Medical terminologies for patients

Elena Cardillo

Institute of Informatics and Telematics UOS Rende (CS), National Research Council, Consenza, Itally

Annex 1
to SHN Work Package 3
Deliverable D3.3

Final version, March 29, 2015

Document description

Deliverable:	Annex 1 to SHN WP3 Deliverable D3.3
Publishable summary:	In this Annex, the state of the art of consumer-oriented medical terminologies is described, together with the tools to evaluate readability of medical texts. An approach is presented to the development of a multilingual consumer-oriented vocabulary, covering the most prevalent concepts in medical communication. Finally, integration of such a vocabulary into health information systems and linking to multiple pertinent international reference terminologies (International Classification of Primary Care, International Classification of Diseases, SNOMED-CT) is discussed.
Status:	Final Version
Version:	1.5
Public:	x Yes
Deadline:	May 29, 2015
Contact:	Elena Cardillo elena.cardillo@iit.cnr.it Robert Vander Stichele robert.vanderstichele@ugent.be
Editors:	Elena Cardillo

Table of content

Note :

This annex was commissioned by Prof. Dr. R. Vander Stichele, Workpackage Leader of SemanticHealth Net WP3, to Elena Cardillo, PhD, Institute of Informatics and Telematics UOS Rende (CS), National Research Council, lo, PhD, Institute of Informatics and Telematics UOS Rende (CS), National Research Council, Via P. Bucci, 17B, 7 floor, 87036, Rende, Cosenza, Italy

Email: elena.cardillo@iit.cnr.it Tel. : +390984494957

1 Introduction

During the last twenty years efforts in research and innovation in the field of Information Technologies for the biomedical domain and in particular for electronic Healthcare (eHealth) have defined new methodologies and standards to improve the management of healthcare data and processes of care. The integration of all the different information systems which need to exchange these data, and finally the creation of healthcare systems useful both for healthcare professionals (Electronic Health Records - EHRs) and for healthcare consumers (Personal Health Records - PHRs[1]) have also been improved. Thanks to the institutional measures and to the European *eGovernment Action Plan 2011-2015*, some European countries have started integrated management of patients' healthcare data, with the creation and the distribution of a federated solution of "Electronic Health Record" to be aligned with the international scenario.

To reach this aim, many projects and localized initiatives have been launched and infrastructures useful to support this solution have been created. These focus in particular on the Patient Summary and on electronic Prescription, searching for the alignment of medical encodings related to these two processes. Few efforts and contributions, on the other hand, are related to other important building blocks such as the Personal Health Records (PHRs), for the autonomous management and organization of healthcare data by the patient and on integrated medical vocabulary for the patient.

Standardized methodologies are developed to encode, retrieve, represent and integrate healthcare data, terminologies and classification systems by and for healthcare consumers, to help them define their personal and familiar clinical history and to ease access to clinical data and communication with professionals. The creation of personal health records meets the international challenge of "Patient Empowerment", giving more power to consumers for managing and organizing their own healthcare data. Patients want to be able to review and contribute to their records, to include their perspectives and priority symptoms for discussion and shared decision making with clinicians. To this end they need support to create the best possible clinical documentation and help with data entry in order to avoid degradation of information. This means investing in interface terminologies, dealing with multiple languages, and with the difference between lay language and medical jargon. Standard international classifications or terminologies may be needed to bring a real semantic interoperability, but their use may too complex even for domain expert, and certainly for patients. It is necessary to help healthcare consumers and all possible end-users (patients, physicians, nurses, etc.) to better understand the available terminologies. In order to allow high quality data entry and useful applications, medical terminologies must reflect the words and phrases of both intended users: clinicians on one hand and patients on the other.

To mitigate the linguistic gap between the lay language adopted by healthcare consumers[2] and the specialized, technical language of physicians and other healthcare providers, lay terminologies or vocabularies need to be created. Using these systems would allow easier and more efficient management and interpretation of patients' healthcare data, and better understanding of medical reports by consumers. In addition, physicians could import terms expressed by the patients into

[1] PHRs (Personal Health Records) are electronic health record managed by the patient, generally available on the web, that differ from EHRs since they are updated, and integrated directly by the patient with data on their clinical history and it is usually used to access to their clinical reports or test results, or for the self-monitoring of specific diseases conditions.

[2] The terms patients and healthcare consumers used to identify any actual or potential recipient of health care.

their EHRs, automatically encoded (e.g. symptoms, administrative procedure requests, adverse events, etc.).

A first step in this direction is the creation of PHR models which could be fully integrated in the EHR framework. A further step is needed for the identification of a contextualized lay language used by consumers for expressing healthcare concepts and consequently for the creation of consumer-oriented lexicons or vocabularies which implies the extraction of semantic and linguistic correspondences between lay language and international standardized medical classification systems, nomenclatures, and thesauri (ICD-10, ICD9-CM, ICPC-2, LOINC, SNOMED CT, UMLS, and MeSH) used all over Europe and in many other foreign countries for the medical documentation and coding and for medical information retrieval. The use of these lay terminologies and their interoperability to standardized terminologies can assure consumers easy access to and interpretation of their data, also in case of emergency, for example during travels abroad.

Concerning this last point, but also referring to those patients such as immigrants which have difficulties first of all with the hosting language and secondly with the technical jargon used in the health care setting, the multilingual aspect needs to be considered by providing the development of consumer-oriented medical vocabularies in various languages. This would be an added value together with semantic interoperability that will facilitate cross-border care.

Finally, the use of new technologies and languages, such as the one used in the Semantic Web [4] to represent these terminological resources and to model the knowledge included in healthcare information systems (e.g. EHRs and PHRs) will facilitate a series of knowledge services and automatic reasoning on the patient clinical data which will allow an easier and more efficient management (by clinicians) or self-management (by patients) and interpretation of healthcare data as well as an easier and quicker retrieval of medical information.

