated (if we want to include, for example, social media or public repositories), together with the standard quality indicators as selection bias, sample size, and measurement noise.

ii) Avoid confusion, over-enthusiasm, or preconception. An increasing number of misunderstandings about big data need to be avoided. For example:
   a. Social networks and big data are not synonyms. While social networks certainly convey big data, the sources of big data in biomedicine can be many others, as reported in this paper.
   b. Big data are not a solution, but an opportunity. They won't cure anybody. They will provide fruitful information if they are integrated in well-designed research studies and proper research cycles, for hypotheses generation, hypotheses confirmation, and for monitoring health care processes [72].
   c. Big data are not evil, generating only noise. Their analysis can be fruitfully included in a fully scientific cycle, provided that their origin and components are well understood and characterized.

iii) Exploit technologies. The big data wave that has already happened in science and business has pushed towards a new generation of IT solutions for data management and parallel computing. Even if not all applications in health care will have to deal with big data, many of them may successfully benefit from such new technologies. Map-Reduce, Hadoop, NoSQL data bases, cloud computing, are likely to be used to improve the performance of health care IT systems very soon.

iv) Work on data processes and regulations. A full exploitation of big data sources would be possible if trustworthy "big-data" systems are put in place. Trust is needed to ensure that science-based approaches are followed and, at the same time, that the data management is respectful of citizens' expectations about their privacy and the proper use of their information.

As technology progresses, it is likely that the term "big data" will slowly fade away or will be surpassed by other advances. It is also possible that it won't be one of Kuhn's paradigm shifts. It is certainly, though, a clear change of perspectives and a challenging opportunity. Such opportunity advocates more than ever the need of data scientists in the biomedical informatics arena [73].

Acknowledgements

The IMIA working group on Data Mining and Big data Analytics is gratefully acknowledged for insightful discussions. I also sincerely thank my collaborators of the Biomedical Informatics Labs "Mario Stefanelli" for their help and partnership when entering into the big data era. I'm in debt with Lucia Sacchi and Davide Capozzi for revising early drafts of this paper. Finally, I sincerely thank the Editors of the Yearbook for their corrections and insightful suggestions.

References

1. Ross JW, Beath CM, Quaadgras A. You May Not Need Big Data After All. Harvard Business Review, Dec 01, 2013.
2. Grossman RL, White KP. A vision for a biomedical cloud. J Intern Med 2012 Feb;271(2):122-30.
3. Smitha T, Suresh Kumar V. Applications of big data in data mining. International Journal of Emerging Technology and Advanced Engineering 2013;7(3) (www.ijetae.com).
4. Peek N, Sun J, Holmes J, Martin-Sanchez F, Bellazzi R. Biomedical and Healthcare Analytics on Big Data. AMIA 2013 Symposium Proceedings, 2013, November, 1116-7.
5. http://www-01.ibm.com/software/data/bigdata/
6. Eaton C, DeRoos D, Deutsch T, Lapis G, Zikopoulos P. Understanding Big Data. McGraw Hill; 2012.
7. http://www.techspot.com/news/52011-one-minute-on-the-internet-640tb-data-transferred-100k-tweets-204-million-e-mails-sent.html
8. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013 Apr 3;309(13):1351-2.
9. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev 2012 Oct;90(10):60-6, 68, 128.
10. Cuzzocrea A, Moussa R, X. Guandong. OLAP*: Effectively and Efficiently Supporting Parallel OLAP over Big Data, Model and Data Engineering, Lecture Notes in Computer Science Volume 8216, 2013. p. 38-49.
11. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. PLoS Med 2013;10(4):e1001413
12. Schultz T. Turning healthcare challenges into big data opportunities: A use-case review across the pharmaceutical development lifecycle. Bulletin of the Association for Information Science and Technology 2013;39(5):34-40.
13. Costa FF. Big data in biomedicine. Drug Discov Today 2013 Oct 29.
14. O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 2013 Oct;46(5):774-81.
15. Shah NH. Translational bioinformatics embraces big data. Yearb Med Inform 2012;7(1):130-4.
16. Lecroq T, Soualmia LF. From genome sequencing to bedside. Findings from the section on bioinformatics and translational informatics. Yearb Med Inform 2013;8(1):175-7.
17. Van Horn JD, Toga AW. Human neuroimaging as a "Big Data" science. Brain Imaging Behav 2013 Oct 10. [Epub ahead of print]
18. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al.; eMERGE Network. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013 Oct;15(10):761-71.
19. Leduc R, Vaughn M, Fonner JM, Sullivan M, Williams JG, Blood PD, et al. Leveraging the national cyberinfrastructure for biomedical research. J Am Med Inform Assoc 2013 Aug 20.
20. Dong X, Bahroos N, Sadhu E, Jackson T, Chukhman M, Johnson R, et al. Leverage Hadoop Framework for Large Scale Clinical Informatics Applications. AMIA Summits Transl Sci Proc 2013 Mar 18;2013:53.
21. Athey BD, Braxenthaler M, Haas M, Guo Y. tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform

