E-patients in Oncology: a corpus-based characterization of medical terminology in an online cancer forum

Antonio Jesús Láinez Ramos-Bossini & Maribel Tercedor Sánchez
Servicio de Radiodiagnóstico, Hospital Universitario Virgen de las Nieves, Granada (Spain), Facultad de Traducción e Interpretación, Universidad de Granada (Spain)
asb_jbossini@ugr.es, itercedo@ugr.es

Abstract

This study aimed to characterize medical terms in an online cancer forum, with particular focus on specialization and semantic features. A three-step analysis was carried out on a 60-million-word corpus to detect and characterize the most typical medical terms used in a cancer forum by means of (1) keywords contrastive, (2) co-text-guided, and (3) semantic analyses. More than half of the 1000 words analysed were medical terms according to the co-text-guided analysis carried out. Most of them (73%) were dictionary-defined medical terms, followed by co-text-defined terms (9%) and medical initialisms (8.5%). The semantic analysis showed a higher number of terms within the fields of Anatomy, Treatment, Hospital and Symptoms. Our findings suggest that medical terms are commonly used in cancer forums, especially to share e-patients’ concerns about treatment, symptoms and hospital environment. The method followed is efficient and could be applied in future studies. Altogether, this article contributes to characterizing medical terms used by e-patients in online cancer forums.

Keywords: medical terminology; online cancer forums; corpus linguistics; keywords contrastive analysis; semantic analysis.

Resumen

Los e-pacientes en Oncología: caracterización basada en corpus de la terminología médica en un foro en línea sobre cáncer

El objetivo de este estudio fue caracterizar los términos médicos de un foro en
Ibérica 39 (2020) : 69-96

ANTONIO JESÚS LÁINEZ RAMOS-BOSSINI & MARIBEL TERCEDOR SÁNCHEZ

línnea sobre cáncer, prestando especial atención a los rasgos semánticos y de especialización. Con el fin de detectar y caracterizar los términos médicos más típicos utilizados en un foro de cáncer, se llevó a cabo un análisis de un corpus de 60 millones de palabras en tres etapas: (a) contraste de palabras clave, (b) estudio del cotexto y (c) análisis semántico. Más de la mitad de las 1000 palabras analizadas eran términos médicos de acuerdo con el análisis guiado por el contexto. La mayor parte de ellas (el 73%) se corresponden con “términos médicos definidos en el diccionario”, seguidos de “términos definidos por el cotexto” (9%) y de “siglas y abreviaturas médicas” (8,5%). En el análisis semántico se encontró un mayor número de términos en los campos de Anatomía, Tratamiento, Hospital y Síntomas. Nuestros hallazgos sugieren que los términos médicos son frecuentemente utilizados en los foros oncológicos, sobre todo para compartir las preocupaciones de los e-pacientes sobre el tratamiento, los síntomas y el medio hospitalario. El método seguido es eficiente y podría aplicarse en estudios futuros. En definitiva, este artículo contribuye a la caracterización de los términos médicos utilizados por los e-pacientes en foros en línea sobre cáncer.

**Palabras clave:** terminología médica; foros en línea sobre cáncer; lingüística de corpus; análisis contrastivo de palabras clave; análisis semántico.

1. Introduction

Traditionally, most of the studies carried out in the context of physician-patient and patient-patient communication have been based on a model of patient with low health literacy and poor medical (terminological) knowledge. According to this model, patients are unable to understand many medical terms (LeBlanc, Hesson, Williams, Feudtner, Holmes-Rovner, Williamson, & Ubel, 2014) leading to a breach in physician-patient communication. This gap is accentuated by the fact that lay health expressions may be difficult to understand (Tse & Soergel, 2003). However, recent studies suggest a shift towards a new model of patient, termed “e-patient” (Ferguson, 2007), with medium-to-high level of health literacy as evidenced by the use of medical terms, resulting from permanent interactions with health information available online (Fage-Butler & Nisbeth Jensen, 2016). In fact, individuals with high health literacy are more likely to use social media platforms to obtain health-related information than those with low health literacy (Kim & Xie, 2017).

Patients’ terminological use has been explored in different online settings, including webpage user queries in medical websites (Mccray, Loane, Browne,
consultation e-mails to cancer information services (Smith, Stavri, & Chapman, 2002) and, most recently, online discussion groups (Fage-Butler & Nisbeth Jensen, 2016; Harvey, Brown, Crawford, Macfarlane & McPherson, 2007; Seale, Charteris-Black, MacFarlane & McPherson, 2010; Seale, Ziebland, & Charteris-Black, 2006), and other social media (Naslund, Aschbrenner, Marsch, & Bartels, 2016). Moreover, several approaches have been adopted to characterize patients’ terminology, such as written questionnaires (Davis et al., 1991), online surveys (Thomas, Walker, Leighton, Yong & Batchelor, 2014), oral interviews or corpus-based analyses. The first studies in this field soon revealed the specific nature of patients’ terminology (Mccray et al., 1999), but further research is still needed to delimit the extent to which e-patients’ medical knowledge should be assumed.

The study of health literacy by means of terminological use may be especially relevant in the context of cancer, given the social, medical, and psychological implications of this group of diseases. The social concern about cancer is well known, and it is justified by determining epidemiological facts, such as the increase in cancer incidence worldwide (Forman, Bray, Brewster, Gombe Mbalawa, Kohler, Piñeros, Steliarova-Foucher, Swaminathan & Ferlay, 2014) or the high mortality rates (8.8 million deaths in 2015) (WHO, n.d.). Furthermore, cancer treatment has significantly improved in the last decades (Miller, Siegel, Lin, Mariotto, Kramer, Rowland, Stein, Alteri & Jemal, 2016), increasing survival rates, quality of life and other prognostic indicators. From a psychological perspective, the study of cancer patients’ communication may contribute to understanding the cognitive processes elicited by illness experience, as being diagnosed with cancer has been suggested as a “psychic trauma” that triggers hidden fears and emotions (Lanceley & Clark, 2013). In sum, a variety of dimensions concerning patients’ illness experience, including worries, concerns, and information validity, can be explored through patients’ linguistic cues.

