Big Data and Health Economics: Opportunities, Challenges and Risks

Diego J. Bodas-Sagi, José M. Labeaga

Universidad Nacional de Educación a Distancia (UNED) (Spain)

Received 15 July 2016 | Accepted 3 February 2017 | Published 27 March 2017

Abstract

Big Data offers opportunities in many fields. Healthcare is not an exception. In this paper we summarize the possibilities of Big Data and Big Data technologies to offer useful information to policy makers. In a world with tight public budgets and ageing populations we feel necessary to save costs in any production process. The use of outcomes from Big Data could be in the future a way to improve decisions at a lower cost than today. In addition to list the advantages of properly using data and technologies from Big Data, we also show some challenges and risks that analysts could face. We also present an hypothetical example of the use of administrative records with health information both for diagnoses and patients.

Keywords

Big Data, Healthcare, Data Science, Machine Learning.

DOI: 10.9781/ijimai.2017.03.007

I. Introduction

According to Edd Dumbill from O'Really Media, "*Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it*" [1]. The challenge is not about dealing with trillions of bytes of streaming data, it is about getting started with a quantitative approach so that you can drive value from your data, whatever size that data is. Data Scientists help understand the value of data to take timely and relevant actions [2].

Our economy depends on data. Data is everywhere, in every sector, in every country. We generate and consume data. Our interaction with machines, people, companies and public institutions produces data. Information allows us to improve the business processes and provide our customers and partners with the best quality standards, services and products. Big Data generates value in several ways, according to McKinsey [3]:

- Creating transparency: making data accessible timely manner.
- Enabling experimentation: collecting more accurate and detailed performance data, setting up controlling experiments.
- Segmenting populations to customize actions, target promotions and advertisement.
- Replacing/supporting human decisions making with automated algorithms.
- Innovating new business models, products and services: UBER, Spotify, LinkedIn, Twitter, Netflix are well-known examples of this.

Big Data is affecting healthcare too. In 2012, worldwide health care data reached 500 petabytes and it is expected that in 2020 there will be more than 25000 petabytes available. The 2011 report by McKinsey Global Institute estimate that the potential value that can be extracted from data in the healthcare sector in US could be more than $300 billion per year.

Again, in the US, several initiatives encouraging the use of Big Data for health, like the Affordable Care Act, a set of health data initiatives by the department of health and human services. The Heritage Provider Network Health Prize (http://www.heritagehealthprize.com) challenge offers a $3 millions prize to improve healthcare avoiding unnecessary hospital admissions. More than 71 million individuals in the United States are admitted to hospitals each year, which approximately implies a $30 billion bill wasted, according to a survey from the American Hospital Association. Medicare penalizes hospitals that have high rates of readmissions among patients with hearth failure, hearth attack and pneumonia, to avoid this loss.

US Government holds other projects like BRAIN (Brain Research through Advancing Innovative Neurotechnologies), bolding $100 million to revolutionize our understanding of the human brain (that generates a huge amount of information). The scientists' goal is to get answers to Alzheimer's disease, epilepsy and new treatments for traumatic brain injury. In March 2012, the Obama Administration launched a $200 million "Big Data Research and Development Initiative", one of whose main aims is to transform the use of big data for scientific discovery and biomedical research.

The European Commission (EU) is not an exception and some projects have been proposed under the EU Research and Innovation Programme Horizon 2020. Other Health 2.0 initiatives are being carried out by different countries with the aim of accelerating innovation and obtaining better ways to manage patients, institutions and establish more convenient policies. Medical institutions, insurance companies and governments are applying healthcare Big Data to cut down medical service costs and to optimize patient's attention.

Many initiatives gaze at or are focused on healthcare data digitalization. An Electronic Health Record (EHR) refers to the systematized collection of patient and population electronically stored health information in a digital format [4]. In 2005, only about 30% of office-based physicians and hospital in the US used EHR. By the end

* Corresponding author.

E-mail addresses: diegobodas@yahoo.es (D. J. Bodas-Sagi), jlabeaga@cee.uned.es (J. M. Labeaga).

of 2011, this figure rose to more than 50% for physicians and 75% for hospitals. We come back to the importance of EHR for healthcare below.

The rest of the paper is structured as follows: First, we explore the relation between healthcare and economy. Next, we try to explain in more detail the opportunities offered by Big Data in the healthcare sector. We also comment on some challenges and risks. Big Data projects require a specific methodological approach commented in section V. Section VII present a hypothetical example based on real although non-public data from administrative records. Finally, we conclude and summarize.

