Generating Big Data Sets from Knowledge-based Decision Support Systems to Pursue Value-based Healthcare

Arturo González-Ferrer1*, Germán Seara1, Joan Cháfer1, Julio Mayol2

1 Unidad de Innovación, Hospital Clínico San Carlos; Instituto de Investigación Sanitaria San Carlos (IdISSC), Madrid (Spain)
2 Dirección Médica, Hospital Universitario Clínico San Carlos (HUCSC), Madrid (Spain)

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

ABSTRACT

Talking about Big Data in healthcare we usually refer to how to use data collected from current electronic medical records, either structured or unstructured, to answer clinically relevant questions. This operation is typically carried out by means of analytics tools (e.g. machine learning) or by extracting relevant data from patient summaries through natural language processing techniques. From other perspective of research in medical informatics, powerful initiatives have emerged to help physicians taking decisions, in both diagnostics and therapeutics, built from the existing medical evidence (i.e. knowledge-based decision support systems). Much of the problems these tools have shown, when used in real clinical settings, are related to their implementation and deployment, more than failing in its support, but, technology is slowly overcoming interoperability and integration issues. Beyond the point-of-care decision support these tools can provide, the data generated when using them, even in controlled trials, could be used to further analyze facts that are traditionally ignored in the current clinical practice. In this paper, we reflect on the technologies available to make the leap and how they could help driving healthcare organizations shifting to a value-based healthcare philosophy.

I. Introduction

HEALTHCARE made a big step towards modernization with the emergence of the Evidence Based Medicine (EBM) concept in the late eighties [1]. EBM is an approach to medical practice that aims to apply the best known scientific evidence into clinical decision-making regarding diagnosis and effective management of specific conditions and diseases. While the EBM concept was generally well received by care professionals, many factors, as their daily work conditions or their high workload, affect putting into practice this approach in the expected way. A recent report from the Institute of Medicine in 2012 revealed that only 10-20% of the decisions clinicians make are evidence-based [2]. This fact reflects the need for medical practitioners, supported by their healthcare organizations, to make a shift in their behavior about the way clinical practice is currently carried out.

The idea of EBM emerged in very different conditions to the current scenario. An explosion of technical possibilities - in nearly thirty years- have come into place to help organizations taking a more modern approach, providing them with support in this regard. Not only epidemiological research can drive EBM, but also new data-oriented approaches. When saying “data-oriented”, we refer to data about the real daily clinical practice: how, when, why and by whom are clinical actions carried out (or not), and what are the health results of those actions. Nonetheless, this might still be hampered by the current design of Electronic Medical Records (EMRs) and by the role and focus that contemporary doctors should adopt. The use of EMRs by physicians could be insufficient, as recognized by studies[2] that expose that, even after post-digitalization of healthcare, they are not utilized to their maximum potential at all.

The fact that the EBM approach was crafted with the goal in mind of pursuing effectiveness in disease management left behind the consideration of organizational and human factors that are crucial in how decisions are truly made. By analyzing data generated by healthcare organizations we could yield information about what are the pitfalls that are hindering evidence-based clinical actions. At the same time, new evidence could be unveiled that is probably not considered in the current production of clinical practice guidelines (CPGs). For example, Touss et al. [3] used data mining techniques to find out how physicians prescribe medications in diverse cases with various clinical conditions, in order to complement existing clinical guidelines where absence of enough evidence occur. Furthermore, specific training actions could be directed to address common failures detected in the management of medical conditions.

Therefore, the problem that healthcare organizations are trying to solve, under the hypothesis that the “Big Data” paradigm will change the way clinical practice is currently carried out, is how can they produce data that help to unveil real clinical behavior and mindlines [4], linked with other organizational data (e.g. costs) and context information that could be behind their actions and decisions. Only making this analysis possible will they be able to change their philosophy to pursue and underpin value, beyond so-called effectiveness. And value here means detecting which actions, later possibly abstracted into policies, could

* Corresponding author.
E-mail address: arturogf@gmail.com

KEYWORDS


DOI: 10.9781/ijimai.2017.03.006
really improve the behavior of the organizations and care professionals for the better care of their main users, the patients.