Corpus linguistics has become one of the most extended methodologies to characterize language use, and has proved useful in health settings given the potential of combining quantitative and qualitative methods (Harvey et al., 2007). Nevertheless, this approach raises practical difficulties in sociolinguistics, as it is usually based on written texts (Andersen, 2010), and assumptions such as the correspondence between familiarity and frequency of use deserve further consideration (Alarcón-Navío, López-Rodríguez & Tercedor-Sánchez, 2016; Sánchez, Rodríguez & Velasco, 2014). This
limitation could be overcome in online discussion forums, where the spontaneity typical of oral communication pervades written formats.

This article presents a corpus-based study of the medical terminology used in an online forum about cancer. Within this context, three specific analyses are carried out in this study to respectively answer the following research questions:

1. Which medical terms are typical of the cancer forum studied?
2. What specialization features do such medical terms have? and
3. What is their semantic nature?

2. Methods

A corpus\(^2\) of more than 60 million words was compiled from two online forums: one about cancer and another one about general topics. The cancer forum subcorpus contains posts from the web “The Cancer Forums” (www.cancerforums.net), and has a total of 31,545,923 words. The generic forum subcorpus contains posts from several sections (e.g. news, music, sports, etc.) from the web “Voat” (www.voat.co), and has a total of 29,383,459 words. Both forums were selected following a search in Google that was guided by the aim of optimizing corpus representativeness in terms of number of users registered, active topics, spam control and moderation. No general guidelines to address ethical issues concerning social media usage in forums are available currently (Denecke, Bamidis, Bond, Gabarron, Househ, Lau & Mayer, 2015). However, the content of the forums was publicly accessible and users’ privacy was preserved throughout the whole study.

The compilation process involved three steps. Firstly, the content was mass-downloaded using an open-source software called Scrapy (available at https://scrapy.org/). Secondly, the content was refined, excluding unwanted elements (e.g. signatures, emoticons, redundant posts). Thirdly, the content was divided into files of appropriate format (plain text) and size (20 MB) to facilitate its processing using WordSmith tools® (Scott, 2012). This lexical analysis software was preferred over other similar tools (e.g. term extractors such as TermoStat or BioTex) because our purpose was to analyse both technical and non-technical words and because WordSmith tools® has been used in similar settings (examples provided below).
To the best of our knowledge, this is one of the largest corpora compiled to research online discussion forums about cancer. Given the massive amount of data and the absence of user-type labels (e.g. *patient, caregiver, health care professional*), no classification according to the type of user was carried out. However, a general overview of the cancer forum posts makes it feasible to assume that the majority of posts were sent by patients and caregivers.

The procedure for the retrieval and classification of the most representative terms in the cancer forum relied on a combination of automatic techniques and manual analysis. The whole process comprised three consecutive steps (Figure 1):

1. **Keywords contrastive analysis.** A contrastive analysis of keywords was applied using WordSmith tools® version 6 (Scott, 2012). This programme offers three basic functions. Firstly, it allows us to obtain *wordlists* from a given corpus where words are sorted by frequency. Secondly, it is possible to generate *keyword lists* by contrasting two wordlists; keywords represent outstandingly frequent words from a source wordlist, as compared to a reference wordlist¹. Thirdly, the programme offers an option to analyse words in their context by means of *concordances*.

This software has been already used to analyse patients’ terminology in different settings (Seale, Boden, Williams, Lowe, & Steinberg, 2007), including health care (Adolphs, Brown, Carter, Crawford & Sahota, 2004) and, specifically, the domain of cancer (Seale et al., 2006; Taylor, Thorne, &

---

¹ Reference wordlist: A wordlist that serves as a comparison point to identify words that are unique to the source wordlist.
Oliffe, 2015). A wordlist from each subcorpus was obtained, and a keyword list of 2569 words was automatically generated using the wordlist from the cancer forum as *source wordlist*, and the wordlist from the generic forum as *reference wordlist*. For efficiency reasons, only the 1000 words with the highest keyness value from the keyword list were considered for the analysis. A rough examination of the keyword list showed several terms belonging in the field of Medicine, warranting a more detailed analysis as follows.

2. **Co-text-guided analysis.** In a recent study (Fage-Butler & Nisbeth Jensen, 2016), Fage-Butler and Nisbeth-Jensen explored the use of medical terms in an online forum through a novel analysis that is worth replicating. To explore the specialization nature of the medical terms found in our keyword list, we classified them following the categories proposed by Fage-Butler and Jensen’s approach (henceforth referred to as “co-text-guided analysis”) with minor modifications: dictionary-defined medical terms, co-text-defined medical terms, medical initialisms, drug brand names, and colloquial technical terms. The remaining words were considered as “non-(medical) terms”. The classification of words was aided by the use of *concordances* in order to take into account their co-text. The limit in the number of n-grams analysed was set to trigrams by default. The criteria chosen for each category were the following:

- **Dictionary-defined medical terms**: Terms defined in the *Stedman’s Medical Dictionary* (SMD) online. The choice of this medical dictionary was made on the basis of the large amount of medical terms included (above 100,000), its online accessibility, and its prior use in the study on which the co-text-guided analysis is based. Examples of these terms are *abdomen, dysplasia* and *sphincter*.

- **Co-text-defined medical terms**: Terms that were not defined in the SMD, but with a specialized meaning acquired in context as shown in the analysis of concordances, e.g. *caregiver, surgical removal, undetectable PSA*.

- **Medical initialisms**: Terms described in medical abbreviation lexicons or explained in the surrounding co-text, such as *BCC* (Basal Cell Carcinoma), *NSCLC* (Non-Squamous Cell Lung Cancer) or *RO* (Radiation Oncologist).

- **Drug (brand) names**: Terms referring to either drug active ingredients, drug types or drug brand names, such as *Adriamycin, Erlotinib* or *Valium®*.
• **Colloquial technical terms.** Terms shortened that are not defined in usual medical resources, such as *Derm* (Dermatologist), *Neuro* (Neurologist), or *Dx* (Diagnosis).

