Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200191

Providing an Integrated Access to EHR Using *Electronic Health Records Aggregators*

Belen PRADOS-SUAREZ [a,1], Carlos MOLINA [b] and Carmen PEÑA-YAÑEZ [c]
[a] *University of Granada, Granada, Spain*
[b] *University of Jaen, Jaen, Spain*
[c] *San Cecilio Hospital, Granada, Spain*

Abstract. The integration of health data systems is an open problem. Most of the active initiatives are based on the use of standards, each one proposed for a concrete use, without integrating other needs or standards allowing on homogenous use. We propose an alternative to get an unified view of health related data, valid for several uses, that can integrate heterogeneous sources. The proposal set the framework to integrate the developments made so far to automatically learn the extraction and conversion of the data. All the sources are integrated under a single access point. We present the main concepts of EHR_{agg} as a middleware between systems that can incorporate several sources giving an unified access following the FAIR principles.

Keywords. Health Data, Interoperability, Data Mining, Natural Language Processing

1. Introduction

Nowadays the need for interoperability between Hospital Information Systems is obvious, as it enables patient mobility, not only geographically, but also between medical services and health care providers.

Many efforts have been made to achieve this *interoperability*, starting with standardization initiatives (e.g. [9,10]). The main issues with these standards are two: first, there is no agreement about which one to choose; and second, it is going to take long before all the systems comply with these standards and become effectively interoperable, due to the huge efforts required to adapt current ad hoc hospital information systems.

This is a problem that also affects data *accessibility*, especially in health research, where data science offers very promising and useful tools ([1,2,3]), but has limited data to work with. This is due to the lack of integrated and standardized datasets for health data. Many initiatives to improve the quality and number of sources for research have been proposed (e.g [4,5,6]). For example, both USA ([7]) and Europe ([5]), have developed catalogues where researchers can add references to datasets and look for others, but

[1]Corresponding Author: Belen Prados-Suarez. University of Granada. e-mail: belenps@ugr.es. Partially supported by PGC2018-096156-B-I00 *Recuperación y Descripción de Imágenes mediante Lenguaje Natural usando técnicas de Aprendizaje Profundo y Computación Flexible* of the Ministerio de Ciencia, Innovación.

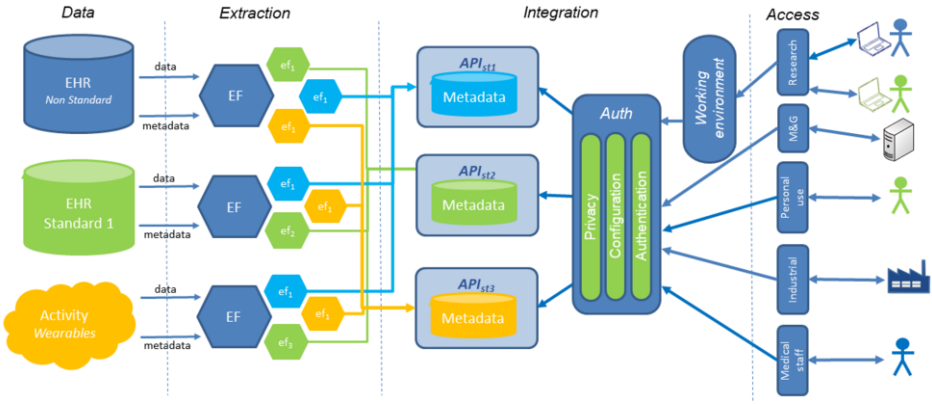

Figure 1. Architecture for EHR_{agg} systems

they need to adapt the data or methods to work with each of the datasets. As Schulz et al. indicate ([8]), although several standards have been developed, they are not generic enough to be valid for different uses.

Moreover, there is a third variable to add to the current landscape. New personal devices are emerging every day, which gather very useful information about people's way of life, but they are developed by private companies under different ad-hoc implementations, and completely disconnected from the rest of the systems and datasets.

In sum, we are faced with the following circumstances: (1) several sources of information, not only from medical institutions, that should be integrated; (2) different access needs with distinct requirements (personal use, medical use, research purposes,...), especially regarding privacy protection; and (3) the need for a homogeneous access point for all this data, with adaptation capabilities (to fit new systems, standards and needs), but without the need to transfer the ownership of the data.

The main idea is to build what we call an *EHR Aggregator* (EHR_{agg}) that integrates the developments made so far (and upcoming ones) to automatically learn how to convert current information systems into standard systems. This system also includes a layer that operates as a homogeneous access point to the information stored in all the sources, without necessarily having to transfer the data, but enabling its processing.

