Je-LKS
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal
PEER REVIEWED PAPERS
STEPS TOWARD THE DIGITAL AGENDA: OPEN DATA TO OPEN KNOWLEDGE
Vol. 10, n.2, 2014
ISSN: 1826-6223 | eISSN: 1971-8829

EXPLOITING BIG DATA FOR IMPROVING HEALTHCARE SERVICES

Massimo Mancini
ASL Bari (Italy)
massimogiuseppemancini@gmail.com

Keywords: e-Health, Open Data, Big Data, Data Analytics, OLAP

Healthcare systems produce an enormous amount of data which can be effectively used to reduce the cost of healthcare while improving quality, and supporting prevention and personalization.
This paper begins with an analysis of the basic characteristics of big data in medicine and the development of open data thinking. It goes on to discuss the potential of big data to improve the efficiency and efficacy of personal care while reducing health costs. During the study carried out at ASL Bari, a model for the continuous monitoring of health data was used and some experimental results are presented. Overall, the results provide useful insights into reducing cost and increasing the effectiveness and efficacy of health services. The experience demonstrated the usefulness of big data and data analytics techniques in health care and highlighted some directions for further research.

for citations:
Mancini M. (2014), *Exploiting Big Data for Improving Healthcare Services,* Journal of e-Learning and Knowledge Society, v.10, n.2, 23-33. ISSN: 1826-6223, e-ISSN:1971-8829

1 Introduction

Innovation in healthcare systems is at the centre of a large debate in developed countries and stakeholders are increasingly aware of the different aspects involved (Srinivasan & Arunasalam, 2013; Silva *et al.*, 2012). Features related to governance comprise legislation and governance models possessing business and funding tools that can measure returns on investment and properly model funding strategy incentives. Information-oriented aspects involve processes and strategies for information management (creation, acquisition, dissemination, use, etc.). Here both intra-enterprise and inter-enterprise processes must be considered as well as those processes involving stakeholders outside the enterprise. Instead, data-oriented aspects address the effective and efficient management of patients' personal data (i.e. Electronic Health Record - EHR), device data, and social media data while ensuring policies for personal health information disclosure according to privacy and security restrictions.

As technology has increased its ability to capture huge amounts of both structured and unstructured data, governments and companies have recognized that this big data can be a fundamental resource for discovering useful information to enhance economic and social growth. In the health care sector, access to big data is enabling stakeholders to exploit the potential of previously unusable data (Khorey, 2012; Srinivasan & Arunasalam, 2013). Indeed, the need to improve health services while reducing costs has brought about a complete rethinking of health organization and processes by using the most suitable available ICT. Hence, big data has emerged as a fundamental tool for the improvement of e-health.

An effective use of big data is rightly considered crucial to reduce the cost of health care while improving quality, prevention and personalization. Its use can produce a wide range of innovations for personalized care as well as engage patients, inform clinical decisions, reduce costs, and improve the quality of health care delivery. Big data is characterized by its large volume, high velocity and increased variety. These characteristics make big data difficult to handle using standard data management technologies and require the definition of new approaches (Frost & Sullivan, 2012; Joseph & Johnson, 2013; Vera-Baquero *et al.*, 2013).

The concept of open data refers to the policies and practices of openness regarding public data. Data is called "open" if anyone has free access to the data itself and has the ability to use it and redistribute it. Open data represents a fundamental element of open government initiatives based on public sector transparency and is a prerequisite for a successful administration model able to collaborate with the public incorporating ICT systems and techniques (Lakomaa & Kallberg, 2013; Srinivasan & Arunasalam, 2013).

The huge amount of e-health data is a very special type of open data which requires specific regulation criteria based on the specific type of health data being considered. This means finding a suitable compromise between transparency/public utility and security/privacy requirements (Hoxha & Brahaj, 2011).