3. **Semantic analysis.** To characterize some semantic aspects of the medical terms from the keyword list, each term was classified by the two authors independently in one out of ten medical domains. These were chosen on the basis of similar classifications that have been carried out elsewhere (i.e. MeSH®, OncoTerm®, SNOMED CT) (Faber, 2002; Kostick, 2012). The choice of domains followed a review of the keyword list in order to select proper categories, avoiding excessive specificity. Cohen’s kappa test was run to assess the agreement between the two authors. Full agreement was achieved by consulting an independent medical expert that decided on a final category in case of discrepancy. The domains chosen are described in Table 1.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>Medical procedures, substances and parameters related to diagnosis</td>
<td>Biopsy, MRI, enema, bloodwork, blood pressure</td>
</tr>
<tr>
<td>Anatomy</td>
<td>Terms related to parts and constitutive elements of the body</td>
<td>Artery, gall bladder, endometrial, neurovascular bundle, pineal gland</td>
</tr>
<tr>
<td>Hospital</td>
<td>Places, processes and agents related to physician-patient interactions</td>
<td>Hospital, GP (General practitioner), patient</td>
</tr>
<tr>
<td>Treatment</td>
<td>Therapeutic procedures or instruments, excluding drugs</td>
<td>Androgen deprivation therapy, chemotherapy, Whipple’s procedure</td>
</tr>
<tr>
<td>Symptoms</td>
<td>Signs and symptoms, including treatment side effects</td>
<td>Anxiety, incontinence, jaundice, seizure</td>
</tr>
<tr>
<td>Disease</td>
<td>Names of diseases and syndromes</td>
<td>Astrocytoma, cancer, leukemia, stroke, small cell lung cancer</td>
</tr>
<tr>
<td>Physiological entity</td>
<td>Physiological processes and substances</td>
<td>Apoptosis, motor function, testosterone, urine</td>
</tr>
<tr>
<td>Disease description</td>
<td>Descriptions of the nature and behaviour of diseases</td>
<td>Biology [of the tumor], chemoresponse, relapse, primary tumor</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>Terms used in Epidemiology and clinical research</td>
<td>Nadir, statistics, prevention, prognosis, follow-up, [clinical] trials</td>
</tr>
<tr>
<td>Drugs</td>
<td>Drug brand names, active ingredients, metabolites, chemotherapy regimens</td>
<td>Aprepitant, Bimix, Gemcitabine, Cannabinoid, CHOP</td>
</tr>
</tbody>
</table>

Table 1. Domains chosen for the semantic analysis.
3. Results

Figure 2 summarizes the results of each step of the analysis. All the words analysed, including both the co-text-guided and the semantic analyses, can be consulted in the Appendices.

![Graph showing the distribution of terms and non-terms from the keyword list; categories in the co-text-guided analysis; and categories in the semantic analysis.]

Figure 2. Global results. Distribution of (1) terms and non-terms from the keyword list; (2) categories in the co-text-guided analysis; (3) categories in the semantic analysis.

3.1. Co-text-guided analysis

The co-text-guided analysis resulted in a total of 503 medical terms (50.3%), most of which were dictionary-defined terms (367 terms). This group was followed far behind by the rest of categories, which presented a homogeneous distribution, including 46 co-text-defined terms, 43 medical initialisms, 29 drug (brand) names and 18 colloquial technical terms. The results are summarized in Figure 3.
3.2. Semantic analysis

Cohen’s Kappa test resulted in $\kappa = 0.859$ (95% CI, 0.826 to 0.891), indicating almost perfect agreement between raters (Hallgren, 2012). As shown in figure 4, the results show that the categories with a higher number of terms in the semantic analysis are TREATMENT (82 terms), ANATOMY (70 terms), SYMPTOMS (66 terms) and HOSPITAL (57 terms), whilst the categories with a lower number of terms are PHYSIOLOGICAL ENTITY (31 terms) and EPIDEMIOLOGY (25 terms).
4. Discussion

In this article we have presented a corpus-based study aimed to analyse several terminological features of the most salient medical terms of an online cancer forum. To do so, a contrastive analysis of keywords has been carried out to select the most outstanding words from a cancer forum. By applying two subsequent qualitative analyses, we have explored the specialization and semantic features of several medical terms. Given the intrinsic difference between the analyses carried out, they will be discussed separately.
Keywords contrastive analysis

Keywords contrastive analysis allows us to easily spot words that are typical of a given milieu throughout millions of words, which is a useful way to characterize a genre (Scott, 2010b). In fact, our keyword list revealed very specific lexical units of the forum, as revealed by the illustrative case of some usernames – which obviously did not appear in the reference corpus. Despite the revealing results obtained, the validity of keywords analysis may be undermined without co-text analysis, which is particularly relevant in a cancer setting (Taylor et al., 2015). In this regard, the addition of a co-text-guided analysis lessens the restraints of examining words in isolation from their co-text, improving term detection. In sum, operationalizing term detection with keywords contrastive analysis is efficient to explore terminological patterns, but the sole consideration of word frequencies constitutes a limitation that needs to be complemented with qualitative analyses.

Co-text guided analysis

The co-text-guided analysis resulted in a large number of medical terms in the keyword list, which suggests that e-patients use medical jargon frequently in the cancer forum studied. In fact, the large amount of dictionary-defined terms detected may be seen as an evidence of high-health literacy as, according to Nation (2001) (cited in (Fage-Butler & Nisbeth Jensen, 2016)), as such terms are considered “the most technical kind of words”. This interpretation seems unquestionable in the case of terms such as adenocarcinoma, dysplasia or neuropathy. However, it must be noted that (relative) frequency of use does not necessarily correlate with familiarity or with specialized knowledge (Alarcón-Navío et al., 2016). Besides, considering dictionary-defined medical terms as “the most technical kind of words” solely on the ground that they are defined in a medical dictionary seems controversial as many words included in a medical dictionary do not present specialized connotations when used in general communication. On the whole, although not all dictionary-defined medical terms found in our corpus are invariably used in a medical sense, our findings are consistent with the hypothesis that cancer forum e-patients are familiar with and repeatedly use medical terms, corroborating the findings of Fage-Butler and Nisbeth Jensen on a thyroid forum (Fage-Butler & Nisbeth Jensen, 2016).

Co-text-defined terms constitute the second most frequent category of the analysis, although far behind dictionary-defined terms. This result reveals that e-
patients give a specialized meaning to everyday words and emphasizes the importance of co-text analysis to improve term detection. However, the subjectivity involved in this category is patent as not all researchers would consider lexical units like *complementary medicine*, *caregiver* or *dietary supplements* as medical terms. The examination of concordances showed that the terms included in this category were consistently used within a medical context, implying field-specificity and, thus, at least some degree of specialized knowledge. The problem lies in the fact that, apparently, there is no obvious connection between co-text-defined terms and their degree of specialization. Nonetheless, a closer look at the definition of co-text-defined terms as “terms which overlap expert and general language” (Fage-Butler & Nisbeth Jensen, 2016) reveals a clear resemblance with the notion of *sub-technical term* stated by Baker: “items which are neither highly technical and specific to a certain field of knowledge nor obviously general” (Baker, 1988).