In this paper, we intend to reflect some existing techniques, beyond current electronic medical records (EMRs) that can help to generate such data sets, considerations to be made, providing some examples of initiatives we are trying to push forward from the Innovation Unit of Hospital Universitario Clínico San Carlos (HUCSC).

II. KNOWLEDGE-BASED DECISION SUPPORT SYSTEMS (KB-DSS)

Gartner\(^\text{TM}\) recently reported [5] a five-stage evolution model for electronic health records (EHRs) where they established a path of characteristics, in terms of eight core capabilities, that EHRs should follow in order to provide the proper support to care professionals. Systems complying with Generation 3 requisites are supposed to be able to bring evidence-based medicine to the point of care, and theoretically coincide with the capabilities of most EHRs currently available. These EHRs have progressed mainly through the core capabilities of ‘system management’, ‘interoperability’ and ‘clinical data models’, even if there is still space for improvement. Generation 4 is expected to improve the core capabilities of ‘decision support’, ‘clinically relevant data analysis’, ‘presentation’ and ‘clinical workflow management’.

Greenes offers his view about the past and future of knowledge-driven Health IT [6], stating that current EHR systems were built for a model that is now old and even inappropriate, supported by proprietary infrastructures and knowledge content. He also mentions the gradual increase in knowledge-based applications during the 2000s, with the creation of computer-interpretable clinical guideline formalisms like GLIF [7] and others [8]. By that time, these systems were having little penetration into real clinical settings, mostly due to the lack of pervasiveness of standards and the use of proprietary tools. Fortunately, this fact is something widely recognized by the current Health IT community and steps have been directed to tackle these problems. From requirements analysis of data standards [9] and development data integration mechanisms [10] for making DSS interoperable, the emergence of new lightweight web services standards like the HL7 Fast Healthcare Interoperability Resources (FHIR) [11], to substantial investments from public bodies that ended up with real deployments and piloting of patient guidance systems. A good example is MobiGuide [10], [12], a project funded by the European Commission under the seventh framework program (FP7). Its goal was to create an intelligent KB-DSS to help physicians and patients taking the most appropriate decisions to manage concrete conditions (atrial fibrillation, gestational diabetes) using a backend server and wearable sensors to monitor patients’ status.

In this context, Figure 1 represents the architecture that represents our view, very aligned to positions already expressed by some research communities [13]. From top to bottom and left to right, physicians and epidemiologists develop CPGs that can be computerized, together with knowledge engineers, into CIG models. With the proper validation mechanisms, using data previously aggregated into clinical data repositories, these models can be trialed, after the corresponding integration into hospital information systems. The execution of CIG models can start generating data sets that are composed of acceptance or denials by physicians of recommendations (e.g. diagnosis, drug prescriptions, therapies, etc.) provided by the knowledge-based DSS developed, and treatments paths followed for different patient profiles. These paths can later on be analyzed by means of process mining techniques [14], [15], unveiling common practices followed while using decision support and comparing the compliance of traditional clinical practice with the one recommended by the evidence-based
DSS. At the same time, normalized clinical data repositories, while ensuring the quality of the data stored, can be used in the traditional view of machine learning and big data research [16], [17]. The results could be complemented by comparing them with the output data sets of the KB-DSS. The output of the research could provide new evidence to be included in new versions of the CPGs (continuous improvement).

III. INNOVATIVE PROJECTS IN HUCSC

The Innovation Unit of Hospital Universitario Clínico San Carlos, being transversal to the healthcare institution, is intended to cover two main aspects of innovation, always pursuing to increase value. On the one hand, it is expected to help hospital professionals to get their research into the market, when there is an opportunity for it. On the other hand, it maintains a technical department to develop innovative products and test their prototypes, driving the Hospital to maximize the possibilities that technological solutions could provide, especially artificial intelligence-based tools.