Along with the growing availability of big and open health data, today, data analytics techniques allow the transformation of data into new knowledge. In fact, analytics is able to examine large amounts of health data, from a variety of data sources and formats, and employ a multitude of strategies from different fields ranging from artificial intelligence to pattern recognition, and data mining to natural language processing (Begoli & Horey, 2012; Pavel *et al.*, 2013). Therefore, data analytics can offer an understanding of the relationships underlying vast amounts of data from different data sets. Consequently, it can provide useful insights for decision makers in healthcare services and institutions and create new business value (Joseph & Johnson, 2013; Pavel *et al.*, 2013; Wu *et al.*, 2014).

This paper presents an on-going collaboration with the "Local Health Provider" in Bari, Italy (hereafter referred to as ASL-Bari). ASL-Bari exploring the use of health data in order to reduce its costs and increase the effectiveness and efficacy of its health service. For the purpose, the KNIME tool was chosen to analyze Hospital Discharge Records (HDR) and provide useful reports to the ASL-Bari decision makers. The paper is organized as follows. Section 2 presents some aspects related to big data and healthcare systems. Section 3 discusses the problem of using data analytics tools for exploiting the value of health data. Section 4 presents some experimental results carried out on data from ASL-Bari that show the validity of the proposed procedure in extracting new knowledge useful for decision makers. Then some conclusions and future directions of this research are presented in Section 5.

2 Big Data and Healthcare Systems

Big data concerns the huge amount of structured and unstructured data an organization creates. There are three main characteristics of big data. One is its enormous volume which exceeds the capability of traditional data management technologies. Another is its high velocity since companies and web-users generate new data with increasing frequency and speed. The final feature is its extensive variety since a wide range of devices connected to the internet produce a proliferation of new data types (Frost & Sullivan, 2012; Katal *et al.*, 2013; Vera-Baquero *et al.*, 2013).

Along with the development of new healthcare systems there is a growing production of data. Recent studies estimate that over 30% of all data stored on earth is medical data and this percentage is expected to increase rapidly.

Medical data are produced at very high speed by a huge quantity of devices. Therefore, they are created in a very large set of different types and formats as they originate from a variety of sources (Joseph & Johnson, 2013).

Within the wide diffusion of open data, medical data are given special attention. This is due to the very special set of issues associated with medical data such as security and privacy. These issues necessitate the definition of a large set of technological solutions in order to maintain data availability while guaranteeing the adequate protection and confidentiality required by stakeholders involved in the complex processes related to healthcare systems (Hoxha & Brahaj, 2011; Pavel *et al.*, 2013; Wu *et al.*, 2014). More precisely, to ensure the principles of open data, health data should be complete, primary, timely, accessible and searchable, readable and in non-proprietary formats, free from licenses restricting use, and reusable. Complete means that the data should include all components (metadata) that allow its exportation, online and offline use, aggregation with other resources, and diffusion on the web. Primary refers to the digital resources which must be structured so that the data is presented in a sufficient grain to allow users to integrate and aggregate it with other data and digital content. Timely indicates that users must be allowed to access and use the data in the network quickly and immediately, maximizing the value and usefulness arising from authorized access and use of these resources. Accessible and searchable means that the data must be made available to as many users as possible without contract subscription, payment or registration, and that the data must be easily identifiable on the network with search engines. Readable and in non-proprietary formats specifies that data should be machine-readable so that a computer can automatically process it and that it must be encoded in open and public formats. Free from licenses restricting their use indicates that data must have open licenses that do not restrict use, distribution or redistribution. Finally, reusable means that users should be allowed to reuse and integrate data themselves.

It is worth noting that the open government data movement promotes the opening of a specific category of data which is identified as Public Sector Information (PSI), non-personal (anonymized or pseudonymized), that administrations produce in the performance of their institutional tasks and that, if made available via web in an open format, are capable of increasing transparency and encouraging collaborative interaction between citizens and government. Indeed, PSI embodies a considerable economic potential which depends on its reusability to create services based on the aggregation and presentation of data (Hoxha & Brahaj, 2011; Joseph & Johnson, 2013; Katal *et al.*, 2013).

