Savana: Re-using Electronic Health Records with Artificial Intelligence

Ignacio Hernández Medrano[1]*, Jorge Tello Guijarro[1], Cristóbal Belda[2], Alberto Ureña[1], Ignacio Salcedo[1], Luis Espinosa-Anke[1,3], Horacio Saggion[3]

[1] Savana, Madrid (Spain)
[2] HM Hospitales, Madrid (Spain)
[3] TALN DTIC, Universitat Pompeu Fabra, Barcelona (Spain)

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

Abstract

Health information grows exponentially (doubling every 5 years), thus generating a sort of inflation of science, i.e. the generation of more knowledge than we can leverage. In an unprecedented data-driven shift, today doctors have no longer time to keep updated. This fact explains why only one in every five medical decisions is based strictly on evidence, which inevitably leads to variability. A good solution lies on clinical decision support systems, based on big data analysis. As the processing of large amounts of information gains relevance, automatic approaches become increasingly capable to see and correlate information further and better than the human mind can. In this context, healthcare professionals are increasingly counting on a new set of tools in order to deal with the growing information that becomes available to them on a daily basis. By allowing the grouping of collective knowledge and prioritizing "mindlines" against "guidelines", these support systems are among the most promising applications of big data in health. In this demo paper we introduce Savana, an AI-enabled system based on Natural Language Processing (NLP) and Neural Networks, capable of, for instance, the automatic expansion of medical terminologies, thus enabling the re-use of information expressed in natural language in clinical reports. This automatized and precise digital extraction allows the generation of a real time information engine, which is currently being deployed in healthcare institutions, as well as clinical research and management.

Keywords

Natural Language Processing, Artificial Intelligence, E-Health, Machine Learning, Electronic Health Records.

DOI: 10.9781/ijimai.2017.03.001

I. Introduction

The information that physicians write in Electronic Health Records (EHRs) during their daily practice generates vast amounts of valuable information. Doctors' notes illustrate the real and practical approach in which they address casuistry *at ground level*, where factors associated to their work environment and to uncertainty conditions come into play [1]. However, only a minor portion of all this information is leveraged today, namely that which "sees the light" in the form of scientific literature or other venues where experts share information (articles, reviews, meta-analyses, opinion pieces, conference submissions, and specialized webs in the medical domain) [2]. A fundamental bottleneck preventing large-scale automatic reuse of this information is that it is mostly encoded in *natural language*, i.e. free text written by medical practitioners in EHRs [3]. The traditional approach for knowledge extraction was, until very recently, to pre-structure certain EHR systems so that only certain type of information is allowed in certain fields. However, today there is an increasing line of thought discouraging this practice, as the complexity of clinical reality cannot be modeled simply by means of splitting information in EHRs via drop-down menus.

* Corresponding author.
E-mail address: ignacio.hernandez@salud.madrid.org

As such, it is widely agreed that comprehensive reuse of information generated daily in every point of care of the Health System is of utmost importance. While individual actions do not generate added value due to lack of statistical significance, all the accumulated information provided by specialists in a medical area is an unequivocal and highly valuable reference for any practitioner. Especially considering that part of their actions is supported by the usage of Evidence Based Medicine [4]. Thus, in the daily reality of a medical professional, it is regular practice that physicians ask others, according to their subarea of expertise, confident that their decisions are generally supported by existing scientific knowledge [5].

Moreover, Spain is one of the world's leading countries in terms of impact of EHRs, which results in a very high availability of informattion. Every 10 minutes, tens of thousands of EHRs are written in Spanish medical institutions, which results in a total of billions, if we consider how long have medical practitioners been writing down their notes in electronic form. An additional factor is the need for real-time accurate information, which is explained by the fact that knowledge (and particularly, medical knowledge) grows exponentially. IBM currently estimates that in 2020 there will be 200 times more medical information than what a single individual would be able to absorb in all his or her life [6]. Additionally, we do know that, today, doctors have on average one doubt every two patients they see [7].

