Big Data - Smart Health Strategies

Findings from the Yearbook 2014 Special Theme

V. Koutkias[1], F. Thiessard[2], Section Editors for the IMIA Yearbook Special Section

[1] INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France; Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France

[2] Univ. Bordeaux, ISPED, Centre INSERM U897-Epidemiologie-Biostatistique, Equipe de Recherche en Informatique Appliquée à la Santé, F-33000 Bordeaux, France

Summary

Objectives: To select best papers published in 2013 in the field of big data and smart health strategies, and summarize outstanding research efforts.

Methods: A systematic search was performed using two major bibliographic databases for relevant journal papers. The references obtained were reviewed in a two-stage process, starting with a blinded review performed by the two section editors, and followed by a peer review process operated by external reviewers recognized as experts in the field.

Results: The complete review process selected four best papers, illustrating various aspects of the special theme, among them: (a) using large volumes of unstructured data and, specifically, clinical notes from Electronic Health Records (EHRs) for pharmacovigilance; (b) knowledge discovery via querying large volumes of complex (both structured and unstructured) biological data using big data technologies and relevant tools; (c) methodologies for applying cloud computing and big data technologies in the field of genomics, and (d) system architectures enabling high-performance access to and processing of large datasets extracted from EHRs.

Conclusions: The potential of big data in biomedicine has been pinpointed in various viewpoint papers and editorials. The review of current scientific literature illustrated a variety of interesting methods and applications in the field, but still the promises exceed the current outcomes. As we are getting closer towards a solid foundation with respect to common understanding of relevant concepts and technical aspects, and the use of standardized technologies and tools, we can anticipate to reach the potential that big data offer for personalized medicine and smart health strategies in the near future.

Keywords

Big data, cloud computing, personalized medicine, smart health strategies

Yearb Med Inform 2014:48-51
http://dx.doi.org/10.15265/IY-2014-0031
Published online August 15, 2014

Introduction

Since the birth of computers, we have been facing a huge increase in the volume of data. The term "big data" appeared in 1997, and was first used by few specialists. In 2001, reference to the data explosion phenomenon was made by using the "3Vs" [1], i.e. increasing *volume* (amount of data), *velocity* (speed of data's generation, integration, sharing, and processing), and *variety* (heterogeneity of data types and sources). These attributes have been widely adopted to define "big data", while in some other definitions a fourth "V" has been also employed referring to *veracity* (to address the quality of data and consequently the quality of evidence that can be derived from these data). Yet, if we look at the evolution of the number of queries that contain the term "big data", e.g. via Google Trends, there is a very sharp increase from 2011 (the number of queries for "big data" is ten times greater in 2013 than before 2011). "Big data" really became a buzzword, sometimes properly used, but many times also as a topic for debate and discussion.

Volumes of biomedical data have also increased significantly and will continue to do so with the use of Electronic Health Records (EHRs), the wider penetration of personal health systems supporting disease and lifestyle management, the production of genetic data especially thanks to next generation sequencing, and the extensive use of social media platforms, to name a few. For a description of the different types of data concerned with big data in biomedicine we refer the readers to the survey paper of the special theme of the 2014 edition of the IMIA Yearbook [2]. In addition to the above data-centric characteristics, it is evident that appropriate technologies and methods have to be in place to enable efficient storage, access, analysis, visualization, and sharing, in order to effectively exploit the data wealth.

In the scope of our review, we were excited to read various interesting contributions in the field spanning from: (a) infrastructures enabling cloud-based access to EHRs [3], real-time ECG monitoring and analysis via the cloud [4], large image set analysis [5], and EHR-based analytics [6] and (b) techniques for dimensionality reduction [7], distributed querying of combined unstructured and structured data [8] and reference modeling of structurally complex data [9], to (c) applications for, e.g. neonatal care [10], healthcare costs reduction [11], analysis of antibiotic resistance trends [12], and pharmacovigilance signal detection [13], [14].

For an overview of big data technologies, challenges, and application examples in health and biomedicine, we refer the readers to the Editorial [15] and the contribution of the IMIA Working group "Data Mining and Big Data Analytics" [16] of this IMIA Yearbook.