for Clinical and Translational Research. AMIA Summits Transl Sci Proc.2013 Mar 18;2013:6-8.
22. Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L. Translational bioinformatics: linking knowledge across biological and clinical realms. J Am Med Inform Assoc 2011 Jul-Aug;18(4):354-7.
23. Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi R. Mining health care administrative data with temporal association rules on hybrid events. Methods Inf Med 2011;50(2):166-79.
24. Colombo GL, Rossi E, De Rosa M, Benedetto D, Gaddi AV. Antidiabetic therapy in real practice: indicators for adherence and treatment cost. Patient Prefer Adherence 2012;6:653-61.
25. Dalle Carbonare S, Cerra C, Bellazzi R. Development and representation of health indicators with thematic maps. Stud Health Technol Inform 2012;180:220-4.
26. Sortsø C, Thygesen LC, Brønnum-Hansen H. Database on Danish population-based registers for public health and welfare research. Scand J Public Health 2011 Jul;39(7 Suppl):17-9.
27. Lippert S, Kverneland A. The Danish National Health Informatics Strategy. Stud Health Technol Inform 2003;95:845-50.
28. Martin Sanchez F, Gray K, Bellazzi R, Lopez-Campos G. Exposome informatics: considerations for the design of future biomedical research information systems. J Am Med Inform Assoc 2013 Nov 1.
29. Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. Genet Med 2013 Oct;15(10):802-9. doi: 10.1038/gim.2013.121. Epub 2013 Sep 5.
30. de Lissovoy G. Big data meets the electronic medical record: a commentary on "identifying patients at increased risk for unplanned readmission". Med Care 2013 Sep;51(9):759-60.
31. Cases M, Furlong LI, Albanell J, Altman RB, Bellazzi R, Boyer S, et al. Improving data and knowledge management to better integrate health care and research. J Intern Med 2013 Oct;274(4):321-8.
32. Restuccia JD, Cohen AB, Horwitt JN, Shwartz M. Hospital implementation of health information technology and quality of care: are they related? BMC Med Inform Decis Mak 2012 Sep 27;12:109.
33. Masoni M, Guelfi MR, Conti A, Gensini GF. Pharmacovigilance and use of online health information. Trends Pharmacol Sci 2013 Jul;34(7):357-8.
34. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol 2013;9(10):e1003256. Epub 2013 Oct 17.
35. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2013 May 1;20(3):413-9.
36. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014 Mar 14;343(6176):1203-5.
37. Malin JL. Envisioning Watson as a rapid-learning system for oncology. J Oncol Pract 2013 May;9(3):155-7.
38. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. N Engl J Med 2012 Feb 9;366(6):489-91.
39. deBronkart D. How the e-patient community helped save my life: an essay by Dave deBronkart. BMJ 2013 Apr 2;346:f1990.
40. Giuse NB, Koonce TY, Storrow AB, Kusnoor SV, Ye F. Using health literacy and learning style preferences to optimize the delivery of health information. J Health Commun 2012;17 Suppl 3:122-40
41. Furlanello C, Emerging data waves in biomedicine: the challenge of reproducibility, IDAMAP 2012 keynote lecture, Pavia, November 22, 2012.
42. Baggerly KA, Coombes KR. Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-throughput Biology. Annals of Applied Statistics2009;3(4):1309-34.
43. Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. Nat Genet 2009 Feb;41(2):149-55.
44 Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010 Oct;11(10):733-9.
45. https://www.scienceexchange.com/reproducibility
46. Sneddon TP, Li P, Edmunds SC. GigaDB: announcing the GigaScience database. Gigascience 2012 Jul 12;1(1):11.
47. Stodden V, Guo P, Ma Z. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. PLoS One 2013 Jun 21;8(6):e67111.