2 The communication gap in medicine

Electronic Health Records necessitate the integration of medical terminologies and coding systems that ease the registration of clinical data by the GPs or other healthcare professionals and that allow for the structured and interoperable reporting of clinical data. Such resources are characterized by a physician-oriented technical language. In the past years, steps forward were taken in the field of e-Health and so-called "Patient Empowerment". Patients or more generally healthcare consumers have taken an active role, both in the consultation of medical information online (thanks to the proliferation of dedicated websites) and in the access to and management of their healthcare data through the use of PHRs, available on the web, on their mobile or tablet. It is evident that consumers need support to read, interpret, and manage their medical data.

A solution, at least from the patient perspective, would be a simplification of the medical language and consequently a cleaning process of the language itself, through the creation of a terminology composed of well-defined and rigorously applied words. This does not imply, however, the abolition of specialized terms, because no specific language can do without its lexical background. Moreover, the medical language is overloaded with obsolete and archaic terms, eponyms, multiple synonyms, and semantic ambiguities [31].

Too often the professional inclination to use highly technical terms makes the patient uncomfortable, implying a strong effort in order to understand. Physicians and healthcare operators do not talk "with" the patient but "to" the patient [2]. Patients need clear and understandable language to communicate effectively [34]. These statements become more relevant when considering the communication (not mediated by the physicians) between patients and the recent tele-health applications.

A solution to the problem is the creation of medical vocabularies, terminologies or ontologies, specifically designed for patients [6]. The challenge is to map these terminological resources for patients to the specialized medical terms.

3 State of the art in the development of consumer-oriented medical terminologies

3.1 Consumer-oriented Vocabularies in the medical domain

Over the past 20 years researchers have worked on the development of lexical resources and terminologies that reflect the way healthcare consumers express and think about health related concepts.

An early example of a consumer-oriented medical terminology which considers multilingualism was commissioned by the European Commission in 1994. The *Multilingual Glossary of Popular and Technical Medical Terms* contained both lay and technical terms and expressions in 9 European languages and was limited to the terminology used in medication leaflets. It is, in fact, composed of the 1,400 most frequent technical terms used in drugs package inserts, with corresponding lay terms or definitions in English, Danish, German, Spanish, Italian, Dutch, Portuguese and Greek [21].

Later Soergel and Tze defined a methodology for the development of a common medical vocabulary, namely Consumer Medical Vocabulary, mapped to two different resources: (i) an intermediate vocabulary named Mediator Medical Vocabulary used by health care operators such as nurses who mediate between patients and clinicians, and (ii) a specialized medical vocabulary – Mediator Medical Vocabulary – used by professionals [43]. In this initiative linked lay and technical terms have also been mapped to the UMLS Metathesaurus to find synonyms and quasi-synonyms, and an intermediate layer has been created to interpret and mediate among the different types of vocabularies. A large number of common expressions (hundreds of thousands of tokens) were examined, leading to the discovery that between 20% and 50% of the lay expressions was not represented in the specialized medical vocabularies [25].

One of the major international efforts in this regard is the Consumer Health Vocabulary Initiative (CHV Initiative), launched in 2006 at the Brigham and Women's Hospital, Harvard Medical School for the development of the Open Access Collaborative Consumer Health Vocabulary (OAC CHV). It is a consumer-oriented medical vocabulary for English defined as: "a collection of forms used in health-oriented communication for a particular task or need […] by a substantial percentage of consumers from a specific discourse group and the relationship of the forms to professional concepts" [47]. In particular, the CHV includes common medical terms and their synonyms in multiple medical subdomains. In 2009 the CHV was officially incorporated in the Unified Medical Language System (UMLS) Metathesaurus. The lay terms of the CHV were linked to technical medical concepts (e.g. *Shortness of breath* linked to *Dyspnea*).

This type of vocabulary can have three possible bridging roles between consumers and health applications or information systems:

- Information Retrieval, because CHV facilitates automated mapping of consumer-entered queries to technical terms, producing better search results;
- Medical Records, since medical records and test results are nowadays available to patients, they frequently contain jargon, so a CHV can represent these terms with consumer-understandable names to help patients better interpret the medical concept;

- Health Care Applications, where patients may enter consumer expressions such as “nose bleeding” or “cluster headache”, receiving help via an integrated CHV, which would facilitate automated mapping of these expressions to technical concepts (in this case “epistaxis” and “histamine cephalalgia”) enabling consequent analysis and response.

In some cases, these vocabularies were applied in concrete use cases [27, 46].

In the United States context, many initiatives promoted by for-profit companies can be found. An example is the Consumer Health Terminology Thesaurus, developed by WellMed Inc., which is based on SNOMED (Systematized Nomenclature of Medicine) and contains more than 20,000 terms familiar to the patient, including many cases of dialectical and cultural lexical variants [51] [35].

In Europe, the *Italian Consumer Medical Vocabulary* (ICMV) was created. In line with the OAC CHV, it provides two main contributions: the creation of an Italian medical vocabulary oriented to consumers and patients developed by applying an hybrid methodology of knowledge acquisition and terminology extraction validated by domain experts [9]; and the integration of this resource with some specialized medical terms in the UMLS Metathesaurus (ICPC-2, ICD-10, SNOMED CT, LOINC) used by healthcare professionals in primary and secondary care, by using Semantic Web technologies and languages[3] [8].

One of the most recent attempts in building lay terminologies is the Mayo Consumer Health Vocabulary (MCV), a taxonomy of approximately 5,000 consumer health terms and concepts partially mapped to SNOMED CT and ICD-9 [40]. The authors here also developed text-mining techniques to expand its coverage by integrating disease concepts from UMLS as well as non-genetic (from deCODEme[4]) and genetic (from GeneWiki+[5] and PharmGKB[6]) risk factors to diseases. A comprehensive review of the literature from different databases (e.g. PubMed MEDLINE, CINAHL) and information sources (Library and Information Science Abstracts, and Library Literature) about consumer and patient language, and controlled vocabularies showed that consumer contributions to controlled vocabulary appear to be seriously under-researched inside and outside of health care [41].