II. Healthcare and the Economy

Health performance is positively correlated with economic performance; wealthier countries have healthier populations [5]. In many countries, the healthcare system has to afford several major challenges including ageing populations, chronic illnesses, ensuring universal access, guarantying equity and raising quality of care. New technologies and data analysis techniques might help to overcome these tasks. According to the Organisation for Economic Co-operation and Development (OECD) and due to the economic crisis, many years of consecutive health expending growth ground to a halt in 2008; health budgets were cut since then and, they are likely to remain tight for a number of years to come [6, 7]. On average, countries devote only 3% of their health budgets to spending on prevention [8]. The OECD recommends to policy makers focus their efforts on building health systems that meet population needs and deliver excellent value for money. Being able to reliably measure and compare health system performance will be crucial to achieve this goal. In this context, it is very important for governments and institutions to obtain timely data. It could help to ensure adequate and sustainable provision of high-quality services at correct administrative costs. In addition, it is necessary to study the occurrence and cost of fraud, abuse and corruption in health systems, as well as the policies to fight them. All these aims require new data, new statistics, better measures of outcomes and more patient-reported measures and it opens the door to the use of big data and suitable methods.

Health systems must adapt to take advantage of the development of new technologies to get personalized medicine. This paradigm tries to overcome the limitations of traditional medicine taking into account the unique genetic map for each individual. It is mandatory to effectively integrate new technologies into health systems to get personalized medicine and move to national aggregate measures of health care quality to more granular measures at hospitals [6, 7]. Healthcare information and advanced analytics may contribute to shift from population-based evidence for healthcare decision-making to the fusion of population and individual-based evidence in healthcare [9]. The effects might be immediate and cover from better treatments and diagnoses to reduce labor force transitions after a health/disability shock [10].

Focusing on the economic aspects of healthcare, we need to improve the state-of-the-art of forecasting models of health spending to develop expenditure projections that explore the impact of different policy scenarios and policies [11]. Some facts confirm this statement. The US has not seen an increase in life expectancy or last-days quality of life to match its huge outlay on health care. Although the US healthcare expenditures are the highest of any developed country, at 17,1% of GDP in 2014, according to the World Health Organization Global Health Expenditure database (http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS), such expenditures does not seem to improve health outcomes. However, the rising cost of medical care and health insurance is impacting the livelihood of many Americans [12].

Other countries have different problems. South Korea has one of the most advanced information technology (IT) infrastructures in the world. The application of IT in health systems is rapidly progressing from computerization to information, ubiquitous and start systems. All in all, the cost of health in terms of GDP is 7,4%, less than half the cost in of the US system. However, a major problem concerning healthcare resources lies in the regional disparities between medical services [13].

There are many researchers and institutions involved in the study of economic and health inequality [14]. Inequalities in health are linked to many factors, including differences in exposure to risk factors, and differences in the access to health care [15]. The economic crisis has had deep consequences in the labor market and public policies of many developed countries. Labor market conditions have deteriorated with increased unemployment rates and wage cuts, reforms in the public pension and the health care systems, among others. Undoubtedly, this may have an impact on the population's health and/or the equity and efficiency of healthcare systems. Several authors provide evidence on the relationship among unemployment rates, business cycle conditions or housing conditions on health variables in the short-run [16, 17, 18]. Budget cuts have an impact on dependent people, for example and in Spain, demand for private long-term care insurance has grown in recent years and this can be attributed to budget cuts affecting the implementation of System of Autonomy and Attention to Dependent People [19].

Of course, we cannot ignore the potential that Big Data Technologies provide to pharmaceutical companies. In this sector spending is declining in real terms, due to top-selling drug patent losses and to fiscal consolidation measures adopted by many OECD countries. Using Big Data, pharmaceutical companies can better identify new potential drug candidates and develop new effective products, approving and reimbursing medicines more quickly [3].

III. Big Data Opportunities

Healthcare needs more efficient practices, research, and tools to harness the full benefits of personal health and healthcare-related data [20]. Many healthcare researches use advanced analytics tools to bring order, understanding data and reduce complexity. Researches, hospitals and physicians have access to rich sources of data that have potential for an increased understanding of disease mechanisms and better reporting. However, the size and complexity of the data present many challenges. There is a recognizable need for scalable tools that can discover patterns without discounting the statistical complexity of heterogeneous data or falling prey to the noise it includes [20]. This data-driven culture together with a share-knowledge attitude can play a critical role in the emergence of personalized healthcare. Numerous diseases have preventable risks factors or at least indicators of risk. Improving the prevention systems is possible and viable; we can consider not only healthcare or genomic variables, but also economic, demographic and lifestyle variables. Healthcare is moving from a disease-centered model towards a patient-centered model [21, 22]. Big Data technologies offer many opportunities to proactive medicine too. From the clinical patient's data, it is possible to find similarities of that patient to millions of other patients. So, this allows physicians to go ahead and predict the likely of new relapses and the effect of drugs.

