Reconciliation of patient/doctor vocabulary in a structured resource

Mike Donald Tapi Nzali, Jérôme Azé, Sandra Bringay, Christian Lavergne, Caroline Mollevi, Thomas Optiz

Health Informatics Journal

Original Article

Reconciliation of patient/doctor vocabulary in a structured resource

Health Informatics Journal
1–13

DOI: 10.1177/1460458217751014
journals.sagepub.com/home/jhi
SAGE

Mike Donald Tapi Nzali and Jérôme Aze
Université de Montpellier (UM), France

Sandra Bringay and Christian Lavergne
Université Paul-Valéry Montpellier III, France

Caroline Mollevi
Institut du Cancer de Montpellier, France

Thomas Optiz
INRA, France

Abstract

Today, social media is increasingly used by patients to openly discuss their health. Mining automatically such data is a challenging task because of the non-structured nature of the text and the use of many abbreviations and the slang terms. Our goal is to use Patient Authored Text to build a French Consumer Health Vocabulary on breast cancer field, by collecting various kinds of non-experts' expressions that are related to their diseases and then compare them to biomedical terms used by health care professionals. We combine several methods of the literature based on linguistic and statistical approaches to extract candidate terms used by non-experts and to link them to expert terms. We use messages extracted from the forum on 'cancerdusein.org' and a vocabulary dedicated to breast cancer elaborated by the Institut National Du Cancer. We have built an efficient vocabulary composed of 192 validated relationships and formalized in Simple Knowledge Organization System ontology.

Keywords

consumer health vocabulary, information extraction, ontology, social media, text mining

Objective

Controlled vocabularies such as SNOMED,[1] MeSH[2] and UMLS[3] play a key role in biomedical text mining applications. These vocabularies contain solely the terms used by health professionals. For the last 10 years, vocabularies dedicated to health care consumers also called Consumer Health

Corresponding author:
Mike Donald Tapi Nzali, LIRMM, UMR 5506, Université de Montpellier (UM), 860 rue de Saint Priest – Bât 5, Cedex 5, Montpellier 34095, France.
Email: mike-donald.tapi-nzali@umontpellier.fr

Vocabulary (CHV) have been created. These CHVs link common terms related to health used by laymen to expert terms used by health professionals. Such CHV can be used to extract relevant information from social media.

In this article, we address the challenge of semi-automatically building a CHV in French. For example, we seek to link the term 'onco' used by patients for the term 'oncologue' (oncologist in English) used by health professionals. The originality of our approach is to use texts written by patients (patient-authored text (PAT)) collected from forums.

The main contributions of this article are twofold. First, we describe a global process from candidate terms extraction to formalization of the relationships in Simple Knowledge Organization System (SKOS). Second, we compare three measures based on different paradigms to link common terms to expert terms. Our method has been validated successfully, automatically and manually on the forum at 'cancerdusein.org', which is dedicated to breast cancer.

Motivations and current situation

According to a survey carried out in 2011 by the Health On the Net (HON) Foundation,[4] the Internet has become the second source of information for patients after consultations with a doctor; 24 per cent of the population uses the Internet to find information about their health at least once a day (this can rise to six times a day) and 25 per cent at least several times per week. While maintaining anonymity, social media also allow patients to freely discuss with other users and also with health professionals. They discuss their medical results and their treatment options, but they also receive moral support. Househ et al.[5] explored the range of social media platforms used by patients and examines the benefits and challenges of using these tools from a patient perspective.

In previous work,[6] we were interested in the study of the quality of life of patients with breast cancer in social media. We captured and quantified what patients express in forums about their quality of life. The main limitation of this work is due to the vocabulary in these types of posts which decrease the performances of automatic methods. Indeed, most patients are laymen in the medical field. They use slang terms, abbreviations and a specific vocabulary constructed by the online community, instead of expert terms that can be founded in medical terminologies used by health professionals such as SNOMED, MeSH and UMLS. Most of the text mining methods have shown their limits because of this particular vocabulary. In this work, we therefore propose to build a French CHV specialized for breast cancer.

Initially, the creation of this CHV has been motivated by the need to reduce the vocabulary gap between patients and health professionals.[7] Indeed, literature shows that patients' understanding of medical terminology is essential to participate in the medical decision-making process.[8] Some researchers used CHV to improve the readability by non-experts of medical documents[9] or patient's electronic records.[10]

Recent methods have been developed to extract consumer health expressions from social media. Doing-Harris and Zeng-Treitler[11] automatically generated candidate terms to be processed by humans for inclusion in a CHV. Jiang and colleagues[12,13] used co-occurrences for consumer health expression extraction from social media. Bouamor et al.[14] used a learning-to-rank method, where statistical and linguistic features are combined to determine whether a term is associated with lay or a specialized audience. Keselman et al.[15] extracted consumer health concepts manually from health-focused Bulletin boards. They mapped these concepts automatically to UMLS concepts using MetaMap and manually for the remaining frequently used terms. Patrick et al.[16] manually extracted common terms from e-mail questions submitted by consumers to a health care institution and from enquiries submitted by consumers to a health care information website. They mapped them to expert terms extracted from a corpus of 25,000 family-medicine progress notes created by

family physicians. Vydiswaran et al. used Wikipedia as corpus for pair extraction using explicit patterns (e.g. *also called, commonly referred to as*). They associate expert terms with common terms and also non-expert terms with expert terms.[17] Despite these initiatives, currently, only one CHV is available: Open Access and Collaborative Consumer Health Vocabulary (OAC CHV).[18] This CHV is included in UMLS. To our knowledge, there is currently no CHV in French.