About the Paper Selection

A comprehensive literature search was performed using two bibliographic databases, namely, Pubmed/Medline from NCBI (National Center for Biotechnology Information) and Web of Science® from

Thomson Reuters. The selection of papers was based on a complex query targeting big data and smart health technologies, which included various search terms appearing in the title or abstract. These terms concerned: (a) big data, e.g. "big data", "open data", "linked data", "large-scale data", "complex data", "high throughput", "data-intensive", "data-driven", "analytics", (b) smart health strategies, e.g. "personalized health", "individualized health", (c) relevant technologies, e.g. "cloud computing", "social networks", "ubiquitous computing" or "pervasive computing", and (d) filtering, aiming to "narrow" the results (especially those obtained from Web of Science®) to the healthcare sector, e.g. "clinical", "medical", "patient", "hospital", or "health". The search was restricted to journal papers that were written in English and published within 2013.

It is important to note that in case of journals publishing printed issues, 2013 was assigned to the date that the paper was published in an issue, and not the early access date (if available). In this respect, for example, the papers that are included in the special issue of JAMIA on big data that were published in vol. 21, no. 2, 2014, were omitted in the selection process, although some of them were available online within 2013 as "online first" (they will be eligible for the 2015 Yearbook selection).

The search yielded a total of 2,174 references. Initially, these references underwent blinded review by the two section editors for their relevance with the special theme. Even though the number was large, remarkably, many references were easily excluded as they contained in their abstract the searched terms (and especially the "big data" term) without being actually relevant or referring to research works. This observation is based on the fact that "big data" is being referred extensively in Editorials, "viewpoint"-type of papers, in various journals, but also used many times as a "buzzword". As a result of this initial filtering and a consensus meeting between the two section editors, we arrived at a short list of 15 papers. These candidate best papers were then been thoroughly reviewed by external reviewers (at least two), the IMIA Yearbook editors, as well as the section editors.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2014 in the section 'Big Data - Smart Health Strategies'. The articles are listed in alphabetical order of the first author's surname.

Section Big Data - Smart Health Strategies
▪ LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, Ferris TA, Shah NH. Pharmacovigilance using clinical notes. Clin Pharmacol Ther 2013 Jun;93(6):547-55.
▪ Mudunuri US, Khouja M, Repetski S, Venkataraman G, Che A, Luke BT, Girard FP, Stephens RM. Knowledge and theme discovery across very large biological data sets using distributed queries: a prototype combining unstructured and structured data. PLoS One 2013 Dec 2;8(12):e80503.
▪ O'Driscoll A, Daugelaite J, Sleator RD. 'Big Data', Hadoop and cloud computing in genomics. J Biomed Inform 2013 Oct;46(5):774-81.
▪ Post AR, Kurc T, Cholleti S, Gao J, Lin X, Bornstein W, Cantrell D, Levine D, Hohmann S, Saltz JH. The Analytic Information Warehouse (AIW): a platform for analytics using electronic health record data. J Biomed Inform 2013 Jun;46(3):410-24.

Table 1 lists the 4 papers finally selected as best papers. A content summary for each of them can be found in the appendix of this synopsis. In summary, the first paper elaborates on the use of large volumes of unstructured data, and specifically clinical notes from Electronic Health Records (EHRs) for pharmacovigilance [13]. The second paper discusses the potential for knowledge discovery by querying large volumes of complex biological data, both structured and unstructured, using big data technologies and relevant tools [8]. The third paper provides a methodological review on cloud computing and big data technologies in the field of genomics [17]. Finally, the fourth paper presents an architecture for developing the "Analytic Information Warehouse", an open source software suite enabling secure, high-performance access to and processing of large datasets extracted from EHRs [6].

Conclusions and Outlook

Big data is a highly evolving field of major interest. Its potential in biomedicine (along with criticism in some cases) has been pinpointed in various viewpoint papers and editorials. The review of current scientific literature illustrated various promising methods and applications in the field, but outcomes are still to come. Nevertheless, the field is constantly improving, as we are moving closer towards the adoption of a common conceptualization of the domain, a clearer understanding of relevant technical aspects, and the use of standardized technologies and more mature tools. We expect that future works advance further to illustrate how big data would be the enabling technology for the realization of successful personalized medicine and smart health strategies.