Here we present the concept of EHR Aggregator and the conditions that each element should comply with in section 2. Our purpose here is to present the general structure. In section 3 we put forward our conclusions about how it aims to solve these issues.

2. General Structure

In this section we cover the concept of Electronic Health Record Aggregator. We first present the idea and the general structure, while the following subsections are dedicated to describing each of the elements in the structure.

An *EHR aggregator* (EHR_{agg}), in Figure 1, is a system that acts as middleware between EHR systems and others with health-related data (e.g. activity, sleep analytics, etc.) and the users that want to access the information (which can be people or other systems). It provides a unified view of the underlying Health Data Systems, in the *Data*

Layer, by learning automatically, and in the *Extraction Layer*, by converting the information stored within. This uniform access point allows the user to query and retrieve information from all of the systems without adapting the access to the architecture of each one, which is possible through the *Integration Layer*. This layer acts as a unique access point to all the collected data for external applications. To improve reusability, there are platforms in this *Access Layer* that collect and update the generic functions to facilitate the implementation of external applications. Each of these layers are discussed in the following sections.

2.1. Data Layer (DL)

One of the main elements of the structure are the sources of information. From each one we have to obtain not only the data, but also the metadata. We then need to convert or translate this information in order to integrate it into the aggregator.

Depending on the type of data, we can use different functions to perform this conversion. For example, the functions to convert metadata may be different from those to convert data, and those used for images may be different from those used for medical analysis information. Each of these functions are called *extraction function* (ef).

During this process it is essential to consider the quality of the conversion. In a perfect situation, we will translate the information without any loss of data or metadata. In this case we will consider a value $c_i^j = 1$, indicating a perfect conversion. However, in most of cases we will not have perfect conversions, so we shall indicate this using a value in the $[0,1]$ range. The nearer to 1 the value, the better the conversion is, and the less information is lost.

2.2. Extraction Layer (EL)

The most critical part is to retrieve the data from the source systems and transform it so that the API (application programming interface, see next section) can access all the information. The components in charge of this, called ef (extraction functions), transform the information stored in the source systems, to allow the API to work with it.

If the source is based on a health standard (e.g. [9,10]), the conversion may be direct but not always is so ([8]). Most of the current health data systems are not completely based on standards, so we need to adapt the data. So far, great advances have been achieved in standards compliance. However, system standardization, in most situations, is very complicated or, under the best circumstances, is costly in time and effort. This is why we propose here to get the most of these advances by using methods that automatically learn how to perform the conversion from any system to one of the standards and between the standards themselves.

This automatic learning process allows us to (1) directly integrate systems that already comply with one standard; (2) reuse the learned transformations to apply them in the integration of other non-standard compliance systems, and (3) keep open to future standards, sources or advances in integration or interoperability.

Depending on the type of data, we may define different ef functions. For non-standard systems, the ef need to learn the structure of the data (and metadata) and transform it to be used by the EHR_{agg}. Most EHR systems are based on semi-structured documents and we can find several proposals in the literature using Text Mining to extract

data from EHR (see [11]). Other approaches are based on Natural Language Processing (NLP) ([12,13,14]). These processes can be applied to extract not only data but also metadata. In both cases, the proposed solutions can be tested to obtain information regarding confidence in the conversion, and the best algorithm for each concrete type of data will be chosen. This value, normally in $[0,1]$, provides the user with information regarding the confidence of the data offered by the aggregator. We can define the ef function combining several proposals according to the data to be transformed.

2.3. Integration Layer (IL)

The *IL* has three components. First, the *API* itself, with the methods to access and/or retrieve the data from *EL* and providing access to the functions and the working environment. The *authentication* module, which controls and registers the accesses, configures privacy restrictions, and ensures compliance with regulations. The last one is the *working environment*, where the access platforms can process all the data without having to leave the system, and avoiding the transfer of the data to other layers.

Through the *API*, the user has access to the information stored in each and all of the data sources integrated in the EHR_{agg}. It maximizes accessibility to information, with the only restrictions being those imposed by regulations and privacy protection rules. Using this API to develop their applications, users can run queries over all the data and metadata available in a homogeneous way, regardless of where it comes from or how it is stored. It is possible because the API makes use of the extraction functions to retrieve the data.

