Measuring the Quality of Data in *Electronic Health Records Aggregators*

Carlos Molina
Computer Science Department
University of Jaen
Jaen, Spain
carlosmo@ujaen.es
ORCID: 0000-0002-7281-3065

Belen Prados-Suarez
Software Engineering Department
University of Granada
Granada, Spain
belenps@ugr.es
ORCID: 0000-0002-3980-102X

***Abstract*—There is an increasing work to integrate health related data for health care services and research purposes. Most of the current proposals adapt the schemas of the data sources to extract automatically the information, but they do not measure the quality of the resulted data. Even more, smart personal devices gather health related data into private non-standard compliant databases. Although these sources could be useful for health care systems and research, to consider the quality of the information they offer is essential. Electronic Health Records Aggregators (EHR_{agg}) are a new concept to integrate this kind of information, that considers the quality of the data. In this paper we present several factors that affect the quality (intrinsic to the data and related to the later use of it) and proposed a fuzzy quality measure to be used inside the EHR_{agg} systems.**

***Index Terms*—EHR, Data Integration, Data Quality, Aggregation**

I. 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. But for our purposes, interoperability standards like FHIR HL7 [7], OpenEHR [16], and European ISO 13606 [8] are especially noteworthy. Many proposals to adapt ad hoc systems to these standards [11] can be found in the literature. 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.

In the side of the data integration, some proposals are based on ”materialized views” or ontologies (e.g [7], [20]). In these proposals, the authors integrate the sources considering a transformation of the schemas to be able to extract the stored data. The processes are automatic so, no validation of the extracted values is carried out. None of them measures the quality of the process, so all the extracted data is considered equally valid. This is problematic because the user that evaluates the data has no information related to the quality. Therefore, he/she may take a decision based on erroneous data. A way to deal with this imprecision is needed to be able to build trustful systems.

If all the systems are based on the same standard, the integration is easier. An example of these systems based on the summarization of UMLS based EHR systems [23] can be found in [10].

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 (i.e. quality of sleep [19], daily physical activity [24] or even emotions [5]), but they are developed by private companies under different ad-hoc implementations. This data may have significant medical value and it is very useful for research (e.g. symptom registering), but it is completely disconnected from the rest of the systems and datasets. Even more the degree of confidence of this data may vary depending on the concrete device or the data collection method (e.g. introduced by the user or automatically taken by sensors).

This is a problem that also affects data *accessibility*, especially in health research, where data science offers very promising and useful tools ([4], [6], [25]), 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 [14], [15]). For example, in USA, the BD2K program ([12]) developed a catalogue where researchers can add references to datasets and look for others. In the proposal of Oliveira et al. [15], users can add data explicitly and query the datasets within the catalogue, but they need to adapt the data or methods to work with each of the datasets, complicating the *reusability* of any new developments. As Schulz et al. indicate in [21], although several standards have been developed, they are not generic enough to be valid for different uses (e.g. as hospital information systems and to support research about related pathologies or interferences between medications). In addition, they often collide with regulations on data ownership, privacy protection and data transfer conditions, making the transfer of data between different institutions very difficult. It lets to the need not only for the integration of the data but also to get an integrated access to the data, independently of the purpose.

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

In sum, we are faced with the following circumstances:

- Several sources of information, not only from medical institutions, that should be integrated.
- Different degrees of confidence depending on the several factors like the source or the data conversion processes.
- Different access needs (personal use, medical use, research purposes, etc.) with distinct requirements (privacy protection,law regulation, etc.), especially regarding the quality.
- But also 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.

Recently we have proposed the *Electronic Health Records Aggregators,* EHR_{agg} ([17]), that set the framework to integrate the developments made so far, and upcoming ones, to automatically learn how to convert current information systems into standard systems. In that paper we pointed the need of a measure of the quality of the data extracted. So using it and the fuzzy logic, the system can work with the imprecision, allowing the user to have the control in all the process. In this paper we define this quality measure, both in the data sources (data quality degree) and the use (access profile weights), and how to compute them, to contemplate the multiple factors that affect it, in order to propose a fuzzy function that has all of them into account.