The similarity between the previous definitions is explained by the existence of two complementary approaches of term specialization. The former aims to directly measure the “degree of specialization” of terms, and it has been traditionally carried out through specific analyses that usually sort terms along a “quantitative” continuum of specialization (e.g. technical, semi-technical and non-technical (Chung & Nation, 2003)). The latter, pioneered by Fage-Butler and Nisbeth Jensen’s model, privileges other qualitative parameters (e.g. grammar, semantics or lexicography), but it is based on the assumption that terms are actually (highly) specialized, which is not necessarily true for categories such as *dictionary-defined* or *co-text-guided terms*. Ideally, both approaches should be carried out together to minimize subjectivity, but even very quantitative perspectives of term specialization lack consensus among researchers (Marín Pérez, 2016). The fact that e-patients use co-text-defined terms frequently adds evidence on their presumable acquaintance with terminological knowledge, but the degree of specialization entailed by these terms should be clarified in future studies.

The number of *medical initialisms* found in the keyword list has significant implications provided that initialisms constitute a form of metonymy, a “process in which a vehicle provides access to a target” (Kövecses & Radden, 1998), that serves as a shortcut to their referents (i.e. specialized medical concepts). As the tendency to shorten multi-word terms increases when the speakers use them often, the connection between frequency of use, familiarity and specialized knowledge becomes neat, explaining why medical experts make use of acronyms and abbreviations frequently
In our keyword list, we have found more than 40 acronyms that clearly have a specialized nature (e.g. BCC [Basal Cell Carcinoma], EBRT [External Beam Radiation Therapy]), which can be regarded as a proof of familiarity with medical jargon, especially considering that many of them correspond to chemotherapy regimens (e.g. CHOP, FOLFIRINOX). Nevertheless, it is important to highlight that acronyms are especially frequent in English (Plag, 2003), and therefore the high prevalence of this type of terms should not be assumed in forums in other languages.

*Drug (brand) names* have been considered as specific-in-nature in previous studies exploring medical terminology (Fage-Butler & Nisbeth Jensen, 2016; Koch-Weser, Rudd, & Dejong, 2010), which would justify an association between their frequency of use and terminological familiarity. In line with this premise, the amount of drug names found in the keyword list can be interpreted as a proof of familiarity with the field of oncology given that many of them correspond to chemotherapy agents, similarly to the case of initialisms. Besides, we found several pairings of drug brand name-active ingredient (e.g. Xeloda®-Capecitabine, Tarceva®-Erlotinib, Avastin®-Bevacizumab), which reinforces the notion of familiarity with cancer medical terms, as even medical experts may not be familiar with brand names that are very specific to the oncology domain. We suggest it would be useful to further explore the terminological role of such pairings as indicators of familiarity.

*Colloquial technical terms* constitute the least frequent category of the co-text-guided analysis. According to Fage-Butler & Nisbeth Jensen (Fage-Butler & Nisbeth Jensen, 2016), these terms represent “not only familiarity with colloquial medical language but also the assumption of shared knowledge”. Interestingly, many terms included in this group refer to medical staff (e.g. doc, onc, endo). This may be explained by close links and familiarity with health care professionals on the ground that they are humans rather than abstract entities. As with the previous categories, colloquial technical terms show significant specialization cues that contribute to gain insight into cancer e-patients’ health literacy, in line with Fage-Butler and Nisbeth-Jensen’s findings on a thyroid forum.

The co-text-guided analysis presents some limitations. The main practical drawback in terms of efficiency is the need for co-text consultation, which is compensated with a keywords contrastive analysis that facilitates the selection of medical terms. Another handicap concerns the inaccuracy when
measuring the degree of specialization of medical terms, which could be managed with complementary analyses in future studies. As concluding remarks, this analysis, which basically reproduces the approach pioneered by Fage-Butler and Jensen, proved its usefulness to depict essential information concerning the specialized nature of medical terms used in a cancer forum.

Semantic analysis

The category TREATMENT accounts for the highest number of terms within this analysis. A qualitative overview of forum posts showed that patients are likely to express their doubts about treatment issues. In most cases, they ask other patients, who feedback their posts providing information and emotional support. Consequently, e-patients acquire knowledge in a context of power balance between peers, as opposed to the context of power asymmetry between physicians and patients. This way, promoting patient-patient online communication may serve researchers as a potential resource to improve shared decision making (Joseph-Williams, Elwyn, & Edwards, 2014). Besides, several medical terms included in this group show a high degree of technicality (e.g. hysterectomy, immunotherapy, resection), which may be explained by patients’ familiarity with this field. In fact, we have found many examples in the forum indicating that patients are acquainted with very specialized terms related to therapy (e.g. “Sounds like the next step is cyberknife”, “I had 3 rounds of CHOP chemo”). On the whole, these outcomes have significant implications because few works have aimed to study cancer patients’ worries regarding treatment (Martin, Fouad, Oster, Schrag, Urmie & Sanders, 2014).

The salient amount of terms related to the field of ANATOMY is another remarkable finding. Several studies have found conceptual and terminological knowledge deficits in different medical domains for a long time (Boyle, 1970). More recent studies reaffirmed such lack of knowledge, both in the general public (Chapman, Abraham, Jenkins, & Fallowfield, 2003; Pieterse, Jager, Smets, & Henselmans, 2012) and in patients (Weinman, Yusuf, Berks, Rayner, & Petrie, 2009). Conversely, our findings suggest that patients’ knowledge of anatomical terms might be greater than expected, which is consistent with other studies (Seale et al., 2006; Tercedor Sánchez & Láinez Ramos-Bossini, 2017). However, it must be pointed out that several anatomical terms found in our analysis may be considered to have a low or medium degree of specialization (e.g. chest, heart, liver), with remarkable exceptions, such as endometrial, peritoneal or thoracic. The latter
usually appear in the context of sharing the results of imaging exams (e.g. *CT scan reports*), suggesting a close link between specialized anatomical terms and the domain of Diagnosis, and also that forum users learn directly from specialized medical reports. In conclusion, we propose to further explore the hypothesis that cancer forum users possess at least moderate anatomical knowledge.

