(wileyonlinelibrary.com) DOI: 10.1002/qre.1853

Published online 24 August 2015 in Wiley Online Library

Identifying Oncological Patient Information Needs to Improve e-Health Communication: a preliminary text-mining analysis

Rosa Falotico,[a*†] Caterina Liberati[a] and Paola Zappa[b]

Extant healthcare literature has demonstrated that providing oncological patients with detailed, continuous, and diversified information on their pathology positively affects patient empowerment and care continuum effectiveness.

Medical information is traditionally conveyed by physicians and professional care givers. In the latest years, however, there has been a rapid rise in the use of web media as a source of health information. Improving cancer care continuum requires also the implementation of a reliable and high-quality oncological web communication system, as well as a deeper understanding of patients' requirements in this respect. To the best of our knowledge, however, no papers have investigated the importance and role of web communication, and of websites, in particular.

This work aims to address these aspects explicitly, proposing the use of text mining as a tool for identifying oncological patient communication needs. For this purpose, we conducted an exploratory study on a sample of rare cancer patients. Text mining techniques were applied on transcriptions of semi-structured interviews on patients' information requirements, and the text content was synthetized. By means of correspondence analysis, the main concepts expressed by patients were identified, and the association between patients and concepts was represented in an appropriate metric space.

Although our study is at an early stage, the findings highlighted by this preliminary text mining analysis could support the design of a patient-centered e-health communication system and be the basis for further analysis. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: oncological patient empowerment; oncological patient communication needs; health communication; web content; text mining; correspondence analysis

1. Introduction

In recent years, cancer research has progressed considerably. Ongoing advancements like early detection, massive screening activities, and effective treatments have allowed oncological patients to live longer.[1] This has led to conceive cancer as a serious chronic condition[2] and has promoted the emergence and formalization of the concept of a cancer care continuum, that is, the delivery of health care to a cancer patient through all phases of the illness – from diagnosis to the end of life.

Improving the cancer care continuum has become the main goal of the National Healthcare Systems (NHSs) in several Western countries. The cancer care continuum involves several types of care, as well as the transition from one type to another. However, to date, health policy makers have focused on physiological aspects of the cancer care continuum (i.e., prevention, screening, and medical treatments). Relatively, little attention has been devoted to other – equally important – aspects of the disease, such as patient empowerment.[3]

Patient empowerment is a process aiming at facilitating self-directed behavioral change[4] and at enhancing patients' abilities to actively understand and influence their own health status.[5] According to the World Health Organization (WHO), patient empowerment is a prerequisite for health and a self-care strategy for chronically ill patients. Empowerment allows for the improvement of health outcomes and quality of life.[6]

A substantial aspect of empowerment is communication. Several studies have outlined that providing oncological patients with complete, up-to-date, and continuous information on their disease and on existing treatments or possible improvements would significantly enhance their empowerment, compliance to therapies, and disease monitoring. Information provision, moreover, could help in decreasing anxiety and fear about cancer and its treatment, correcting erroneous beliefs, increasing adherence to medical advice, and improving coping mechanisms, psychological well-being, and doctor–patient relationships.[7]

[a] *Università degli Studi di Milano-Bicocca, Milan, Italy*
[b] *Università della Svizzera Italiana, Lugano, Switzerland*
**Correspondence to: Rosa Falotico, Dipartimento di Economia, Metodi Quantitativi e Strategie d'Impresa, Università degli Studi di Milano-Bicocca, via Bicocca degli Arcimboldi 8, 20126 Milan, Italy.*
[†] *E-mail: rosa.falotico@unimib.it*

Strengthening cancer care continuum and patient empowerment requires solid support for oncological communication, which can be made possible only by the development of a patient-centered information system. The hub of this system will primarily be the clinician, who is the most reliable source of information for the patients.[8] Besides the clinician, however, the Internet has also become more and more important, so that nowadays, web-user patients consider online media as a reliable source of medical information.[9,10] As a consequence, the future of empowerment will imply targeting web communication with respect to patient requirements.[7]

Designing and developing customized websites involve a broad awareness of the oncological patient communication needs. Policy makers can find suitable suggestions on this issue in health literature. However, as far as we know, existing studies suffer two major limits:

- They do not provide a synthetic and up-to-date framework that summarizes the main aspects of oncological communication;
- They adopt qualitative (in depth) or quantitative (extensive) analytical tools only, while not trying to integrate them both.