Next section is dedicated to explain the structure of the Electronic Health Records Aggregators. Section III presents our proposal for measuring the quality of the data extracted from the sources considering multiple factors, and how to incorporated the quality requirements for different access profiles. An example of use is explained in Section IV. The paper ends with the main conclusions an future works.

II. Electronic Health Records Aggregators (EHR_{agg})

In this section we briefly present the structure of the EHR_{agg} to be able to focus the proposals of this paper. In [17] a detailed definition is presented. Figure 1 presents the general structure of the system.

We can identify four layers in the EHR_{agg}:

- *Data layer*: this layer includes all the sources of data that are aggregated in the EHR_{agg}. From these sources we extract the data and the metadata. Later, in Definition 1 we formalize these elements.
- *Extraction layer*: inside this layer we can find the methods used to extract the data, and metadata, form each source. These processes are performed using automatic methods based on Ontologies (e.g. [13]), data mining (e.g [22]) and NLP techniques (e.g. [1], [3], [9]).
- *Integration layer*: we can differentiate three elements:
  - The *API* itself, with the methods to access and/or retrieve the data from the conversion layer 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 *working environment*, where the access platforms can process all the data in the system without the data having to leave the system, and avoiding the transfer of the data to the upper layers.
- *Access layer*: 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.

In the first layer each source of information is formalized as follows:

Definition 1: Let S_i be one health data source. We define on this source the following structure:

$$S_i = (D_i, M_i, EF_i) \tag{1}$$

where:

- D_i is the data stored in the source i.
- M_i is the metadata of the source.
- $EF_i = \{ef_i^1, ef_i^2, .., ef_i^m\}$ is a set of functions that transform the data and metadata of the source. Each of the functions would be defined as

$$ef_i^j : D_i x M_i \rightarrow (d_i^j, m_i^j, c_i^j) \tag{2}$$

where d_i and m_i are the data and metadata respectively, and c_i^j stands for a value in $[0, 1]$ giving the quality of the conversion.

According to this definition, from each source the aggregators extract the data (D_i) and the metadata (M_i). These extraction process is carried out by means of several methods, called *extraction functions* (EF). During the process it is essential to consider the quality of the conversion. In a perfect situation, we will translate the information without any loose of data or metadata. In this case we will consider a value $c_i^j = 1$, indicating the perfect conversion. However, in most of the cases, we will not be in this perfect conversion, so we will indicate it using a value in $[0, 1)$. The nearer to 1 the value, the better the conversion is, and the less information is lost.

Once we have defined each of the information sources, we can face the incorporation of them into the *aggregators*. Each component is built based on the data and metadata extracted from the sources.

Definition 2: An *Aggregator* is a structure EHR_{agg} defined as

$$EHR_{agg} = (S, D_a, M_a) \tag{3}$$

where:

- $S = (S_1, S_2, ...S_n)$: a set of sources from where to extract the health data.

