Revista de Bioética y Derecho
Perspectivas Bioéticas
www.bioeticayderecho.ub.edu - ISSN 1886-5887

DOSSIER BIG DATA

Healthcare Data Analysis and Return of Investment for Patient Care: Challenges for a University Hospital

Análisis de datos sanitarios y retorno de la inversión para el cuidado de los pacientes: los retos para un hospital universitario

Anàlisi de dades sanitàries i retorn de la inversió per la cura dels pacients: els reptes per un hospital universitari

XAVIER PASTOR-DURAN *

OBSERVATORI DE BIOÈTICA I DRET DE LA UNIVERSITAT DE BARCELONA

La Revista de Bioética y Derecho se creó en 2004 a iniciativa del Observatorio de Bioética y Derecho (OBD), con el soporte del Máster en Bioética y Derecho de la Universidad de Barcelona: www.bioeticayderecho.ub.edu/master. En 2016 la revista Perspectivas Bioéticas del Programa de Bioética de la Facultad Latinoamericana de Ciencias Sociales (FLACSO) se ha incorporado a la Revista de Bioética y Derecho.

* Xavier Pastor Duran. Jefe de Informática Médica del Hospital Clínic, Profesor Titular de la Facultad de Medicina de la Universidad de Barcelona. España. E-mail: xpastor@ub.edu.

Abstract

Ethical and scientific reasons sustain concerns when applying Big Data analytics in Healthcare. Patient's data must be anonymized and must have enough quality. But the extreme variety of data and their multiple sources combined with many different stakeholders registering data in the Electronic Patient Record without a semantic standardization makes the process harder than in other business. Majority of patient's data are in natural text or are only communicated verbally without a formal registration. To take a real benefit of the Healthcare data we must ensure their quality, high integrity, explicit meaning, context recording and complete anonymization. Three technologies must be of great help: Standardization for Semantic interoperability, Knowledge representation using Ontologies and advanced Natural Language Processing. Ethical conscience regarding confidentiality and ecological impact of ICTs must be empowered and practiced by individuals and organizations.

Keywords: healthcare; medical informatics; electronic patient record; big-data analysis; biomedical research; return of investment; patient care; text mining.

Resumen

La aplicación de la analítica de datos masivos en salud levanta preocupaciones éticas y científicas. Los datos de los pacientes deben ser anónimos y de calidad. Pero la gran variedad de datos, sus múltiples fuentes y los diferentes actores que registran datos en la historia clínica del paciente sin un estándar semántico, dificulta más el proceso que en otras áreas. La mayoría de los datos de los pacientes se dan en texto natural o se comunican sólo verbalmente, sin registro formal. Para beneficiarnos efectivamente de los datos de salud tenemos que asegurar su calidad, integridad, contenido explícito, contexto de registro y su completo anonimato. Existen tres tecnologías que pueden ser de gran ayuda: estandarización para la interoperabilidad semántica, representación del conocimiento usando ontologías y el procesamiento avanzado de lenguaje natural. La conciencia ética sobre la confidencialidad y el impacto ecológico de las tecnologías de información y comunicación (TIC) deben ser practicadas y potenciadas por los individuos y las organizaciones.

Palabras clave: salud; historia clínica del paciente; informática médica; analítica de datos masivos; investigación biomédica; retorno de la inversión; cuidado del paciente; minería de texto.

Resum

L'aplicació de l'analítica de dades massives en salut aixeca preocupacions ètiques i científiques. Les dades dels pacients han de ser anònimes i de qualitat. Però la gran varietat de dades, les seves múltiples fonts i els diferents actors que enregistren dades en la història clínica del pacient sense un estàndard semàntic en dificulta el procés. La major part de les dades dels pacients es donen en text natural o es comuniquen només verbalment, sense cap registre formal. Per beneficiar-nos efectivament de les dades de salut hem d'assegurar la seva qualitat, integritat, contingut explícit, context de registre i anonimat. Existeixen tres tecnologies que poden ser de gran ajuda: estandarditzación per a la interoperabilitat semàntica, representació del coneixement usant ontologies i el processament avançat de llenguatge natural. La consciència ètica sobre la confidencialitat i l'impacte ecològic de les tecnologies d'informació i comunicació (TIC) han de ser practicats i potenciats tant pels individus com per les organitzacions.