To overcome these limits, we propose to analyze oncological patient communication needs using text mining techniques. They would allow us to reach an overall understanding of patient requirements by collecting textual data (by means of narrative interviews) and exploiting their potential – thus combining the benefits of qualitative and quantitative analysis.

Text mining alleviates drawbacks of qualitative studies without significant loss of information because it makes available a neutral treatment of data. This technique is not bound to subjective analyst interpretation but uses semi-automatic tools to synthesize information. In addition, textual analysis on an extensive set of documents is, proportionally, moderately time-consuming, so researchers can deal with a considerable amount of data in a reasonable amount of time.

Extensive management literature proposes the use of text mining techniques mainly in new product development processes (DPP) and in identifying customer requirements.[11–13] In our work, we propose to apply text mining in order to uncover rare cancer patient communication needs. We focus, in particular, on web communication and try to derive indications useful for its implementation.

The remainder of this paper is organized as follows. Section 2 illustrates the theoretical background, summarizing existing literature on communication to oncological patients as well as possible insights from management research methodology on the topic. Section 3 introduces the methodological approach adopted in order to uncover patients' communication needs. A case study on rare oncological pathologies is presented in Section 4. We describe the sample, the data collection process, and present the main results. Finally, Section 5 concludes.

2. Theoretical background

Medical literature has amply demonstrated the importance of an effective health communication. An extensive study conducted more than a decade ago on a large sample of UK oncological patients[14] has shown that 87% of respondents preferred to have as much information as possible on their pathology, either positive or negative; 98% of patients needed to know the exact nature of their diagnosis, while 95% wished to be informed about their chances of care and survival. In spite of this empirical evidence, until the early 2000s, providing oncological patients with information on their pathology and on treatments did not represent a priority for the NHSs and the clinicians in several countries.

The recent diffusion of a model of care based on patient empowerment has led to a shift in the management of care from a central health care system toward a patient-centered system.[7] As a consequence of the increased participation of patients in cancer care continuum, there has been also an upsurge in patients' expectations about provision of information, especially from clinicians, and a focus on patient-centered interactions. They are interactions in which the patient's point of view is actively sought by the clinician and by other stakeholders, who must behave in a way that facilitates the patient to express herself or himself, to speak openly, and ask questions.[8]

Policy makers are increasingly aware of the need of an information system focused on patient requirement that can support and enhance the standard care delivery. The pivotal element of this system is necessarily the doctor. Existing papers have highlighted that most cancer patients are likely to receive information personally from clinicians, who are, therefore, the main source of information. Communication between clinician and patient is relevant to treatment outcomes. In addition, patients evaluate clinician communication skills as important as their expertise.[8]

In the latest years, as the availability of online tools has increased, also health communication has increasingly exploited them.[7,9,10] Hence, patients have become more open to bypassing traditional sources of information and to considering online tools as reliable sources of information.[‡] The literature has defined e-health this phenomenon and e-patients those who use the Internet to obtain information on the disease or online diagnoses and to interact with other stakeholders of the cancer care continuum system.[10] Patients search online information because of the desire for reassurance, for a second opinion on their disease, or for the need of additional information. Also, patients rely on the Internet when they perceive external barriers to accessing information through traditional sources.[15]

This increased interest of patients toward the web world has led also to a growing number of interventions aimed at patient empowerment that have been delivered online.[6] Nevertheless, the use of web communication in health care seems to raise some critical issues. Since the Internet is an unsupervised source of information, it is almost impossible to assess the quality of information available online.

[‡] *A survey conducted in Europe in 2007 shows that the percentage of users searching medical information online is quite large (46.8%) and percentage of Europeans considering Internet as an important source of health information seems to be growing.*[14]

National Healthcare Systems can overcome the e-health challenges in setting up a reliable and high-quality web communication, primarily based on the provision of patient-oriented websites. Policy makers pursuing this goal have to carefully choose the contents of communication with special attention to patients' requirements. The analysis of patients' communication needs is a starting point for an effective intervention.