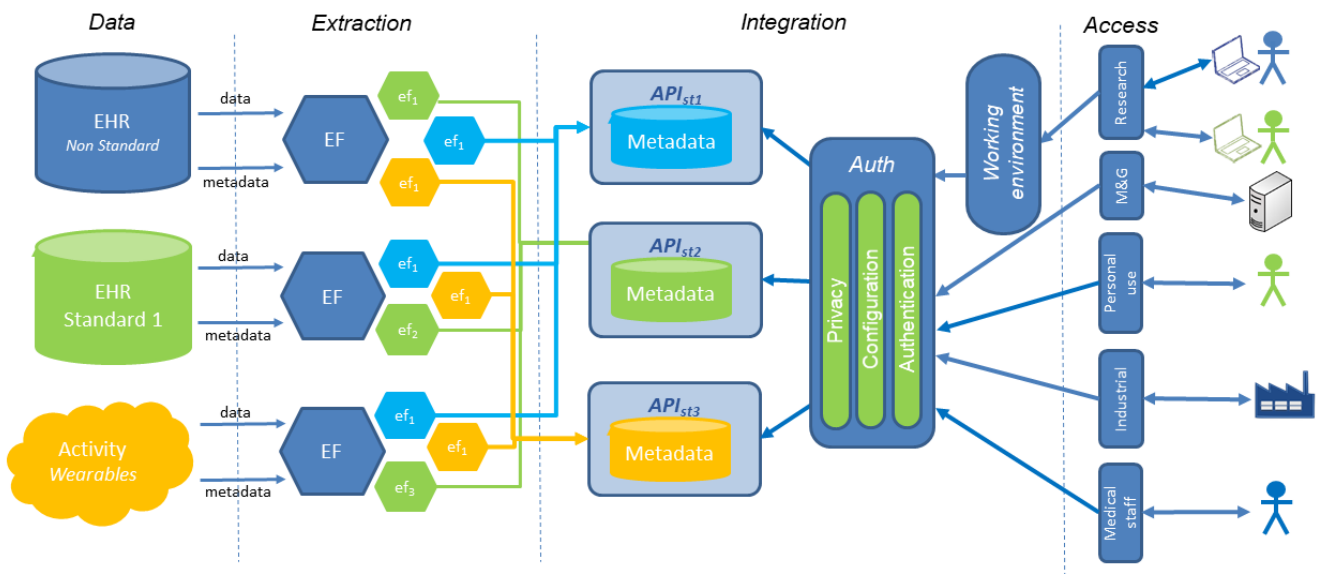

Fig. 1. Architecture for EHR_{agg} systems

- D_a: it represents the data that can be accessed from the EHR_{agg} as a tuple (d_i, c_i) where c_i gives the information regarding the quality of the data d_i. This information is not really stored in the system.

$$D_a = \bigcup_{i=1..n} (d_i, c_i) \text{ such that } ef_i^j(D_i, M_i) = (d_i, m_i, c_i) \quad (4)$$

- M_a: the metadata of the EHR_{agg} as the result of the union of the metadata of the sources. As in the case of the data, each element will be a tuple (m_i, c_i) with the value $c_i \in [0, 1]$ representing the quality of the value d_i. The set is define as

$$M_a = \bigcup_{i=1..n} (m_i, c_i) \text{ such that } ef_i^j(D_i, M_i) = (d_i, m_i, c_i) \quad (5)$$

The c_i are defined in $[0, 1]$ so we can use the fuzzy logic to manipulate the data. By means of fuzzy aggregators we are able to summarize the data, including the quality in the process, so the best values have a higher influence in the results than the worst ones.

In next section we analyse the different factors that affect the quality of the data and how to model them inside a fuzzy function to get the c_i values.

III. Fuzzy Data Quality Function

As we have commented in Definition 2, we need a value $c_i \in [0, 1]$ to control de confidence of the extracted data. This value represents the quality or validity of this piece of information. If we define it in a fuzzy manner, then we can access, manipulate or aggregate the data using fuzzy operators (e.g. [18]).

A. Data quality

To start, we have identified three factors that have influence on c_i: $c_i = (c_s, c_{ef}, c_A)$, where

- Quality factor based on the source (c_s): the type of source is very important. We can have data directly from an hospital information system that can be considered as trusted or from a smart device developed by a private company. In the later case, if the device is medically certified (like *holters monitor* for heart rate) the data can be considered reliable. However in personal device (like a *smartband* with no medical supervision) the data, although useful in some cases, should be considered unreliable.
- Quality factor based on the *Extraction function* (c_{ef}): The extraction is made in an automatic way, so we have to take into account the quality of the process. All the automatic processes have a confidence in the results than can be extracted from the testing process. We can have several measures (e.g. confidence, mean square error, AUC, etc.). We can use these measures to control de quality regarding the data extraction, if the measure m verifies:
  - $m \in [0, 1]$

– If $m = 0$, then the method has the worse quality possible (non of the extractions are valid).
– If $m = 1$ then the method makes a perfect extraction without information lost.

- Quality factor based on the Age (c_A): In medical praxis normally the data has a limited time validity. This value gives a measure of the age of data. The values are in $[0, 1]$, where a value near 1 means the data is really updated, meanwhile a value near 0 means a very old data. A possible function to measure the age of a data d can be formulated as follows:

$$c_A(d) = \frac{1}{1 + (now - d_{date})} \tag{6}$$

where $now - d_{date}$ indicate the number of months between the date when the data was taken (d_{date}) and the current date (now).