Past attempts to apply Artificial Intelligence (AI) to medical decision support systems have traditionally encountered a strong limitation in the complexity of human language [8]. Today, the state of the art of Natural

Language Processing, along with the availability of the computational power needed to perform large scale text understanding, results in a mature field for performing cutting-edge exploitation of text data in domain-specific scenarios. A viable system, however, must simplify its routines as much as possible, and leverage the statistical exploitation of semantic concepts (and not simply words) by combining NLP [9] and data aggregation techniques.

Savana's starting point, in 2013, was motivated by the goal to maximize the huge amount of information contained in EHRs, which up to today had only been used to follow individual patients' progress. Likewise, other associated issues such as defining a correct medical usage for such information, surmounting legal requirements (data protection, for example), or technical considerations, had to be accounted for.

In this context, Savana is born as a platform for clinical decision support, based on real-time dynamic exploitation of all the information contained in EHRs corpora. Savana performs immediate statistical analysis of all patients seen in the platform (which can be queried either searching all the available EHRs, or those belonging only to a single hospital, depending on the institution's interests), and offers results relevant to input variables provided by the user.

II. Methodology

In order to take advantage of the information contained in EHRs, it is necessary to combine computational skills with NLP (a research area which specializes in processing and understanding text written in natural language). EHRs are a paramount example of unstructured information sources: they are incomplete, contain lexical and semantic ambiguities, acronyms, named entities (e.g. commercial names of pharmacological products), and are frequently not properly structured in sections. In addition to these challenges, there are other issues related to the digital exploitation of medical data, among which we find the following:

- There is currently very high sensitivity towards how EHRs are used. While the Organic Law of Protection of Personal Data[1] states that an anonymized clinical record loses its condition of personal data, several stakeholders are of the opinion that despite not possessing them, it should potentially be possible to maliciously locate specific individuals by performing an inverse association from records to patients.
- A system of such characteristics must by definition exist in the cloud, as it requires constant and on-line training.
- Different EHR systems are incompatible, and hence interoperability is seriously hindered, and data sparsity becomes an additional issue to deal with.

For the above reasons, in Savana we decided to address the technical design with the following priorities.

- The source should not matter, as long as there is access to written text. Savana had to detach itself from formatting issues, and be capable to encode any input in text format as its own 'language'.
- It was essential to ensure that individual (single patient) information was irrelevant. In fact, we purposely randomly tamper each record, so that if a third party with malign purposes would breach into this information, it would never know which of it was accurate, and which was not (not even the team in Savana should know).
- However, information should be correct at aggregation time. Statistical approaches would be expected to automatically and reliably clean any false information the very moment in which a doctor, a manager or a researcher asked a question or performed a query.
- Records would not leave the hospital or the institution's data center. They would be processed there in situ, and the cloud would only contain clinical concepts codified according to a predefined custom terminology.

In addition to the above concerns, we faced the challenge posed by current medical terminologies, which are not designed for the reuse of EHRs, and thus constitute a starting point, but not a long term solution. Thus, in Savana we created our own terminology, a process which, for obvious reasons, had to be done automatically. The techniques followed for automatic terminological expansion were designed in-house, and are the content of a recently published paper authored by the authors signing this article [11].

In sum, by combining Big Data with AI approaches, we designed a *robot* that "didn't read well, but excelled at summarization", which surmounted existing shortcomings and allowed us to advance with real use cases, where the goal was to reuse information linked to clinical experience, which had been traditionally limited. The usual approach had always been to implement systems that encoded information on the physician's side (structured systems for inputting information, by means of e.g. dropdown menus). These approaches did not have much success due to, among others, the fact that clinical experience is very complex, and the time available to practitioners to document it, very limited.