Paraules clau: salut; història clínica del pacient; informàtica mèdica; analítica de dades massives; recerca biomédica; retorn de la inversión; cura del pacient; mineria de textos.

1. Introduction

Health is one of the most appreciated values by the human beings. The care of health has been a professional activity well documented since thousands of years ago, but the formalization of healthcare activities is a matter of the last 150 years. The development of a unique document, containing the patient's data and the opinions of physicians and nurses about their health problems and how to solve them was developed in 1907 by Root and Plummer at Mayo Clinic in Rochester, Minnesota and quickly was adopted all over the world.

The Hospital Clinic of Barcelona has more than one hundred years of existence. Since the beginning it's a University Hospital closely related with the Faculty of Medicine of the University of Barcelona. The hospital is fully devoted to offer clinical services for the public healthcare system which contracts and pays for the activity done. Teaching and research are common activities among its healthcare professionals and the clinical care has a maximum level of innovation and quality. International scores classify the hospital among the top ten institutions in Biomedical Research in Europe. The hospital began the digital transformation in 1984 implementing a Hospital Information System (HIS) to register the main activities done over the patient. The deployment of an Electronic Patient Record (EPR) based on clinical documentation was initiated in 1996 and completed around 2001 with the introduction of medical images. In 2003 the EPR system was fully merged with the Enterprise Resource Planning (ERP) of the Institution and the backbone of the main process of care was established. In November 2011 HIMSS certified the EMRAM stage 6 and since then the progressions done has led Hospital Clínic to become a "paperless" hospital. Along this long journey of 33 years a huge amount of patient data have been collected and stored with the protective measures required according the legal regulations in Spain.

Big Data and its proper analytics methodology has appeared in the arena since a couple of years. Marketing departments of the technological companies and software sellers have made also big promises, as big as many data could you offer for the analysis under their tools and consultancy. Healthcare organizations were among the commercial targets. They conform to the three "Vs" that define Big Data: Volume of data, Variety of data and Velocity because there is interest to get out results as soon as possible. Healthcare organizations also have two additional "Vs": Veracity and Value of data because they represent a treasure: the health data of the citizens. In my opinion, the marketing departments of the companies that offer their services to the Healthcare organizations make a mistake assuming a similarity with other business whose data model is more simple, accurate and homogeneous.

Once arrived to that point several questions become interesting to answer in a proper way.

First question: How much big is "Big"? Are we talking about as many patients as the system has? Is it also related with as many data the system can keep for each individual patient? Or, are we thinking in the many (maximum) values of the each independent data? The immediate answer could be: "*the combination of the three: (many)*[3]" and still, once achieved the highest number, another question should appear: "*and some more?*".

Second question: Are Big Data by themselves so relevant? Traditional businesses like Finances, Telecommunications or Sales & Retail have successful stories about the use of Big Data Analytics. Imagine all the economical transactions performed by a bank every day, the number of messages interchanged daily among the users of social network using mobile devices or the goods sold in a day by the stores of a big brand company. Huge volumes of data very simple and homogeneous in their meaning are available in an easy computable format because they are mainly numbers (items, codes of products, prices, time, etc.). The human beings and the care of their health can provide millions of data. Some of them are persistent over the time, but many others can change and the sense of this change can be very informative especially when data are analyzed in their context and according their evolution over the time. Therefore, it's not so easy to validate any hypothesis with a simple analytical approach without context consideration like Big Data Analytics proposes.

Third question: Are all the Healthcare Data computable? The right answer is that it depends on the source of data. If data are collected in a prospective study previously designed and with a specific goal, then it's most likely to record them using electronic forms to collect numeric or logic values, dates or codes and the answer is definitively yes. That's the way we have been doing research since many years of scientific research in the Biomedical field. But if data comes from the EPR the answer is no. More than 80% of data recorded at the EPR for clinical purposes is recorded in a very free manner in a narrative way. To make it computable it's required the use of Natural Language Processing (NLP) technologies which still are very immature in general and moreover in the healthcare domain.