B. Quality requirements for different access profiles

Depending on the use of the data, each quality aspect may has more or less influence for the result. If the user wants the data for diagnosis process (e.g. a patient visiting a doctor) then the quality of the source, the conversion and the age may be equally very important. But if, for example, the data is used in a demographic research study, the age may not be important. So, depending on the $access$ module that asks for the data, we may define a different weight vector for the data quality factors $W_p = (w_s, w_{ef,w_A})$ where w_s,w_{ef} and w_A are defined in $[0, 1]$ and stand for the weight for the source, extraction function and age factor respectively. We call W_p the *profile weights vector*.

We want the quality function as a combination of the data quality by itself and the weights regarding the access profile. Considering the above discussion, we define the quality function as follows:

Definition 3: Let be S_i a source of information and W_p the *profile weights vector* of the access. We define the

$$Q(c_i, W_p) = 1 - max(w_s(1 - c_S), w_{ef}(1 - c_{ef}), w_A(1 - c_A)) \tag{7}$$

This function acts as an aggregator of the fuzzy quality measures, giving as result the *quality of the piece of data d_i for the access profile W_p*. As mentioned before, if we need to aggregate data, we can use the $Q(c_i, W_p)$ values as membership values an apply fuzzy aggregators.

IV. Example

In this section we present a simple case to exemplify the presented quality function. We consider three sources with information regarding the heart rate of the patients (Figure 2).

The first one is the EHR Information System of a hospital and the values have been introduced by medical staff during revisions. Examples of values are shown in Table I. The second one, has values measured by a *holter monitor* during 24 hours (Table II). The last one, collects the data from a *smartband* (Table III). In the three tables, the columns entitled c_A present the quality based on the age according the Equation 6.

In this example, we consider two possible accesses. One is made by a doctor at the consultation room with the patient and needs the data for a diagnosis process. In this situation, the quality of the source (c_s) is very important, so the diagnosis is based on the best quality data. In that situation, the extraction process is important too, because if the source is good but we loose the quality in the extraction, then the data is not valid at all. The age of the data has a high influence because the diagnosis has to be based on the actual situation of the patient. So, a possible *profile weights vector* for this access would be $W_p^1 = (w_s = 1, w_{ef} = 0.9, w_A = 1)$.

The other access module corresponds to research activity. Suppose in that situation, the aggregates values for tendency

TABLE I
HEART RATE IN MEDICAL CONSULTATION

Date	**HR**	C_A	$Q(C_i, W_P^1)$	$Q(C_i, W_P^2)$
01-01-2019	75	0.08	0.62	0.87
06-10-2019	60	0.13	0.64	0.87
12-15-2019	55	0.50	0.77	0.87
01-20-2020	65	1.0	0.94	0.87

TABLE II
HEART RATE BY *holter*

Date	**HR**	C_A	$Q(C_i, W_P^1)$	$Q(C_i, W_P^2)$
01-21-2020 09:00	70	1.0	0.8	0.9
01-21-2020 11:00	75	1.0	0.8	0.9
01-21-2020 13:00	90	1.0	0.8	0.9
01-21-2020 15:00	75	1.0	0.8	0.9
01-21-2020 17:00	65	1.0	0.8	0.9
01-21-2020 19:00	60	1.0	0.8	0.9
01-21-2020 21:00	55	1.0	0.8	0.9
01-21-2020 23:00	60	1.0	0.8	0.9
01-22-2020 01:00	54	1.0	0.8	0.9
01-22-2020 03:00	60	1.0	0.8	0.9
01-22-2020 05:00	56	1.0	0.8	0.9
01-22-2020 07:00	58	1.0	0.8	0.9
01-22-2020 09:00	65	1.0	0.8	0.9