Other works have been focused on the extension of consumer health vocabularies for specific medicine subdomains [36], where terms and the expressions used by lay persons speaking French to talk about *breast cancer* are identified and organized in a concept-based terminology.

A computer assisted update (CAU) system of the open access and collaborative (OAC) CHV is presented in [15], as a system consisted of three main parts (i.e. a Web crawler, an HTML parser, and a candidate term filter) that identifies new candidate terms from live corpora for inclusion in the (OAC) CHV. The CAU system was applied, for evaluation, to the health-related social network website PatientsLikeMe.com identifying 237 valid terms not yet included in the OAC CHV or in UMLS, among 774 candidate terms selected by the term filter from 300 crawled webpages.

[3] In particular the use of languages such as RDF and OWL for the formal representation of the terminological resources and of the vocabulary itself, and the use of SPARQL as query language for the extraction of the semantic correspondences. Finally, the use of collaborative tools such as *Semantic Media Wiki* for the terminology acquisition directly from users and for the clinical mapping by domain experts, as well as for the final publication of the Vocabulary.

[4] http://www.gene-tests.org/decodeme

[5] http://genewikiplus.org/wiki/Main_Page

[6] https://www.pharmgkb.org/

3.2 Techniques and tools for evaluate and enhance readability of medical information

The reviewed literature suggests that while giving patients access to medical documents has many benefits, lay people have difficulty in understanding medical information and in most cases this causes problems that only with the help of supporting tools can be solved. This requires first of all in-depth understanding of the nature and causes of comprehension errors that lay people make when dealing with clinical documents (e.g. misunderstanding of clinical concepts, misreporting of physician's findings, confusion or misspelling of clinical terms) [26]. In recent years research has focused to this end on the assessment of the readability of clinical documents as well as techniques useful to improve readability and guarantee the easier access by lay people to medical information. In [53] the problem has been approached by contrasting the 'readability' of two types of clinical documents: referral letters (76,012) vs. other genres of narrative clinician notes (2,118,463), using as a baseline a corpus of MedlinePlus articles—exemplars of fine patient education materials crafted for lay audiences. The readability has been quantified using three different measures: Flesch-Kincaid Grade Level (FKGL); Simple Measure of Gobbledygook (SMOG); Gunning-Fog Index (GFI). This work presented some limitations in the medical documents analyzed by the authors, retrieved from a single institution and patient care service (hematology/oncology), with the implication of the difficulty in the generalization of the results to medical content produced by other institutions or care setting. Another limitation recognized by the authors is the use of only computational measures to estimate readability. Other works are present in the literature aimed at the assessment of readability of clinical texts for lay people (see for instance [42], and [49]), but only few of them propose text simplification methods and tools to improve patients' electronic health record comprehension. One example is described in [24], where a simplification tool has been developed to simplify health information, that addresses semantic difficulty by substituting difficult terms with lay synonyms or through the use of hierarchically and/or semantically related terms (e.g. hyponyms or hyperonyms). The described tool also simplifies long clinical sentences by splitting them into shorter grammatical sentences, and has been tested to simplify electronic medical records and journal articles with good outcomes (e.g. for electronic medical records a statistically significant improvement has been shown in the cloze test score from 35.8% to 43.6%). Another good example of a text simplification tool is the BioNLP system NoteAid developed by [38], a system composed of a Concept Identifier module that performs typical NLP tasks and also matches concepts in other resources and a Definition Locator module that looks for concepts definitions from UMLS, MedlinePlus and Wikipedia. This system integrated in patients' electronic health records automatically recognizes medical concepts and links these concepts with consumer oriented, simplified definitions from external resources. An evaluation of the system showed that Wikipedia significantly improves EHR note readability, while MedlinePlus and the UMLS need to improve their content coverage for consumer health information in order to be useful as resources for NoteAid. A similar method has been proposed in [33], by computing in this case term familiarity to help estimate text difficulty.

All these studies highlight the necessity of informatics support tools to be used by health care professionals, to make clinical information understandable to patients or on the contrary used by healthcare consumers to better understand their clinical documents or healthcare information on the web.

4 Combined approach for the development of consumer-oriented medical vocabularies

Most of the approaches reviewed in this report share some peculiar steps that need to be performed when building vocabularies or terminologies oriented towards healthcare consumers. In particular it is possible to identify 4 main steps for the development of consumer medical vocabularies (CMV):

1) Identification of consumer-friendly terms used to indicate medical concepts in daily life and during their encounters with health care professionals (in particular to express symptoms and complaints, medical procedures and diseases), that can be performed by means of different elicitation techniques and usability studies and using automatic term extraction techniques from different corpora or websites oriented towards consumers;
2) Review of the consumer-friendly terms and selection of the candidate terms for the CMV;
3) Mapping of the selected terms to the corresponding terms/concepts in standardized medical terminologies, classification or ontologies used in the domain of application (in particular primary and secondary care) by means of a semi-automatic approach, including validation of the mappings by domain experts;
4) Evaluation of the feasibility of the vocabulary by means of its application within personal health records or its application to search engines. A brief description of these tasks is given in the next sections, in particular steps 1) and 2) in Section 4.1.; step 3) in Section 4.2. and step 4) in Section 4.3.

4.1 Generation of the vocabulary

One of the crucial steps in developing consumer medical vocabularies (CMV) is the identification of consumer health expressions. In order to identify consumer-friendly terms, usability studies can be performed, as shown in [22], where a usability study of the patient-friendly terms used in an ambulatory electronic medical record and associated patient web portal has been carried out investigating the usage patterns of consumer health vocabulary and evaluating the mapping to controlled terminologies used in the electronic medical records (e.g. UMLS). Other usability studies have been executed by [29] to find out the differences between consumer and medical vocabulary.