In this article, our goal is to use PAT published on social media to build a French CHV in the field of breast cancer. The volume of texts written by patients in social media, such as forums, is becoming more and more important.[19] Such PAT provides access to many descriptions of patients' experiences and a wide range of topics. In the past 5 years, there has been a growing interest in the exploitation of these PAT as a tool for public health, for example, to analyse the spread of infectious diseases.[20] In this work, we propose to mine forums to extract common terms by collecting various kinds of patients' expressions, such as abbreviations, frequently misspelled terms or common terms used by non-experts to talk about their diseases.

The originality of our approach is to combine three types of measurements based on three paradigms in order to reconcile non-expert and expert terms.

The two first measurements are based on co-occurrence paradigm. The envisaged hypothesis is that the patients' vocabulary evolves when exchanging with the community. Consequently, we retrieved common and expert terms in the same threads in forums or in the same documents indexed by Google. We have implemented traditional measurements to calculate the degree of association between non-expert and expert terms. Such measurements (Dice, Jaccard, Cosine or Overlap) were recently discussed in Pantel et al.[21] and Zadeh and Goel.[22] These measurements are used in many fields such as ecology,[23] medicine[24] and language processing.[25] After preliminary experimentations, we present a modified version of the Jaccard measurement which compares the number of occurrences of two terms independently and together in forum threads. Similarly, we implement the Normalized Google similarity which is based on the number of occurrences of the two terms together and independently in the documents indexed by the web search engine Google.[26]

The last measurement is based on the architecture of the collaborative universal encyclopaedia Wikipedia.[27] We exploit the network of links between Wikipedia pages to calculate a metric. The French version of Wikipedia from 25 February 2016 contains more than one and a half million articles. Wikipedia articles have been used successfully in knowledge acquisition,[28] question/answer applications[29] and text categorization.[30] Chernov et al.[31] used links between Wikipedia categories to extract semantic information. Witten and Milne[32] used links between Wikipedia articles to determine the semantic proximity between terms.

In order to benefit from the advantages of the CHV, we formalized the obtained relationships as an ontology. We decided to use SKOS. SKOS is a concept diagram representation language, like thesauri, taxonomies and controlled vocabularies.[33] It allows expressing and managing easily interpretable models by machines in the perspective of the semantic web. SKOS itself being an OWL ontology, the representation of SKOS is based on Resource Description Framework (RDF) graphs. Increasingly, vocabularies are implemented SKOS for health and audiovisual applications,[34] for education and culture,[35] for Food and Agriculture (Agrovoc),[36] for activities of the European Union (Eurovoc),[37] for the environment (GEMET)[38] and for the economy (STW).[39]

Material and methods

Relationship extraction

Figure 1 shows the proposed method, structured in six steps. This method uses a medical resource containing expert terms we are willing to match with non-expert terms. We have chosen

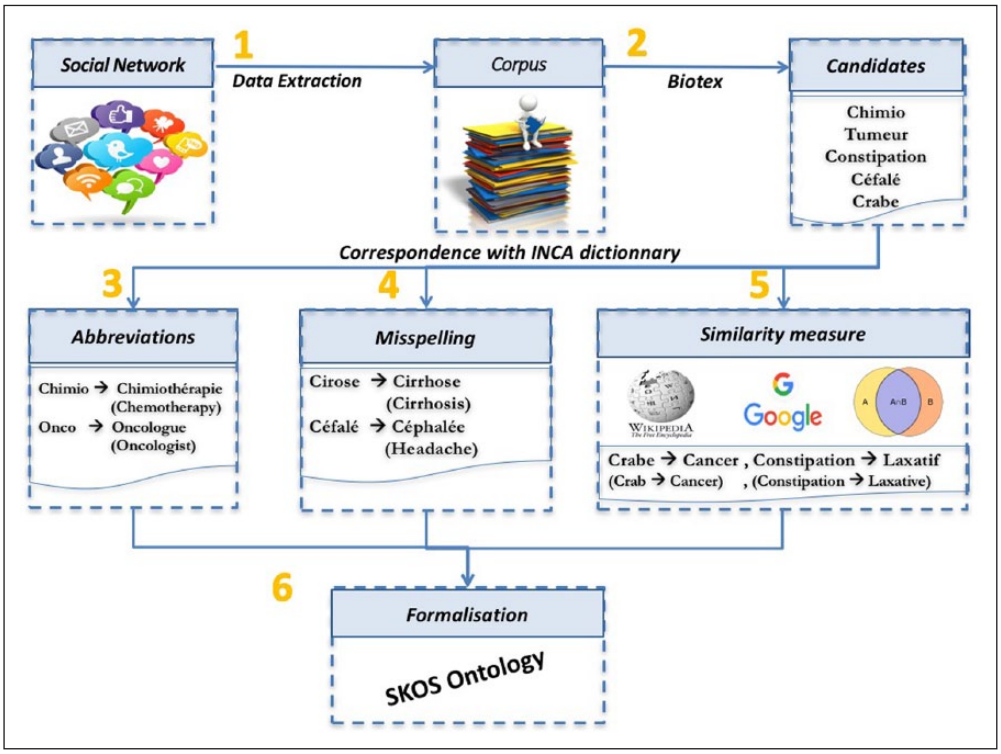

Figure 1. Semi-automatic building of a CHV in six steps.

as a reference resource denoted by the INCa dictionary the vocabulary given on the INCa[40] website composed of 1227 terms, all included in the French version of the MeSH.