Getting into further detail, Big Data Analytics will impact healthcare in several ways [23, 24]:

- Right living: data can help patients to take an active role in their own health (i.e. practicing some sports).
- Right care: data can improve outcomes and reduce medical errors.
- Right provider: hospital and patients can select the best provider based on data.
- Right value: data analytics has potential to eliminate fraud, waste and abuse.
- Right innovation: a sharing-knowledge culture and data-driven

networks allow more flexible, efficient and innovative ways of working.

- Providing patient centric services: provide faster relief by providing evidence-based medicine, reducing readmissions and reducing costs.
- Detecting spreading diseases earlier.
- Monitoring the hospital's quality.
- Improving the treatment methods.

Frequently, Big Data and machine learning go together. Most of the challenges previously cited require a machine learning approach to obtain suitable models. Some applications can be found in [25]. Data from a variety of sources can be used to improve the accuracy of determining which chemical compounds would be effective drug treatments for a variety of diseases. Machine learning at scale has significant potential to boost drug discovery [26]. Some medical specializations like radiology need to deal with different formats like image or text, working with Big Data technologies, researchers can process together different data structures obtaining knowledge [27, 28].

A. Data Sources and Big Data

The majority of healthcare data are structured rather than semi-structured or unstructured. On the one hand, data refers for relational database records, clinical notes, clinical images, statistical data, electronic healthcare records (EHR) and so on. On the other, researchers in the field of applied health economics often use survey data. For example, the Health and Lifestyle Survey (HALS) in Great Britain requires 1 hour face-to-face interview plus several questionnaires (physiological + cognitive + functional). In the 1984 – 1985 edition, only 53.7% on a sample of 12,672 people provided a complete answer to the full questionnaires. In addition, researchers take into account other information from economic and demographic surveys, reporting socioeconomic status, household income, education, marital status, ethnic, children, ages, etc. These surveys provide aggregate instead of personal data [29].

On the contrary, Big Data analytics are frequently based on individual (anonymized) and dated data. This kind of data comes from 'real' and individual actions (usually administrative records), not from surveys. A real action can be a visit to a physician, a surgery, a treatment, a clinical checkup…. Although this data is also often aggregated due to privacy requirements, both in its individual or its aggregate forms, it offers more possibilities to researchers. Now, scientists do not completely depend on complex surveys designs, they have access to timely and relevant information based on individual records. This increase in data make easier to adopt the machine learning methodology. Moreover, access to sufficient data provide advantages as reduction of problems to obtain our training sample, more validation possibilities and datasets for additional testing.

Many works evaluate lifestyle data in conjunction with health data, smoking, drinking and related behaviors that have a direct impact in health [30, 31]. The HALS is commonly used as primary dataset. This survey compiles questionnaire answers related to this subject. Around 10,000 individuals are interviewed each year with a low rate of complete responses. It is not trivial for surveys to take into account health and lifestyles models or health related behavior due to technical restrictions or attrition bias. With surveys, researchers cannot resample and obtain other values due to the required time to perform the survey and the technical complexity. Biased data is not an accurate reflection of reality. If data reflects biases, the obtained models can be wrong.

Surveys have also problems when researchers like to study the evolution over time of some variables. One problem is the frequency of the data. For some research questions it would be important to have daily data at hand but it is not usual in survey data, where it is common to employ monthly or annual data. However, data democratization is coming to help researchers all over the world to commit their objectives. For example, Banco Bilbao Vizcaya Argentaria bank (BBVA) is pioneering a new service generation providing anonymazed transaction data and offering forms of collaboration with research institutions and universities. Transaction data is a valuable information source including data about expenses using Point of Sale (PoS), pharmacies, gyms, etc. Again, this information can be completed by open data, including air quality, climatic parameters, etc. When we like to answer questions about public health arising from environmental problems, individual or aggregate health records can be complemented both with previous climatic variables and also with variables from smart cities (see, for instance, the decumanus project -http://www.decumanus-fp7.eu/home/).