In order to tackle these and other technological challenges, we take advantage of current technologies such as, but not limited to:

- Supervised Machine Learning. We have designed and registered algorithms for the different stages of processing, so that, for instance, our system is able to determine that a given paragraph belongs to the 'Background' section, and not 'Diagnosis', due to certain morphologic cues (appearance of adverbs, for instance). Note that, while a traditional approach to such problem could be the development of an expert or rule-based system, in this case the output of the system is based on a statistical model which optimizes a function defined at training time.
- Unsupervised Machine Learning: These techniques are aimed at designing statistical models sensitive to data distribution *without a priori knowledge about the class or label associated to each data point*. We took advantage of neural models for NLP (which imitate the way human brain works) for building a computational model (known in the NLP community as *word embeddings models*) for determining the semantic content of words [12]. For instance, the algorithm learns autonomously, i.e. without predefined semantic relations to be looked up, that Alzheimer's and Parkinson have similar meanings, very different to e.g. Naproxeno and Ibuprofeno, which in addition are themselves semantically similar (see Figure 1

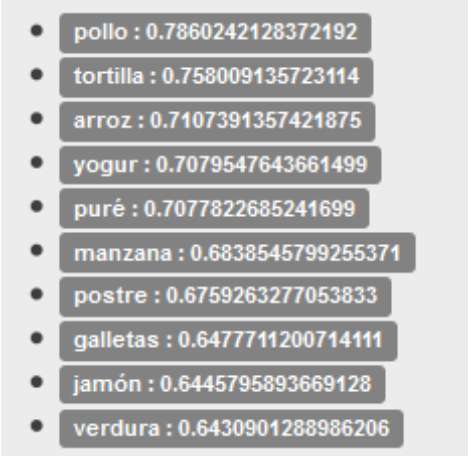


Fig. 1. Example of Savana's unsupervised learning model. It shows the result when asked for words semantically related to *dieta sana*.

1 https://www.boe.es/buscar/act.php?id=BOE-A-1999-23750

Fig. 2. Example of the control panel of Savana Manager.

for the output of the algorithm for a given query). Savana's model, which is being used in several modules of our infrastructure, has been trained with over 500M Spanish words coming from EHRs, and enables the robot to decide, for instance, when 'no' refers to the negation adverb, and when it is an abbreviation of the medical concept 'neuritis óptica', depending on the contextual content. To the best of our knowledge, this is the largest embeddings model trained exclusively with EHRs.

III. Results

In this section, we cover the main functionalities and products Savana offers for healthcare professionals.

A. Functionalities

Savana's technology can be leveraged in different use cases. Today, there are three available applications already implemented and with real-world users, as well as three additional systems in development.

Once the service is deployed in an institution, usage tracking is incorporated, so that additional functionalities can be adapted, which allows Savana to develop improvements and new related services, depending on the actual use of the tool. This makes it possible to adapt the product to the users' requirements (for instance, if its usage is more interesting in certain areas or clinical situations).

In what follows, we describe currently available applications, and their usage.

1) Savana Manager

This application is designed to learn about clinical practice and resource consumption, by computing data in a single institution, and comparing its data and trends with the average of Savana users (Figure 2). The user can also design intuitively custom tables depending on the type of information desired. In addition, a control panel is available where classic management indicators can be found, which again, can be adapted depending on the needs of each individual institution (Figure 3).

This application can be used to measure quantitatively, among others: How much variability there is in an institution's practice; which are the average costs per intervention, which patients are more likely to take part in a clinical trial; the quality of clinical records; when is it likely that clinical tests have been duplicated; what is an institution's position with respect to others of its kind; and in sum, any managerial question solvable with standard metrics.

2) Savana Consulta

This is the world's first application for real-time clinical decision support in Spanish, and is designed to be used at the time of the patient's visit, in front of him/her (Figure 4).

Fig. 3. Home screen of Savana Manager, all the information and configuration options appears in a simple way in only one screen.

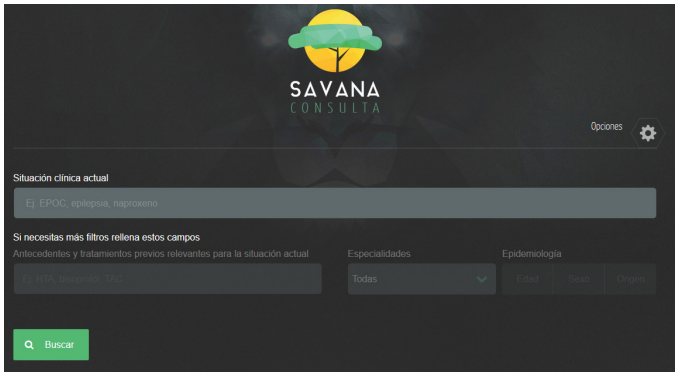


Fig. 4. Home page of Savana Consulta.