When carrying out these usability studies it has to be considered that the use of patient-friendly terms for some types of medical concepts would not always help to bridge the language gap between providers and consumers. In fact, considering diagnoses, the professional terms are used more frequently than their patient-friendly counterparts, typically in cases where the professional terms are more simple or common than the patient-friendly terms (e.g. in the case of Diabetes, Hypothyroidism, etc.). Consequently a low level of usability of lay terms for diagnoses has not to be seen as a negative result. What is helpful in this case is the identification of lay expressions or descriptions to clarify the meaning of a diagnosis.

Generally, combined approaches of semi-automatic analysis of text corpora and manual revisions performed by domain experts are preferable for the identification of lay linguistic forms used by healthcare consumers [48].

It needs to be highlighted that most of the methods used by existing studies in identifying consumer health expressions involve human efforts (e.g. by organizing interviews or focus groups with different samples of people), which is very time-consuming. Frequent is also the use of clinical guidelines as terminological resources, but being professional-oriented do not seem completely

adequate to the objective of this task. These issues can be reduced by considering as a terminological source online health social media sites that provide consumers with healthcare information sources as well as communication platforms for social interactions such as discussion forums and online social groups. These platforms collect an enormous amount of evolving consumer-contributed healthcare content. A great deal of information can be found on these social media sites before they are reflected to health professionals or recorded by some other means (health consumer opinions on new medical treatments, drugs, vaccine, etc.). So it turns out to be a great resource to harvest the timely consumer health expressions, which are not available through other channels. As these sources include both patients' comments and questions and answers posted by healthcare professionals, it is important to find methods to categorize posts produced by patients and those produced by health professionals in order to extract exactly consumer-friendly terms. A similar study can be found in [19] and [**Error! Reference source not found.**], where a supervised approach based on n-grams (vocabulary), emotion markers, uncertainty markers and misspellings has been evaluated to distinguish the two categories of posts on a French health forum (AlloDocteurs.fr).

One example of the automatic identification of relevant lay terms or expressions from consumer-contributed content on the web is the execution of co-occurrence analyses on consumer-based corpora such as messages on forums dedicated to specific medical subdomains, as done in [30], where relevant lay terms related to Adverse Drug Reactions (ADRs) have been extracted from a corpus of 120,393 discussion messages on a forum. Another example is the hybrid approach proposed in [9], that combines traditional knowledge acquisition techniques with the automatic collection of consumer-friendly terms from various sources on the web, in particular forum postings written by healthcare consumers on medical-consultation websites (i.e. medicitalia.it[7]), by applying NLP techniques and tools for term extraction, parsing, tagging and normalization. By means of NLP tools for term extraction it is possible to detect single and multiword medical terms and a basic semantic structure defining relations between the extracted terms (BT, NT, RT). Some of these tools (e.g. Text 2 Knowledge –T2K, or the tool used in the Terminology Extraction for Semantic Interoperability and Standardization project - TExSIS) developed specific adaptation modules to the biomedical domain [14, 32], being in this way more reliable in terms of precision and recall and for the extraction of semantic relations between terms. In fact, extracted single and multiword medical terms can be structured into fragments of taxonomical chains reconstructed from the internal linguistic structure of the terms itself (e.g. *corneal abrasion* is automatically associated with the relation IS-A to the term *abrasion*. Clusters of semantically related terms (RT) can be inferred through dynamic distributionally-based similarity measures using a context-sensitive notion of semantic similarity (computed with respect to the most relevant co-occurring heads). Doing so it is possible to extract relations between the term "contusion" and the terms sprain, and injury.

Concerning the selection of the candidate terms to be included in a CMV, it can be divided into two main tasks: (i) a statistical analysis based on term frequency and on the degree of familiarity of the extracted terms (which represents the level of understandability and use of a certain medical term for healthcare consumers) possibly assigned by a sample of healthcare consumers; and (ii) a clinical and semantic review of the extracted terms performed both by one or more domain experts (e.g. physicians, nurses and pharmacists) and by terminologists. The manual review by physicians serves principally for quality assurance, while the review by terminologists can be useful to find mistakes and incongruities in categorization and synonymy. As mentioned in [26], possible and categorizing types of mismatches from automated mechanisms include: misspellings (e.g., "erpes"

[7] http://www.medicitalia.it/

instead of “herpes” or “celebral ictus” instead of “cerebral ictus”), truncations (e.g., “Down” for “Down Syndrome”), acronyms, abbreviations and word fragments.

The review step include a conceptualization of the extracted terms and consequently the association to each term of a standardized definition in order to make explicit their meaning as well as perform a general categorization of the concepts (e.g. is the concept a disease, a symptom, an anatomical concept, etc.). This task is important to disambiguate such terms that can have different meaning, so homonyms (terms that have the same orthographical form and phonology but can express more than one meaning). In fact, there can be terms such as the term “mark” or “spot” used in medicine to indicate a visible sign on the skin (defined respectively as “a visible impression on a surface, as a line or spot”, and “a small blemish or other mark on the skin”), but also having the other “non-medical” meanings (e.g. respectively “a symbol used in writing or printing”, a punctuation mark, “a position in an organization or hierarchy”, etc.). It is important, then to recognize the medical meaning of the concepts in this phase in order to allow ahead a semantically correct alignment of the selected patient-friendly terms and those included in international medical classifications or terminologies.

Some of the studies mentioned in Section 3.1., in order to find standardized definition of the patient-friendly terms, performed automated extraction of the definitions, in English, from the UMLS Metathesaurus (see for instance [47], [43] and [10]), so considering the so-called “Concept Unique Identifier” (CUI), a code that represents a concept in the UMLS Metathesaurus. The possible ‘senses’ of a term can be considered as the set of CUI’s which list this term as a possible realisation (e.g. the term *trauma* in UMLS is a possible realisation of the two concepts: C0043251 *Injuries and Wounds* in the sense of traumatic injury and C0021501 *Physical Trauma*. Since in UMLS this term can be used to express more than one UMLS concept, a disambiguation process is needed to find out which of its possible senses is actually being used in each particular context where there term trauma is used [52].