The need for semantically interoperable EHR is now a well-established tenet [32, 33, 34]. Market mobility of the population (changes of residence, job changes, tourism) and its demand to have access to services of similar quality to those of their place of origin are factors that set in motion the creation of information systems based on interoperable EHR. For researchers, EHR implies the possibility of access to detailed information about individual patients, clinical histories, clinical notes, family histories, treatments and results, etc., in short, all the patient's medical information. The spread of this standard remains difficult and challenging. The adoption of EHR implies a slow process. The integration if this kind of data with hospital information systems is a tough task. We consider that it is necessary to promote dissemination campaigns on the use of the standard to avoid errors and to make users aware of the importance of completing the requested information.

IV. Challenges and Risks

Despite the benefits of the Big Data, some resistances have to be solved [3]:

- Resistance to change: providers used to make treatment decisions independently using their own clinical judgment rather than protocols based on big data. However, Big Data technologies and algorithms are not intended to replace physicians, they just try to support the decision-making process.
- Resistance to uncertain returns: many Big Data projects should be viewed from an experimental and research side. It not possible to determine in advance the accuracy of the developed algorithms.
- Resistance to face new challenges related with privacy and deal with many players, technologies and data sources.

To solve these resistances, it is necessary to develop and spread talent transformation initiatives involving physician, managing positions and technical staff. The objective is to show the benefits linked to Big Data and, at the same time, to raise awareness of the risks associated with the management of expectation, privacy, security and technical challenges.

Data anonymization is a mandatory step to comply with the current legislation [35]. The available of open health data for secondary use is fundamental for advance in the medical knowledge. The use of public datasets by researchers has effects on the acceleration of scientific advances as well as improvements in both the efficiency and efficacy of health processes [36]. A responsible use of individual's data must be guaranteed, but it is possible to reconcile individual data privacy with socially valuable uses [37].

Many initiatives are trying to evolve the security and privacy standards. Some proposals come from sectors others than the healthcare system. For example, Blockchain systems were first developed for finance applications. Blockchain is paramount to realize the benefits of improved data integrity, decentralization and disintermediation of trust, and reduced transaction costs. It can offer a promising new distributed framework to amplify and support integration of health care

information across a range of uses and stakeholders. Blockchain relies on established cryptographic techniques to allow each participant in a network to interact without preexisting trust between the parties without in a model where there is not a central authority [38].

Clinical notes are a common piece of information in the daily routine of physicians, hospitals and laboratories. Understanding clinical notes in the right context is a great challenge. Clinical notes introduce mistakes, ungrammatical and short phrases, abbreviations, misspellings, semi-structured information, inconsistencies, which do not reflect all the information.

The state-of-the-art in text mining applications is evolving quickly and some successful use cases have been achieved. Watson from IBM, for instance, is able to understand questions and context, and to analyze through 200 millions pages of data and provide precise responses in seconds to physicians.

We can list in a non-exhaustive way a list of other technical challenges:

- Select risk factors [39]. Big data technologies and machine learning techniques made possible to obtain scalable models using large amount of data.
- Patient similarity: researcher uses graphs theory and similarity measures to obtain patient similarity patters and improve the prevention system. For example, collaborative filtering methods allow us to leverage similarities across a large group of patient pool in real-time to deliver a personalized treatment (taking into account all available demographics and previous medical history). Using Big Data Science, we can generate predictions focused on other diseased that are based on data from similar patients.
- Medical Image Retrieval: analyzing huge image databases imply dealing with high dimensional and complex data. Dimensionality reduction techniques are useful to process this information.
- Genetic data: Using genetic data for treatment optimization. Single human genome is about 3Gs. In order to get a complete genome the use of cloud technologies implies a cost around $5000.
- Public health: it is interesting to understanding disparities related to race, social condition, age, gender for epidemics or illnesses, using both clinical and socio-economic data.

V. Methodologies and models

Big Data and Data Science terms are strongly related. Data Science is about how to extract knowledge or insights from data in various forms, either structured or unstructured [40]. This scientific field implies the use of statistics, machine learning, data mining, predictive analytics, coding, etc. The Data Science process has been explained in [41] and is shown in the next figure.

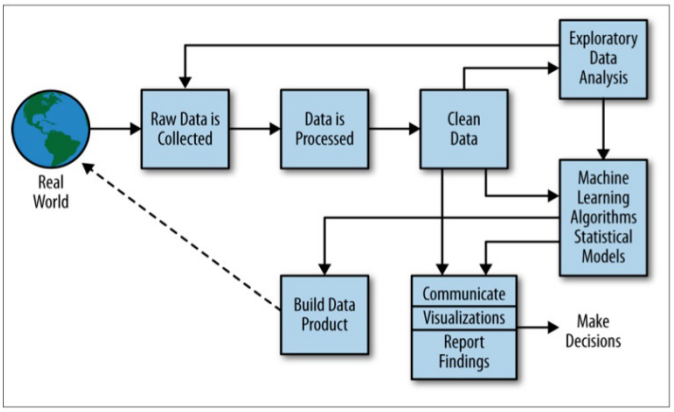


Fig. 1. The Data Science Process [41].

