pHealth 2021
B. Blobel and M. Giacomini (Eds.)

doi:10.3233/SHTI210587

Digital Phenotypes for Personalized Medicine

Carlos MOLINA[a,1] and Belén PRADOS-SUAREZ[a]
[a] *Software Engineering Department, University of Granada, Spain*

Abstract. In this paper we propose a new definition of *digital phenotype* to enrich the formulation with information stored in the Electronic Health Records (EHR) plus data obtained using wearables. On this basis, we describe how to use this formalism to represent the health state of a patient in a given moment (retrospective, present, or future) and how can it be applied for personalized medicine to find out the mutations that should be introduced at present to reach a better health status in the future.

Keywords. Personalized medicine, Digital phenotype, Wearables, Artificial Intelligence, Electronic Health Record

1. Introduction

Personalized medicine ([10]) arises from the differences between the results of treatments on distinct patients, depending on their performance at molecular level, which is conditioned by the patient's genotype. In this sense, the presence or not of a concrete gen can make a patient more or less receptive to one treatment or another ([11]). The application of this principle is giving promising results, but so far it is limited to a small number of samples and very concrete diseases, because of the high cost of genetic sequencing ([3]).

In parallel, there is an increasing interest on storing phenotype related data ([15]), but usually considering the classic focus of the observation based-medicine for diagnosis, so only symptoms are stored. As an example, in the literature we can find some recent researches using mobile devices with sensor (*wearables*) ([2]) for medical purposes. Most of them design especial devices to detect concrete pathologies (e.g. heart attack, [12], stress [19][23]). Other approaches look for patterns in patient behavior (e.g. in the care of elder people [21], scoliosis patients [8]). All these proposals are for concrete cases providing ad-hoc solutions, but do not consider high level concepts (like life habits).

Recently, the data records from the interaction of patients with mobile devices has been used in Psychology ([9]) to obtain three patients' characteristics: the behaviour, the conscience and the mood. This approach has proven to be useful for identifying some mental disorders ([20]). In this proposal, only a very limited piece of information is considered and it could be quite enriched by incorporating the great data source that is the patient's Electronic Health Record (EHR). In the previous presented Psychology approach ([9]) the authors use these high level concepts (e.g the patient mood) but do not integrate with EHR data.

On the other hand, the data stored in the EHR has been used in combination with data mining methods, mainly oriented to diagnosis purposes ([22]). Among these methods, there are some interesting proposals considering simple temporal relations

[1] Corresponding Author: Carlos Molina, University of Granada; E-mail: carlosmo@ugr.es. ORCID: Carlos Molina (0000-0002-7281-3065), Belén Prados-Suarez (0000-0002-3980-102X)

([4][7]). These approaches find temporal relations by means of rules (looking for temporal relations between the antecedent and consequent) or by performing predictions with a fixed future window (e.g. one month). To our best knowledge, there are no proposals that take into account the complex temporal relations between elements in the EHR like sequences of facts with non lineal relations, different intervals depending on the patients, parallel evolutions of pathologies, interactions between drugs in multi-pathologies patients, etc.

The integration of both health information sources (wearables and EHR) is currently under study ([6]), mainly from the point of view of data integration than with the focus on patient care solutions; since to achieve this goal the first step is the integration of data from different sources. Solutions as the *Electronic Health Records Aggregators* (EHR_{agg},[17]) and the *European Health Data Space* ([5]) are the perfect starting points to solve this issue.

In this paper we redefine the concept of *digital phenotype* to incorporate information from EHR and wearables, to find out complex relations that opens another way to personalize treatments, parallel to the current genetic via. Our proposal thus is framed into the Smart Data area ([18]) because it aims to deal not only with the volume of the Big Data but also with the correct representation of the information to be able to model and learn more complex relations.

2. Methods

In this section we present our proposal to define the enriched basic concepts related to digital phenotypes.

