# Patient Standardization Identification as a Healthcare Issue

Mario Macedo[1] and Pedro Isaías[2]

[1] IPT-Escola Superior Tecnologia Abrantes, Rua 17 de Agosto de 1808, 2200-370 Abrantes, Portugal
mario.macedo@mail.telepac.pt
[2] Universidade Aberta, Rua da Escola Politécnica, 141-147, 1269-001 Lisbon, Portugal
pisaias@univ-ab.pt

**Abstract.** Healthcare organizations use information systems with several different types of data and user interfaces. The lack of standardization means loss of efficiency and effectiveness. It limits the expected quality of Healthcare services. Some difficulties for this standardization are known. However there are models that can respond to the complexity of this area of science and evolve with the development of knowledge. A problem which is common to several organizations is the lack of automatic identification of patients. Another one is how to solve the problem of having information duplicated in different databases.

The purpose of this paper is to show the importance of the standardization of clinical data and the development of unique models of identification that will enable setting unique access keys and the interconnection between all the clinical data.

The empowerment of systems that support clinical decision and the use of workflows for treatment plans that involve more than an organization of Healthcare will only be possible if they use standard models, open technologies and unique patient identification.

**Keywords:** HER, Medical Guidelines, Healthcare Plan Workflow.

## 1 Background

The prime objective of having a unique ID for identification of a patient and access to his/her clinical data is to avoid clinical records becoming sidelined and to ensure the correct corroboration among each individual's data. Each individual's historical records, along with those of his/her forebears, constitute essential background information for the evaluation of his/her state of health and the likelihood of future pathologies.

The storage, integration and standardization of clinical data also make it possible to provide personalized healthcare.

The supply of personalized clinical data makes it possible to make more accurate diagnoses and prescribe the treatment most suitable for each pathology and each individual.

In order to assist with diagnosis it is possible to develop systems to assist in clinical decisions. There exist three levels of system to assist in clinical decisions.

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part II, HCII 2009, LNCS 5618, pp. 566–575, 2009.

According to HL7 CDS Project Update, (2008) [1], those levels are information, rules and computer-interpretable guidelines.

At the information level only information is provided. At the rules level alarms, data interchange and data validation become available.

In addition, according to Shabo (2005) [2], having the possibility to include genetic data in the electronic record of clinical data for each patient increases the amount of knowledge on which to base health care provision decisions.

According to that author Shabo (2007) [3], there are three essential hurdles in the way of complete recording of all of a patient's clinical data:

- Because of data protection legislation each hospital generates its own policies for data security and filing methods compatible with its preserving the privacy and confidentiality of clinical information. For this reason it is impossible for a patient who attends different hospitals on different days to have all his/her data integrated.
- Another hurdle is a time-based one. It is a simple fact that an individual's average life span is far greater than the maximum time that data can be/is stored, So, if an individual lives for 70 years it is very unlikely that the hospital will be able to keep records that long.
- Another hurdle derives from the fact that, even if clinical terminology were all standardized between various Health Care Units, it would be extremely difficult to maintain semantic compatibility over a period of several years because the terminology itself is also in permanent evolution.

To those three hurdles can be added the question of genetic data, which has evolved in structure and complexity at one and the same time as science itself has evolved.

Those hurdles aside, it is self-evident that genetic information needs to be included in the electronic record of clinical data.

The SNOMED standard already includes genetic terminology, thus opening the door to the creation of genetic data archetypes.

For the HL7 standard a working group was formed to develop a limited model for the storage of chromosome data. That data is referenced by a set of metadata stored in a RIM platform (Reference Information Model). This model is still used in only a limited fashion to communicate data between hospitals and the pharmaceutical industry.

OpenEHR works in this area but no defined genotype model as yet exists.

Meanwhile another question has to be raised. If clinical data needs to be kept for a long time, and if it needs to retain all data concerning genetics, pathologies and treatment given to every individual what will the storage infrastructure need to be like? There will have to exist either distributed data bases or clinical data banks. The FEHR (Federation Electronic Health Record) concept.

The domain of the data is another highly important aspect. What to be the nature and type of data to constitute the identification of an individual and what data quality frameworks will need to be put in place.

Some types of data can identify a specific individual unequivocally, whereas other data are secondary or of less importance. Characterization and definition of models is rather complex.