3 Big Data: Exploiting its Value in a Healthcare System

Mining large amounts of data and looking at patterns and models to understand actual conditions and forecast future scenarios has become a reality with the development of advanced health infrastructure combined with other data sources based on new medical devices and social media. This paper divides the areas of improvement into three categories (Hoxha & Brahaj, 2011; Taleb *et al.*, 2010):

Policy, financial and administrative tasks – through the use of well-defined key parameters, extracted from big and open data, data analytics techniques can support decision makers in defining policy, and financial and administrative tasks.

Population-oriented and public tasks – social media provide a huge amount of data that can be useful both to support clinicians in understanding disease trends in populations, and as a way of obtaining feedback from users on the quality of healthcare received.

Clinical tasks – clinical data analysis is fundamental to understand trends in diseases on a personal and company scale. In fact, clinical data analysis can support real time identification of best practice treatments to patients on the basis of specific pattern analysis strategies able to enhance pre-diagnosis and early disease detection. An outcome-based analysis allows the definition of the optimal treatment for specific patients by analyzing comprehensive patient and outcome data to compare the effectiveness of various procedures. Clinical data are useful to monitor treatment adherence with the aim of reducing hospitalization time, emergency room visits and long-term health complications.

The net result is that data analytics can provide higher-quality patient treatment and more cost effective new treatments while providing for better public health surveillance and disease response (Srinivasan & Arunasalam, 2013). Yet to really make data analytics useful in health care requires a substantial rethinking of the way health care is directed and implemented.

With this in mind, Figure 1 schematically sketches data flow as it was modelled in this work. A closed-loop strategy was used for continuous monitoring of big/open health data. A seen in the figure, data originates from ASL (internal data) and from the other sources (external data) and is then divided into four types (Pavel *et al.*, 2013):

Organization-oriented data: mainly administrative data that support the medical organization (billing, scheduling, etc.);

Social-oriented data: user data obtained also with web-based applications (on-line reservations, social networks, forums, etc.);

Clinical-oriented data: the main part of medical data involving primarily

unstructured clinical documents, medical images and signals (i.e. X-ray computed tomography, PET, ECG, etc.) as well as data from sensor-based devices for remote health monitoring systems;

Scientific-oriented data: scientific publications and medical reference material (i.e. PubMed, ClinicalTrials, Genomic data, etc.).

A Data Analytics Engine (DAE) was used to extract new knowledge from data consistent with Data Analysis Monitor (DAM) strategies. These include tactics specifically devoted to continuous data monitoring in order to automatically detect critical conditions (in terms of costs or quality of service) and report this information to Decision Makers (DMs). Alternatively, DMs can also interact directly with the DAM through a non-standard request of analysis. The DMs can then make the decisions that are used to tune ASL in order to achieve a continuous performance improvement.