Step 1: corpus constitution. We use the forum at 'cancerdusein.org' dealing with breast cancer. This forum facilitates sharing with other patients. Patients publish updates, photos or documents and send messages to all group members. The dataset contains 16,868 posts from 675 members, which have been collected between 2010 and 2014. Most members are the patients. Just a few posts belong to health professionals or relatives of the patients.

Step 2: candidate extraction. We sought terms from the corpus that have a high probability of belonging to the medical field using a BioTex tool.[41] BioTex implements state-of-the-art measurements for automatic extraction of biomedical terms from free text in English and French. In our study, more than 200 patterns have been used to identify the candidates. The candidates are filtered according to LIDF-value (Linguistic patterns, IDF and C-value information).[42] If BioTex has been trained for biomedical literature, our preliminary experiments show its efficiency on PATs because patients use similar constructions to experts with substitutions and misspelling. These expressions follow the same construction rules and are captured by the patterns (e.g. Noun–Adjective matches 'Echo mammaire' – Breast ultrasound in English). As an outcome of this step, we obtain a set $T = t_1, \ldots, t_N$ of N n-grams ($n \in [1, \ldots, 4]$) that are not listed in the INCa dictionary. We use them at steps 3, 4 and 5 as explained below.

Table 1. Equivalent between patient and medical terms (abbreviations and spelling errors).

Non-expert terms	Expert terms	Types of error
Chimio	Chimiothérapie (chemotherapy)	Abbreviation
Onco	Oncologue (oncologist)	Abbreviation
Mammo	Mammographie (mammography)	Abbreviation
Cyrose	Cirrhose (cirrhosis)	Spelling error
Abcé	Abcès (abscess)	Spelling error
Metastase	Métastase (metastasis)	Spelling error

Step 3: spelling correction. In the set of terms identified at step 2, we search for those corresponding to common spelling errors. We seek to match all the terms $t_i \in T$, with a correctly spelled term in the INCa dictionary. For this, we use the Aspell software[43] in order to obtain a set $M = \{m_1, m_2, \ldots, m_m\}$ of m propositions of corrections for the term t_i and we only keep the propositions found in the INCa dictionary. Levenshtein distance is used to compare the term t_i and each term $m_j \in M$. Only the terms whose distance to t_i is lower than or equal to 2 are pairing to t_i. Three additional conditions are necessary: (1) paired terms must begin with the same letter; (2) the length of matched terms is more than 3 characters and (3) the comparison is case-insensitive. If all these conditions are validated, the term t_i is paired with the term m_j with a $weight(m_j,t_i) = 1/|M|$. Examples of frequent spelling errors are listed in Table 1.

Step 4: abbreviations. Biomedical expressions are often truncated by patients. In the set of terms identified at step 2, we search for those corresponding to abbreviations. For this, we adapted the stemming algorithm Carry[44] using a list of the common suffixes used in the biomedical field (e.g. logie, logue, thérapie and thérapeute in French, respectively, logy, logue, therapy and therapist in English). For a term $t_i \in T$, we obtain a set $A = \{a_1, a_2, \ldots, a_k\}$ of k propositions of abbreviations included in the INCa dictionary. The term t_i is paired with each abbreviation $a_j \in A$ with a $weight(a_j, t_i) = 1/|A|$. Examples of abbreviations are listed in Table 1.

Step 5: similarity between two terms. We focus here on all terms produced at step 2 which are neither spelling errors (step 3) nor abbreviations (step 4). We try to match these terms with three approaches: (1) considering a semantically structured resource (Wikipedia), (2) considering the generalists co-occurrences on the web (Normalized Google similarity) and (3) considering the co-occurrences in messages from patients (Jaccard).

Wikipedia similarity. The hypothesis is to use the network of the links between pages of the Wikipedia resource. For this, we query this resource through its Application Programming Interface (API).[45] In this encyclopaedia, a referenced term is described by a page[46] and is linked to other terms which are described by other pages. The terms linked to the considered term are found on a dedicated page.[47] Given $W_t = (w_1, \ldots, w_n)$, the set of terms linked by Wikipedia to a term t_i that belongs to the INCa dictionary. A term t_i is associated to a term $w_j \in W$ according to the two measurements calculated using equation (1). Note that the set W solely contains terms present in the INCa dictionary

$$Wiki(w_1, w_2) = \frac{AvgNW(w_1, w_2)}{\sum_{k=1}^{|W|} AvgNW(w_k, w_2)} \quad (1)$$

$$AvgNW(w_1, w_2) = \frac{NW(w_1, w_2) + NW(w_2, w_1)}{2} \quad (2)$$

where $NW(w_{i,}, w_j)$ is the frequency of matching terms w_i on the Wikipedia page of the term w_j.

Wikipedia provides a knowledge base for computing word relatedness in a structured fashion and with more coverage than WordNet.[48] The intuition is the following one. Two close terms occur in linked pages.