Acknowledgment

We would like to acknowledge the support of Mrs. Martina Hutter and of the independent reviewers, who were engaged in the paper selection process for the special theme of the current IMIA Yearbook.

References

1. Laney D. 3D data management: controlling data volume, velocity and variety. Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Retrieved: May 2014.
2. Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. Yearb Med Inform 2014:14-20.
3. Bahga A, Madisetti VK. A cloud-based approach for interoperable Electronic Health Records (EHRs). IEEE J Biomed Health Inform 2013;17(5):894-906.
4. Xia H, Asif I, Zhao X. Cloud-ECG for real time ECG monitoring and analysis. Comput Methods Programs Biomed 2013 Jun;110(3):253-9.
5. Bourgeat P, Dore V, Villemagne VL, Rowe CC, Salvado O, Fripp J. MilxXplore: a web-based system to explore large imaging datasets. J Am Med Inform Assoc 2013 Nov-Dec;20(6):1046-52.
6. Post AR, Kurc T, Cholleti S, Gao J, Lin X, Bornstein W, et al. The Analytic Information Warehouse (AIW): A platform for analytics using electronic health record data. J Biomed Inform 2013 Jun;46(3):410-24.

7. Song M, Yang H, Siadat SH, Pechenizkiy M. A comparative study of dimensionality reduction techniques to enhance trace clustering performances. Expert Syst Appl 2013 Jul;40(9):3722-37.
8. Mudunuri US, Khouja M, Repetski S, Venkataraman G, Che A, Luke BT, et al. Knowledge and theme discovery across very large biological data sets using distributed queries: a prototype combining unstructured and structured data. PLoS One 2013 Dec 2;8(12):e80503.
9. Alonso F, Lara JA, Martinez L, Pérez A, Valente JP. Generating reference models for structurally complex data. Methods Inf Med 2013;52(5):441-53.
10. McGregor C. Big data in neonatal intensive care. IEEE Computer 2013 Jun;46(6):54-9.
11. Srinivasan U, Arunasalam B. Leveraging big data analytics to reduce healthcare costs. IEEE IT Professional 2013 Nov-Dec;15(6):21-8.
12. Teodoro D, Lovis C. Empirical mode decomposition and k-nearest embedding vectors for timely analyses of antibiotic resistance trends. PLoS One 2013 Apr 25;8(4):e61180.
13. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clin Pharmacol Ther 2013 Jun;93(6):547-55.
14. Simpson SE, Madigan D, Zorych I, Schuemie MJ, Ryan PB, Suchard MA. Multiple self-controlled case series for large-scale longitudinal observational databases. Biometrics 2013 Dec;69(4):893-902.
15. Bellazzi R. Big data and biomedical informatics: a challenging opportunity. Yearb Med Inform 2014: 8-13.
16. Peek N, Holmes J, Sun J. Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. Yearb Med Inform 2014:42-7.
17. O'Driscoll A, Daugelaite J, Sleator RD. 'Big Data', Hadoop and cloud computing in genomics. J Biomed Inform 2013 Oct;46(5):774-81.

Correspondence to:
Vassilis Koutkias, PhD
INSERM, U1142, LIMICS
Campus des Cordeliers
15 rue de l' École de Médecine
75006 Paris, France
E-mail: vasileios.koutkias@inserm.fr

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2014, Special Section "Big Data – Smart Health Strategies"

LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, Ferris TA, Shah NH

Pharmacovigilance using clinical notes

Clin Pharmacol Ther 2013 Jun;93(6):547-55

Medication safety is an important priority worldwide. In the context of active, post-market drug surveillance, it is necessary to explore all types of available data sources, e.g. besides spontaneous reports, observational healthcare databases, either structured or unstructured, the literature and social media content. This paper investigates the opportunity to use clinical notes included in EHRs for pharmacovigilance.

The study employed a data source spanning 18 years of patient data from 1.8 million patients, containing 19 million encounters, 35 million coded ICD-9 diagnoses, and over 11 million unstructured clinical notes, combining pathology, radiology, and transcription data. The authors presented an approach for annotating this large volume of clinical notes via biomedical ontologies (notably, these annotations comprise ~3.75 billion records), generating this way a de-identified patient–feature matrix encoded using standardized medical terminologies. This matrix was used for detecting drug–adverse event associations and adverse events associated with drug–drug interactions.