4.2 Integration of the vocabulary with standardized terminologies and classification systems

After creating the consumer-oriented vocabulary, a crucial step is its semantic integration with standardized technical and physician-oriented medical terminologies or classification systems. This consists in finding for each selected lay term the corresponding technical ones in resources that are widely distributed and used worldwide in the domain of application (e.g. ICD-9-CM, ICD-10, ICPC-2, SNOMED CT, UMLS, etc.). By means of the mapping process it is possible to reconstruct the meaning inherent in the lay usage of a term, and consequently to show that compatibility between lay and professional terms exists on the basis of this deeper meaning, rather than on the basis of the lexical form. In order to have an integration that could be feasible at a local level, it is important to map CMV not only with standardized international terminologies and classification systems but also with reference terminologies and end-user terminologies that represent the concepts and terms used in the daily practice by physicians in a specific local context. These terminologies are generally created by extracting concepts and terms directly from EHRs, by guidelines and also by online medical consultations. Many studies have been published on analysis of physicians language extracted from EHRs and compared to standardized coding systems [7] and also on the development and semantic integration of reference and end-user terminologies (see for instance [28**Error! Reference source not found.**] and [10]). The correspondence between consumer-friendly terms in the CMV and the terms in the selected

standardized resources can be of four different types: exact mapping between the pairs, when the lay term has exact correspondence to a term in the other resource and both have the same meaning; synonymy relation, when the lay term does not exist in the professional resource, but corresponds to a technical term that denotes the same concept, as in the case of *nosebleed* and *epistaxis*; hyponymy relation, if the lay term is considered as term of inclusion of one or more concepts in the other resource (e.g. the relation between *absence of voice* and the broader concept *voice symptom*); hypernymy relation, when the lay term includes in its meaning one or more terms in the other resource (e.g. the relation between *bronchitis* and *chronic bronchitis*.

Mappings can be performed manually or automatically. When CMV does not include a high number of terms and it needs to be mapped only to one standardized resource that is not very granular either (e.g. ICPC), then manual mapping can be performed. In this case, domain experts are generally called to manually find a “one-to-one” or “one-to-n” mapping of lay terms with the corresponding medical concepts in the other resource, to define explicit relationships among them.

On the contrary, in the presence of a large vocabulary and considering to align more than one standardized resource to the CMV, than an automatic or semi-automatic mapping process is preferred. Some semi-automatic methods for detecting matching between biomedical vocabularies are based on similarity functions and measures applied both to single words and n-grams, as proposed in [37].

In [6] both the approaches have been used. First, a clinical manual mapping between the Italian Consumer-oriented Medical Vocabulary (ICMV) and the International Classification of Primary Care, 2nd revision (ICPC-2) has been performed by a sample of general practitioners, who identified correspondences between Italian lay expressions or terms in the ICMV (e.g. “sentirsi il cuore in gola” - feel your heart your throat) and the ICPC-2 rubrics (e.g. in this case K04 Palpitazioni - Palpitations). Second, the automatic mapping between the ICMV and selected international classifications and terminologies in the UMLS Metathesaurus (more precisely SNOMED CT, ICD-10, MeSH and LOINC), has been executed by using semantic web technologies [8]. Here, in particular, after the encoding of the ICMV and of the selected international resources in RDF (Resource Description Framework), the resulted graphs have been collected in an RDF triple store (e.g. Virtuoso, or Sesame), and SPARQL queries performed on the stored graphs to extracted semantic mappings between them using ICPC-2 as a pivot to access UMLS vocabularies. At the end of this process, manual mapping between ICMV and ICPC2 and the automatic mapping to UMLS vocabularies via ICPC2 was compared, in order to evaluate the best approach.

The use of Semantic Web technologies has led to promising results in the area of information integration across heterogeneous resources. For example the work of Bodenreider (2008) can be mentioned [5], where the Resource Description Framework (RDF) is used for comparing formal definitions between LOINC and SNOMED CT and in SNOMED CT and the NCI Thesaurus. However, the major use of these technologies for healthcare has been tested for the formalization of existent medical terminologies or classification systems in ontologies (one of the latest works in this sense is the collaborative ontological development of ICD-11 revision, coordinated by the Stanford Centre for Biomedical Research (BMIR) [45]. These technologies are very useful in healthcare, especially considering that the treatment of a patient may involve several practitioners from different healthcare institutes, and that there is an increasing need to access patients' healthcare records electronically wherever they are stored. Knowledge representation techniques provide, for example, suggestions on how to manage a patient's condition; tests that have to be

carried out; medications or treatment to take into account, etc. In this case ontologies[8] become relevant if integrated into EHRs or PHRs, since they manage an increasing volume of narrative data, in order to allow: structuring and semantics of the recorded information; and references to concepts from terminologies such as ICD 10/9 or SNOMED CT [11].

After performing the definition of mappings between consumer-oriented medical vocabularies and other international medical resources an evaluation of the quality assurance of the mappings needs to be performed.

The integration of consumer-oriented medical vocabularies with international terminologies and classification systems is needed to supply knowledge services to support the development of semantic-based healthcare information systems which implies interchanges with patients and healthcare consumers in general and consequently to guarantee semantic interoperability.

4.3 Evaluation of consumer-oriented vocabularies and application in healthcare information systems

Older efforts to improve consumer-friendliness of PHR and EHR information have focused on user interface design and/or the links to references or educational materials (i.e., infobuttons) [13], [3].

New methods for evaluating the feasibility of consumer-oriented vocabularies consider their integration in PHRs for supporting healthcare consumers in data entry of medical terms (e.g. symptoms, diseases, interventions, allergies, and other relevant information of their clinical history), but above all for supporting them during medical information searching for the comprehension of their clinical reports, test results, discharge letters, etc., available on their PHR. In this case the CMV and its integration framework are used as knowledge sources to be queried when consumers edit terms on their searching panel on their PHR or need to add terms on specific fields of the application.