Fourth question: Who ask for "Big Data" in a hospital like Hospital Clinic? There is, of course, an internal demand since the beginning. A classical complain of the stakeholders is "*we are continuously feeding with data that monster but we cannot retrieve information at all. Just only data obtained by pre-formatted queries using traditional filters: patient identity, range of dates, or very simple conditions based on values "lower, equal or greater than…*". On the top of the Organization, the Board of Directors needs data analysis to compute the critical management scores and make comparisons with historical data and projecting the trends to observe if the results of the hospital fit in the strategic planning. The main instrument for that purpose is the Balance Scorecard and the methodology used consists in a proper extraction of administrative, financial and clinical data followed by an aggregation and a descriptive analysis with a comparative study of the same data in previous periods. Clinical and administrative managers of different areas of the hospital need data analysis to check the activity and

performance of their units, the quality of care, the epidemiological trends of diseases and patient conditions, etc. Another area of interest is the Clinical Trials (CT) Unit. The clinical assays to test the effects and benefits of new therapies follow a very strict regulation and methodology. Always are prospective and traditionally data are recorded in specific databases designed and built with that intention. Physicians, nurses, pharmacists and other healthcare employees at the organization can request data analysis for research purposes of their own interests and the better understanding of the health problems that they are dealing daily.

The external demand has different origins. One of the requesters is the Health authorities of the Public Administration. By law, it's compulsory to provide regularly Minimum Datasets of Clinical activity to the Catsalut, the public insurance organization of Catalonia. Those data are used for monitoring the activity contracted at the beginning of the year, to make the monthly payments according the reported activity. Also they are delivered to the Catalan Agency of Quality and Evaluation in Healthcare (AqUAS) who use them to analyze and make comparisons among the different Healthcare providers in a yearly benchmarking report ("Central de Resultats") adjusted by several factors that influence the results like complexity, comorbidity, etc. Also by official regulations, the hospitals and the primary care centers must deliver some clinical documentation to a unique repository for sharing them with the patient ("La meva Salut") and healthcare professionals of other organizations which contribute to the public healthcare system (Shared Medical Record of Catalonia). There is an ongoing project lead by the AqUAS, called PADRIS, whose goal is to offer all this clinical documentation for public biomedical research once patient identity has been fully removed. Other groups of interest in clinical data are academic and research teams involved in multicenter studies. These are in general, biomedical research projects which get funds from a public or a private source after a competitive process. They follow a strict methodology and they have to be previously approved by an ethical committee. Also, they must be fully auditable along all their life-cycle. At last, external for-profit companies like Biomedical Publishers, Pharmaceutical or Technological enterprises are requesting the access and use of Biomedical Data to do business with them.

2. Why do they ask for Big Data Analysis in Healthcare?

The reasons why all the previous persons or organizations ask for "Big Data" in a hospital and their expectations are diverse. Some propose to use massive biomedical data analysis to obtain models or patterns of diseases that could, hopefully, be applied to individual cases for prediction. In other cases, the interest is to compare patient outcomes among different circumstances and try to improve the results. More academic requests are addressed to confirm or reject previous

knowledge or even to "discover" unexpected relationships. Finally, many companies try to gain a competitive advantage and to do business and trade with data and analytics. Personally, I have some concern in such approach because the proposal of a data-driven knowledge discovery is based upon a blind faith in the computer power to make any computation without restrictions and have immediate and applicable results. In summary, a "Discovery Engine" that only needs to be feed with as much data as possible. And the point is: does the Healthcare Big Data Analysis ensure the Return of Investment by itself?

3. Does the quality of Healthcare data ensures the Return of Investment?

Healthcare data analysis has been conducted since many years ago. A classical milestone was the Claude Bernard's book entitled "An Introduction to the Study of Experimental Medicine", published in Paris in 1865. He proposed to apply the scientific method to discover the reason of the diseases to explain the clinical picture and the evolution of the patients. The classical approach included the testing of a Hypothesis (design, collect and analyze), establishing relationships (mapping and relating concepts), inferring new knowledge (trying to know: why?) and applying to do something (to solve a problem: what?). Once finished this cycle it's possible to evaluate and balance between the expected and the real benefit of the investment in research.