Working with raw data involves spending time processing and cleaning data. The data acquisition phase requires up to 80% of the project time. Exploratory Data Analysis covers ways to summarize and visualize important characteristics of a data set. This step allows us to describe our data and generate hypotheses. After building our models (using machine learning or alternative algorithms) it is necessary to communicate the achieved results in an effective way. Finally, we can build a data product that adds value to companies, physicians and potential patients. As we can see in Figure 1, this is not a straightforward path.

Collecting, processing and cleaning data correctly (the data acquisition phase) are tough tasks. This is why is useful to standardize some steps in this phase. Figure 2 shows some recommended steps according many authors' experience (and also ours).

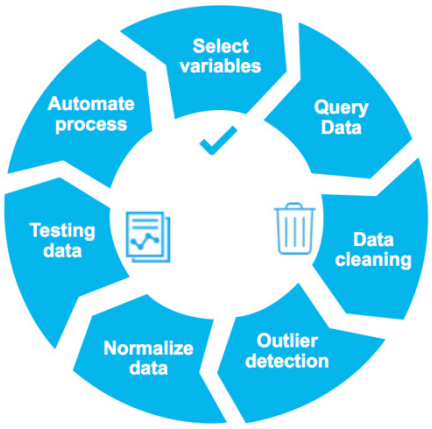


Fig. 2. The data acquisition phase.

We can start selecting variables or data sources to collect and querying the data from a distributed file system, relational database management system or non-relational database. Next, cleaning data is a common task. Outliers need to be identified and treated properly. Data normalization allows us to seek for relations and compare results. Before automate the process, we need to test data in order to analyze whether variables and data source are useful and enough to accomplish our goals.

Researches in this area need to consider training and test set or, better, training, validation and test set if necessary for model selection using Big Data. We use the training set to discover relations and train the model. The validation set is used for tuning model parameters and, finally, the test set is used for performance evaluation. All these phases involve the use of some models, which normally must be specific to health-related variables. It is common in these cases to have counts (visits to physicians, episodes of illness, etc.), discrete variables (decisions made by physicians-patients, either in an agency-principal model or in any other context), data reflecting dynamics (duration of an illness-episode, duration of a treatment, etc.) and many other. Dealing with these specific data requires specific methods when we try to extract causal relationships. All these issues are of course out of the scope of this paper.

VI. An EmpiricalExample

This section is devoted to describe an example of the use of Big Data and Big Data methods applied to a specific health problem, which also involves economic issues. Data could be in a repository and it could correspond to administrative records of a health institution (a hospital, a diagnosis center, etc.). The availability of such a repository containing medical diagnoses could allow doing some analysis based on the text introduced by the practitioners. Let us assume, for instance, that we have a large database containing image-diagnostic reports

(we assume a sufficiently large sample size to avoid problems of non-representativeness). In these cases, we usually have data concerning patient ID, radiologist ID, hospital, section, date / time of the clinical test, date / time of the diagnosis, room in which the test has been performed, type of test (X-ray, CT, ultrasound, magnetic resonance, etc.). In addition, records usually contain a comment written by the physician who ordered the imaging test. The information is normally completed with patient-related data such as birthdate, address, birthplace, sex and, finally, some material with clinical diagnoses.

A problem of interest faced in reality and developed in some research papers may be to automatically recognize medical diagnoses that report an allergic reaction to the contrast provided to make the test. In this way, it is possible to compare incidence rates between different hospitals to test if there is any effect of the procedures used at the different units or even if any effect arising from clinical materials of the different suppliers could be identified. This question is not trivial because, in some cases, the doctor will note that the patient "reports that he has previously suffered an allergic reaction or is allergic to contrast". On the other hand, if the patient suffers an allergic reaction, the doctor should write it indicating the necessity (or not) to provide any treatment.