Access to clinical data is limited to specific users. There will be a need for various levels of access interconnected with temporal windows. Access to personalized data will be available only for the purpose of providing care to the patient.

Another consideration raised is whether only public entities shall have access to a patient's data or whether, on the contrary, private entities will also have access to these data.

Identification of those users permitted access to clinical data needs to be protected with secure authentication, and in no way to permit one user's identification to be used by any other person. In addition, access to the system by non-identified users should not be possible.

Legal protection relating to the use and communication of clinical data needs to prevent unauthorized use and transfer of data to third parties. As an example let us examine the case of prescriptions to each individual. From the medication prescribed it will be possible to deduce what each individual's pathologies and their frequency of occurrence are. Is this information, which is available to pharmacies (chemist's), actually protected?

The storage systems for each individual citizen's identification are also extremely significant in relation to the architecture of the entire system. Clearly, each patient's identification will need to be stored in a central data base available to all players in the health system. However, if there are public entities, private entities and entitled entities what will need to be the nature of the central file identifying all users?

There are writers who argue that clinical data should be de-identified. What this means is that after being used in a medical episode they should be removed from the individual identification of each person.

But how and where would this function be carried out? In the event of it being necessary again to access the patient's historical data what should the data personalization process be like?

## 2 Security

The security has some dimensions like privacy and confidentiality, identity verification, users identification and authentication. These concepts can have different meanings.

**Privacy**

According to Kent (2002) [4], *privacy is the right of an individual to decide for himself or herself when and on what terms his or her attributes should be revealed.*

According to Department of Health (2007) [5], *Patient information is generally held under legal and ethical obligations of confidentiality. Information provided in confidence should not be used or disclosed in a form that might identify a patient without his or her consent. There are a number of important exceptions to this rule but it applies in most circumstances.*

**Identity**

According to Kent (2002) [4], The identity of X according to Y is a set of statements believed by Y to be true about X.

According to The Department of Health (2007) [5], *Patient Identifiable Information includes name, address, full post code, date of birth, pictures, photographs, video, images, NHS number and anything else that may be used to identify a patient directly or indirectly.*

**Identification**
According to Kent (2002) [4], is the process of determining to what identity a particular individual corresponds.

According to The United Kingdom Parliament (n.d.) [6], citing The *Data Protection Act 1998, personal data is defined as:*

Data which relate to a living individual who can be identified from those data, […]

**Authentication**
According to Kent (2002) [4], is the process of confirming an assert identity.

The Patient identification and data archives should be compliant with all these issues. Our proposed model for a Federation of Electronic Health Record should include the necessary features to overcome these issues.

## 3 The Patient Identification Domain

For any individual there exist several possible IDs. For example, NHS, Medicare, Health Care number, Identity Card, passport number, driving licence number, Inland Revenue, IRS number or even just a number generated for the specific purpose.

There are, however, some considerations to be taken into account.

The first question is that not all of the above IDs are available at the time of the individual's birth.

For this reason, only a code generated for each individual will act continuously and without fail throughout an individual's life. The genetic code is, a priori, an element unique to, and permanently present in, every individual.

The principal advantage of using the DNA code as a key to access each individual's clinical data is that it is unique and works across all existing systems. In addition, analysis of gene mutations can help in the identification of pathologies or the likelihood of pathologies occurring.

For these reasons, the use of genetic data to assist in clinical decisions is of the utmost importance.

The HL7 organization has introduced a standard called Clinical Genomes Level 7, (Clinical Genomics, 2009) [5]. The model put forward by the HL7 includes a layer of associations between genotype and phenotype entitled Clinical Genomics Standard.

The models for recording genetic data are somewhat more complicated than the archetypes for recording other clinical data. The main reasons for this are:

- The quantity of data
- The complexity of representing the DNA molecule and its variants
- The semantic transcription of the genotype/phenotype association.

Accessibility to genetic data even makes it possible to develop genomic-oriented applications to assist in clinical decisions. These applications can possess parsers for identifying sequences of significant genes for any study taking place.

The use of DNA data in the electronic records of clinical data represents an unprecedented advance in medicine and in the provision of medical care. It will be possible not only to identify patients unequivocally and access their entire history but also to take preventative action. It is even possible to observe genetic changes through systems based on artificial intelligence.