This application was developed from its inception considering first general practitioners, as well as emergency physicians (which have high patient load and very limited time), and then, specialists.

It improves the corroboration potential, as in practice using Savana Consulta means to query in real-time all the specialists, and hence incorrect data (statistical anomalies) is factored out from the aggregated response. These common features constitute the content of the answer (which may have not been considered a priori by the practitioner), and can be relevant for decision-making. The vision behind Savana Consulta is that of a helper or second opinion when a medical question is asked (an example can be found in Figure 5).

From a social standpoint, it means that patients are provided with a new type of clinical resource, accessible from any medical institution, and with a very low cost as compared with regular clinical technology. It improves the accuracy in diagnoses and treatments given to patients by any practitioner, thus having a direct impact in their overall health.

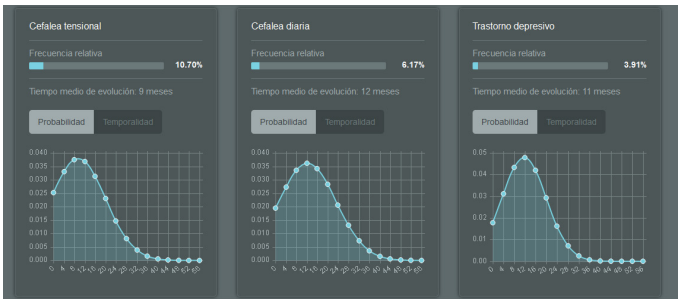


Fig. 5. Example of a question to Savana Consulta about the most frequent evolution of a patient with migraine, and their most probable timespan. This information can be obtained with just one click.

Savana Consulta can be implemented either in a national (interoperable) EHR system, or in more delimited system (e.g. an autonomous community, a set of hospitals or one single medical institution). However, let us highlight that the higher the amount of data, the more significant the results become. Information is shared among all users of the network, without being possible to trace back which hospital provided which bit of information. Moreover, each user can decide whether they are interested in sharing their own information or not. In the latter case, information only becomes available to users in the same institution.

The main contributions of this tool are: Suggestions for each specific clinical case, with non existent precision in current scientific literature; evidence coming from the system itself, with its own resources and population; as well as suggestions for better practices in which there is no Evidence-Based Medicine data available.

3) Savana Research

Our third working product has its usefulness in clinical research, by performing time-sensitive analyses of the behavior of certain patient typologies. It analyzes the e volution of each individual case, and is capable of performing predictions based on existing data.

For a given patient typology, the system can determine how many cases there are (prevalence in an institution), estimate the next cases of a certain set of events in the institution (for instance, a patient with a certain illness comes back for further assistance), as well as defining evolutions according to a set of input tests and treatments, by detecting typical lines of treatment for prototypical patients.

The system analyzes a patient's timeline (illustrated in Figure 6), and hence it is possible to compute the most likely timespan of an occurring event, or if evolutions span a short period, it enables detection of incorrect actions. The main goal of this application is to quickly guide research hypotheses. In addition, Savana Research provides an exponential speed up of a physician's capacity to provide answers to research questions, or guide work hypotheses, without requiring data extraction from EHRs via the traditional, slow methods based on (semi)manual processing.

As an overall conclusion, in Table 1 we provide a listing of interventions carried out in real-world cases thanks to specifically taking advantage of the information encoded by Savana.

B. Current Implementation State

Savana is so far the result of 20,000 hours of computational development. Savana is currently providing service in 24 Spanish hospitals, distributed across three autonomous communities and two private groups. Today, more than 3000 queries have been delivered to the different applications, by a total of 216 users.

Fig. 6. Example output of Savana Research: It shows the most likely admittance of patients with diabetes mellitus (again, this information can easily be obtained with just one click).