A classic and standard approach to address this type of problem consists in selecting a broad sample where researchers manually label reports with and without allergic reaction occurred during the test (of course, researchers must read and interpret the recorded text). Therefore and for this exercise, we only need the clinical diagnosis issued by the practitioner, in a raw text format that must be vectorized prior to the elimination of stop-words. This just constitutes a classification problem. The goal is to obtain an algorithm that allows detecting the occasions where the patient has suffered an allergic reaction during the test. The original sample is separated into training and checking sets. We use a very simple example of the potential of some approaches based on previous works by the authors using text containing Spanish language. We can report some conclusions from empirical evidence obtained: when the text contains the words (in Spanish) "alérgica" and "refiere" there is a 97% probability that the doctor refers to a previous problem with the contrast or he is reporting the patient is contrast-allergic. On the contrary, if only the word "alérgica" appears, there is a high probability that an allergic episode is being reported.

Classification algorithms can be used jointly with association rules. Association rules algorithms allow us to obtain the relevant item sets (those ones with support values close to or higher than the chosen reference). In this context, an item set is a bag of words indicating common word association in radiology reports. Furthermore, rules allow us to find associations between words. For example, in our dataset we have found that if the report contains the (Spanish) words "conclusión" and "compara" it is highly likely (+96%) the appearance of the word "previo". This association computed for a real problem is presented in Figure 3. A very simple conclusion can be inferred: radiologists are frequently studying the progress of diseases and they need to compare different tests performed on different dates. But also, radiologists can associate different histories attending common socio-demographic variables in a way such patients can benefit from better reports. Even using very simple examples, we hope the reader can realize the potential of these techniques using a huge dataset for saving time and costs.

How can these methods help to reduce cost or to advance in individual diagnosis (personalized medicine)? Text vectorization also makes it possible to analyze the most similar diagnoses (using similarity measures) to a given one. In this way and, after receiving a new clinical diagnosis, it is possible to find similar diagnoses in the historical to check if there is a relationship between patients already treated and the new one. If this relationship exists, as pointed out in previous paragraphs of this paper, the patient can be given a better preventive service based on the evolution of similar cases. Finally, this same technique also allows the execution of topic modeling algorithms that permit grouping documents into different topics. This is useful for launching new research hypotheses based on documents that are grouped in an unexpected way. Classification algorithms can also help to infer associations among clinical terms in relationships sometimes unknown. They also can help association among patients who share similar demographics with benefits for the quality of the diagnosis and treatments. All in all, the health services can avoid in many cases spending because of trial-error in the treatments to new patients and, in many other cases, they can shorten the duration of the treatment when the historical information results in an adequate treatment for the new patients. In both cases, the health institutions can get important budget cuts and they can also optimize the quality of the of the care provided.

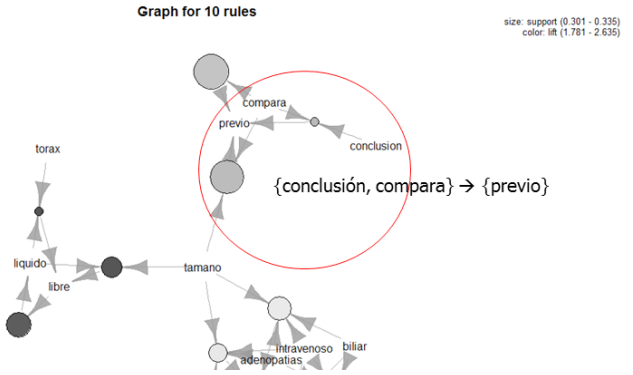


Fig. 3. Example of association rules

VII. Conclusions

Any economy depends on using data. Data is everywhere, in every sector, in every institution, in every country. The potential value that can be extracted from data in the healthcare sector is considerable and promising. Big Data offers many opportunities but there are some associated risks too. To ensure patient's privacy is paramount. Solving privacy frontiers will allow researchers to share data and accelerate the availability of results. Researchers need to consider Data Science methodologies to be able to successfully deal with huge amount of data.

The current state of public finances in many countries could help decision-makers in adopting some measures or policies concerning the use of all available information in a more effective and efficient ways to reduce both costs of administration and production in the healthcare sector. Each day the users of public or private institutions providing health services produce thousands of administrative records, which can be used by analysts and researchers to inform the policies as a way of ex – ante or ex – post evaluation of them. Here is where Big Data and Big Data technologies have opportunities and challenges, but also risks.

The collaboration of public and private institutions with experts and researchers in various fields (privacy, anonymization, machine learning...) is required, when we like to take full advantage provided by Big Data and Big Data technologies. All agents must work together transparently. In addition, it is necessary to inform and train all those involved effectively.

Acknowledgment

We acknowledge very useful comments from an editor of the journal.

References

[1] McKinsey & Company, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Glob. Inst.*, no. June, p. 156, 2011.