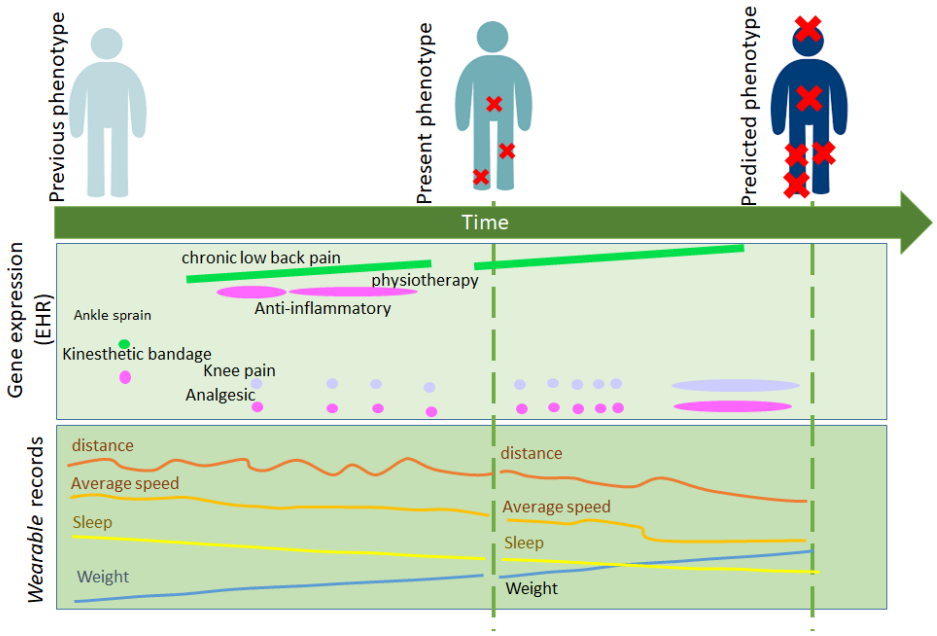

Figure 1. Structure of *Digital Phenotype* for patient *Mr. Smith*.

Definition 1. A *Digital Nucleotide* (*dn*) is each one of the elements stored in the EHR (a diagnosis, a treatment, an analysis result, a surgical intervention, a symptom, etc.) or from wearables devices (activity, sleep habits, mood, diet data, etc.).

In Figure 1, we have some examples of *dn*. Some are from the patient's EHR (e.g. a diagnosis like *ankle sprain*, symptoms like *chronic low back pain* or *knee pain*, and different treatments). Other *dn* are high level concepts defined over the data from wearables like *decreasing activity*, *increasing of weight* or *bad sleep routines*.

Definition 2. A *Digital Fen* (*df*) is a set of $j \in [1, +\infty)$ nucleotides that includes *dn*'s from the EHR, wearables, or both; and a relation $<_t$ that establishes the temporal ordering between them.

$$df\_i = (\{dn_1, \dots, dn_j\}, <_t) \tag{1}$$

To exemplify, as shown in Figure 1, we may define a *Digital Fen df* modelling *lumbar facet syndrome* as the *digital nucleotides* dn_1={*chronic low back pain*}, dn_2={*Anti-inflammatory*}, and dn_3={*physiotherapy*}, with the relation $<_t$ defined as:

$$\begin{aligned} dn_1 &<_t dn_2 \\ dn_1 &<_t dn_3 \end{aligned} \tag{2}$$

meaning that the dn_1 appears before dn_2 and dn_3, but there is no order between these two later nucleotides; so they can be applied before, after or concurrently but always after dn_1 appears.

These *digital fens* are the result of applying specific data mining methods over the data from the EHR databases and the wearable records, so these hidden relations could be discovered.

With these definitions, we are able to present the concept of *Digital Phenotype*.

Definition 3. The *Digital Phenotype* (DP_t^X) of a patient *X* at a specific time *t* is the set of *digital fens* that the patient presents:

$$DP_t^X = \{df_1, \dots, df_n\} \tag{3}$$

As an example, the *DF* at present time of the patient *Mr. Smith* shown in Figure 1 would be:

$DP_{present}^{Smith}$= {*lumbar facet syndrome, decreasing activity, increasing weight, bad sleep routines, knee pain*}