TABLE III
HEART RATE BY *smartband*

Date	**HR**	C_A	$Q(C_i, W_P^1)$	$Q(C_i, W_P^2)$
01-21-2020 10:00	80	1.0	0.5	0.64
01-21-2020 12:00	85	1.0	0.5	0.64
01-21-2020 14:00	78	1.0	0.5	0.64
01-21-2020 16:00	60	1.0	0.5	0.64
01-21-2020 18:00	55	1.0	0.5	0.64
01-21-2020 20:00	121	1.0	0.5	0.64
01-21-2020 22:00	65	1.0	0.5	0.64
01-22-2020 00:00	110	1.0	0.5	0.64
01-22-2020 02:00	62	1.0	0.5	0.64
01-22-2020 04:00	58	1.0	0.5	0.64
01-22-2020 06:00	55	1.0	0.5	0.64
01-22-2020 08:00	47	1.0	0.5	0.64
01-22-2020 10:00	65	1.0	0.5	0.64

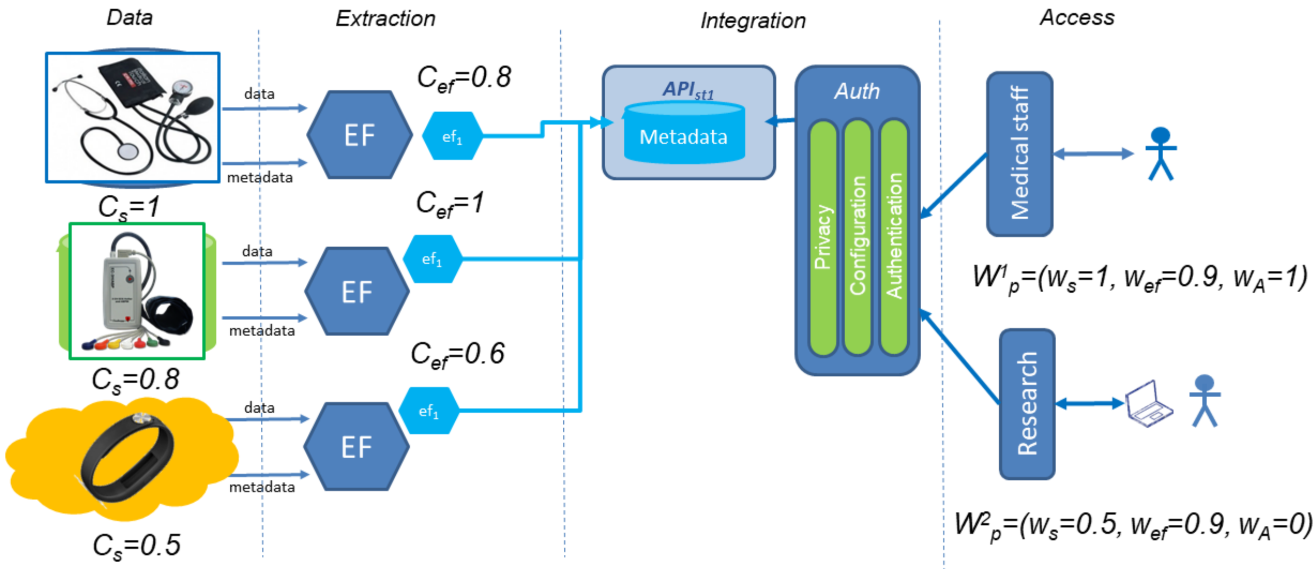

Fig. 2. Integration schema of the example with the quality factor of each source, ef and user profile

studies are preferred to concrete values. Hence, the age of the data is not a decisive factor. Regarding the source, we prefer to have as much data as possible, so the results are representative, therefore we can tolerate lower quality sources. The quality of the *extraction functions* is important so the processed data is as similar as possible to original date stored in the sources. In this profile, the weights are $W_p^2 = (w_s = 0.5, w_{ef} = 0.9, w_A = 0)$.

In the tables I, II and III the columns entitled $Q(c_i, W_P^1)$ and $Q(c_i, W_P^2)$ give the quality values of each measure applying the *profile weights vector* presented.

These quality values can be used to consider the data as fuzzy, and to be able to aggregate the information considering this quality in the process. Table IV collects the results of applying the maximum, minimum and average, following the proposals in [18]. The results are presented using fuzzy sets.

Even with this small example, we can see in the results there are significant differences for both queries. For example, in the diagnosis access considering the maximum we see that the value with higher membership degree is 65, meanwhile in the research query that value is 90 and even with a higher membership degree (0.9 against 0.82 in the first case). In the minimum we have a similar situation (65 in the W_p^1 and 54 in W_p^2) because with the higher restriction in the first profile imposes that only the data from the Hospital Information System is considered really trusted. In the case of W_p^2 we relax the restriction and other values affect the results.