Google similarity. The hypothesis is to exploit the co-occurrences in the documents indexed by the Google search engine. Google search is restricted to French language documents. We use the measurement proposed by Cilibrasi and Vitanyi[26] to compute the semantic similarity between two terms, based on the number of results returned by a Google query. This standard similarity is obtained as follows

$$NGD(w_1, w_2) = \frac{\max\{\log NG(w_1), \log NG(w_2)\} - \log NG(w_1, w_2)}{\log M - \min\{\log NG(w_1), \log NG(w_2)\}} \quad (3)$$

where $NG(w_i)$ is the number of 'hits' for the term w_i, $NG(w_i, w_j)$ is the number of 'hits' for the pair of terms w_i and w_j and M is the number of web pages indexed by Google.

The Normalized Google similarity between two terms is derived from Park et al.[49] This similarity is based on the number of results (hints) returned by Google search engine. The intuition is the following one. The more two terms are close, the more the two terms appear together in the documents returned by Google.

Jaccard similarity. The hypothesis is to exploit the co-occurrences not in the Web document as the previous measurements but in the text of the corpus produced by patients. If we consider all the messages of a patient, we often found common terms associated with expert terms. We use a formula similar to the Jaccard measurement to compute the similarity between w_1 and w_2

$$JAC(w_1, w_2) = \frac{NJ(w_1, w_2)}{NJ(w_1) + NJ(w_2) - NJ(w_2, w_1)} \quad (4)$$

where $NJ(w_i)$ is the number of occurrences of the term w_i in the corpus and $NJ(w_i, w_j)$ is the number of co-occurrences of the terms w_i and w_j.

Step 6: formalization in SKOS. We use the relationships obtained in steps 3, 4 and 5 to create an SKOS ontology. This ontology links an INCa term to different patient terms: preferential terms are used to define the MeSH term representing the expert term, alternative terms are used to represent abbreviations and hidden terms are used to represent spelling errors. An example of the obtained graph is presented in Figure 2.

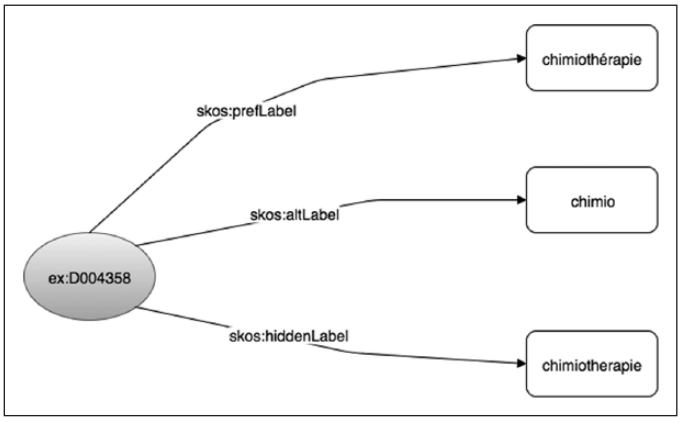

Figure 2. SKOS ontology extract for the term 'chimiothérapie' (chemotherapy). 'Chimiothérapie' is the expert term, 'chimiotherapie' is a spelling error and 'chimio' is an abbreviation.

Relationship validation

At the end of step 5, we have k relationships r_i with $i \in [1, k]$. Each relation r_i connects a common term pat_j (the term from the corpus) with an expert term bio_l (the term of the dictionary provided by the INCa). Each relationship is associated with a category $type \in \{spelling\ error, abbreviation, association\}$. In this section, we present two methods of validation (automatic and manual). The final manual validation is important to present weaknesses of the associations obtained with the quantitative methods.

Automatic validation. We automatically validate a relationship r_i defined by the pair pat_j–bio_l, if r_i exists in the dictionary of relations with a high confidence level provided by the contributory game – 'http://www.JeuxDeMots.org'.[50] This Game With A Purpose (GWAP) is used by Internet users not specialized in the medical field. The goal of this game is to build an extensive network of lexical semantic. The graph is composed of nodes. The nodes are linked by different types of relations including 179.578 occurrences of the synonymy relationship and medical vocabularies.[51] This tool has been used efficiently for disambiguation of medical terms and is a reliable source of gold data for medical vocabularies.[52]

Manual validation. All relationships r_i which could not be validated automatically were presented to five people, including an expert from the medical field. We propose relationships in the form 'pat_j – bio_l – type' in order to validate the association and its label. Two choices are proposed to annotators: (1) Yes: to validate the relationship and (2) No: to invalidate the relationship. We kept a relation if at least three annotators (including the medical expert) had validated it.

Global validation. As a result, we obtained a set of labelled relationships (pat_j – bio_l – type). Because we do not know in advance how many relationships exist for a common term pat_j, we cannot compute the recall. For this reason, we decided like Doing-Harris and Zeng-Treitler[11] to evaluate our results in terms of precision and used equation (5)

$$P = \frac{|R_a| + |R_m|}{|R|} \text{ and } R_a \bigcap R_m = \varnothing, \quad R_a \subseteq R, R_m \subseteq R, |R| \succeq |R_a| + |R_m| \tag{5}$$

Table 2. Examples of terms.