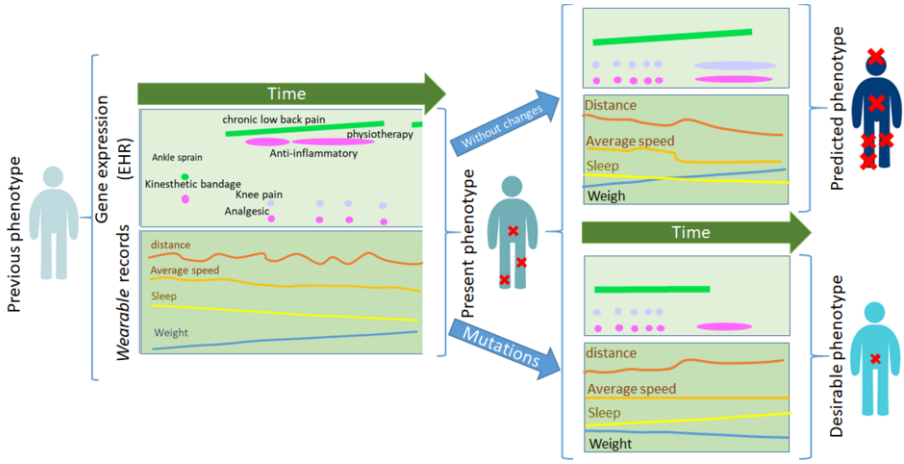

Figure 2. Evolution of the *Digital Phenotype* for *Mr. Smith*.

3. Results

According to this definition above, a patient has different *DPs* evolving from one to another over time. For example, Figure 1, applying prediction methods we can estimate the patient's future digital phenotype, if the patient does not change his habits. Then his *DP* for the future (what we can call *Predicted Digital Phenotype*) would be

$$DP^{Smith}_{predicted} = \{lumbar\ facet\ syndrome,\ obesity,\ depression\} \quad (4)$$

However the idea behind the *DP*'s is not only to be able to model in a better way the patient's health and predict his/her evolution, but to be able to identify what we should be done to redirect the patient's health towards a *Desirable Digital Phenotype* ($DP_{desirable}$)} (Figure 2). To follow the *genome simile* we want to find which *mutations* (changes in the *Digital Phenotype*) must be applied to achieve the $DP_{desirable}$.

Our proposal, to find the mutations is to identify which *df*'s should be modified (added or removed) in the patient's $DP_{present}$. To be able to do it, we need to learn how the *DP* changes depending on the *digital fens* considered:

$$DP_t + \{df_1, \dots, df_m\} \rightarrow DP_{t+1} \quad (5)$$

This task can be achieved by means of machine learning methods capable to work with the presented structure. By learning these *transformation functions* (→), we can predict how a patient's health status can change and identify the *df*'s to induce a better health evolution.

4. Discussion

Once we have the basic concepts, we define the stages of the process that complete the framework, indicating the methods that should be developed at each step:

- *Identification stage.* This step identifies the *dn*'s from the EHR and wearable records. We are currently developing data mining methods capable of working over multi-source data. As shown in [13], it is essential that these methods take into account the reliability and quality of the data.
- *Evolution stage.* This step learns the mechanisms though which the DP evolves; i.e. which mutations results in which health state. Techniques in this step must be explainable, so medical staff and patients can understand the evolution process and results. This is the case of methods such as temporal association rules [1] or gradual dependencies [14] able to deal with imprecision and interpretable to the user (e.g [16]).
- *Prediction stage.* At this phase, the knowledge discovered in previous steps is used to estimate the future health state. Here estimation and prediction methods should be applied to figure out the $DP_{Predicted}$.
- *Mutation stage.* Methods in this final step will be able not only to predict the evolution but also to identify the changes to introduce in patient's treatments or life habits, so the $DP_{Desriable}$ can be achieved. This process will involve data mining methods to learn these changes and their effects in the DP's.

5. Conclusions

We have proposed and enriched concept of *Digital phenotype* as a data model able to represent not only the data from the patient's wearable records but also the data from his/her EHR and their complex temporal relations. This framework, represents a step into the real SmartData modelling to give an holistic view of the patient. With the *digital phenotypes* we can model the patients evolution, learn hidden patterns and relations.

We are working on the definition of methods to automatically identify the relations between *df*, not only frequent, but also those existing in rare diseases (less frequent). They can help in early diagnosis and in the improvement of the expected evolution of patients.

The integration of the *digital phenotypes* with patients' genetic information will be another step to improve the *personalized medicine* as a global concept.