PHRs can communicate with physicians' EHRs, for example to receive medical reports, test results, and other documents which use specialized medical terminologies (e.g. ICD-9, ICD10) so having integrated a CMV could give the possibility to foster the readability of data deriving from EHRs but also to send to physicians PHR content filled out in lay terms and then to translate this in technical language.

During the last five years, many experimental uses of PHR systems were promoted to improve patient empowerment. The scientific community launched standardization initiatives to develop common formats for allow consumers and patient collecting and managing personal healthcare data and to solve the problem of data interoperability with EHRs and other healthcare information systems [18]. In some EU countries normative measures have been adopted in this field, such as the e-Government 2012 plan in Italy, which scheduled, beyond the simplification and digitalization of digital prescriptions and disease certifications, online reservation systems, the creation of infrastructures to supply healthcare services which meet consumers' needs. In this national context a technological infrastructure for a federated electronic patient record, namely Fascicolo Sanitario Elettronico (FSE), that enables citizens and authorized health professionals to access the health data wherever they are located in the national territory or abroad, preserving the citizen privacy, and facilitating the management of the evolution of the process of care has been developed (within the InFSE - Infrastruttura del Fascicolo Sanitario Elettronico - and the OpenInFSE - Evoluzione e

[8] Ontologies, defined as "explicit specification of a conceptualization" [12], capture the meaning of a particular subject domain that corresponds to what a human being knows about that domain. They are typically represented as classes, properties, attributes and values. Furthermore, ontologies allow sharing knowledge between people, agents, and software; enable reuse of domain knowledge; and enable automated reasoning on data.

interoperabilità tecnologica del Fascicolo Sanitario Elettronico - projects) [12, 16]. Within the FSE project some attempts have been made on the development of a specific section of FSE, namely *Taccuino personale del Cittadino*, addressed to healthcare citizens for the easy registration of personal and familiar healthcare information and data, on their habits concerning food and physical activities, as well as for the registration of clinical documents issued by healthcare facilities that are not affiliated to the National Healthcare System and to maintain a daily diary for relevant events (e.g. visits, diagnostic exams, observable parameters monitoring, etc.). Different solutions have been proposed by Regions and in the context of national research projects, but the most advanced solution in terms of functionalities, interoperability and usability has been experimented and funded by the Autonomous Province of Trento, where the TreC (Cartella Clinica del Cittadino) PHR has become a concrete integrated service for the citizens of this territory and is the sole relevant case in Italy where consumer-oriented vocabularies have been applied [44].

Regarding the possibility to use a CMV as support tool for the retrieval of health care information and literature during web searching both performed by lay persons or by physicians, some studies proposed methods providing reformulation of consumers' queries for better medical information search returns [30]. Results improves if the CMV is mapped to UMLS and in particular to MeSH, as used for searching purposes and health-related literature indexing) [23].

5 Conclusions

In this annex the problem of the linguistic gap between “lay” and “specialized” terminology in the medical domain has been treated. The reviewed studies have shown that patients often create semantic ambiguities in using lay terms for expressing clinical concepts, in fact they use some terms interchangeably as synonyms (as in the case of headache and migraine) making errors. Above all they find it difficult to understand technical medical terms used by physicians and other professionals in clinical documents and EHRs as well as in informative webpages, etc. This can wrongly influence medical information comprehension and consequently decision making. In order to avoid these consequences researchers have started to address this issue through the creation of consumer-oriented medical vocabularies by also providing their integration with standard and international medical terminologies and classification systems used in the healthcare domain, in contrast to traditional approaches which proposed the use of specialized medical terminologies to be integrated in consumer-oriented healthcare application such as personal health records or mobile application for the monitoring particular health conditions.

The present document highlights the need for mapped consumer-oriented vocabularies in order to guarantee an integration framework that can be reused in different consumer-oriented healthcare applications, taking advantage of Semantic Web technologies.

The application of consumer-oriented medical vocabularies can contribute to the support of healthcare consumers and laypersons as well as physicians in different scenarios:

1) translating and interpreting clinical notes or test results containing medical jargon (e.g., mapping between a standardized medical vocabulary used in the physician’s EHR to a consumer-oriented vocabulary could be useful in providing consumer-understandable names to help patients interpret these documents);

2) searching for healthcare information (e.g., facilitating automated mapping of consumer-entered queries to technical terms – if the term queried is mapped to a thesaurus such as MeSH it would produce better search results during the search in a bibliographic database such as PubMed);

3) helping patients in the description of their clinical history, symptoms and complaints both in their PHR and in online medical consultations; helping physicians and other healthcare providers in the process of encoding reasons for encounters (symptoms, diseases, diagnoses and procedures); helping physicians to automatically interpret their patients clinical history stored in their PHR, and to automatically produce clinical notes understandable by healthcare consumers and patients.

Finally, many potential applications in the patient empowered health care (e.g. Home care) are possible.