Health literature provides much empirical evidence on the communication requirements of oncological patients. Warren *et al.*[7] have identified 79 cognitive requirements classified according to three categories: spectrum of cares, themes, and nature of questions asked by patients. Treatment is the most common topic in spectrum of care, procedure and side effects are the main themes, and issues of a practical nature are the most frequent questions. Rutten *et al.*,[16] conducting a meta-analysis of the most relevant articles published in English from 1980 to 2003 about information need and source of information among cancer patients, have found 64 distinct subcategory needs classified in 10 categories. They cover a variety of issues and seem to imply that patients need to interact with different stakeholders and to exploit different communication sources, which could fulfill the emerging requirements. While treatment-related (38.1%) and cancer-specific (12.8%) information covers more than a half of the information needs, the study reveals that other issues, not directly related to the care paths, are critical for patients. These are, for instance, rehabilitation information (12.2%) like self-care or home-care issues during recovery, coping information (8.8%), namely, emotional reactions and support and, finally, interpersonal/social information (6.0%), that is, effects of cancer on family, friends, or caregivers. In line with these findings, a survey conducted in Italy in 2004[17] has shown that 85% of respondents claimed to have received insufficient information, especially on the aforementioned aspects on the disease, and to have actively sought for them using a variety of sources.

The aforementioned contributions to the analysis of patients' communication requirements are fundamental, but there is still much room for improvement. Some studies date back to many years ago.[16,17] Indeed, also the most recent ones do not apply advanced statistical tools[7], nor are they able to provide detailed information on patient-oriented web communication. To the best of our knowledge, then, methodologies employed in health research to date are unsatisfactory for the purpose of this study.

Management literature can offer some suggestions in this respect, providing theoretical as well as methodological indications on how to design and implement patient-oriented web communication – and patient-oriented websites in particular.

The success of a patient-oriented website, like in case of user-oriented website, depends on the satisfaction of its user, and the effectiveness of website contents is a primary factor for user satisfaction.[18] While management literature proposes several measurement tools for the a posteriori assessment of web-user satisfaction,[18–20] as far as we know, no tools for a priori investigation of web-user requirements are provided.

Indications on methodologies for the analysis of patients/web-user needs can be borrowed from studies on consumer needs. The identification of consumer requirements is a primary issue in management literature, mainly in DPP.[21] It has been demonstrated that taking into account the 'voice of consumer' during the stages of design and test of a new product is a key element for its success.[22] The recent exponential advancements in technology have contributed to the expansion of consumer needs, making the processes of product development progressively more complex. To better identify these requirements, especially in cases of the technological offer, the most advanced statistical tools have been adopted.

The availability of unstructured data, coming from feedback left on the Internet by consumers, is constantly growing[23] and has pulsed the use of data mining on textual data.[24] In the last decade, the application of this innovative technique in DPP has definitely increased.[11–13]

In particular, in order to elicit consumer needs, statistical tools were proposed, aiming to extract keywords from textual data and to employ them to automatically classify original documents in main categories.[11,25] The extraction of keywords requires performing an association analysis on the different concepts present in the documents, but every attempt of textual analysis is particularly problematic because of the difficulty in managing the wide number of associations produced.[12] Binary Correspondence Analysis (CA) is a statistical tool commonly used to overcome this problem.[26]

Performing CA on textual data allows a graphic representation in a reduced space of the association among the documents and among the terms stored in a textual database.[26] In fact, CA provides an approximation of high-dimensional multivariate matrices with a reduced space[27] allowing the highlighting of the latent semantic structure of a corpus of documents.

3. Methods and assessment

Text mining is one of the cutting-edge topics, to which many statistical researchers have been devoting their attention. The link between statistics and major disciplines concerning texts (linguistics, discourse analysis, content analysis, and artificial intelligence) has been clinched thanks to extraordinary speed-up in information retrieval and machine learning realized in the recent years.