The results indicate the potential to flag adverse events in most cases before official alerts, filter spurious signals by adjusting for potential confounding, and compile prevalence information. Using a reference set of known drug–event pairs, the paper concluded that when exposure data are numerous enough, the use of quite simple text mining with standard association strength tests for signal detection can still provide important insights.

Mudunuri US, Khouja M, Repetski S, Venkataraman G, Che A, Luke BT, Girard FP, Stephens RM

Knowledge and theme discovery across very large biological data sets using distributed queries: a prototype combining unstructured and structured data

PLoS One 2013 Dec 2;8(12):e80503

This paper elaborates on the potential of deriving meaningful information by querying large volumes of complex biological data, both structured and unstructured. In this respect, big data technologies and relevant tools were employed like Hadoop and MapReduce.

The data that were explored consisted of 20 million literature abstracts obtained from PubMed in XML format, mRNA expression data and miRNA expression data from a single Glioblastoma patient downloaded from TCGA, along with a gene and disease lexicon, using EntrezGene and NCI Thesaurus, respectively. Queries were performed using both a commodity hardware-software cluster and a commercial Big Data Appliance, and targeted (a) concept/theme discovery and (b) identification of differentially expressed genes.

The results suggest that the available technologies within the big data domain can reduce the time and the efforts needed to utilize and apply distributed queries over large datasets in practical clinical applications of the life sciences domain. In terms of methods and technologies employed, the paper sets the stage for a more detailed evaluation that investigates how various data structures and data models are best mapped to the proper computational framework.

O' Driscoll A, Daugelaite J, Sleator RD.

'Big data', Hadoop and cloud computing in genomics

J Biomed Inform 2013 Oct;46(5):774-81

This paper provides a methodological review on cloud computing and big data technologies in the field of genomics. By referring to the huge volumes of data generated by next generation sequencers, the authors illustrate the scale that both storage and processing methods have to address. An overview of cloud computing and big data technologies is provided, with explicit reference to Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS), along the "big data" perspective. The paper refers also to parallelized big data technologies and their application to genomics, providing a categorization of Hadoop-based bioinformatics implementations.

The authors identified major challenges in the field, such as the fact that big data

technologies are still in their infancy, the significant effort and the expertise that are currently required to develop parallelized programs and visualization tools, the requirement to transfer large amounts of data from/to the cloud, addressing interoperability issues and, of course, addressing privacy concerns, to name a few.

The paper provides also an outlook of the future of the big data area, foreseeing major contributions in the fields of system biology as regards models' development and validation, improved protein function prediction, metagenomics, and personalized medicine.

Post AR, Kurc T, Cholleti S, Gao J, Lin X, Bornstein W, Cantrell D, Levine D, Hohmann S, Saltz JH

The Analytic Information Warehouse (AIW): a platform for analytics using electronic health record data

J Biomed Inform 2013 Jun;46(3):410-24

The paper presents the construction of an analytics platform enabling quality improvement investigations via the specification and detection of clinical phenotypes (and other derived variables) in Electronic Health Record (EHR) data.

The paper introduces the architecture of the so-called Analytic Information Warehouse (AIW), which supports: (a) data transformation from different physical schemas into a common data model, (b) specification of derived variables in terms of the common model to enable their reuse, (c) computation of derived variables while enforcing invariants and ensuring correctness and consistency of data transformations, (d) long-term curation of derived data, and (c) export of derived data into standard analysis tools. The above features are implemented in a software suite and offered via a computing environment that enables secure high-performance access to and processing of large datasets extracted from EHRs. A specific deployment of AIW has been used as part of hospital operations in a project to reduce rates of hospital readmission within 30 days. The project examined the association of over 100 derived variables representing disease and co-morbidity phenotypes with readmissions in 5 years of data from the authors' institution's clinical data warehouse and an administrative database originated from over 200 hospitals from academic medical centers or affiliated with such centers.

The paper concludes that such a platform could accelerate the use of EHR data for quality improvement and comparative effectiveness studies. An important contribution of this work is availability of the corresponding software as open source.