The ultimate intention is to disseminate the existence of these techniques while facilitating its understanding, create a culture of innovation within the Hospital and, when possible, get external companies to finalize these prototypes, or collaborate in the development, if they are demonstrated relevant and close to a market possibility. The following are several ongoing projects aligned with the goal expressed before and contributing several methods and artifacts to the architecture presented:

A. Computer-interpretable Guideline for Diagnosis and Treatment of Hyponatremia

The Endocrinology Department demanded a process-based solution to help new residents to improve their ability to diagnose and manage the hyponatremia condition (presenting low levels of serum sodium). Hyponatremia is the most frequent electrolyte disorder, however, according to some studies, it has proved to be very difficult to comprehend by physicians in general [18]. To address this project, we developed a CIG model [19], [20] using the PROForma set of tools [21], [22], covering the diagnosis of hyponatremia, classifying it into thirteen different subtypes. During a retrospective validation of the system with the data from 65 patients, we compared the system’s output to the diagnosis consensus of two experts, obtaining a very high agreement (kappa=0.86). The agreement found was also higher than a previous experiment found in the literature [23], carried out by comparing the performance of a resident physician -using the original paper guideline- with the diagnosis of senior physicians. Nonetheless, the most relevant advance of using such a system, beyond its successful diagnosis performance, was the identification and recording of data cases that were contrary to the consensus of international hyponatremia experts, specifically regarding hypoaldosteronism, where concrete markers thresholds were thought to be associated to its diagnosis. The application of our model found several cases where this hypothesis did not apply, showing the lack of real evidence and the need for further research. This is a concrete demonstration of how putting into practice these knowledge-based systems can help detecting where evidence is failing and focusing new research directions ahead.

B. Unsupervised Learning of Discharge Data (Big Data)

The syndrome of inappropriate antidiuretic hormone secretion (SIADH) represents around one-third of all cases of hyponatremia. We carried out a project [23] to identify clusters of hospitalized SIADH patients sharing diagnosed pathologies (comorbidities), where the results coincided and extended previous research identifying individual comorbidities.

Our methods included testing of two different distance measures and hierarchical agglomerative clustering. We used similarity profile analysis for determination of the number of significant clusters and membership of individuals [25] (by means of the SIMPROF method included in the clustsig R package). The method provides also the members of each proposed cluster, where validation of the clusters produced is assessed by iteratively carrying out hundreds of permutations tests. Analyzing the data from around 650 patients, it unveiled 8 clusters, where the most significant ones were five: cancer patients, urinary tract infection patients, patients with renal failure, patients with respiratory problems, and patients with atrial fibrillation and other heart conditions.

We found a main problem; this process is very costly to be carried out in a personal computer, especially when having thousands of columns in the data (variables). We are evaluating the use of the Cloudera big data framework along with Apache Mahout [26] to build a next stage of scalable algorithms that are able to cope with big data sets. If successful, this should be accompanied by the deployment of a private cloud infrastructure [27] able to provide a machine learning as a service (MLaaS) platform, due to the characteristics of patient sensitive data.

C. Hikari: a Case Study of Mental Health (Big Data)

In June 2015, Fujitsu Laboratories of Europe Ltd. and Fujitsu EMEIA in Spain signed a strategic research collaboration agreement with the Foundation for Biomedical Research of Hospital Clínic San Carlos (FIBHSC). Mental health was selected as a key target for the initial project for several reasons: 1) the high levels of disability and morbidity associated to mental illness; 2) the important burden that mental illness imposes on patients, both at individual and social level, and on the use of healthcare resources; and 3) the virtual impossibility to analyze results and its value, despite an apparently perfect design and theoretical structure of mental health services [28].