A first attempt to give a quantitative framework to textual study is related to content analysis,[28,29] developed in the United States at the beginning of the last century. Defined as a statistical formulation of a document, content analysis has aimed to characterize the meanings in a given body of discourse in a systematic and quantitative fashion. Categorization of a text into relevant segments/themes, defined a priori, makes difficult the application of such method;[30] thus, the modern research of artificial intelligence in this field of investigation has devoted the heart of activities to representation of the meaning of sentences in order to improve the man–machine dialogue (Natural Language Processing).[31]

3.1. Text preprocess

Traditionally, natural language processing has tended to view the process of language analysis as being decomposable into a number of stages, mirroring the theoretical linguistic distinctions drawn between syntax, semantics, and pragmatics. Thus, sentences of a text are first analyzed in terms of their syntax; this provides an order and structure that is more amenable to an analysis in terms of semantics, or literal meaning. This phase is followed by a stage of pragmatic analysis whereby the meaning of the utterance or text in a context is determined.[32] Such view is partially borrowed by statistical approach to textual analysis, which has been more focused on methods derivation for adapting qualitative discrete variables to model texts complexity.[33,34] Statistical paradigm generally applied to solve empirical problem also in this field is defined as problem identification, data collection, processing, and results interpretation. Of course, more than in other applications, the processing phase, here, plays a crucial role because results of exploratory analysis depend on checking and cleaning data that influence the process of data reduction. Much of the challenge of such phase involves resolving text ambiguities and normalizing units of the corpus, that is, the task of converting a raw text file into a well-defined sequence of linguistically meaningful graphical forms (words). Statistical methods rely on measurements and counts of the identical elements (occurrences) either of more general objects added together treated as identical. In practice, two steps are involved here: text segmentation (tokenization), which parcels out a well-defined text corpus into its component words, and sentences and text normalization, which performs merging all word variations into a single representative form (lemmatization).[26]

3.2. *Spatial representation of key concepts*

According to the preprocess scheme, it is possible to reduce the number of terms present in the corpus's dictionary, counting only relevant terms, and a high-dimensional contingency table T (documents-by-terms) is obtained, where each cell represents the frequency of a word in a document. The geometric representation of a rectangular table that describes the similarities among the rows (and among the columns) has to be computed according to chi-square distance.[27,33,35,36] Variability of a frequency matrix as T, defined by chi-square metric, has to be represented in a continuous space generated by a vector decomposition that holds all the relative distances among rows. The binary CA provides such decomposition.

In a two-way contingency table as T ($k \times V$), we denote f_{ij} the joint relative frequency or proportion referred to the i-th row ($i = 1, \ldots, k$) and associated to column j ($j = 1, \ldots, V$) such that $\sum_{i=1}^{k} \sum_{j=1}^{V} f_{ij} = 1$. In CA, a general row i of a matrix T ($k \times V$), is considered as a point in the space[§] $\Re^V$ with coordinates $\{f_{ij}/f_{i\cdot}, j = 1,.., V\}$, weights $\{f_{i\cdot}, i = 1,.., k\}$, and the centroid of the rows set is the point $\{f_{\cdot j}, j = 1,.., V\}$. As already motivated in this section, the proximities between rows (documents in our contest) have to be measured using the χ^2 distance. Thus, the square distance between two row-points i and i' is given by the equation:

$$d^2(i, i') = \sum_{j=1}^{V} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 \tag{1}$$

Consistently with the geometric approach, the dispersion of the set of rows (and symmetrically, of the set of columns) around its centroid is measured through the inertia:

$$\phi^2 = \sum_{i=1}^{k} f_{i\cdot} d^2(i, centroid) = \sum_{j=1}^{V} f_{\cdot j} d^2(j, centroid) = \frac{\chi^2}{k} \tag{2}$$

being k the total counted units and V the vocabulary size. CA inspects the distances between row profiles as a whole or, equivalently, the distances between each profile and the mean profile. It also describes the discrepancy of the observed system from the independence model, by displaying approximations between rows onto the axes of maximum dispersion (factorial axes). The factorial axes can be obtained by performing a principal component analysis (PCA) on the table $\widetilde{T}$, whose general term is the residual with respect to the independence model (weighted by the inverse of the cross product of the row and column marginal sums), here shown as follows:

$$\widetilde{t}_{ij} = \frac{f_{ij} - f_{i\cdot} \cdot f_{\cdot j}}{f_{i\cdot} \cdot f_{\cdot j}} \tag{3}$$