TABLE I.
EXAMPLES OF INTERVENTIONS TAKEN THANKS TO THE INFORMATION GENERATED BY SAVANA

Avoid usage of unnecessary elastic packs, after analyzing parts of the operating room.
Discovering that the most frequent point of care after the diagnosis of the Alzheimer's disease is Traumatology.
Ascertaining that new oral anticoagulants are safer than acenocoumarol in atrial fibriliation.
Detecting candidates for undergoing Parkinson surgery, which had been wrongly discarded.
Correct a 2x error in the foresight of beds and salbutamol for bronchiolitis.
Identify patients with refractory essential tremor which were treated with ultrasound.
Call in patients with family aortic myocardiopathy (CIE code unavailable) for a clinical trial.
Knowing how many women who give birth come back to the same hospital in the future.
Listing how many debulking procedures a specific surgeon performed.
Counting how many cases of bronchiolitis were incorrectly derived to pediatric ICU
Anticipating how many spinal surgeries can actually be prevented thanks to the back school
Quantifying the number of cases of suspected apendicits in which computerized tomography + abdominal ultrasound were carried out
Detecting nosocomial infections
Finding out how many breast cancers were treated with lapatinib

IV. CONCLUSIONS

A large scale query, submitted to a vast number of practitioners, and supported by a computational tool, facilitates and speeds up the clinician's task. This is a disruptively new concept, which we call Evidence Generating Medicine, and which constitutes a novel layer of knowledge. On the other hand, in addition to the assistance activity, having all the information contained in EHRs readily available is highly useful for obtaining epidemiological information. This technique is framed within the data mining paradigm, aimed at efficiently exploiting big data. An area destined to revolutionize many areas, including healthcare.

The main avenues where our platform could undergo improvements are: (1) number of referrals to specialists; (2) fitness of diagnostic tests and treatments to recommendations issued in clinical practice guides; (3) number of subsequent visits; (4) reduction of hospitalizations; and (5) improvement of diagnosis.

In the case of Savana Consulta, this application allows patients without access to the best specialists to benefit from their collective knowledge. With the data we have today, the picture at 10 years sight is that we would be leveraging input from hundreds of millions of specialists, always depending on the number of patients under consideration. With Savana Research, we make the research process grow up to 15 times, enabling doctors to focus on interpreting information, rather than extracting it.

The Savana project has an almost universal potential impact, as it can be used in any healthcare point. It is known that technologies related to Internet access and EHR are exponential, and therefore they will become globally available in a few years to the majority of the population.

References

[1] Dawes M and Sampson U. Knowledge management in clinical practice: a systematic review of information seeking behavior in physicians International journal of medical informatics. 2003; 71(1), 9-15.

[2] Bravo R. La gestión del conocimiento en medicina: a la búsqueda de la información perdida. Anales del Sistema Sanitario de Navarra (Vol. 25, No. 3, pp. 255-272).

[3] Gonzalez-Gonzalez AI, Escortell Mayor E, Hernandez Fernandez T, Sanchez Mateos JF, Sanz Cuesta T and Riesgo Fuertes R. Necesidades de información de los médicos de atención primaria: análisis de preguntas y su resolución. Atención Primaria. 2005;35(8): 419-22.

[4] Lopez-Torres Hidalgo J. Hábitos de lectura de revistas científicas en los médicos de Atención Primaria. Atención Primaria. 2011;43(12): 636-37.

[5] Brassey J, Elwyn G, Price C and Kinnersley P. Just in time information for clinicians: a questionnaire evaluation of the ATTRACT project. Bmj. 2001;322: 529–30.

[6] Ferrucci D, Levas A, Bagchi S, Gondek D and Mueller ET. Watson: Beyond Jeopardy! Artificial Intelligence. 2013;93(105): 199–200.

[7] Louro Gonzalez A, Fernandez Obanza E, Fernandez López E, Vazquez Millan P, Villegas González L and Casariego Vales E. Análisis de las dudas de los médicos de atención primaria. Atención Primaria. 41(11), 592-597.

[8] Weiskopf NG, Hripcsak G, Swaminathan S and Weng C. Defining and measuring completeness of electronic health records for secondary use. Journal of Biomedical Informatics 2013;46(5): 830–6.

[9] Geissbuhler A, Haux R and Kulikowski C. Electronic patient records: some answers to the data representation and reuse challenges findings from the section on patient records editors. IMIA Yearbook of Medical Informatics 2007. Inf Med Methods. 2007; 46(1): 47-9.