Regarding the category SYMPTOMS, it is not surprising to find a high number of terms given the impact of cancer and its treatment on patients’ lives. Beyond the physical signs (e.g. *bleeding, cough, spasm*), experiencing cancer as a life-threatening condition may cause permanent psychological distress. In fact, it has been associated with mental health problems such as anxiety and depression (Benioudakis et al., 2016). In this regard, it is not surprising to find medical terms related to emotions in this category, whether referring to psychological entities connected with emotion (e.g. *anxiety, depressed, stress*), or as symptoms that are typically affected by the emotional status (e.g. *tired, appetite, discomfort*). This finding is consistent with the richness of emotion expressions found in other studies on cancer patients (Lanceley & Clark, 2013), and in e-patients in online forums (Tercedor Sánchez & Láinez Ramos-Bossini, 2017).

Finally, HOSPITAL could be to some extent regarded as a miscellaneous category, as it includes two subcategories that could be considered independent: agents implied in doctor-patient relationships (e.g. *GP, oncologist, patient*), and places where doctors and patients interact (hospital, clinic, cancer centre). Interestingly, the former group accounts for many generic terms alluding to physicians (e.g. *docs, doctors, MD*), as well as terms referring to specialists (e.g. *oncs, pulmonologist, urologist*), reflecting the large amount of health professionals and specialists that are involved in the care of cancer patients (Brar, Hong, & Wright, 2014). As discussed previously, the presence of colloquialized terms referring to medical staff -revealing a high degree of familiarity- is consistent with the high frequency of these terms.

The semantic analysis presents a drawback in terms of consistency, which is caused by the dynamicity of conceptual activation, that is, the same term may elicit different semantic features of a concept depending on co-textual and cognitive factors (Rogers, 2004; Tercedor, 2011). This variability calls for subjective assessment, which may give rise to discrepancies between researchers. This problem was tackled by the examination of concordances to take co-textual factors into account, together with an inter-rater reliability.
test (i.e. Cohen’s kappa test) to secure a high degree of consistency. However, the high value obtained for this coefficient might be biased to a certain degree since both researchers were familiarised with the criteria used for term classification and this and similar corpora (Tercedor Sánchez & Láinez Ramos-Bossini, 2017, 2019 and 2020). In all, the semantic information provided by this approach is of undeniable help to gain insight into patients’ use of medical terminology.

**Other limitations of the study**

Apart from the limitations of each approach, it should be noted that this study was carried out by non-native English speakers, which could bias the classification of some terms. Besides, despite the familiarity of the authors with the medical field, none of them were specialists in Oncology. Also, the hypothesis that most forum users were patients and caregivers was based on a qualitative overview of the forum, but absence of quantitative data supporting this supposition (e.g. use of user-type labels) limits the extent of the assumptions made regarding the terminological usage of patients. Finally, the experimentation on a single cancer forum could lead to a lack of representativeness, despite the effort made during the selection phase of the study. Similarly, the use of one single dictionary, despite being one of the largest sources of information in the medical field, constitutes a limitation. In sum, in future studies more lexicographic resources as well as other online settings and social media shall be explored to complement the results of the present study, including different medical specialties as well.

**5. Conclusions**

Important aspects of our research questions have been elucidated. The use of medical terms is representative of the communication that takes place in the cancer forum studied, as evidenced by the fact that more than half of the keyword lists obtained are medical terms. Most of them are defined in a medical dictionary, whilst co-text-defined terms, medical initialisms, drug brand names and colloquial technical medical terms show a similar distribution, revealing interesting specialization facts. The semantic analysis showed a higher number of medical terms within the domains ANATOMY, TREATMENT, SYMPTOMS and HOSPITAL, which may indicate that e-patients are familiar with and concerned about topics related to these fields.
Our findings have significant implications to understanding and delimiting the model of e-patient in a cancer setting, although other social media and medical specialties shall be explored to complement the results of the present study.

The characterization of medical terms used in online forums provides relevant information about e-patients’ health literacy. In line with previous studies, our results are consistent with the hypothesis that e-patients are familiar with medical terms in the domain of cancer, indicating potential high health literacy. Given its efficiency and comprehensiveness, the methodology followed might be applied in future research in this area as it allows to (1) easily spot medical terms typical of a target corpus; (2) describe their specialization features; and (3) classify them according to different medical domains adapted to particular settings.

Altogether, this study sheds light on specialization and conceptual features of the most representative medical terms used in online cancer forums, and optimizes a method that could be easily replicated in similar studies.

Acknowledgements

We thank David Jiménez Sequero for his assistance to download and filter the messages analysed in the study.

References


Antonio Jesús Láinez Ramos-Bossini holds a degree in Translation and Interpreting (English) and a degree in Medicine from the University of Granada (UGR). He is a radiologist trainee at the Hospital Universitario
Virgen de las Nieves (Granada, Spain), and collaborates with the Faculty of Translation and Interpreting of the UGR.

**Maribel Tercedor Sánchez** is full professor in Translation at the Faculty of Translation and Interpreting of the University of Granada, where she teaches multimedia and scientific and technical translation. Her main research interests are in the fields of lexical and cognitive aspects of scientific translation, terminology and accessibility in translation with the Faculty of Translation and Interpreting of the UGR.

**NOTES**

1 Keywords are calculated by a parameter that measures how outstanding a word is in the keyword list, called “keyness value”. This parameter was calculated on the basis of a log-likelihood test applied for each word in the source and reference wordlists compared (Scott, 2010).

2 The term “corpus” is used here to refer to the whole content analysed in the study, but it must be noted that each online forum was treated independently throughout all analytical procedures. For consistency purposes, we refer to the whole content from each forum as “cancer/generic forum subcorpus”, but they could be considered as independent corpus as well.

**Appendices**

Appendix A: List of the 1000 lexical units analysed.