The second element is the *Authentication* module. It has the following functions: To register all of the data retrievals, regardless of the API used (*authentication module* in *Auth* block (*AB*)); to control privacy protection restrictions imposed by the source that provides the data or according to how the data will be used (*privacy* in *AB*); to adapt the access point to different regulatory frameworks (*regulation configuration* in *AB*).

Finally, we have the *working module*, that enables direct processing of data without really having the data. The idea is to offer a work space where, for example, researchers can use the API to launch their methods and perform calculations on the data, without retrieving the data or transferring it from the source institution.

2.4. Access Functions (AF)

The access layer is used to run queries in the system. Since we are referring to a large-scale system, we have taken into account that there will be different access requirements, depending on how it will be used. For example, research activities may require anonymized data and statistical methods, while a medical EHR system needs access to detailed information of the EHR of a concrete patient homogeneously, when the data in the original sources is spread out and fragmented. Therefore, as a starting point, we have identified at least five types of data use with distinct needs. These are: personal use of data, medical use, use for research, for management and governance purposes, and industrial use. This is not a closed classification, since new needs may arise in the future.

Here we propose to build an *access platform* per use or purpose where, in a similar way to software as a service (SaaS) platforms, the most frequent operations will be implemented for each given use. This way, as shown in figure 1, they can be reused by *external functions*, applications and systems of users with similar needs.

Finally, *external functions* can only access the system through the *access platforms*, which in turn have access from the *working module* or directly through the *API*. Let us remark that in any case, all of the access points pass through the *authentication module*.

3. Conclusions

With the EHR_{agg} we propose to address the interoperabity and accessibility problem using the same pragmatic approach: instead of trying to have all the systems agree with the same standard, we propose a translation between standards, and of systems to any standard, reducing effort and time. It includes an integration layer that acts as a single access point, offering an unified view of the underlying data sources. This layer also ensures access control, privacy protection and adaptation and compliance with regulations. In addition, it has an access layer especially designed to facilitate the development of external access functions, based on the principle of reusability.

It is an ambitious proposal that aims to centralize current efforts to accelerate understanding between health-related systems. Here we present the general structure of the *EHR aggregator*, establishing the first steps to build a unified framework, but there is still a lot of work to be done. Some of the current and future lines of work focus on the study of access time requirements, feedback on the ef used to improve the system, the classification of access functions and their assessment to be integrated in the access platforms. Nevertheless, we believe this practical and open approach can give great results in the short and medium term.

References

[1] Xu Lina, Simjanoska Monika, Koteska Bojana, et al. What Clinics are Expecting From Data Scientists? A Review on Data Oriented Studies Through Qualitative and Quantitative Approaches *IEEE Access.* 2019;7:641-654.
[2] Health EU . Health EU – Human avatars to prevent and cure diseases https://www.health-eu.eu/ 2019.
[3] European Union . Future Health https://www.futurehealtheurope.eu/ 2016.
[4] National Institutes of Health (USA) . NIH Strategic plan for data science (https://datascience.nih.gov)
[5] Oliveira José Luís, Trifan Alina, Silva Luís A. Bastião. EMIF Catalogue: A collaborative platform for sharing and reusing biomedical data *Int. J. Med. Informatics.* 2019;126:35 - 45.
[6] NCBC (USA) . Informatics for Integrating Biology and the Bedside (https://www.i2b2.org)
[7] Margolis Ronald, Derr Leslie, Dunn Michelle, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative *J. Am. Med. Inform. Assoc.*. 2014;21:957–958.
[8] Schulz Stefan, Stegwee Robert, Chronaki Catherine. *Standards in Healthcare Data*:19–36. Springer 2019.
[9] standard HL7. http://www.HL7.org 2011.
[10] ISO-13606 . ISO 13606: Electronic health record communication 2008.
[11] Simmons Michael, Singhal Ayush, Lu Zhiyong. *Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health*:139–166. Springer 2016.
[12] Dreisbach Caitlin, Koleck Theresa A., Bourne Philip E., Bakken Suzanne. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data *Int. J. Med. Informatics.* 2019;125:37 - 46.
[13] Kaur Rajvir. *A comparative analysis of selected set of natural language processing (NLP) and machine learning (ML) algorithms for clinical coding using clinical classification standards*. PhD thesis 2018.
[14] Chen Liang, Song Liting, Shao Yue, Li Dewei, Ding Keyue. Using natural language processing to extract clinically useful information from Chinese electronic medical records *Int. J. Med. Informatics.* 2019;124:6 - 12.