According to Marko (2005) [8], the challenges of creating an HER that integrates an organization's clinical record system with a biorepository and a genomic information system involve complex organizational, social, political, and ethical issues that must be resolved.

In fact, if, on the one hand, it is going to be possible to analyze the likelihood of a patient succumbing to a particular illness, on the other hand, that patient's privacy must be guaranteed lest society discriminate against certain individuals.

According to Nakaya(2007) [9], The elemental techniques of the data collection platform are the information model, the ontology and the data format.

According to this author, the Genomic Sequence Variation Markup Language (GSVML) is a Markup language and is the data exchanging format of genomic sequence variation data to use it mainly in human health. This norm should be standard in the near future.

## 4 Proposed Technologies

The proposal model uses some technologies that should be compliant with standards and industry best practices.

**Communications**

The IETF (Internet Engineering Task Force) (n.d.) [10] develops norms and standards for communication on the Internet. The standardization documents are designated as RFC, Request for Comments. RFC 2821 defines the SMTP (Simple Mail Transfer Protocol) and RFC 2616 the http (Hypertext Transfer Protocol)

RFC 3335 specifies how EDI (Electronic Data Interchange) messages can be transmitted securely via a peer to peer link. This standard, in addition, ensures communication of messages according to the protocols *for Electronic Data Interchange, (EDI – either the American Standards Committee X12 or UN/EDIFACT, Electronic Data Interchange for Administration, Commerce and Transport), XML or other data used for business to business data interchange,* (Request for Comments: 3335, Network Working Group,) (2002) [11].

This standard specifies several messages such as the format of the message delivery receipt with or without digital signature, the non-repudiation of receipt message, the format of the message envelope (MIME), with or without signature, and the body of the EDI message with or without cryptography.

Using this technology it is possible to define a peer-to-peer archetype communication relationship.

These archetypes can contain the clinical data necessary for the HER.

**Archetypes**

The word "archetype" comes from Greek and means "original pattern".

According to Soley (2004) [12], an archetype is a primordial thing or circumstance that recurs consistently and is thought to be a universal concept or situation.

The concept of "archetype" defined in this way makes it possible to define business objects suitable for any and every activity. These business objects can be any kind of data model stereotype.

Object-oriented (OO) information technology reflects the archetype application domain.

In this way, an archetype model can be constructed and this model applied to cases with real data.

The archetypes define for each type of data the various possible dimensions and methods available. Archetypes can even contain rules for coherence and inter-association. Archetypes also have the property of pleomorphism, which enables different instances of each archetype to be created.

Archetype models are specified in UML (Unified Modeling Language) (2009) [13], language, for which several modeling tools exist. Some of these tools even enable UML models to be transposed into physical models.

Even archetype patterns can be defined. An archetype pattern contains optional elements that can be implemented or not implemented. The name "pattern configuration" is attributed to each instance of an archetype pattern. Both well-formed and ill-formed configurations can exist.

In order to avoid ill-formed configurations there has been created a set of rules to which the name "Pattern Configuration Rules" has been applied .

According to Soley (2004) [12] a Pattern Configuration Rule is a formal language for expressing the rules for well-formed pattern configuration.

Some party archetype patterns are standardized. For instance, ISO 3166 contains country codes and country names and ISO 5218 contains a representation of the human sexes.

In the health area there exist two different approaches to information system architectures, HL7 (Health Level 7) (2009) [14] and OpenEHR (OpenEHR) (2009) [15].

Both approaches present both a model designed for object programming and a reference model. OpenEHR also puts forward a language called "Archetype Definition Language" for defining archetype models.

## 5 Proposed Model

The correct registration, treatment and integration of clinical data are of utmost importance for the provision of health care.

Integration of clinical data makes it possible to watch out for public health indicators and carry out epidemiological research and scientific investigation.

It is of the greatest importance to develop systems that enable patients to be treated collaboratively and that simultaneously provide data for other levels of tactical, strategic and scientific management.