<table>
<thead>
<tr>
<th>ABDOMEN</th>
<th>DOCS</th>
<th>MALIGNANT</th>
<th>SCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABDOMINAL</td>
<td>DOCTOR</td>
<td>MANY</td>
<td>SCHEDULED</td>
</tr>
<tr>
<td>ABLE</td>
<td>DOCTORS</td>
<td>MARCH</td>
<td>SECOND</td>
</tr>
<tr>
<td>ABNORMAL</td>
<td>DOCTOR’S</td>
<td>MARGINS</td>
<td>SECONDARY</td>
</tr>
<tr>
<td>ACHING</td>
<td>DOESNT</td>
<td>MARIA</td>
<td>SEDATED</td>
</tr>
<tr>
<td>ACS</td>
<td>DONE</td>
<td>MARK</td>
<td>SEIZURE</td>
</tr>
<tr>
<td>ACTIVITY</td>
<td>DOSE</td>
<td>MARROW</td>
<td>SENDING</td>
</tr>
<tr>
<td>ADENOCARCINOMA</td>
<td>DOUG</td>
<td>MASS</td>
<td>SENSITIVE</td>
</tr>
<tr>
<td>ADJUVANT</td>
<td>DOWNS</td>
<td>MAY</td>
<td>SENT</td>
</tr>
<tr>
<td>ADMITTED</td>
<td>DR</td>
<td>MAYO</td>
<td>SEPTEMBER</td>
</tr>
<tr>
<td>ADRENAL</td>
<td>DRAINING</td>
<td>MD</td>
<td>SESSIONS</td>
</tr>
<tr>
<td>ADRIAMYCIN</td>
<td>DRE</td>
<td>MDS</td>
<td>SEVERAL</td>
</tr>
<tr>
<td>ADVICE</td>
<td>DRENCHING</td>
<td>ME</td>
<td>SEVERE</td>
</tr>
<tr>
<td>ADVISE</td>
<td>DRS</td>
<td>MEASURING</td>
<td>SHARING</td>
</tr>
<tr>
<td>ADVISED</td>
<td>DRUG</td>
<td>MEDICAL</td>
<td>SHAWW</td>
</tr>
<tr>
<td>AFTER</td>
<td>DRUGS</td>
<td>MEDICATION</td>
<td>SHE</td>
</tr>
<tr>
<td>AFTERWARD</td>
<td>DUE</td>
<td>MEDICINE</td>
<td>SHOCK</td>
</tr>
<tr>
<td>AGAIN</td>
<td>DULCIMER</td>
<td>MEDS</td>
<td>SHORTLY</td>
</tr>
<tr>
<td>AGE</td>
<td>DULCIMERALG</td>
<td>MELANOMA</td>
<td>SHORTS</td>
</tr>
<tr>
<td>AGGRESSIVE</td>
<td>DULL</td>
<td>MELANOMAS</td>
<td>SHOWED</td>
</tr>
<tr>
<td>AGGRESSIVENESS</td>
<td>DURING</td>
<td>MEMORY</td>
<td>SIBLINGS</td>
</tr>
<tr>
<td>AGO</td>
<td>DWELL</td>
<td>MERRY</td>
<td>SICK</td>
</tr>
<tr>
<td>AGONY</td>
<td>DX</td>
<td>METABOLIC</td>
<td>SICKNESS</td>
</tr>
<tr>
<td>ALIMTA</td>
<td>DYSPLASIA</td>
<td>METASTASIS</td>
<td>SIDE</td>
</tr>
<tr>
<td>Word</td>
<td>Synonym 1</td>
<td>Synonym 2</td>
<td>Synonym 3</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>ALL</td>
<td>EACH</td>
<td>METASTASIZE</td>
<td>SIG</td>
</tr>
<tr>
<td>ALLERGIES</td>
<td>EARLY</td>
<td>METASTASIZED</td>
<td>SIGNATURE</td>
</tr>
<tr>
<td>ALOE</td>
<td>EASE</td>
<td>METASTATIC</td>
<td>SIGNS</td>
</tr>
<tr>
<td>ALOT</td>
<td>EBRT</td>
<td>METS</td>
<td>SINCE</td>
</tr>
<tr>
<td>ALSO</td>
<td>EFFECTS</td>
<td>MG</td>
<td>SISTER</td>
</tr>
<tr>
<td>ALTERNATIVE</td>
<td>EGFR</td>
<td>MICROSCOPE</td>
<td>SISTER'S</td>
</tr>
<tr>
<td>ALTHO</td>
<td>EKG</td>
<td>MIKE</td>
<td>SITUATION</td>
</tr>
<tr>
<td>ALTHOUGH</td>
<td>ELES</td>
<td>MINE</td>
<td>SKING</td>
</tr>
<tr>
<td>AM</td>
<td>ELEVATED</td>
<td>MINOR</td>
<td>SKIN</td>
</tr>
<tr>
<td>AMANDA</td>
<td>ENCORDING</td>
<td>MISDIAGNOSED</td>
<td>SLEEP</td>
</tr>
<tr>
<td>AND</td>
<td>ENDODMETRIAL</td>
<td>MM</td>
<td>SLEEPING</td>
</tr>
<tr>
<td>ANDERSON</td>
<td>ENDO</td>
<td>MOHS</td>
<td>SLOWLY</td>
</tr>
<tr>
<td>ANEMIA</td>
<td>ENDOMETRIUM</td>
<td>MOLE</td>
<td>SO</td>
</tr>
<tr>
<td>ANESTHETIC</td>
<td>ENDOTHELIAL</td>
<td>MOM</td>
<td>SOFTENER</td>
</tr>
<tr>
<td>ANGER</td>
<td>ENEMA</td>
<td>MOM'S</td>
<td>SOFTENERS</td>
</tr>
<tr>
<td>ANGIogenesis</td>
<td>ENLARGED</td>
<td>MONDAY</td>
<td>SOMETIMES</td>
</tr>
<tr>
<td>ANJ</td>
<td>ENLARGEMENT</td>
<td>MONITORING</td>
<td>SON</td>
</tr>
<tr>
<td>ANNE</td>
<td>ENT</td>
<td>MONTH</td>
<td>SONS</td>
</tr>
<tr>
<td>ANNIVERSARY</td>
<td>ER</td>
<td>MONTHS</td>
<td>SOON</td>
</tr>
<tr>
<td>ANOTHER</td>
<td>ERECTILE</td>
<td>MOOD</td>
<td>SORE</td>
</tr>
<tr>
<td>ANSWERS</td>
<td>ERLOTINIB</td>
<td>MORNING</td>
<td>SORRY</td>
</tr>
<tr>
<td>ANTIBIOTIC</td>
<td>EVALUATED</td>
<td>MORPHINE</td>
<td>SPASMS</td>
</tr>
<tr>
<td>ANTIBIOTICS</td>
<td>EVALUATION</td>
<td>MOTHER</td>
<td>SPECIALIST</td>
</tr>
<tr>
<td>ANXIETY</td>
<td>EXCRUCIATING</td>
<td>MOVABLE</td>
<td>SPHINCTER</td>
</tr>
<tr>
<td>ANY</td>
<td>EXPECTANCY</td>
<td>MRI</td>
<td>SPLEEN</td>
</tr>
<tr>
<td>ANYHOW</td>
<td>EXPERIENCE</td>
<td>MUCOUS</td>
<td>SPOTS</td>
</tr>
<tr>
<td>ANYONE</td>
<td>EXPERIENCED</td>
<td>MUCUS</td>
<td>SPREAD</td>
</tr>
<tr>
<td>APEX</td>
<td>EXPERIENCES</td>
<td>MUM</td>