References

1. *Abdaoui A., Azé J., Bringay S., Grabar N., and Poncelet P. Analysis of forum messages written by health professionals and patients In Proceedings of MIE 2014. Istanbul, Turkey. Poster.*
2. *Baldini M., Parlare al paziente, parlare "col" paziente, in Proceedings of the Conference L'arte medica tra comunicazione, relazione, tecnica e organizzazione, Scriptorium, Torino, pp. 9-25, 1996.*
3. *Baorto D. M. and Cimino J. J.. An "infobutton" for enabling patients to interpret on-line pap smear reports. In Proceedings of AMIA Symposium - AMIA2000, pages 47–50, 2000.*
4. *Berners-Lee T., Hendler J., and Lassila O.. The Semantic Web. Scientific American, 284(5), 34—43, 2001.*
5. *Bodenreider O.. Issues in Mapping LOINC Laboratory Tests to SNOMED CT. In AMIA Annual Symposium, AMIA2008, pp. 51—55, 2008.*
6. *Cardillo E.. A lexi-ontological resource for consumer healthcare. The Italian Consumer Medical Vocabulary, Doctoral Thesis, Fondazione Bruno Kessler, University of Trento, Italy, April 2011.*
7. *Cardillo E., Chiaravalloti M. T., Pasceri E.. Assessing ICD-9-CM and ICPC-2 Use in Primary Care. An Italian Case Study. In Proceeding of the 5th International Conference on Digital Health (Digital Health 2015), Florence, 18-20 May 2015 (accepted).*
8. *Cardillo E., Hernandez G., and Bodenreider O.. Integrating consumer-oriented vocabularies with selected professional ones from the UMLS using Semantic Web Technologies. Proceedings of the 3rd International Conference of Electronic Health (eHealth2010), 12-15 December, Casablanca, 2010.*
9. *Cardillo E., Tamilin A., and Serafini L.. A Methodology for Knowledge Acquisition in Consumer-oriented Healthcare. In Knowledge Discovery, Knowledge Engineering and Knowledge Management: IC3K Revised Selected Papers, pp. 64 -71, Springer Berlin, 2010.*
10. *Cardillo E., Warnier M., Roumier J., Jamoulle M., Vander Stichele R.. Using ISO and Semantic Web standards for creating a Multilingual Medical Interface Terminology: A use case for Heart Failure, in Guadalupe Aguado de Cea, Nathalie Aussenac-Gilles, Terminologies and Terminology and Artificial Intelligence, Paris, IRIT, 2013, pp. 27-34, (10th International Conference on Terminology and Artificial Intelligence - TIA2013, Paris, 28 - 30 October).*
11. *Ceusters W., Smith B. and De Moor G.. Ontology-Based Integration of Medical Coding Systems and Electronic Patient Records. In MIE2005, 2005.*
12. *Chiaravalloti M. T., Ciampi M., Pasceri E., Sicuranza M., De Pietro G., Guarasci R.. A model for realizing interoperable EHR systems in Italy, in Proceedings of the 15th International HL7 Interoperability Conference, pp. 13-22.*
13. *De Clercq P. A., Hasman A., and Wolffenbuttel B. H.. A consumer health record for supporting the patient-centered management of chronic diseases. Medical Informatics and the Internet in Medicine, 28(2):117–127, June 2003.*
14. *Dell'Orletta F., Venturi G., and Montemagni S.. Unsupervised Linguistically-Driven Reliable Dependency Parses Detection and Self-Training for Adaptation to the Biomedical Domain. In Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), Association for Computational Linguistics, pages 45–53, Sofia, Bulgaria, August 4-9, 2013.*

15. *Doing-Harris K. M. and Zeng-Treitler Q.. Computer-assisted update of a consumer health vocabulary through mining of social network data. J Med Internet Res. 2011 May 17;13(2):e37.*
16. *Esposito A., Sicuranza M., Ciampi M.. A patient centric approach for modeling access control in EHR systems. 13th International Conference on Algorithms and Architectures for Parallel Processing, ICA3PP 2013. Springer International Publishing, pp. 225-232.*
17. *Euzenat J. and Shvaiko P.. Ontology Matching. Springer, 2007.*
18. *Flatley B. P. et al.. Project HealthDesign: Rethinking the power and potential of personal health records. Journal of Biomedical Informatics , Volume 43 , October 2010, Issue 5 , S3 - S5.*
19. *Grabar N. and Hamon T.. Automatic extraction of layman names for technical medical terms. In Proceedings of ICHI 2014. Pavia, Italy.*
20. *Gruber M., Thomas R.. Toward principles for the design of ontologies used for knowledge sharing. Padua workshop on Formal Ontology, 1993.*
21. *Heymans Institute of Pharmacology EEC. Multilingual Glossary of Technical and Popular Medical Terms in nine European Languages: Final report, Technical report, University of Ghent, Mercator College, Department of Applied Linguistics, Ghent, Belgium, 1995.*
22. *Hong Y., Ehlers K., Gillis R., Patrick T., and Zhang J.. A Usability Study of Patient-friendly Terminology in an EMR System. In Studies in Health Technology and Informatics, Volume 160: MEDINFO 2010; pp. 136 – 140. doi=10.3233/978-1-60750-588-4-136.*
23. *Hong Y., Gillis R. D., Donnell R. F. .Use of consumer health vocabularies in online physician directory to improve physician search. In Proceedings of AMIA Annu Symp 2008, Nov 6:974.*
24. *Kandula S., Curtis D., and Zeng-Treitler Q.. A Semantic and Syntactic Text Simplification Tool for Health Content. In Proceedings of AMIA Annual Symposium, pp. 366-370, 2010.*
25. *Keselman A., Logan R., Smith C.A., Leroy G., and Zeng Q., Developing Informatics Tools and Strategies for Consumer-centered Health Communication, in «Journal of the American Medical Informatics Association», 2008, vol. 14, n. 4, pp. 473-483.*
26. *Keselman A. and Smith C.A.. A classification of errors in lay comprehension of medical documents. J Biomed Inform. 2012 Dec;45(6):1151-63.*
27. *Kim H., Zeng Q., Goryachev S., Keselman A., Slaughter L., and Smith C.A.. Text Characteristics of Clinical Reports and Their Implications for the Readability of Personal Health Records, in Kuhn K. A., Warren J. R., Leong T.-Y. (eds.): MEDINFO 2007 - Proceedings of the 12th World Congress on Health (Medical) Informatics. Building Sustainable Health Systems, 20-24 August 2007, Brisbane, Australia, in «Studies in Health Technology and Informatics», 2007, vol. 129, pp. 1117-1121.*
28. *Jamoulle M., Cardillo E., Roumier J., Warnier M, Vander Stichele R.. Mapping French terms in a Belgian guideline on heart failure to international classifications and nomenclatures: the devil is in the detail. In Informatics in Primary Care (2014), 4, pp. 189-198. DOI=10.14236/jhi.v21i4.66*
29. *Jeongeun K., Jeeyoung J. and Yoonju S.. An Exploratory Study on the Health Information Terms for the Development of the Consumer Health Vocabulary System. In Studies in Health Technology and Informatics; Volume 146: Connecting Health and Humans; Ios Press, 785 – 785, 2009. doi=10.3233/978-1-60750-024-7-785.*