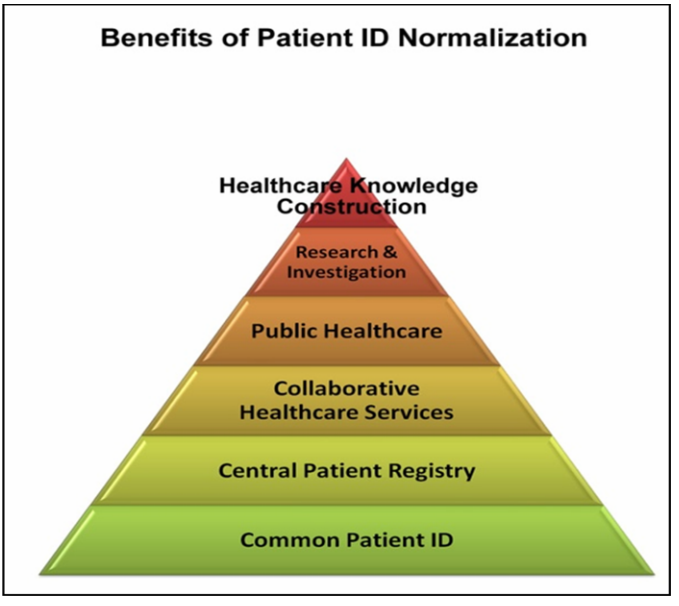


**Fig. 1.** Benefits of Patient ID Normalization, (Authors Proposal)

The model proposed is intended to create an integration framework for all the clinical data for each patient.

Clinical data can be integrated into repositories called data pools. These data pools are in their turn filed in a data base called Master Patient Index where all data are stored.

The de-identification process enables data to be depersonalized once each episode has been closed. In this way the data from each closed-episode Data Pool are guaranteed not to contain any data comprising personal information, However, via the Master Index data can be personalized.

Access to de-personalized data is controlled by search filters that possess no authentication or access authority. Data relating to episodes still open is only available to be consulted by the service that opened the episode, and this authority can be passed on only if the patient has been transferred to another service.

The policies relating to access and personalized data search procedures will be approved by a privacy and data protection commission, and will need to be relieved of authorization case by case.

With this model the various actors involved in health care provision will be able to share data about each episode.

Messages will have to be transmitted under AS1 or AS2 protocol with digital signature and data encryption.

In this way authentication, confidentiality and interoperationality between the various information systems within each organization can be ensured.

The Master Index will even act as a Federation of Electronic Health Record. This Master Index will control the relationship between the various keys, (DNA, NHS,

Healthcare Service number, ID, passport number, and Tax Number) and for each system will establish which keys are necessary for indexing the various systems.

In addition, it is proposed that there be created an onthology language which will set out the search rules to be enacted in order to ensure the citizen's privacy and security of their personal data.

Interoperationality among the various systems is ensured via communication protocols that allow online and offline communication between systems. At the same time encryption and authenticity of data must be guaranteed.

The protocol proposed is AS1 on smtp. The advantage of this protocol is that it is an asynchronous message protocol in xml.

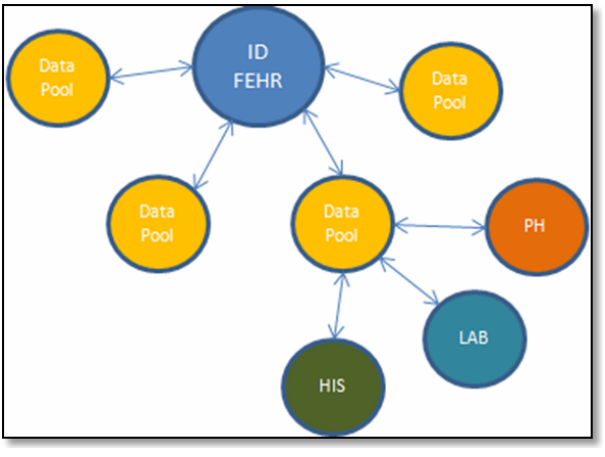


**Fig. 2.** The Proposed Model, (Authors Proposal)

This protocol can be used to communicate among systems of various technologies and in addition it employs a message technology, smtp, which is already well distributed around the market.

When a patient is presented to the system, the Hospital Information System Queries the ID FEHR (Federation Electronic Health Record), to find an identification, an associate open episode and all clinical data related with the patient.

The Patient Identification and data network are resolved with data mining algorithms.

The identity resolution is intended to find who is who and create links between data that belongs to the same patient.

The data used is demographic data and background clinical history. If there are some proximity of data attributes around a cluster centric it could be possible to say that all data belongs to the same patient.