48. Dreyer NA, Schneeweiss S, McNeil BJ, Berger ML, Walker AM, Ollendorf DA, et al. GRACE Principles: Recognizing High-Quality Observational Studies of Comparative Effectiveness. Am J Manag Care 2010:16(6):467-71
49. Schadt EE. The changing privacy landscape in the era of big data. Mol Syst Biol 2012;8:612
50. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: a transnational perspective. Int J Med Inform 2013 Jan;82(1):1-9.
51. Andersen MR, Storm HH; on behalf of the Eurocourse Work Package 2 Group. Cancer registration, public health and the reform of the European data protection framework: Abandoning or improving European public health research? Eur J Cancer 2013 Oct 10.
52. Di Iorio CT, Carinci F, Oderkirk J. Health research and systems' governance are at risk: should the right to data protection override health? J Med Ethics 2013 Dec 5.
53. Zou Q, Li XB, Jiang WR, Lin ZY, Li GL, Chen K. Survey of MapReduce frame operation in bioinformatics. Brief Bioinform 2013 Feb 7. [Epub ahead of print].
54. Dong X, Bahroos N, Sadhu E, Jackson T, Chukhman M, Johnson R, et al. Leverage Hadoop Framework for Large Scale Clinical Informatics Applications. AMIA Summits Transl Sci Proc 2013 Mar 18;2013:53.
55. Nordberg H, Bhatia K, Wang K, Wang Z. BioPig: a Hadoop-based analytic toolkit for large-scale sequence data. Bioinformatics 2013 Dec 1;29(23):3014-9.
56. Manyam G, Payton MA, Roth JA, Abruzzo LV, Coombes KR. Relax with CouchDB-into the non-relational DBMS era of bioinformatics. Genomics 2012 Jul;100(1):1-7.
57. Lee KK, Tang WC, Choi KS. Alternatives to relational database: comparison of NoSQL and XML approaches for clinical data storage. Comput Methods Programs Biomed 2013 Apr;110(1):99-109.
58. Triplet T, Butler G. A review of genomic data warehousing systems. Brief Bioinform 2013 May 14.
59. Wolfe PJ. Making sense of big data. Proc Natl Acad Sci U S A 2013 Nov 5;110(45):18031-2.
60. Zilberstein S. Using Anytime Algorithms in Intelligent Systems. AI Magazine1996;17(3):73-83.
61. Žliobaitè, Indrè. Learning under concept drift: an overview. arXiv preprint arXiv:1010.4784; 2010.
62. Ryan Hoens T, Polikar R, Chawla NV. Learning from streaming data with concept drift and imbalance: an overview. Prog Artif Intell 2012;1(1):89-101.
63. Stella F, Amer Y. Continuous time Bayesian network classifiers. J Biomed Inform 2012 Dec;45(6):1108-19.
64. Wall DP, Kudtarkar P, Fusaro VA, Pivovarov R, Patil P, Tonellato PJ. Cloud computing for comparative genomics. BMC Bioinformatics 2010 May 18;11:259.
65. Lin CW, Abdul SS, Clinciu DL, Scholl J, Jin X, Lu H, et al. Empowering village doctors and enhancing rural healthcare using cloud computing in a rural area of mainland China. Comput Methods Programs Biomed 2013 Nov 9.
66. Lin YC, Yu CS, Lin YJ. Enabling large-scale biomedical analysis in the cloud. Biomed Res Int 2013;2013:185679.
67. Kaur PD, Chana I. Cloud based intelligent system for delivering health care as a service. Comput Methods Programs Biomed 2014 Jan;113(1):346-59.
68. Zhou S, Liao R, Guan J. When cloud computing meets bioinformatics: a review. J Bioinform Comput Biol 2013 Oct;11(5):1330002.
69. Ohno-Machado L, Farcas C, Kim J, Wang S, Jiang X. Genomes in the Cloud: Balancing Privacy Rights and the Public Good. AMIA Summits Transl Sci Proc 2013.
70. http://bd2k.nih.gov/
71. Jaulent MC. Personal communication.
72. Neff G. Why big data won't cure us. Big data 2013 Sep;1(3):117-23.
73. Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. Harv Bus Rev 2012 Oct;90(10):70-6, 128.

Correspondence to:
Riccardo Bellazzi
Biomedical Informatics Labs "Mario Stefanelli"
Department of Electric, Computer and Biomedical Engineering
University of Pavia
Tel: +39 0382 985720, +39 0382 985059, +39 0382 985981
Fax: +39 0382 985373
E-mail: riccardo.bellazzi@unipv.it