30. *Jiang L. and Yang C. C.. Using Co-occurrence Analysis to Expand Consumer Health Vocabularies from Social Media Data. IEEE International Conference on Healthcare Informatics (ICHI), 2013.*
31. *Iandolo C., Parlare col Malato. Tecnica, Arte ed errori della comunicazione. Armando, Roma, 1983.*
32. *Lefever, E., Macken, L., & Hoste, V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. The Association for Computational Linguistics, Athens, Greece.*
33. *Leroy G., Endicott J. E., Mouradi O., Kauchak D., and Just M. L.. Improving perceived and actual text difficulty for health information consumers using semi-automated method. AMIA Annu Symp Proc. 2012; pp. 522-531, 2012.*
34. *Lucchini A. (eds.). Il linguaggio della Salute. Come migliorare la comunicazione con il paziente, Sperling & Kupfer, Milano, 2008.*
35. *Marshall P. D., Bridging the Terminology Gap Between Health Care Professionals and Patients with the Consumer Health Terminology (CHT), in Proceedings of AMIA Annual Symposium, AMIA2000, p. 1082, 2000.*
36. *Messai R., Simonet M., Bricon-Souf N., and Mousseau M.. Characterizing Consumer Health Terminology in the Breast Cancer Field. In Studies in Health Technology and Informatics, Volume 160: MEDINFO2010, pp. 991 – 994, 2010. doi=10.3233/978-1-60750-588-4-991.*
37. *Ofoghi B., Lopez-Campos G. H., Martin Sanchez F. J., and Verspoor K.. Mapping Biomedical Vocabularies: A Semi-Automated Term Matching Approach. In Studies in Health Technology and Informatics, Volume 202: Integrating Information Technology and Management for Quality of Care, pp. 16 – 19. 2014. doi=10.3233/978-1-61499-423-7-16*
38. *Polepalli R. B., Houston T., Brandt C., Fang H., and Y. H.. Improving patients' electronic health record comprehension with NoteAid. Stud Health Technol Inform. 2013;192:714-8.*
39. *Plovnick R. M. and Zeng Q. T.. Reformulation of Consumer Health Queries with Professional Terminology: A Pilot Study. In J Med Internet Res 2004;6(3):e27. DOI=10.2196/jmir.6.3.e27*
40. *Seedor M., Peterson K. J., Nelsen L. A., Cocos C., McCormick J. B., Chute C. G., and Pathak J.. Incorporating expert terminology and disease risk factors into consumer health vocabularies. Pac Symp Biocomput. 2013:421-32.*
41. *Smith C. A.. Consumer language, patient language, and thesauri: a review of the literature. J Med Libr Assoc. 2011 Apr;99(2):135-44. DOI= 10.3163/1536-5050.99.2.005.*
42. *Smith C. A., Hetzel S., Dalrymple P., and Keselman A.. Beyond readability: investigating coherence of clinical text for consumers. J Med Internet Res. 2011 Dec 2;13(4):e104.*
43. *Soergel D., Tse T., and Slaughter L., Helping Healthcare Consumers Understand: an "Interpretative Layer" for Finding and Making Sense of Medical Information, in Proceedings of the international conference IMIA2004, San Francisco, California, pp. 931-935, 2004.*
44. *TreC – Cartella Clinica del Cittadino. The project is consultable at: http://trentinosalute.net/trec*
45. *Tudorache T., Nyulas C. I., Noy N. F., Redmond T., and Musen M. A.. iCAT: A Collaborative Authoring Tool for ICD-11. In Proceedings of the ISWC 2011 Workshop*

Ontologies Come of Age in the Semantic Web (OCAS), CEUR-WS Vol-809, Bonn, Germany, October 24th, 2011.

46. *Zeng Q., Goryachev S., Keselman A., and Rosendale D.. Making Text in Electronic Health Records Comprehensible to Consumers: A Prototype Translator, in Proceedings of the AMIA Annual Symposium – AMIA2007, 2007, pp. 846-850.*
47. *Zeng Q. and Tse T.: Exploring and Developing Consumer Health Vocabulary. In JAMIA 2006;13:24-29*
48. *Zeng Q., Tse T., Divita G., Keselman A., Crowell J., Browne A. C., Goryachev S., and Ngo L., Term Identification Methods for Consumer Health Vocabulary Development, in «Journal of Medical Internet Research», 2007, vol. 9, n. 1, p. 4.*
49. *Zeng-Treitler Q., Kandula S., Kim H., and Hill B.. A Method to Estimate Readibility of Health Content. ACM SIGKDD Workshop on Health Informatics (HI-KDD 2012), Beijing, China, 2012.*
50. *Zhang S., and Bodenreider O.. Experience in aligning anatomical ontologies. International Journal on Semantic Web and Information Systems, 2007, 3(2):1-26.*
51. *WellMed Medical Management, <https://www.wellmedmedicalgroup.com/>.*
52. *Widdows D., Peters S., Cederberg S., Chan C-K.. Unsupervised Monolingual and Bilingual Word-Sense Disambiguation of Medical Documents using UMLS. In Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine - Volume 13 (BioMed '03), Vol. 13. Association for Computational Linguistics, Stroudsburg, PA, USA, 9-16. doi=10.3115/1118958.1118960*
53. *Wu D. T., Hanauer D. A., Mei Q., Clark P. M., An L. C., Lei J., Proulx J., Zeng-Treitler Q., and Zheng K. Applying multiple methods to assess the readability of a large corpus of medical documents. Stud Health Technol Inform. MEDINFO2013, vol. 192, pp.647-51, 2013. doi:10.3233/978-1-61499-289-9-64*