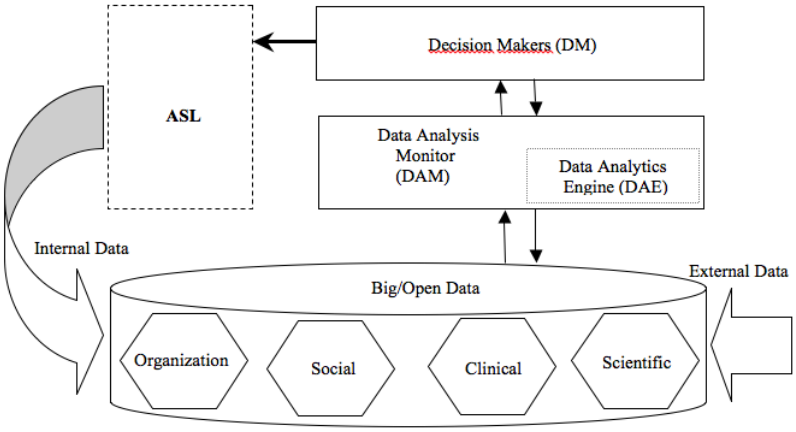

Fig. 1 - Data Flow System

The huge amount of structured and unstructured data in the system required the application of advanced tools in order to extract knowledge that was useful for improving the e-health system and reducing costs. Among others, new technological solutions were necessary in several fields ranging from standard definition data visualization to analytics (Keim *et al.*, 2013; Zheng *et al.*, 2012). In particular, data analytics provided advanced techniques for examining the large amounts of health data from a variety of data sources and in different formats. This was possible because data analytics was able to adopt a multitude of strategies from different fields ranging from artificial intelligence to pattern recognition and data mining to natural language. Hence, one of the aims of the proposed approach was to use data analytics to understand the relationships

underlying this vast amount of data and derive new knowledge useful for healthcare service decision makers (Begoli & Horey, 2012; Keim *et al.*, 2013).

4 Experimental Applications and Results

ASL-Bari is a regional company for e-health that serves more than 1.2 million people living in 41 cities and grouped in 14 health districts. Due to the recent availability of advanced systems for data acquisition and collection, ASL-Bari has been collecting huge amounts of data. Hence, the health service has set out to exploit the potential of this available big data to improve its efficacy and efficiency as well as provide high quality health services to an increasing number of patients. This study focuses primarily on the Hospital Discharge Form (HDF) available at ASL Bari. The HDF is an information tool for the collection of information about each patient discharged from hospital (Silva-Ferreira *et al.*, 2012; Taleb *et al.*, 2010). It is compiled by the doctor that took care of the hospitalized patient and represents a concise and faithful medical record with the aim of enabling a systematic collection of key information. The HDF contains an explanation of several facts like Diagnosis, Surgery, etc. Figure 2 shows the Fact Table for Diagnosis (Pavel *et al.*, 2013; Silva-Ferreira *et al.*, 2012).

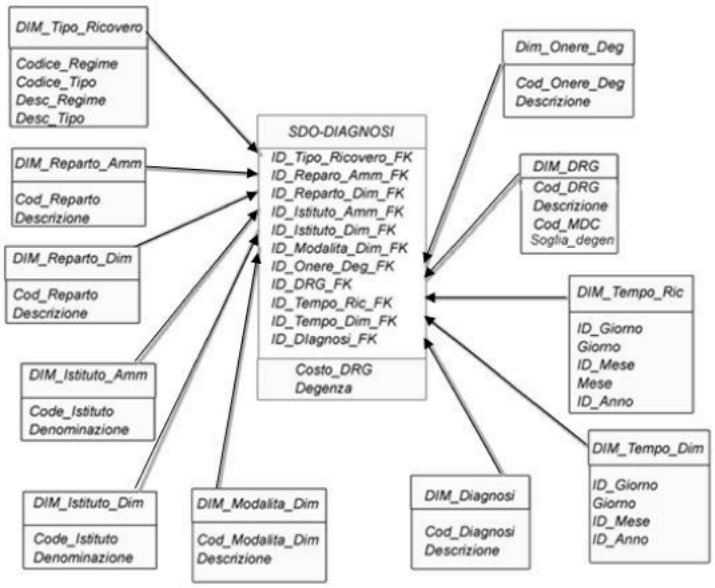

Fig. 2 - Diagnosis - Fact Table

For data analytics and reporting, KNIME software was used. KNIME can

carry out several routines for data analytics in many fields ranging from data mining to time series analysis, from text mining to network analysis and social media analysis. KNIME allows data from various sources and can work in either interactive or batch mode, enabling the data analysis process to be easily integrated into a local job management run on a periodic basis (Begoli & Horey, 2012; Bernd *et al.*, 2013; Keim *et al.*, 2013).