The fuzzy sets can be difficult the understand by non-expert users, so we can defuzzify the results to improve the interpretability like in Table V.

V. Conclusions

There are several proposals to integrate medical data systems, but none of them consider the quality of the extracted data for later analyses. In the case of the integration of other health related data source (e.g. smart devices) this problem is even more important. It is due to the lost of quality in the data conversion can be higher since they are stored in non-standard databases.

Recently we have proposed the EHR_{agg}, a system that integrates all these types of sources considering the quality of the processes. In this paper we analyse the factors than have influence in the quality of the data: three that are intrinsic to the source and other related to the type of access. We have formalized these factors and defined a fuzzy quality function ($Q(c_i, W_p)$) that summarizes the quality into a single fuzzy value. This way, we can apply fuzzy operators to access and aggregate the data. All these developments are being integrated in a prototype.

Under the scheme proposed here, other quality factors could be easily introduced if needed (like quality of the input process: manual or automatic). As well as new profiles (like education or industry).

In this paper we associate the weights to the kind of access, but it would be interesting to let the user to decide the limitations related to quality to impose to the data to be

TABLE IV
AGGREGATION RESULTS (FUZZY SETS)

User profile	MAX	MIN	AVG
W_P^1	{65/0.82, 90/0.8, 121/0.5}	{65/0.82, 54/0.8, 47/0.5}	{65/0.82, 64.9/0.8, 68/0.5, 67.7/0.13}
W_P^2	{90/0.9, 121/0.64}	{54/0.9, 68/0.64}	{64.8/0.9, 64.6/0.82, 68/0.64}

TABLE V
AGGREGATION RESULTS (DEFUZZIFIED)

User profile	MAX	MIN	AVG
W_P^1	95.94	54.21	66.22
W_P^2	98.41	52.10	65.60

used in the queries. As a future work we want to include this feature in the EHR_{agg} allowing the user to stablish a *threshold* to the quality. To choose a concrete value is not easy and the edge problem may occur, so we plan to do it using fuzzy linguistic labels. Another line is trying to integrate other aggregation schemas that take into account the imprecision with no information lost (e.g. [2]).

ACKNOWLEDGEMENTS

This research is partially supported by PGC2018-096156-B-I00 *Recuperación y Descripcion de Imágenes mediante Lenguaje Natural usando técnicas de Aprendizaje Profundo y Computación Flexible* of the Ministerio de Ciencia, Innovación, and the University of Jaen under *Accion 1 (2019-2020)* with reference *EI_TIC01*.

REFERENCES

[1] L. Chen, L. Song, Y. Shao, D. Li, and K. Ding, "Using natural language processing to extract clinically useful information from chinese electronic medical records," *International Journal of Medical Informatics*, vol. 124, pp. 6 – 12, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S138650561830594X

[2] M. Collan and P. Luukka, *Using Innovation Scorecards and Lossless Fuzzy Weighted Averaging in Multiple-criteria Multi-expert Innovation Evaluation*. Cham: Springer International Publishing, 2019, pp. 323–339. [Online]. Available: https://doi.org/10.1007/978-3-319-78075-7_18

[3] C. Dreisbach, T. A. Koleck, P. E. Bourne, and S. Bakken, "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data," *International Journal of Medical Informatics*, vol. 125, pp. 37 – 46, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1386505618313789

[4] European Union, "Future Health," https://www.futurehealtheurope.eu/, 2016.

[5] H. W. Guo, Y. S. Huang, J. C. Chien, and J. S. Shieh, "Short-term analysis of heart rate variability for emotion recognition via a wearable ecg device," in *2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Nov 2015, pp. 262–265.

[6] Health EU, "Health eu – human avatars to prevent and cure diseases," https://www.health-eu.eu/, 2019.

[7] HL7. (2020) Fhir. [Online]. Available: https://www.hl7.org/fhir/

[8] ISO-13606, *ISO 13606: Electronic health record communication*, International Organization for Standardization Std., 2008.