[2] E. Dumbill. What is Big Data? In Introduction to the Big Data Landscape. Available: https://www.oreilly.com/ideas/what-is-big-data Accesed: Dec 2016.

[3] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken, "The 'big data' revolution in healthcare," *McKinsey Q.*, no. January, p. 22, 2013.

[4] T. D. Gunter and N. P. Terry, "The Emergence of National Electronic Health Record Architectures in the United States and Australia," *J. Med. Internet. Res.*, vol. 7, no. 1, pp. 1-e3, 2005.

[5] W. Hersh, J. A. Jacko, R. Greenes, J. Tan, D. Janies, P. J. Embi, and P. R. O. Payne, "Health-care hit or miss?," *Nature*, vol. 470, no. 7334, pp. 327–9, 2011.

[6] OECD, *Work on health*. 2015. Available at: https://www.oecd.org/health/Health-Brochure.pdf. Accessed: Dec 2016.

[7] OECD, *Health at a Glance 2015*. 2015. Available at: http://www.patients-rights.org/uploadimages/FULL_REPORT.pdf Accessed: Dec 2016.

[8] F. Koechlin, P. Konijn, L. Lorenzoni, and P. Schreyer, "Comparing Hospital and Health Prices and Volumes Internationally," *OECD Heal. Work. Pap. No. 75*, pp. 1–63, 2014.

[9] M. A. Hamburg and F. S. Collins, "The Path to Personalized Medicine - Perspective," *N. Engl. J. Med.*, vol. 363, no. 4, pp. 301–304, 2010.

[10] A. M. Jones, N. Rice, and J. Roberts, "Sick of work or too sick to work? Evidence on self-reported health shocks and early retirement from the BHPS," *Econ. Model.*, vol. 27, no. 4, pp. 866–880, 2010.

[11] R. Astolfi, L. Lorenzoni, and J. Oderkirk, "Informing policy makers about future health spending: a comparative analysis of forecasting methods in OECD countries," *Health Policy*, vol. 107, no. 1, pp. 1–10, 2012.

[12] R. A. Cohen, M. G. Renee, and W. K. Kirzinger. "Financial burden of medical care: early release of estimates from the National Health Interview Survey, January-June 2011." *Natl. Cent. Heal. Stat.*, no. March, 2012.

[13] Y. Lee and H. Chang, "Ubiquitous Health in Korea: Progress, Barriers, and Prospects," *Healthcare Informatics Research*, vol. 18, no. 4. pp. 242–251, 2012.

[14] A. Sen, *On economic inequality*. Oxford University Press, 1974.

[15] J. Frenk, "Health and the economy: A vital relationship - OECD Observer," May, 2004. Available at: http://www.oecdobserver.org/news/archivestory.php/aid/1241/Health_and_the_economy:_A_vital_relationship_.html. Accesed: Dec. 2016.

[16] P. García-Góm ez, S. Jiménez-Martín, and J. M. Labeaga. "Consequences of the Economic Crisis on Health and Health Care Systems", *Health Econ.*, vol. 25, no. 2, pp. 3–5, 2016.

[17] C. Navarro, L. Ayala, and J. M. Labeaga, "Housing deprivation and health status: evidence from Spain," *Empir. Econ.*, vol. 38, no. 3, pp. 555–582, 2009.

[18] Dale, Angela, Malcolm Williams, and Brian Dodgeon. Housing Deprivation and Social Change: A Report Based on the Analysis of Individual Level Census Data for 1971, 1981 and 1991, Drawn from the Longitudinal Study and the Samples of Anonymised Records. HM Stationery Office, 1996.

[19] Jiménez-Martín, Sergi, José M. Labeaga and Cristina Vilaplana-Prieto. "Interactions between Private Health and Long-term Care Insurance and the Effects of the Crisis: Evidence for Spain," *Health Econ.,* vol.25, no. 2, pp. 159-179, 2016.

[20] K. Jee and K. Gang-Hoon. "Potentiality of big data in the medical sector: focus on how to reshape the healthcare system."*Healthc. Inform. Res.*, vol. 19, no. 2, pp. 79-85, 2013.

[21] M. W. Stanton, "Expanding patient-centered care to empower patients and assist providers," 2002. Available at: https://archive.ahrq.gov/research/findings/factsheets/patient-centered/ria-issue5/ria-issue5.html Accesed: Dec. 2016.

[22] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: A patient-centered framework," *Journal of General Internal Medicine*, vol. 28, no. SUPPL.3. 2013.

[23] A. K. Roy. "Impact of Big Data Analytics on Healthcare and Society." *J. of Biom. Biostat*., April, 2016.