According to such framework, in the row space, the inertia axis with rank s corresponds to the eigenvectors u_s $\left(\|u_s\|_{D_V} = 1 \right)$ of the matrix $\widetilde{T}' D_k \widetilde{T} D_V$[¶] associated with the eigenvalues λ_s (in decreasing order). In the column space, the inertia axis with rank s corresponds to the eigenvectors v_s $\left(\|v_s\|_{D_k} = 1 \right)$ of the matrix $\widetilde{T} D_V \widetilde{T}' D_k$ associated with the same eigenvalues λ_s (in decreasing order).

Therefore, the vectors of the row scores are

$$r_s = \widetilde{T} D_V u_s = \sqrt{\lambda_s} v_s \tag{4}$$

and the vectors of column scores are

[§] *Without loss of generality, we can assume that the number of words composing the vocabulary is smaller than the number of documents $V < k$.*

[¶] D_k is the diagonal matrix with general term $f_{i\cdot}$, and D_V is the diagonal matrix with general term $f_{\cdot j}$.[36]

$$c_s = \widetilde{T}' D_k v_s = \sqrt{\lambda_s} u_s \qquad (5)$$

These scores, or principal coordinates,[27] lead to both sets of distances (those between rows and those between columns) corresponding to the χ^2 distances defined in (1). Now, distances between row (column) points are easier to visualize in a reduced factor space. Instead, it is not possible to interpret these cross-proximities between a row-point and a column-point, because these two points do not come from the same initial space. Nevertheless, it is possible to interpret the position of a single row-point with respect to the set of column-points or of a single column-point with respect to the set of row-points. The main reason for this simultaneous representation is given by the transition relationships that link the coordinates of one point in one space (the row-space for example) to those of all the points of the other space (the column-space in our example). Transition formulae (also known as quasi-barycentric coordinates), actually, allow the transition from the set of row coordinates to the set of column coordinates (and vice versa).

$$\begin{aligned} r_s &= \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{V} \frac{f_{ij}}{f_{i\cdot}} \cdot c_s(j) \\ c_s &= \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{k} \frac{f_{ij}}{f_{\cdot j}} \cdot r_s(i) \end{aligned} \qquad (6)$$

Such relationships are really useful because they link the coordinates of one point in one space (the row-space for example) to those of all the points of the other space (the column-space).

The simultaneous representation of rows and columns on the same plot aids the interpretation of the researcher together with other measures as the absolute contributions, which describe the proportion of variance explained provided by each element (row or column) in building a principal axis.

$$\begin{aligned} Contr_s(i) &= \frac{f_{i\cdot} \cdot c_s(i)}{\lambda_s} \\ Contr_s(j) &= \frac{f_{\cdot j} \cdot c_s(j)}{\lambda_s} \end{aligned} \qquad (7)$$

Correspondence analysis is used for finding subspaces to represent proximities among profiles, but it can also be used for positioning supplementary rows and columns of the data matrix in this subspace. Once the principal axes u_s and v_s and the eigenvalue λ_s have been computed, supplementary rows or columns can be obtained easily:

$$\begin{aligned} r_s^+ &= \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{V} \frac{f_{ij}^+}{f_{i\cdot}^+} \cdot c_s(j) \\ c_s^+ &= \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{k} \frac{f_{ij}^+}{f_{\cdot j}^+} \cdot r_s(i) \end{aligned} \qquad (8)$$

where r_s^+designates the coordinate of a supplementary row *i* whose profile is ($f_{ij}^+/f_{i\cdot}^+$) on axis *s*, whereas c_s^+designates the coordinate of a supplementary column *j* whose profile is ($f_{ij}^+/f_{\cdot j}^+$) on the same axis. Thus, we can illustrate the configurations with supplementary information that did not participate in the construction of the planes. This has very significant consequences in terms of interpreting the results.

4. Case study

The benefits of focused, supervised, and reliable communication for oncological patients are well known and would be more evident in cases of rare cancer. In fact, patients affected by a rare disease are more in need of empowerment and have fewer chances to find specific information.