Hikari, the Japanese word for light, is a part of the Fujitsu’s Zinrai Artificial Intelligence technologies focused on people that includes data analytics and semantic modeling. In this project we have used relevant dissociated clinical data from the Psychiatric Department, obtained during the last ten years, including patient discharge records and the specific registries of psychiatric emergency care, in order to generate a very simple and friendly tool that allows clinicians to have access to information related to the main diagnosis, comorbidities and associated health risks, and also the possibility of analysis at the population level. It has been also useful to track the pathways through the healthcare system followed by patients, and to analyze the impact on the use of resources and costs.

At the present time, the database includes approximately 30,000 emergency care records and 6,500 hospitalizations, however we expect that by the time this paper will be published, it will include data from more than 370,000 outpatients and 38,000 records of day hospital care. This will help us to establish patterns of behavior of the different pathologies and conditions, both in terms of comorbidities, pathways and use of resources.

D. Clinical Data Repository for Secondary Use

Health Observatories, regardless of regional, national or supranational level, rely for their reports on data that will inform on healthcare structure and compliance with programs or pathways. However, data on health outcomes and results are very few or close to none. This is very closely related to the incoherence and fragmented evolution of health care information systems.

In the last decades it has become increasingly evident the demographic and social change in western societies that has brought the concepts of chronicity, fragility and complexity of patients. This makes the continuity of care centered on patients an absolute necessity
if we are to keep our health systems sustainable. Probably one of the main factors involved in this kind of transformation is the access to daily care data that will enable patients, professionals, managers and health policy makers to address these challenges.

If we consider the previous lines, it becomes more and more evident the desirability of having repositories of relevant dissociated clinical data that will allow to evaluate the procedures and results of the real clinical practice, to compare them with recommendations based on evidence and, at the same time, to generate new evidence from the stored data. It is essential to standardize data structure, context (actors, themes, time), continuity of care (such as UNE-EN-13940), generic reference models (such as UNE-EN-ISO 13606, part1), understandable archetypes for clinicians (such as UNE-EN-ISO 13606, part2), terminologies (such as SNOMED-CT) and ontologies for knowledge representation [29]. And obviously, to fulfill the criteria of privacy and data security provided in the legislation, recently renewed in Europe with a new regulation [30].

### IV. Discussion

The application of KB-DSS in healthcare can provide very diverse information. One of the most useful can be the detection of mistakes incurred frequently by professionals when comparing to evidence-based guidelines. Other outputs can be more research-oriented, identifying situations that were thought to be good recommendations but in fact they could be not, according to decisions and reasons explicitly provided by physicians while using the system.

The reader may have noted that we are not stressing from the very beginning that the requirements of the data sets generated by our approach include being of considerable size (the V for “volume”). The reason for this is that we are convinced that the data generated will be eventually growing. However, there is an increasing need to prioritize the V for “value”. We think this value is closely linked to ensuring the V for “veracity” in big data approaches in Healthcare, beyond the rest of Vs (velocity, variety), that are certainly depending on technological capabilities and solutions. This means that we need to ensure mechanisms to guarantee the quality and completeness of the data collected [31], [32] in normalized repositories, if we want to have success in applying these techniques and obtaining valuable healthcare results.

### V. Conclusion

Decision support systems might be able to facilitate the autonomy of citizens when choosing their health options and the ability of professionals to make the most appropriate decision at the right moment. It may also help health policy makers and managers to prioritize the most needed actions in an environment with increasing health needs and resource constraints. But this will be very difficult without the development and maintenance of repositories of dissociated and normalized relevant clinical data from the daily clinical practice, the contributions of the patients themselves and the fusion with open access data of the social environment. Furthermore, this should be quickly accompanied by a proper regulation [33] by the qualified bodies in Europe and the FDA in the US) that make clearer for entrepreneurs the requirements for the development, testing and validation of these new models.

### References


[23] W. Fenske, S. K. G. Maier, A. Blechschmidt, B. Allolio, and S. Störk, ...


Arturo González-Ferrer


Germán Seara


Julio Mayol