Non-expert terms	Expert terms	Relation	Validation
Chir	Chirurgie (surgery)	Abbreviation	Automatic
Chimio	Chimiothérapie (chemotherapy)	Abbreviation	Automatic
Hopital	Hôpital (hospital)	Spelling error	Automatic
Cheuveux	Cheveux (hair)	Spelling error	Automatic
Tumeur	Cancer (breast)	Association	Automatic
Chute des cheveux	Alopécie (alopecia)	Association	Automatic
Psy	Psychologue (psychologist)	Abbreviation	Manual
Onco	Oncologue (oncologist)	Abbreviation	Manual
Libido	Sexologie (sexology)	Association	Manual
Morphine	Douleur (pain)	Association	Manual

where R_a is the set of relationship automatically validated, R_m is the set of relationships manually validated and R is the set of relationships provided as output by our method.

Results

We evaluated our results in terms of precision on the $k = 1900$ first terms found at step 2; 96 relationships were automatically validated with 'JeuxDeMots'. Experts evaluated the remaining relationships. A kappa Fleiss coefficient *kf* was computed to measure the inter-annotator agreement. We obtained *kf* equal to 0.25. This low agreement is explained by the variability of judgement annotators on the medical value terms. The expert in breast cancer keeps only correct associations in relation with breast cancer and other experts keep all correct relationships even those relationships not linked to breast cancer (e.g. 'orbite – terre – association', 'orbit – hearth' in English). Consequently, we choose to keep only associations validated by three experts including the breast cancer expert. Examples of validated relationships are presented in Table 2. Statistics about the validation are presented in Figure 3. We discuss these results in section 'Discussion'.

For each common term, we keep in the resource the related term having the most important similarity. For example, with the Google similarity, for the non-expert term 'crabe', the related terms are (in French) zodiaque, cancer, tabou, hémorragie and biopsie (zodiac, cancer, taboo, haemorrhage and biopsy in English). We keep the closest term to 'crabe' which is listed in the INCa dictionary, here 'cancer'. Thus, we create the relationship 'crabe–cancer' which is included in the ontology. However, in the case where two relationships have the same weight, the common term may be linked to several expert terms (e.g. onco – oncology, onco – oncologist) and vice versa.

Figure 3 shows the number of validated relationships on our corpora for each measurement. For spelling errors, we obtain an overall precision *P* equal to 76 per cent. We have validated 22 relationships on the 29 obtained at step 3; 15 relationships were obtained by automatic validation and 7 by manual validation. For abbreviations, we obtain an overall precision *P* equal to 100 per cent. We have validated five relationships on the five obtained at step 4; three relationships were obtained by automatic validation and two by manual validation. For the Wikipedia similarity, we obtained an overall precision *P* equal to 88 per cent. We have validated 94 relationships on the 107 obtained; 78 relationships were obtained by automatic validation and 16 by manual validation. With the Google similarity, we obtained an overall precision *P* equal to 52 per cent. We have validated 151

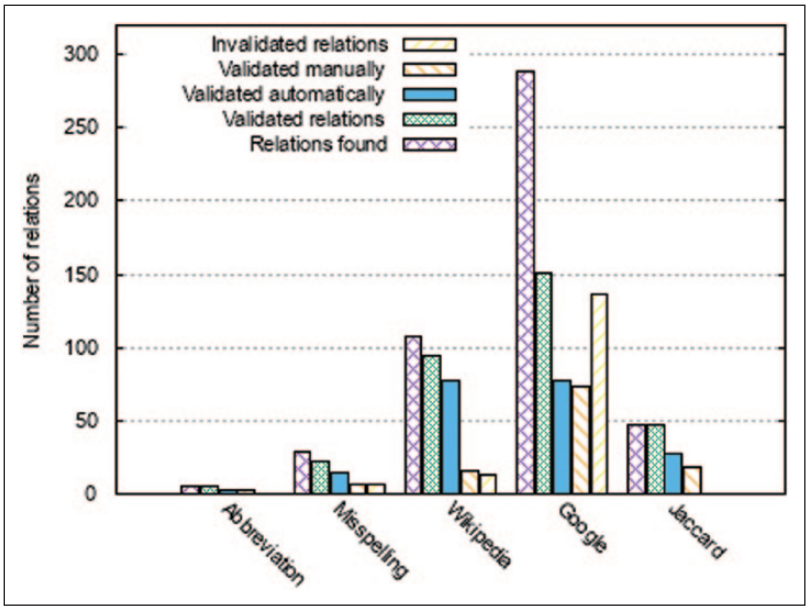

Figure 3. Number of relationships validated automatically, manually and those not validated at all.

relationships on the 288 obtained; 77 relationships were obtained by automatic validation and 74 by manual validation. With the Jaccard similarity, we obtained an overall precision *P* equal to 100 per cent. We validated the 47 relationships obtained; 28 relationships were obtained by automatic validation and 19 by manual validation.

Finally, considering all types of relationships, we obtain an overall precision *P* equal to 55 per cent. We have validated 192 relationships out of 346 that were obtained at step 5.

At step 5, we observed an overlap of the relationships obtained with the three similarities. The 47 relationships obtained by the Jaccard similarity are included in the set of relationships obtained by Google and Wikipedia similarity measure. We also found 80 relationships in common between Google and Wikipedia measurements. Excluding the duplicates among the relationships, we keep 165 relationships to be included in the SKOS ontology.