Policy makers have been taking several initiatives to overcome this aspect of the rare cancer problem. In Italy, some oncological centers (specializing in sarcoma treatments) and sarcoma patient advocacy groups have implemented a project funded by the Italian Ministry of Health, aimed at improving the quality and reliability of the web communication on sarcoma[‖]: PASSnetwork (network of PAtients with SarcomaS).

The primary goal of the PASSnetwork project is to improve the cancer care continuum outcomes of sarcoma patients conveying specific information on this disease. The main activity of the project focused on the planning and development of a patient-oriented website. Targeting web communication requires a broad and up-to-date understanding of patient communication needs. Therefore, the proposed application aims to analyze a sample of sarcoma patients and identify their communication needs in terms of content with respect to the information source and to investigate possible relations with the stages of disease and time from diagnosis. The preliminary findings of our analysis can be considered in order to structure the website and fill it with effective contents.

[‖] *Sarcomas are rare malignant tumors of the connective tissue. Their causes are frequently unknown, and accurate data about the incidences of sarcoma are hard to find. For these reasons, sarcoma patients have difficulties in accessing appropriate pathologic diagnoses, surgical, and medical therapies.*

4.1. Setting and data

In our preliminary study, we extracted a nonrandom sample of 12 rare cancer patients treated at a specialized research center in Northern Italy. The sample size is to be evaluated considering that these diseases are not common (a rare cancer has a prevalence of 5 or less out of 100 000 cases), and our interview model requires much involvement.

The survey was conducted using a semi-structured interview, organized in two parts. The first part aimed to collect information on individual characteristics (i.e., age, educational level, stage of the disease, presence and identification of caregivers), via a short, multiple-choice questionnaire.

The second part was structured as a clinical interview, using a narrative approach. Hence, the interviewer followed the traditional scheme of a 'question and answer' to set the agenda but allowed patients to freely express their ideas and suggestions about oncological communication as well as their propensity toward using new media. In particular, the interviewer encouraged respondents to describe the evolution of their requirements, in terms of communication content, during the care continuum. Also, the interview, conducted by psychologists of the research center, investigated the main sources of information used by patients.

The sample observed is balanced for gender; half of respondents are people aging '30–59 years', one-third are 'more than 60 years' old, whereas 16.7% are young people (under 30).

The interviews were conducted in a leading research center in the treatment of rare oncological diseases in Italy. Although this center is located in Milan, it is not surprising that the sample is equally distributed in northern and central-southern areas because of the high level of care commuting in Italy. The majority of the interviewed patients (75.0%) come from small towns. The respondent residents in big cities mostly live in the northern area of the country (66.7%).

The educational level of the interviewees is quite high: only 16.7% has less than a high school degree, 50.0% has a high school degree, and 33.3% has an academic degree.

Because of its excellent reputation, the research center collects patients whose diagnoses vary quite a lot in terms of time span: the oldest ones date from 6 years ago, but 58.3% of respondents were diagnosed within less than 2 years. The majority of the patients recently diagnosed have a localized disease (85.7%).

The whole sample states to have family support during illness, mainly from partners (58.3%). On the contrary, the relational and informational aid, external to the family nucleus, is not actively searched. Only 16.7% of respondents know at least one of the Italian main associations of rare cancer patients for their pathology (AIMaC, Aig, and GIST) and one-third do not know any. Plus, just one quarter of respondents, in addition to knowing, took part in conferences or forum activities of these associations.

The familiarity with the Internet is largely attested: all interviewees generally visit websites and social network; 91.7% surf the web, while 8.3% delegate web research to younger people in the family. Also, the frequency of Internet use is fairly high: 63.6% surf the web more than once a week while 36.4% chats also on social networks, but nobody reads blogs or forums daily.