Everybody recognizes that Data Quality is a big issue in Big Data analysis in general and particularly in Healthcare. Find below some of the characteristics that must be accomplished by healthcare data to be considered for their analysis in research.

- Completeness: the data set must contain all the elements possibly related with the aspect to be investigated and all the cases to be considered according the inclusion criteria in the data set and the methodology proposed for the analysis.
- Consistency: the same data element recorded in different data sets has the same value.
- Validity: data are valid according the method and metrics used in the recording of each element. This refers to data format, values, range, etc.
- Correctness: Each data value is true. It's also known as accuracy.
- Uniqueness: There is only one data recorded in the same conditions for one patient. There are no replications of the same data / condition.

- Timeliness: each data recorded has a well-defined time-stamp. It's possible to know what data in a data set is the most recently updated.
- Stability: in absence of any change in the conditions a data value of an element in a data set must be the same independently when accessed.
- Relevance: data recorded for one element is the appropriate for the purposes of its ulterior use.
- Contextualization: data is related with all the other data which are required to fully understand it: recording data, stakeholder who records the data, date of last modification, etc.
- Trustworthiness: degree of confidence about the correctness of the recorded data. Usually it depends on the values of the context.

Essentially the EPR is a database. Usually is built according the relational model, and it's oriented toward a systematic data collection of patient data. One of the main characteristics is a very strong patient identification to ensure the integrity of data, that's to say that all data of one patient belongs only to him or her and not to another different from him or her. Demographic data try to define the patient (birth date, gender, name and surnames, official identifiers, etc.) and to register data elements to get in touch with the patient for administrative purposes (home address, phone, mail, insurance details, etc.) The structure and design of the EPR follows, right now, a similar aspect like the traditional paper-based clinical record: a list of the health problems, data about personal past diseases or conditions (like allergies) and relevant health problems of the relatives, detailed description of the actual problem, physical examination, a summary about the likely diagnoses and an action plan which can include medical orders (analytics, imaging, etc.) and therapeutic recommendations (medication, rehabilitation, etc.). Clinical record is a history of the patient health and diseases. In that sense the time-stamps are crucial to make assumptions about the causes of the disease, its evolution and the possible outcome in a short time. In some cases the time stamp is recorded exactly (i.e.: 09:51:24 represents hh:mm:ss). In other circumstances the reference to the time is fuzzier (i.e.: "…approximately two years and a half ago"). Successively more data are incorporated in the EPR: the results of the examinations, the schedule of drug administered, new problems, and a summary of each clinical encounter done between the patient and the healthcare professional.

It's supposed healthcare professionals share the same semantics about all the concepts used to record the different data elements, but as much as the EPR increases the number of data elements, external sources of data, and the EPR is shared by many different stakeholders, the assumption

about "to share the same semantics about the data elements" weakens because the knowledge about the meaning of the concept is more complex and heterogeneous.

The main goal of current EPRs is to help the healthcare professionals to take the right decisions in the right moment offering the relevant and applicable data and recording properly the actions for legal and audit purposes. In that sense must be a good tool in the patient's data management to prevent mistakes and offer data of quality to satisfy the needs of information.

4. How are the health data in the electronic patient record?

The first consideration about patient's data and the EPR is that only 30% of them are recorded in so call "Information space" while the 70% of it are not recorded and used in the "Communications space" which is plenty of data and information shared among healthcare professionals, patients and relatives. This fact represents a breach in the first characteristic of data quality for analysis, the completeness.

When analyzing in depth the "Information space" the results show that nearly 80% of the data recorded are non-structured that is to say they are written in natural language. Natural Language Processing (NLP) technologies are still very immature. Their efficiency is directly related to a very specific environment. A good example is the conversion of a written diagnosis in a diagnosis code of a terminology. This is very successful if the written expressions is in an specific field intended and designed to collect this concept in a final report about an episode of patient's care. The reason is that in such example the semantics and the context are clear and unambiguous.