Discussion

Building a CHV can be done in several ways (manually or semi-automatically). In this study, we built a CHV semi-automatically for the breast cancer field in French. The proposed method allows connecting the terms used by patients with those used by health professionals. To build this vocabulary, we only used data extracted from forums and a connection to the Wikipedia API and to the web search engine Google.

We have compared three similarity measures based on different paradigms. We observe that the Wikipedia similarity provides little work for the expert to validate the candidate relationships. However, the Google similarity gives the highest number of validated relationships, but it requires additional effort for manual validation. With the Jaccard similarity, although the set of relationships found is very small and is included in the ones found with the other similarity measures, the relationships are all validated.

Considering the noisy nature of the biomedical textual data we used, the obtained results are very encouraging. Doing-Harris and Zeng-Treitler[11] have conducted a similar work. They built a general CHV in English. Out of 88.994 terms, they found 774 relationships and validated 237, thus a precision of 31 per cent. In our work, out of the 1900 terms, we found 346 relationships and validated 192, thus the global precision is 55 per cent.

One limitation is the number of matches issued from the initial resource. The INCa resource is composed of 1227 terms. Only 117 expert terms (that correspond to 10% of the initial resource) found corresponding common term with our method. This can be explained by the fact that we do not consider the 470 acronyms (that correspond to 38% of the initial resource). Moreover, we have projected the other 640 terms in the forums and have observed that these terms are frequently used by the patients in forum posts and therefore do not have specific substitutes. Finally, we have created a specific resource dedicated to the field of breast cancer and not a general resource.

A second limitation is the type of users, which produced the PAT exploited in this study. Indeed, unless a group has formal gatekeeping of members, it is difficult to know for sure whether people posting to a forum are patients, health care professionals, care providers, family or friends of patients. Consequently, terms extracted at step 2 may have been generated by users who are not suffering from breast cancer. In particular, it has been known for decades that health information seeking is done principally by friends or family members and then after that by patients.[53] In this work, we made the assumption that the vocabulary of relatives is similar to patients vocabulary and must be included in the CHV. However, in a previous work,[54] we have proposed a method to automatically deduce the role of forum user. This method can be used at the beginning of our chain to exclude the posts of health care professionals and care providers.

Finally, we only use 'cancerdusein.org' forum because French physicians and INCa recommend this forum to patients. However, there are certainly many other online communities related to breast cancer (e.g. in Facebook and Twitter), and of course, the community studied in this article is not necessarily representative of all patients suffering from breath cancer. In particular, we have used in previous work other social networks such as Facebook[55] to capture and compare the topics expressed by patient in forums and in Facebook. We note that most of the messages in forums focus on medical questions (e.g. secondary effects of treatment, non-medical treatment and observance). We retrieve these topics in Facebook, but the focus is also on encouragement and support request. It is important to note that our method can be easily applied to others corpus. As we do not pre-treat patient corpus, we only need to collect the different types of messages. This would allow us to know how patients express themselves in both types of social media.

Conclusion and perspectives

In this article, we present a method for linking the terms used by patients in social media to those used by health care professionals, which are present in controlled vocabularies. An advantage of this method is that aligned expressions can be composed of several terms. We only solicit experts for validation if the relationships are not retrieved using our gold standard.

We applied this method to the breast cancer field, but it can be applied to many other areas. For this, we need to replace the INCa list with a list of terms specific to the domain of the studied disease and to collect a new corpus of messages dealing with the disease.

We have also experimented with this method for French, but the method can be adapted to other languages with appropriate resources.

We compared three similarity measures to link common and expert terms. Such a resource will be an essential step in the automatic processing of social media content in the medical field.

The resource is now freely available to download for the community at the following address: http://bioportal.lirmm.fr/ontologies/MUEVO. We transformed this resource into human readable and machine-readable formats. To do this, we created an ontology in SKOS format to embed the platform BioPortal.[56]

In the long term, we plan to re-use data used to study the quality of life of patients with breast cancer and thus improve our processes similar to the one presented in previous work.[6] We could measure the impact of the resource, for example, on annotation[57,58] and classification tasks.[59] Similarly, we will apply our method to social media in English to extend existing CHV. We will also study the development of the vocabulary of patients over time, using a Latent Dirichlet Allocation (LDA) model.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and /or publication of this article: This work was supported by the ANR SIFR (Semantic Indexing of French Biomedical Data Resources) and by a grant from the French Public Health Research Institute (http://www.iresp.net) under the 2012 call for projects as part of the 2009–2013 Cancer Plan.