Acknowledgements

This research is 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.

References

[1] Ale JM, Rossi GH. An approach to discovering temporal association rules. In Proceedings of the 2000 ACM Symposium on Applied computing. 2000; Volume 1: 294–300.

[2] Banaee H, Ahmed MU, Loutfi A. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. Sensors 2013; 13 (12): 17472–17500.

[3] Borisov N, Tkachev V, Muchnik I, Buzdin A. Individual drug treatment prediction in oncology based on machine learning using cell culture gene expression data. In Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics (2017), pp. 1–6.

[4] Concaro S, Sacchi L, Cerra C, Fratino P., Bellazzi R. Mining healthcare data with temporal association rules: Improvements and assessment for a practical use. In Conference on Artificial Intelligence in Medicine in Europe (2009), Springer, pp. 16–25.

[5] European Commission. European health data space, https://ec.europa.eu/health/ehealth/dataspace_en. Project, July 2021.

[6] Giordanengo A, Bradway M, Muzny M, Woldaregay A, Hartvigsen G, Arsand E. Systems integrating self-collected health data by patients into ehrs: a state-of-the-art review.

[7] Hanauer DA, Ramakrishnan N, Seyfried LS. Describing the relationship between cat bites and human depression using data from an electronic health record. PLoS One 2013; 8(8): e70585.

[8] Iftikhar O. Designing an Automated System using Wearable Devices for Compliance Monitoring and Activity Detection in Scoliosis Patients. PhD thesis, University of Michigan-Dearborn, 2018.

[9] Insel TR. Digital phenotyping: technology for a new science of behavior. Jama 2017; 318(13): 1215–1216.

[10] Jain KK. Personalized medicine. Current opinion in molecular therapeutics 2002; 4(6): 548–558.

[11] Katsios C, Roukos DH. Individual genomes and personalized medicine: life diversity and complexity. Personalized Medicine 2010; 7(4): 347–350.

[12] Manisha M, Neeraja K, Sindhura V, Ramaya P. Iot on heart attack detection and heart rate monitoring. International Journal of Innovation in Engineering and Technology (IJIET) (2016).

[13] Molina C, Prados-Suarez B. Measuring the quality of data in electronic health records aggregators. In 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (2020), pp. 1–6.

[14] Molina C, Serrano J-M, Sanchez D, Vila MA. Measuring variation strength in gradual dependencies. In EUSFLAT Conf. (1) 2007; 7: 337–344.

[15] Oellrich A, Collier N, Groza T, Rebholz-Schumann D, Shah N, et. al. The digital revolution in phenotyping. Briefings in bioinformatics 2016; 17(5): 819–830.

[16] Prados de Reyes M, Molina C, Prados-Suarez B, et al. Interpretable associations over datacubes: application to hospital managerial decision making. Stud Health Technol Inform. 2014;205:131-5.

[17] Prados-Suarez B, Molina C, Peña Yañez C. Electronic health records aggregators (EHRagg). Methods of Information in Medicine 2020; 59 (2,3): 96–103.

[18] Sheth A. Smart data-how you and i will exploit big data for personalized digital health and many other activities. In Proc. IEEE Int. Conf. Big Data (2014).

[19] Simões L, Gonçalves J, Silva J. Mobile application for stress assessment. In 2017 12th Iberian Conference on Information Systems and Technologies (CISTI) (2017), IEEE, pp. 1–7.

[20] Torous J, Staples P, Barnett I, Sandoval LR, Keshavan M, Onnela J-P. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. NPJ digital medicine 2018; 1(1): 1–9.

[21] Tsai C.-H, Chu C-H, Liu S-W, Hsieh S-Y, Tseng VS. Mining life patterns from wearable sensors data for elderly anomaly detection. In 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI) (2017), IEEE, pp. 66–71.

[22] Yadav P, Steinbach M, Kumar V, Simon G. Mining electronic health records (EHRs) a survey. ACM Computing Surveys (CSUR) 2018; 50(6): 1–40.

[23] Zubair M, Yoon C, Kim H, Kim J, Kim J. Smart wearable band for stress detection. In: 5th International Conference on IT Convergence and Security (ICITCS) (2015), IEEE, pp. 1–4.