Only the 20% is structured that means easily "computable". This represents a reduction of the real world because in many circumstances healthcare professionals need to use the natural language to describe a certain condition or situation. But moreover, a deeper analysis demonstrated in the structured data a very heterogeneous typology by themselves what makes harder the data extraction for the analysis. Also, they are the result of the contribution of many human users or machines (more and more with the advancement of Biomedical engineering) without the same semantics about the concept of the data elements.

Clinical images are another chapter. They are native digital. This allows a highly specific processing but only over the data of the image by itself. Context data about the patient is required to make an inference and establish a final conforming diagnosis.

The last point is the personal identification of patient's data in the EPR. As has been told before this is an essential point. Thus, for research purposes and because the legal regulations, a

sophisticated process of "anonymization" must follow the extraction of patient's data from the EPR. This process need to erase any direct or indirect data pointing to a possible recognition of patient identity but at the same time must keep the full integrity of the case. That a real challenge.

5. How to process the Health data?

Text data can be easily computed in terms of counting, sorting, parsing or concatenating. More sophisticated tools are those related with pattern recognition followed by machine learning algorithms for specific purposes, mainly oriented to codify text expressions (like the example of diagnoses explained above) or obtain tags related to full images or parts of it. Text mining to extract data and their associate knowledge is still a field for research.

Structured data allows arithmetic operations (math, basic statistics), logic operations (Sorting, Boolean, Time-lapses, etc.), advanced statistical analysis, clustering for probabilistic advice (Bayesian theorem) and other Artificial Intelligence technologies like neural networks.

6. How to get as much as possible and understand the results of Big Data analysis?

Several recommendations can be of help in the effort to understand better the results of biomedical data analysis. A good database design is crucial to ensure the maximum data quality when the EPR is used in the real scenario. The necessary and sufficient structured data elements to avoid the tiredness and the repulse by the user, leaving free text fields with a clear instruction about their exact meaning and establishing all the double-check internal mechanisms to avoid mistakes and inconsistencies. The incorporation on knowledge in the EPR is a priority. The normalization of the EPR and the clinical process is now possible thanks to the existence of standards like the ISO 13940 and ISO 13606 that enable respectively the standardization of the clinical process and the semantic interoperability among heterogeneous systems. The use of controlled and normalized vocabularies like SNOMED CT (Systematic Nomenclature for Medical Terms), the WHO – ICD (International Classifications of Diseases) among others, is a way to lesser the data entropy among Organizations. Finally, explicit Knowledge Representation in the EPR using ontologies and archetypes to represent data and their knowledge are a big promise to take major advantages of the exploitation of clinical data through different technologies as well as apply to the automatic reasoning to transform the EPR into a true clinical assistant in the healthcare work. This "upgrade" will allow a better

understanding the relationships in both: knowledge-driven and data-driven research in Biomedicine.

7. Is there a real return of the investment in the Big Data analysis in healthcare?

The answer to that question is definitively yes. From a better contribution to Biomedical Knowledge with more sound arguments to ascribe causality, the development of new business rules in healthcare and to build and effective P4 Medicine: Personalized, Predictive, Preventive and Participatory, This requires investing a lot evolving the technology and promoting the professional empowerment to bring and integrate in the healthcare business a new generation of EPRs.

A last aspect I want to point-out is about some ecological considerations. Computing power is not costless. Each Google® query produces 0,2 grams of Carbon dioxide. The use of ICTs for exploiting data must follow legal regulations but also need to be ethically used. We must preserve our lovely planet and there is no ethical justification to use powerful machines that store and process billions of data just to try to find any unexpected and probably non-understandable signal. As the scientist and mathematician John W. Tukey said in 1980: "Finding the question is often more important than finding the answers".

8. Conclusions

- To take a real benefit of the Health care data existing in the EPR/EHR we must ensure quality of Data, high integrity, explicit meaning, context recording, and complete anonymization.
- Three technologies must be of great help: Standardization for Semantic interoperability, Knowledge representation using Ontologies and advanced Natural Language Processing.
- Ethical conscience regarding confidentiality and ecological impact of ICTs must be empowered and practiced at individual level.

All of them are a big challenge for Healthcare Organizations.

Fecha de recepción: 2 de junio de 2017
Fecha de aceptación: 25 de junio de 2017