References

1. SNOMED CT, http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html (accessed 28 February 2017).
2. MeSH, http://mesh.inserm.fr/mesh/ (accessed 28 February 2017).
3. Unified Medical Language System (UMLS), http://www.nlm.nih.gov/research/umls/ (accessed 28 February 2017).
4. Pletneva N, Vargas A and Boyer C. *How do general public search online health information?* Health On the Net Foundation, http://www.hon.ch/Survey/khresmoi_general_public_survey_results.html (2011, accessed 28 February 2017).
5. Househ M, Borycki E and Kushniruk A. Empowering patients through social media: the benefits and challenges. *Health Informatics J* 2014; 20: 50–58.
6. Opitz T, Azé J, Bringay S, et al. Breast cancer and quality of life: medical information extraction from health forums. In: *Proceedings of the medical informatics Europe 2014*, Istanbul, 31 August–3 September 2014, pp. 1070–1074. IOS Press.
7. Zeng QT, Tse T, Divita G, et al. Term identification methods for consumer health vocabulary development. *J Med Internet Res* 2007; 9: e4.
8. Fiscella K, Meldrum S, Franks P, et al. Patient trust: is it related to patient-centered behavior of primary care physicians? *Med Care* 2004; 42: 1049–1055.
9. Wu DT, Hanauer DA, Mei Q, et al. Applying multiple methods to assess the readability of a large corpus of medical documents. In: *Proceedings of the world congress on health and biomedical informatics*, Copenhagen, 20–23 August 2013, pp. 647–651. IOS Press.
10. Ramesh BP, Houston TK, Brandt C, et al. Improving patients' electronic health record comprehension with NoteAid. In: *Proceedings of the world congress on health and biomedical informatics*, Copenhagen, 20–23 August 2013, pp. 714–718. IOS Press.
11. Doing-Harris KM and Zeng-Treitler Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. *J Med Internet Res* 2011; 13: e37.
12. Jiang L and Yang CC. Expanding consumer health vocabularies by learning consumer health expressions from online health social media. In: Agarwal N, Xu K and Osgood N (eds) *Social computing, behavioral-cultural modeling, and prediction*. Cham: Springer, 2015, pp. 314–320.

13. Jiang L, Yang CC and Li J. Discovering consumer health expressions from consumer-contributed content. In: Agarwal N, Xu K and Osgood N (eds) *Social computing, behavioral-cultural modeling, and prediction*. Cham: Springer, 2013, pp. 164–174.
14. Bouamor D, Llanos LC, Ligozat A-L, et al. Transfer-based learning-to-rank assessment of medical term technicality. In: *Proceedings of the 10th international conference on language resources and evaluation (LREC 2016)*, Portorož, 23–28 May 2016. European Language Resources Association (ELRA).
15. Keselman A, Smith CA, Divita G, et al. Consumer health concepts that do not map to the UMLS: where do they fit? *J Am Med Inform Assoc* 2008; 15: 496–505.
16. Patrick TB, Monga HK, Sievert MC, et al. Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. *J Med Internet Res* 2001; 3: E24.
17. Vydiswaran VV, Mei Q, Hanauer DA, et al. Mining consumer health vocabulary from community-generated text. *AMIA Annu Symp Proc* 2014; 2014: 1150–1159.
18. Consumer Health Vocabulary Initiative, http://library.ahima.org/doc?oid=67615#.Wk40XFSdU6g (accessed 2 January 2018).
19. MacLean DL and Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *J Am Med Inform Assoc* 2013; 20: 1120–1127.
20. Sadilek A, Kautz HA and Silenzio V. Modeling spread of disease from social interactions. In: *Proceedings of the international conference on weblogs and social media*, Dublin, 4 June 2012.
21. Pantel P, Crestan E, Borkovsky A, et al. Web-scale distributional similarity and entity set expansion. In: *Proceedings of the 2009 conference on empirical methods in natural language processing*, Singapore, 6–7 August 2009, vol. 2, pp. 938–947. Stroudsburg, PA: Association for Computational Linguistics.
22. Zadeh RB and Goel A. Dimension independent similarity computation. *J Mach Learn Res* 2013; 14: 1605–1626.
23. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945; 26: 297–302.
24. Lu K, Mao J and Li G. Enhancing subject metadata with automated weighting in the medical domain: a comparison of different measures. In: *Proceedings of the international conference on Asian digital libraries,* Seoul, South Korea, 9–12 December 2015, pp. 158–168. Cham: Springer.
25. Islam A, Milios EE and Keselj V. Comparing word relatedness measures based on Google n-grams. In: *Proceedings of the international conference on computational linguistics*, Mumbai, India, 8–15 December 2012, pp. 495–506. Association for Computational Linguistics (ACL).
26. Cilibrasi RL and Vitanyi P. The Google similarity distance. *IEEE T Knowl Data En* 2007; 19: 370–383.
27. Wikipedia, http://fr.wikipedia.org/wiki/Wikipédia:Accueil_principal (accessed 28 February 2017).
28. Hovy E, Navigli R and Ponzetto SP. Collaboratively built semi-structured content and Artificial Intelligence: the story so far. *Artif Intell* 2013; 194: 2–27.
29. Buscaldi D and Rosso P. Mining knowledge from Wikipedia for the question answering task. In: *Proceedings of the international conference on language resources and evaluation*, Genoa, 22–28 May 2006, pp. 727–730. European Language Resources Association (ELRA).
30. Wang P, Hu J, Zeng H-J, et al. Using Wikipedia knowledge to improve text classification. *Knowl Inf Syst* 2009; 19: 265–281.
31. Chernov S, Iofciu T, Nejdl W, et al. Extracting semantic relationships between Wikipedia categories. In: *Proceedings of the 1st international workshop: 'SemWiki2006 – from Wiki to semantics' (SemWiki 2006),* Budva, 12 June 2006, vol. 206, pp. 153–163. Berlin Heidelberg: Springer-Verlag.
32. Witten I and Milne D. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: *Proceedings of the AAAI workshop on Wikipedia and artificial intelligence: an evolving synergy,* Chicago, IL, 13–14 July 2008, pp. 25–30. Palo Alto, CA: AAAI Press.
33. Miles A and Bechhofer S. SKOS simple knowledge organization system reference. In: *W3C recommendation, World Wide Web consortium*, 2005, http://www.w3.org/TR/skos-reference/ (accessed 18 August 2009).
34. Van Assem M, Malaisé V, Miles A, et al. A method to convert thesauri to SKOS. In: *Proceedings of the European semantic web conference,* Budva, 11–14 June 2006, pp. 95–109. Berlin; Heidelberg: Springer.
35. Solomou G and Papatheodorou T. The use of SKOS vocabularies in digital repositories: the DSpace case. In: *Proceedings of the 2010 IEEE international conference on semantic computing,* Pittsburgh, PA, 22–24 September 2010, pp. 542–547. New York: IEEE.