From the data, a wide set of analyses was performed. For instance, Figure 3 shows the admissions for "laparoscopic appendectomy" surgery by sanitary district. It is worth noting that the reported sanitary districts are identified by an integer number from 1 to 10 since this surgery is performed in only 10 of the 14 sanitary districts. Figure 4 reports the average number of hospitalization days for each sanitary district.

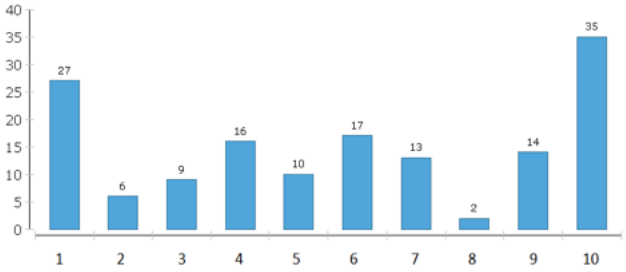

Fig. 3 - Admissions in sanitary districts

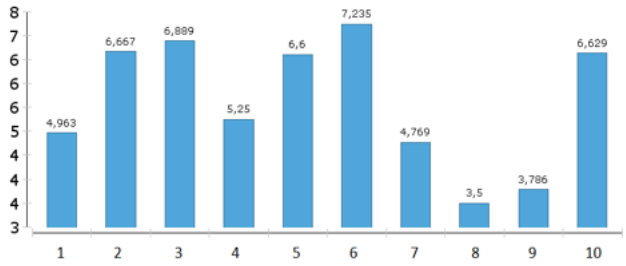

Fig. 4 - Average days of hospitalization per district

Of course, more useful results can be achieved when different information is combined. The scatter plot in Figure 5 shows the number of admissions versus the average days of hospitalization in each sanitary district. In relation to the plot parameters, the most cost efficient/timely sanitary districts are those in the lower-right part of the plot (marked by "H"), whereas the less cost efficient/timely sanitary districts can be found in the high-left part of the plot (marked

by "L").

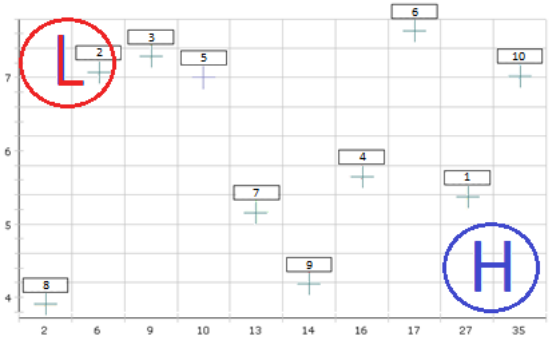

Fig. 5 - Admissions vs. Average days of hospitalization

In addition, the analysis provided an automatic and continuous monitoring of the sanitary districts, in order to verify their levels for specific quality criteria. A simple and intuitive visualization was used to signal critical conditions, as Figure 6 shows (Bernd *et al.*, 2013). In this case, a simple coloured circle indicated, in runtime mode, the sanitary districts with critical performance. More specifically, Figure 6 indicates that a critical condition is occurring at sanitary districts n. 2 and 3, standard conditions are occurring at sanitary districts n. 4, 5, 6, 7, 8 and 10, whereas very good conditions are in progress at sanitary districts n. 1 and 9.

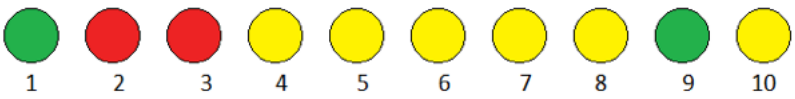

Fig. 6 - Performance Visualization

Discussion and Conclusion

The development of e-health is currently at the centre of a large debate in all developed countries. E-health is rightly considered a critical sector where the expectation of health care and well-being of a large population is in contrast with spending reviews and the cost saving requirements.