[10] Espinosa-Anke L, Tello J, Pardo A, Medrano I, Ureña A, Salcedo I, Saggion H. Savana: un entorno integral de extracción de información y expansión de terminologías en el dominio de la Medicina. Procesamiento del Lenguaje Natural. 2016; 57: 23-30.

[11] Mikolov T, Sutskever I, Chen K, Corrado G, and Dean J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

Ignacio Hernández Medrano

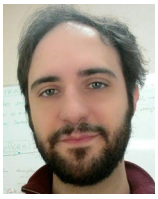
Ignacio Hernández Medrano is a nuerologist in the Ramon y Cajal hospital. He has a long career in healthcare management, where he has coordinated teaching and the research strategy. He holds a Master's degree in Healthcare Management, and a Master's degree in R&D management in health sciences (Spanish National School for Healthcare-ISCIII). He teaches in areas related to innovation and digital health at postgraduate level, e.g. clinical research master's courses, health management or MBAs. Ignacio received a degree from the Singularity University (NASA-Silicon Valley) in 2014 in entrepreneurship with exponential technologies, is TED speaker and the CEO-founder of Savana, a startup focused on the application of AI to Electronic Health Records.

Jorge Tello

Jorge Tello received his Bachelor of Science and Master of Science in Industrial Egineering from the Universidad Pontificia de Comillas (ICAI) in 2006, where he also obtained postgraduate studies in Project Management in 2011. Since 2014 he is Founder and CTO of Savana. His research and work topics include Biomedical data mining, Natural Language Processing and Machine Learning.

Cristóbal Belda

Cristóbal Belda is a medical oncologist and current CEO of HM Hospitales Foundation for Research, an organization involved in the assistance of more than 2 millions of patients every year all over Spain. PhD in Medicine from UAM and former CEO of the Spanish National School of Public Health at NIH "Carlos III". He has developed his career in biomarkers of cancer and, recently, how health economics may help new biomedical advances to be implemented in real life, publishing more than 80 peer-reviewed, JCR- indexed, international papers and international patents for new approaches on biomarker analysis and leading more than 100 clinical trials mainly in lung and brain cancer.



Alberto Ureña

He was born in Madrid, Spain in 1989. He obtained his Msc (2012) in Computer Science from the Complutense University of Madrid. He is currently working at Savana, developing algorithms to extract information from medical records with the goal of improving health system efficiency and future medical breakthroughs. His current interests include NLP and Machine Learning methods, as well as logic programming.

Ignacio Salcedo Ramos

Ignacio Salcedo Ramos (1989, Cuenca, Spain) received his Msc inComputer Science from the Complutense University of Madrid in 2012. He is currently working as R&D engineer in Savana. His research interests include NLP and Machine Learning.

Luis Espinosa-Anke

Luis Espinosa-Anke (Elche, Spain, 1983) received his BA in English Philology from the University of Alicante in 2006. He obtained an MA in English for Speecific Purposes in the same institution, and a second MA in Natural Language Processing and Human Language Technologies in a joint Erasmus Mundus program provided by Universitat Autònoma de Barcelona (Spain) and the University of Wolverhampton (UK). His research interests lie on knowledge-based approaches for semantics and knowledge acquisition and modeling.

Horacio Saggion

Horacio Saggion holds a PhD in Computer Science from Université de Montréal, Canada. He obtained his Bsc in Computer Science from Universidad de Buenos Aires in Argentina, and his MSc in Computer Science from UNICAMP in Brazil. Horacio is an Associate Professor at the Department of Information and Communication Technologies, Universitat Pompeu Fabra (UPF), Barcelona. He is head of the Large Scale Text Understanding Systems Lab and a member of the Natural Language Processing group where he works on automatic text summarization, text simplification, information extraction, sentiment analysis and related topics.His research is empirical combining symbolic, pattern-based approaches and statistical and machine learning techniques. He is currently principal investigator for UPF in several EU and national projects. Horacio has published over 100 works in leading scientific journals, conferences, and books in the field of human language technology.