36. Agricultural Information Management Standards (AIMS). AGROVOC Multilingual agricultural thesaurus, http://aims.fao.org/website/AGROVOC-Thesaurus/sub (accessed 28 February 2017).
37. Eurovoc, the EU's multilingual thesaurus, http://eurovoc.europa.eu/ (accessed 28 February 2017).
38. Themes list, http://www.eionet.europa.eu/gemet (accessed 28 February 2017).
39. STW thesaurus for economics: home, http://zbw.eu/stw/version/9.0 (accessed 28 February 2017).
40. Institut National Du Cancer. Dictionnaire des termes du cancer, http://www.e-cancer.fr/cancerinfo/ressources-utiles/dictionnaire/ (accessed 28 February 2017).
41. Lossio-Ventura JA, Jonquet C, Roche M, et al. BioTex: a system for biomedical terminology extraction, ranking, and validation. In: *Proceedings of the 2014 international conference on posters & demonstrations track,* 2014, vol. 1272, pp. 157–160, http://ceur-ws.org/Vol-1272/
42. Lossio-Ventura JA, Jonquet C, Roche M, et al. Integration of linguistic and web information to improve biomedical terminology extraction. In: *Proceedings of the 18th international database engineering & applications symposium,* Porto, 7–9 July 2014, pp. 265–269. New York: ACM.
43. GNU Aspell, http://aspell.net/ (accessed 28 February 2017).
44. Paternostre M, Francq P, Lamoral J, et al. Carry, un algorithme de désuffixation pour le français. Technical report, Paul Otlet Institute, Brussels, July 2002.
45. MediaWiki, http://fr.wikipedia.org/w/api.php? (accessed 28 February 2017).
46. Wikipédia, l'encyclopédie libre, http://fr.wikipedia.org/wiki/ (accessed 28 February 2017).
47. Pages liées – Wikipedia, http://fr.wikipedia.org/wiki/Spécial:Pages_liées/ (accessed 28 February 2017).
48. Strube M and Ponzetto SP. WikiRelate! Computing semantic relatedness using Wikipedia. In: *Proceedings of the 21st national conference on artificial intelligence (AAAI'06),* Boston, MA, 16–20 July 2006, pp. 1419–1424. Menlo Park, CA: AAAI Press.
49. Park J, Gao X and Andreae P. Query directed web page clustering using suffix tree and Wikipedia links. In: *Proceedings of the international conference on advanced data mining and applications,* Nanjing, China, 15–18 December 2012, pp. 91–99. Berlin; Heidelberg: Springer.
50. Lafourcade M and Joubert A. Increasing long tail in weighted lexical networks. In: *Cognitive aspects of the lexicon: proceedings of the international conference on computational linguistics,* Mumbai, India, 15 December 2012, pp. 5–20. Association for Computational Linguistics (ACL).
51. Lafourcade M. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In: *Proceedings of the 7th international symposium on natural language processing (SNLP'07),* Pattaya, Thailand, 13–15 December 2007, p. 7.
52. Lafourcade M and Ramadier L. Semantic relation extraction with semantic patterns experiment on radiology report. In: *Proceedings of the 10th international conference on language resources and evaluation (LREC 2016),* Portorož, 23–28 May 2016, pp. 4578–4582. Paris: European Language Resources Association (ELRA).
53. Zeng QT and Tse T. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc* 2006; 13: 24–29.
54. Abdaoui A, Azé J, Bringay S, et al. Analysis of forum posts written by patients and health professionals. In: *Proceedings of the medical informatics European (MIE),* Istanbul, 31 August–3 September 2014, p. 1185. IOS Press.
55. Tapi Nzali M, Bringay S, Lavergne C, et al. What patients can tell us: topic analysis for social media on breast cancer. *JMIR Med Inform* 2017; 5: e23.
56. Noy NF, Shah NH, Whetzel PL, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009; 37: W170–W173.
57. Hochheiser H, Ning Y, Hernandez A, et al. Using nonexperts for annotating pharmacokinetic drug-drug interaction mentions in product labeling: a feasibility study. *JMIR Res Protoc* 2016; 5: e40.
58. Gobbel GT, Garvin J, Reeves R, et al. Assisted annotation of medical free text using RapTAT. *J Am Med Inform Assoc* 2014; 21: 833–841.
59. Zhang S, Grave E, Sklar E, et al. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *J Biomed Inform* 2017; 69: 1–9.