<td>SRT</td>
</tr>
<tr>
<td>APOPTOSIS</td>
<td>EXPERIENCING</td>
<td>MUTATED</td>
<td>STAGE</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>EXPERT</td>
<td>MY</td>
<td>STAGES</td>
</tr>
<tr>
<td>APPETITE</td>
<td>EXPERTS</td>
<td>NADIR</td>
<td>STAGING</td>
</tr>
<tr>
<td>APPOINTMENT</td>
<td>EXTEND</td>
<td>NAGOURNEY</td>
<td>STAGED</td>
</tr>
<tr>
<td>APPOINTMENTS</td>
<td>EXTENSIVE</td>
<td>NANCY</td>
<td>STARGED</td>
</tr>
<tr>
<td>APPRECIATE</td>
<td>EXTERNAL</td>
<td>NAPS</td>
<td>STAPLED</td>
</tr>
<tr>
<td>APPRECIATED</td>
<td>FACING</td>
<td>NASAL</td>
<td>STARTED</td>
</tr>
<tr>
<td>APPROX</td>
<td>FACTOR</td>
<td>NAUSEA</td>
<td>STARTING</td>
</tr>
<tr>
<td>APPT</td>
<td>FAMILY</td>
<td>NAVELBINE</td>
<td>STATISTICS</td>
</tr>
<tr>
<td>APRICOT</td>
<td>FATHER</td>
<td>NECK</td>
<td>STAY</td>
</tr>
<tr>
<td>APRIL</td>
<td>FATIGUE</td>
<td>NEEDED</td>
<td>STAYING</td>
</tr>
<tr>
<td>AREA</td>
<td>FEB</td>
<td>NEEDLE</td>
<td>STEROIDS</td>
</tr>
<tr>
<td>ARM</td>
<td>FEEDBACK</td>
<td>NEG</td>
<td>STICKY</td>
</tr>
<tr>
<td>ARMPIT</td>
<td>FEEL</td>
<td>NEGATIVE</td>
<td>STOMACH</td>
</tr>
<tr>
<td>ARMPITS</td>
<td>FEELING</td>
<td>NERVE</td>
<td>STOOL</td>
</tr>
<tr>
<td>ARRIVED</td>
<td>FEELINGS</td>
<td>NERVES</td>
<td>STRENGTH</td>
</tr>
<tr>
<td>ARTHRITIS</td>
<td>FEELS</td>
<td>NEURO</td>
<td>STRESS</td>
</tr>
<tr>
<td>AS</td>
<td>FELT</td>
<td>NEUROENDOCRINE</td>
<td>STRESSING</td>
</tr>
<tr>
<td>ASK</td>
<td>FEMUR</td>
<td>NEUROLOGIST</td>
<td>STUDIES</td>
</tr>
<tr>
<td>ASKED</td>
<td>FEVER</td>
<td>NEUROPATHY</td>
<td>SUGGESTED</td>
</tr>
<tr>
<td>ASPIRIN</td>
<td>FIBROID</td>
<td>NEWLY</td>
<td>SUGGESTION</td>
</tr>
<tr>
<td>ASSAYS</td>
<td>FIGHTING</td>
<td>NEWS</td>
<td>SUMMER</td>
</tr>
<tr>
<td>ATYPICAL</td>
<td>FINAL</td>
<td>NEXAVAR</td>
<td>SUPPLEMENTS</td>
</tr>
<tr>
<td>AUGUST</td>
<td>FIND</td>
<td>NEXT</td>
<td>SURGEON</td>
</tr>
<tr>
<td>AVAILABLE</td>
<td>FINE</td>
<td>NHS</td>
<td>SURGEON'S</td>
</tr>
<tr>
<td>AVASTIN</td>
<td>FINGERS</td>
<td>NODE</td>
<td>SURGERIES</td>
</tr>
<tr>
<td>AWAITING</td>
<td>FIRST</td>
<td>NODES</td>
<td>SURGERY</td>
</tr>
<tr>
<td>AWAKE</td>
<td>FLUID</td>
<td>NODULE</td>
<td>SURGICAL</td>
</tr>
<tr>
<td>AWHILE</td>
<td>FMLA</td>
<td>NODULES</td>
<td>SURVIVAL</td>
</tr>
<tr>
<td>BACK</td>
<td>FNA</td>
<td>NOMOGRAM</td>
<td>SURVIVING</td>
</tr>
<tr>
<td>BAG</td>
<td>FOLLIRINOX</td>
<td>NOMOGRAMS</td>
<td>SWEATS</td>
</tr>
<tr>
<td>BALD</td>
<td>FOLLOW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BASAL</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CELLULAR  HEMOGLOBIN  PERMANENT  TRIAL
CENTER  HEMORRHOIDS  PET  TRIALS
CERVICAL  HER  PHARMACIST
CERVIX  HERBS  PHLEG
CHANCE  HESITATE  PILLS
CHECKED  HERE  PLAN
CHECKUP  HI  PLEASE
CHEMOTHERAPY  HIDEE  POLYPS
CHEST  HIM  PORT
CHOP  HIS  POSITIVE
CHRISTMAS  HODGKINS  TUMOR
CINDY  HODGKINS'  TUMORS
CLEAR  HOLIDAYS  TRIMIX
CLINIC  HOLISTIC  TROUBLE
CLINICAL  HOME  TRUS
CLUB  HOPE  TUBES
CM  HOPEFULLY  ULTRA
CNS  HOPES  ULTRASONIC
COASTER  HOPING  ULTRASOUND
CODEINE  HORMONE  UNDERGO
COLD'S  HOSPICE  UNDERGOING
COLON  HOSPITAL  UNDERSTANDING
COLONOSCOPY  HOSPITALS  UNCOMMON
COLPOSCOPY  HOURS  UNCOMMON
COMA  HOWEVER  UNCOMMON
COMBINATION  HUBBY  UNDETECTABLE
COMFORT  HURTS  UNDERTAKEN
COMFORTABLE  HUSBAND  UNDERWAY
COMFORTING  HUSBAND'S  UNDERWAY
COMPASSIONATE  HYSTERECTOMY  UPDATES
COMPZINE  I  UPDATES
COMPLEMENTARY  IM  UPDATED
CONCERNED  IMMEDIATE  UPDATE
CONCERNING  IMMUNE  UCE
CONCERNS  IMMUNOTHERAPY  UNDERWAY
CONDITION  IMPROVEMENT  UNDERWAY
CONJUNCTION  IMRT  UNDERWAY
CONSECUTIVE  INFECTION  UNDERWAY
CONSTIPATION  INCISIONS  UNDERWAY
CONSULT  INCOMPLETE  UNDERWAY
CONTAINED  INCONTINECE  UNDERWAY
CONTINUES  INDICATE  UNDERWAY
CONVENTIONAL  INDICATION  UNDERWAY
COPD  INFECTIOUS  UPSET
COPE  INFLAMMATION  URENE
CORES  INFO  URINARY
COUGH  INFLAMMATIONS  UROLOGIST
COUNTS  INFORMATION  UROLOGISTS
COUPLE  INFUSION  VACATION
CRAMPS  INJECTION  VANDERBILT
CT  INSTITUTE  VARIOUS
CURABLE  INSURANCE  VEINS
CURE  INTRAVENOUS  VICKIE
CURED  IRRITATED  VISION
CUTANEOUS  ITCH  VISIT
CYBERKNIFe  ITCHY  VISITED
CYCLES  IV  VIVO