[24] J. Archenaa and E. A. Mary Anita, "A survey of big data analytics in healthcare and government," *Procedia Comput. Sci.*, vol. 50, no. 1, pp. 408–413, 2015.

[25] G. D. Magoulas and A. Prentza, *Machine Learning in Medical Applications*, Machine Learning and Its Applications, Springer, vol. 2049, pp. 300–307, 2001.

[26] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande, "Massively Multitask Networks for Drug Discovery," *arXiv*, no. Icml, 2015.

[27] J.E. Bibault, P. Giraud, and A. Burgun. "Big Data and machine learning in radiation oncology: State of the art and future prospects." *Cancer lett.,* 2016.

[28] I. El Naqa, "Perspectives on making big data analytics work for oncology," *Methods*, vol. 111, pp. 32–44, 2016.

[29] Jones, Andrew M., et al. *Applied health economics*. Routledge, 2nd Ed. 2013.

[30] P. Contoyannis and A. M. Jones, "Socio-economic status, health and lifestyle," *J. Health Econ.*, vol. 23, no. 5, pp. 965–995, 2004.

[31] S. Balia and A. M. Jones, "Mortality, lifestyle and socio-economic status," *J. Health Econ.*, vol. 27, no. 1, pp. 1–26, 2008.

[32] A. Muñoz et al. "Proof-of-concept design and development of an EN13606-based electronic health care record service." *J. Am. Med. Inform. Assoc*., vol. 14, no. 1, pp. 118-129, 2007.

[33] Commission of the European Communities-COM. "e-Health - making healthcare better for European citizens: an action plan for a European e-Health Area", 2004. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52004DC0356&from=EN. Accessed Dec 2016.

[34] J. Walker, E. Pan, D. Johnston, J. Adler-Milstein, D. W. Bates, and B. Middleton, "The value of health care information exchange and interoperability," *Health Aff.*, vol. Suppl Web, pp. W5-10-W5-18, 2005.

[35] Somolinos, Roberto, et al. "Service for the pseudonymization of electronic healthcare records based on ISO/EN 13606 for the secondary use of information." *IEEE J. Biomed. Health. Informs*., vol. 19, no. 6, pp. 1937-1944, 2015.

[36] Fienberg, Stephen E. "Sharing statistical data in the biomedical and health sciences: Ethical, institutional, legal, and professional dimensions." *Annu. Rev. Public Health,* vol. 15, no.1, pp. 1-18, 1994.

[37] W. Lowrance, "Learning from experience: privacy and the secondary use of data in health research," *J. Health Serv. Res. Policy*, vol. 8, no. suppl 1, pp. 2–7, 2003.

[38] R.J. Krawiec, D. Housman, M. White et al. "Blockchain: Opportunities for health care". Deloitte. August 2016. Available at: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/public-sector/us-blockchain-opportunities-for-health-care.pdf Accessed: Dic 2016.

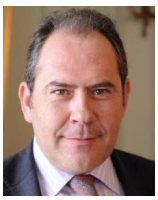
[39] F. P. Machado, "SOR: scalable orthogonal regression for non-redundant feature selection and its healthcare applications." Proceedings of the 2012 SIAM International Conference on Data Mining, 2012.

[40] V. Dhar, "Data Science and Prediction," *Commun. ACM*, vol. 56, no. 12, pp. 64–73, 2012.

[41] C. O'Neil and R. Schutt, *Doing Data Science: Straight Talk from the Frontline*, vol. 1. 2015.

Diego J. Bodas-Sagi

Diego J. Bodas-Sagi is an Associate BBVA Data & Analytics. He holds a PhD from the Complutense University of Madrid in Computer Science. His research interests include Big Data, Data Science, Computational Economics, modelling and e-Health. Contact address is: Universidad Nacional de Educación a Distancia. Departamento de Análisis Económico II. C/ Senda del Rey, 11 28040, Madrid (Spain).



José M. Labeaga

José M. Labeaga is Professor of Economics at the Open University in Madrid and Research Affiliated at UNU-MERIT (Maastricht University) and at Economics for Energy. He is Ms. and PhD. in Economics by Universitat Autónoma de Barcelona. He has served for the Spanish Government as General Director of the Institute for Fiscal Studies during the period 2008-2012. His main research interests rely on applied microeconometric models, microsimulation and ex-ante evaluation of programs as well as ex-post or impact evaluation of public policies in several fields as health, energy or taxation. Contact address is: Universidad Nacional de Educación a Distancia. Departamento de Análisis Económico II. C/ Senda del Rey, 11 28040, Madrid (Spain).