The relationship resolution is intended to find all correlation between the data of different patients. The clusters can be built using a data mining algorithm- Inside the cluster all the data that is less than a $\delta$ distance from the cluster center belongs to individuals relationships.

The type of relationships are clinical data such as

- Pathologies and diagnostics
- Drugs and treatments prescribed
- Hospitals where patients were treated

And demographic such as:

- Nationality, Gender, date of birth and race
- Family relationships
- Living habitats
- Professions

The relevant clinical and demographic data are presented to the clinician as far as the treatment episode is open and would be uploaded to the data pool when the episode is closed.

In the data pool there is a hash algorithm that processes a de-identification of clinical data.

The process of de-identification is intended to overcome the privacy and confidentiality of clinical data. The clinical data primary key is substituted by a hash key data and can only by decrypted by master index algorithm. This master index algorithm is one of the functionalities of ID FEHR.

The master index in ID FEHR can be addressed by all sorts of patient identification keys including genome coding, National Security Number, among others. Besides the master index keeps track of nearby identification data and coded primary key of data pool clinical records.

The ID FEHR is also responsible for users' authentication and retrieval onthologies. These onthologies are used each time a query of data pools is needed. When a patient does not belong to an ID FEHR a negotiation with other ID FEHR is initiated.

## 6 Conclusions

The model proposed is founded on three fundamental aspects:

- An architecture already well distributed around the market
- Use of existing technology allowing interconnection of heterogeneous systems that incorporate privacy and security guarantees
- Use of alternative search keys and onthologies with data access rules

The reasoning behind this proposal is that it is inconceivable to render obsolete the many existing systems, all with their own different characteristics, and to develop one single, global information system.

Additionally, the fact that only one single data repository exists potentially increases the vulnerability of the data.

Development via existing technologies also potentially reduces the development lead-time necessary and reduces the cost.

Further research is required to find out:

How much dada will be needed to store in the Master Index to identify unequivocally a patient with a high degree of confidence?

What algorithm should be implemented to refine different patient matches?

## References


1. HL7 CDS Project Update:Virtual Medical Record (vMR). In: Clinical Genomics (2008), http://www.hl7.org/library/committees/clingenomics/HL7%20Phoenix%20-%20May%2008%20-%20CDS%20Genomics%20Jt%20Session.pdf
2. Shabo, A.: The Implication of Electronic Health Records for Personalized Medicine. Future Medicine (2005), http://www.hhs.gov/healthit/documents/Tab3part2Implications103106.pdf
3. Shabo, A.: Health Record Banks:Integrating clinical and genomic data into patient-centric longitudinal and cross-institutional health records. Future Medicine (2007), http://www.futuremedicine.com/doi/pdf/10.2217/17410541.4.4.453?cookieSet=1
4. Kent, S.T., Millet, L.I.: IDs-Not That Easy: Question About Nationwide Identity Systems. Committee on Authentication Techonologies and Their Privacy Implications, National Research Council (2002)
5. The Department of Health: Patient confidentiality and Access to Health Records (2007), http://www.dh.gov.uk/en/Managingyourorganisation/Informationpolicy/PatientConfidentialityAndCaldicottGuardians/DH_4084181
6. The United Kingdom Parliament (2009), http://www.parliament.uk
7. Clinical Genomics. HL7 (2009), http://www.hl7.org/Special/committees/clingenomics/docs.cfm
8. Marko, P.G., Wine, M., Joanne: Genomic Information Systems and Electronic Health Records (EHR). In: Virtual Medical World (2005), http://www.hoise.com/vmw/05/articles/vmw/LV-VM-10-05-1.html
9. Nakaya, J.: Clinical Genome Informatics (CGI) and its Social. IJCSNS International Journal of Computer Science and Network Security 7(1) (January 2007), http://paper.ijcsns.org/07_book/200701/200701A08.pdf
10. The Internet Engineering Task Force (2009), http://www.ietf.org/
11. Request for Comments: 3335, Network Working Group, MIME-based Secure Peer-to-Peer. In: Network Working Group (2002), http://www.ietf.org/rfc/rfc3335.txt
12. Soley, R.M.: Enterprise Patterns and MDA. Addison-Wesley, USA (2004)
13. Unified Modeling Language. UML Resource Page (2009), http://www.uml.org/
14. Health Level 7 (2009), http://www.hl7.org/