E-PATIENTS IN ONCOLOGY: A CORPUS-BASED CHARACTERIZATION OF MEDICAL TERMINOLOGY IN AN ONLINE CANCER FORUM

Ibérica 39 (2020): 69-96
AGE
AGGRESSIVE
AGGRESSIVENESS
AGO
AGONY
ALIMTA
ALL
ALLERGIES
ALOE
ALOT
ALSO
ALTERNATIVE
ALTHO
ALTHOUGH
AM
AMANDA
AND
ANDERSON
ANEMIA
ANESTHETIC
ANGER
ANGIOGENESIS
ANJ
ANNE
ANNIVERSARY
ANOTHER
ANSWERS
ANTIBIOTIC
ANTIBIOTICS
ANXIETY
ANY
ANYHOW
ANYONE
APEX
APOPTOSIS
APPENDIX
APPETITE
APPOINTMENT
APPOINTMENTS
APPRECIATE
APPRECIATED
APPROX
APPT
APRICOT
APRIL
AREA
ARM
ARMED
ARMPIT
ARMPIES
ARRIVED
ARTHRTIS
AS
ASK
ASKED
ASPIRIN
ASSAYS
ATYPICAL
AUGUST
AVAILABLE
AVASTIN
AWAITING
DULCIMERGAL
DULL
DURING
Dwell
DX
DYSLASIA
EACH
EARLY
EASE
EBRT
EFFECTS
EGFR
EKG
ELES
ELEVATED
ENCOURAGING
ENDO
ENDOMETRIAL
ENDOMETRIUM
ENDOTHELIAL
ENEMA
ENLARGED
ENLARGEMENT
ENT
ER
ERECTILE
ERLOTINIB
EVALUATED
EVALUATION
EXCRUCIATING
EXERCISES
EXPECTANCY
EXPERIENCE
EXPERIENCED
EXPERIENCES
EXPERIENCING
EXPERT
EXPERTS
EXTEND
EXTENSIVE
EXTERNAL
FACING
FACTOR
FAMILY
FATHER
FATIGUE
FEB
FEEDBACK
FEEL
FEELING
FEELINGS
FEELS
FELT
FEMUR
FEVER
FIBROID
FIGHTING
FINAL
FIND
FINE
FINGERS
MELANOMA
MELANOMAS
MEMORY
MERRY
METABOLIC
METASTASIS
METASTASIZE
METASTASIZED
METASTATIC
METS
MG
MICROSCOPE
MIKE
MINE
MINOR
MISDIAGNOSED
MM
MOHS
MOLE
MOM
MOM'S
MONDAY
MONITORING
MONTH
MONTHS
MOOD
MORNING
MORPHINE
MOTHER
MOVABLE
MRI
MUCOUS
MUCUS
MUM
MUTATED
MY
NADIR
NAGOURNEY
NANCY
NAPS
NASAL
NAUSEA
NAVELbine
NECK
NEEDED
NEEDLE
NEG
NEGATIVE
NERVE
NERVES
NERVOUS
NEURO
NEUROENDOCRINE
NEUROLOGIST
NEUROPATHY
NEWLY
NEWS
NEXAVAR
NEXT
NHS
NIGHT
SHOTS
SHOWED
SIBLINGS
SICK
SICKNESS
SIDE
SIG
SIGNATURE
SIGNS
SISTER
SISTER'S
SITUATION
SKING
SKIN
SLEEP
SLEEPING
SLOWLY
SO
SOFTENER
SOFTENERS
SOMETIMES
SON
SONS
SOON
SORE
SORRY
SPASMS
SPECIALIST
SPHINCTER
SPLeen
SPOTS
SPREAD
SRT
STAPLES
STARTED
STARTING
STATISTICS
STAY
STAYING
STEROIDS
STICKY
STOMACH
STOOL
STRENGTH
STRESS
STRESSING
STRONG
STUDIES
SUGGESTED
SUGGESTION
SUMMER
SUPPLEMENTS
SURGEON
SURGEONS
SURGEON'S
SURGERIES