Towards the Exploitation of Data 4.0 in Health Environments

Carlos Tellería[1], Sergio Ilarri[2], Enrique Bernal-Delgado[1]

[1] Instituto Aragonés de Ciencias de la Salud, Grupo ARiHSP
Calle de San Juan Bosco, 13, 50009 Zaragoza, Spain.
Tel. +34- 976-715895, e-mail: telleria@unizar.es
[2] I3A, Universidad de Zaragoza, Grupo COSMOS

Abstract

Designing appropriate techniques and methodologies to exploit *Real-World Health Data* is essential to improve the performance and sustainability of health systems, the well-being of patients, and the value of the data collected. In this short paper, we briefly describe the main goals that we intend to tackle in this area.

Introduction

Real-World Data (RWD) has become a widely used term in health sciences. Real-World Data is commonly defined as data not collected in conventional randomized controlled trials, but in daily clinical activity, with a view to be reused for research and decision-making purposes. The constant digitalization of health systems (eHealth), has led to huge health RWD availability. This data is larger in volume, as well as in diversity and meaning depth. This opens up new opportunities for the development of systems and applications that can assist users in these purposes. However, many use case scenarios usually impose demanding requirements, which could be covered only if we are able to extract useful and good-quality knowledge from the raw data available.

The use of suitable data management techniques and well-design methodologies may have a major impact, particularly, in health environments. For example, on the application of data mining, process mining and decision mining techniques to improve medical pathways, detect drug incompatibilities and adverse effects, help doctors to diagnose patients and prescribe optimal treatments, or even characterize patients that deviate from common behavior (outlier patients), as they represent higher costs for a healthcare system, to cite just a few illustrative examples. In the end, this means improving existing health systems and the well-being of patients. However, for this to fully come true, ensuring a suitable collection of high-quality data and the application of methodologies oriented towards the management of sensitive and high-value data is essential. Efficiency, effectiveness, safety, and privacy, are key elements that must be considered for the design of the data management processes.

Limitations of Existing Approaches

The exploitation of data in health environments is obviously not new (e.g., see [1]). Traditional approaches usually imply applying classical data mining tasks (such as clustering, multi-level analysis, regressions, etc.) on top of patient data. However, this is subject to two main limitations: a reduced scope of data (usually gathered in randomized controlled trials) and a disconnection from the processes that generate them.

The extensive digitalization of health systems enables nowadays the access to a large amount of pseudonymised health RWD. This paves the way to exciting opportunities that can arise by mining large volumes of heterogeneous data coming from a variety of data sources. This includes data owned by the health system itself as well as external data coming, for example, from global health warnings or recommendations, mobile devices such as smartwatches worn by people, or genetic data obtained from New Generation Sequencers (NGS). Accordingly, ensuring the semantic interoperability of data becomes critical. A major issue will be how to effectively exploit non-structured data (e.g., textual information available in clinical guides, hand-written prescriptions, opinions and informal annotations, etc.). In a way, the Big Data era has yet to be fully deployed in the health area.

Besides, traditional approaches have analyzed the data from a process-agnostic perspective, thus implicitly assuming that the clinical process has no impact on the data. But clinical processes are not a collection of isolated events. On the contrary, it is clear that the same health technique applied over different clinical pathways (sequences of acts, techniques and assistance events, with a given order and timing), or applied on different patient

conditions, may have very different results. By disregarding this dependency, we miss a very important aspect that should be considered and we fail to emphasize the importance of enhancing the existing processes as a mean to also improve the final care outcomes. Therefore, along with traditional data focused on the entities of interest (patients, diseases, etc.), it is of major importance to collect and analyze data about the processes and procedures applied and use both types of data in an integrated way instead of as independent elements.

Ongoing Work

We intend to contribute to the development of techniques and methodologies for the exploitation of "data 4.0" in health environments. For that purpose, our plan is to tackle the aforementioned limitations.

Firstly, to improve the scope of the data, we will apply and develop solutions for the analysis of textual information, which has not been fully exploited so far in health domains. The goal is to obtain smart/elaborated data from raw textual data available, for example by identifying keywords, topics, relevant entities, or different sections in the text. This will enable the retrieval of structured data that could help users (e.g., doctors) to find the information they need or provide new inputs for data mining. The analysis of textual information is also essential to explain medical decisions in real processes, as they are rather captured in a structured way. The automatic anonymization of texts is another important aspect that could be considered.

For this purpose, text mining and natural language processing techniques can be applied. The use of semantic techniques, based on the use, and the eventual building and/or translation of ontologies for the medical domain, as well as semantic reasoners, will support and facilitate this analysis.

Secondly, to avoid limiting implicit assumptions, we will consider the key role of health processes on the outcomes, thus highlighting the fact that the data are linked to (and have a validity regarding) specific processes. This will enable, for example, the analysis of different patient flows that are similar regarding the sequence of steps and their timing but with different outcomes in terms of patient's survival or life quality. By linking both elements we will be able to identify issues that are relevant in a health context; for example, we could detect processes that enable quick patient discharges (which at first sight may seem positive) but that are followed by premature hospital re-admissions (which may suggest a premature discharge).

For this purpose, we are analyzing existing process mining techniques [2, 3] and we will propose extensions to enhance traditional process models with variables that may condition the process flow (clinical conditions for decision making) as well as consider different health metrics or outcomes (e.g., survival, autonomy of the patient, risk of re-hospitalization) that go beyond traditional process mining metrics focused on the performance of the processes (in terms of throughput and/or delay).

Conclusions

The effective, efficient, and safe exploitation of data in health environments requires the development of new data management techniques and methodologies. In this short paper, we have indicated the main limitations of most existing health data analysis approaches and we have outlined the work that we are performing to tackle these issues. Our final goal is to facilitate the obtention of actionable data that can assist decision makers in their daily work.

Acknowledgments

This research is supported by the project TIN2016-78011-C4-3-R (AEI/FEDER, UE) and the Government of Aragon (Group Reference T35_17D, COSMOS group) and co-funded with Feder 2014-2020 "Construyendo Europa desde Aragon".

REFERENCIAS

[1]. JOTHI, Neesha, RASHID, Nur'Aini Abdul and HUSAIN, Wahidah. Data Mining in Healthcare – A Review. Procedia Computer Science [online]. 2015. Vol. 72, p. 306–313. DOI 10.1016/j.procs.2015.12.145. Available from: http://dx.doi.org/10.1016/j.procs.2015.12.145.

[2]. ROJAS, Eric, MUNOZ-GAMA, Jorge, SEPÚLVEDA, Marcos and CAPURRO, Daniel. Process mining in healthcare: A literature review. Journal of Biomedical Informatics [online]. June 2016. Vol. 61, p. 224–236. DOI 10.1016/j.jbi.2016.04.007. Available from: http://dx.doi.org/10.1016/j.jbi.2016.04.007.

[3]. VAN DER AALST, Wil. Process Mining [online]. Springer Berlin Heidelberg, 2016. Available from: http://dx.doi.org/10.1007/978-3-662-49851-4