When used according to open data philosophy, big data has enormous potential in health fields. It allows for the acquisition and dissemination of huge amounts of knowledge useful for health care system improvement and cost reduction. In fact, data analytics techniques can be used to transform large amounts of previously unusable data, generated by both e-health organizations and social network users, into useful predictive insights and new knowledge.

This paper presents some results of the exploitation of big data obtained by and for the ASL-Bari health agency. In particular, data analytics techniques were considered for the analysis of hospital discharge forms while advanced reporting tools were used for result visualization. The results demonstrate the effectiveness of data analytics and reporting tools to support decision

makers as well as highlight their role in improving the quality of services while reducing costs.

REFERENCES

Begoli E., Horey J. (2012), *Design Principles for Effective Knowledge Discovery from Big Data*, European Conference on Software Architecture (ECSA), pp. 215-218.

Bernd B. W., Kötter T., Silipo R. (2013), *Analyzing the Web from Start to Finish Knowledge Extraction from a Web Forum using KNIME*, white paper, http://www.knime.org/files/knime_web_knowledge_extraction.pdf

Frost and Sullivan (2012), *Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations*, website: http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technology-complexities-ar.pdf

Hoxha J., Brahaj, A. (2011), *Open Government Data on the Web: A Semantic Approach*, 2011 International Conference on Emerging Intelligent Data and Web technologies (EIDWIT), pp. 107 - 113

Joseph R.C., Johnson N.A. (2013), *Big Data and Transformational Government*, IT Professional, Vol. 15, N. 6, pp. 43 - 48.

Katal A., Wazid M., Goudar R.H. (2013), *Big data: Issues, challenges, tools and Good practices*, Sixth International Conference on Contemporany Computing (IC3), pp. 404-409.

Keim D., Huamin Q., Kwan-Liu M. (2013), *Big-Data Visualization*, IEEE Computer Graphics and Applications, Vol. 33, N. 4. pp. 20 - 21.

Khorey L. (2012), *Big Data, Bigger Outcomes*. Journal of American Health Information Management Association (AHIMA), Vol. 83, n.10, pp. 38-43.

Lakomaa E., Kallberg J. (2013), *Open Data as a Foundation for Innovation: The Enabling Effect of Free Public Sector Information for Entrepreneurs*, IEEE Access, Vol. 1, pp. 558 - 563.

Pavel M., Jimison H.B., Wactlar H.D., Hayes T.L., Barkis W., Skapik J., Kaye J. (2013), *The Role of Technology and Engineering Models in Transforming Healthcare*, IEEE Reviews in Biomedical Engineering, Vol. 6, pp. 156 - 177.

Silva-Ferreira P.R., Patriarca-Almeida J.H.;Vieira-Marques P.M., Cruz-Correia R.J. (2012), *Improving expressiveness of agents using openEHR to retrieve multi-institutional health data: Feeding local repositories through HL7 based providers*, 7th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-5.

Srinivasan U., Arunasalam B. (2013), *Leveraging Big Data Analytics to Reduce Healthcare Costs*, IT Professional, Vol. 15, N. 6, pp. 21-28.

Taleb T., Bottazzi D., Nasser N. (2010), *A Novel Middleware Solution to Improve Ubiquitous Healthcare Systems Aided by Affective Information*, IEEE Transactions on Information Technology in Biomedicine, Vol. 14, N. 2, pp. 335 - 349.

Vera-Baquero A., Colomo-Palacios R., Molloy, O. (2013), *Business Process Analytics Using a Big Data Approach*, IT Professional, Vol. 15, N. 6, pp. 29 - 35.

Wu X., Wu G. Ding W. (2014), *Data Mining with Big Data*, IEEE Transactions on Knowledge and Data Engineering, pp. 97-107.

Zheng Z., Zhu J., Lyu, M.R. (2013), *Service-Generated Big Data and Big Data-as-a-Service: An Overview*, IEEE Int. Congress on Big Data, 2013, pp. 403-410.