[9] R. Kaur, "A comparative analysis of selected set of natural language processing (nlp) and machine learning (ml) algorithms for clinical coding using clinical classification standards," Ph.D. dissertation, 2018.

[10] G.-W. Kim and D.-H. Lee, "Personalised health document summarisation exploiting unified medical language system and topic-based clustering for mobile healthcare," *Journal of Information Science*, vol. 44, no. 5, pp. 619–643, aug 2017.

[11] R. Lozano-Rubí, A. M. Carrero, P. S. Balazote, and X. Pastor, "Ontocr: A cen/iso-13606 clinical repository based on ontologies," *Journal of Biomedical Informatics*, vol. 60, pp. 224 – 233, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046416000290

[12] R. Margolis, L. Derr, M. Dunn, M. Huerta, J. Larkin, J. Sheehan, M. Guyer, and E. D. Green, "The national institutes of health's big data to knowledge (bd2k) initiative: capitalizing on biomedical big data," *Journal of the American Medical Informatics Association*, vol. 21, no. 6, pp. 957–958, 2014.

[13] C. Martínez-Costa, M. Menárguez-Tortosa, and J. T. Fernández-Breis, "An approach for the semantic interoperability of iso en 13606 and openehr archetypes," *Journal of Biomedical Informatics*, vol. 43, no. 5, pp. 736 – 746, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046410000821

[14] National Institutes of Health (USA). (2018) Nih strategic plan for data science (https://datascience.nih.gov).

[15] J. L. Oliveira, A. Trifan, and L. A. B. Silva, "Emif catalogue: A collaborative platform for sharing and reusing biomedical data," *International Journal of Medical Informatics*, vol. 126, pp. 35 – 45, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S138650561830830X

[16] Open-EHR Open Electronical Health Rercords, *http://www.openehr.org*, Std., 2011. [Online]. Available: www.openehr.org

[17] B. Prados-Suares, C. Molina, and C. Pena-Yanez, "Providing an integrated access to ehr using electronic health records aggregators," in *Proceedings of 30th Medical Informatics Europe Conference*, 2020.

[18] E. Rundensteiner and L. Bic, "Aggregates in posibilistic databases," in *Proceeding of the 15th Conf. in Very Large Databases (VLDB'89)*, Amsterdam (Holland), 1989, pp. 287–295.

[19] A. Sathyanarayana, J. Srivastava, and L. Fernandez-Luque, "The science of sweet dreams: Predicting sleep efficiency from wearable device data," *Computer*, vol. 50, no. 3, pp. 30–38, Mar 2017.

[20] M. Savonnet, E. Leclercq, and P. Naubourg, "eclims: an extensible and dynamic integration framework for biomedical information systems," *IEEE journal of biomedical and health informatics*, vol. 20, no. 6, pp. 1640–1649, 2016.

[21] S. Schulz, R. Stegwee, and C. Chronaki, *Standards in Healthcare Data*. Cham: Springer International Publishing, 2019, pp. 19–36. [Online]. Available: https://doi.org/10.1007/978-3-319-99713-1_3

[22] M. Simmons, A. Singhal, and Z. Lu, *Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health*. Singapore: Springer Singapore, 2016, pp. 139–166. [Online]. Available: https://doi.org/10.1007/978-981-10-1503-8_7

[23] U. M. L. S. (UMLS), *https://www.nlm.nih.gov/research/umls/*, Std., 2004. [Online]. Available: https://www.nlm.nih.gov/research/umls/

[24] J. Vuong, Z. Qiao, and W. Zhang, "Smart shoes with adaptive sampling for outpatient daily health monitoring," in *ASME. Frontiers in Biomedical Devices, 2019 Design of Medical Devices Conference*, no. 41037, 2019. [Online]. Available: http://dx.doi.org/10.1115/DMD2019-3213

[25] L. Xu, M. Simjanoska, B. Koteska, V. Trajkovikj, A. M. Bogdanova, K. D. Starivc, and F. Lehocki, "What clinics are expecting from data scientists? a review on data oriented studies through qualitative and quantitative approaches," *IEEE Access*, vol. 7, pp. 641–654, 2019.