We integrated the two sections of the interviews with quantitative analysis. The first part did not require particular cleaning and was stored in a sociodemographics dataset. Conversely, it was challenging to transcribe the interviews and prepare them for textual analysis. The initial corpus included about 20,400 words, grouped in more than 3000 different terms. We cleansed the text of unnecessary parts of speech, eliminating all the articles, prepositions, and less significant adjectives. We normalized the remaining words with semi-automatic techniques for lexical and grammatical analysis. Moreover, to further reduce the corpus size, we (1) aggregated the words with overlap meaning in noun groups (e.g., chance_of_survival); (2) grouped terms that, in different forms (nouns, adjectives, and verbs), indicated the same concept (e.g., Internet = Internet, web_site, etc.; doctor = specialist, oncologist, etc.). The reduced corpus included about 3000 occurrences, grouped in a vocabulary of little more than 650 terms, but we could proceed with a further reduction when the informational contribution for the analysis purpose of low frequency words is very small, hapaxes** and rarely occurring words are commonly ignored.[26] To choose a suitable frequency threshold, we followed Bolasco's[37] scheme. His approach, first, orders terms present in the corpus with respect to their occurrences, then, splits such sorted corpus into two bands (low-frequency and high-frequency terms). The rule of thumb suggested by the author set to the 20th percentile of the low frequency words the threshold for the terms' selection.

According to that, we determined a final corpus including the most frequent 38 terms of reduced corpus. These terms represent the 0.58% of reduced vocabulary but cover the 55.1% of reduced corpus. The final corpus was stored in a *T* documents x terms matrix, which was, then, used to perform textual analysis.

4.2. Results

Before illustrating the evidences detected by performing CA on our data, we plot the concepts by means of word cloud, in order to uncover the words occurrences. The word cloud is a standard way to represent the distribution of terms in a corpus: it provides some insights into the structure of the texts. This graphical technique is essential in textual analysis because it substitutes a bar or pie chart in case of a high number of modalities. In a word cloud, the font size of the words is proportional to the frequencies associated with each term in the corpus, providing a synthetic and understandable visualization of the importance of each concept.

Figure 1 summarizes the distribution of the 38 key terms included in the final corpus. The main source of information for the respondents is the 'doctor' (183 occurrences) that addresses them ['says'(74)]. 'Internet' (141) is almost equally important. This latter result is particularly relevant, because it points out the independence of patients from experts. They primarily 'search' (76) 'simple_information' (69) to share 'same_experiences' (57). In addition, the 'state_of_mind'(81) of the patients affects their information needs.

** *Hapaxe is a word that occurs only once within a context, either in the written record or in a single text.*

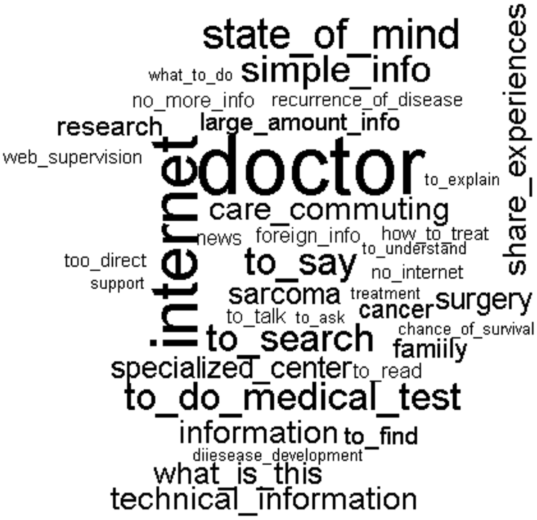

Figure 1. Word cloud of the key concepts in final corpus

Further, and more detailed, information on the needs expressed by respondents can be obtained through textual analysis. In order to identify similarities among patients with different information requirements, we performed a CA on the final corpus. Then, we provided a simultaneous representation of patients and key concepts in a low-dimensional map, as illustrated in previous section.

Aiming to reach a good percentage of inertia, we chose to retain the first four dimensions resulting from CA. These factors explain more than the 70% of total inertia. Four dimensions, however, can be difficult to visualize. For brevity, we limit our results description to the evidences highlighted by the principal plane (Figure 2), which is usually considered the best graphical solution (it reproduces about 55% of total inertia), because it is clear and easy to interpret.

It shows the simultaneous representation of terms and individuals on the correspondence map of the first (Dimension 1) and second factors (Dimension 2).

Figure 2 also shows a typical pattern of CA: the triangular form,[35,38] as we uncover later. Such configuration usually appears when an axis discriminates between two phenomena and the second one among alternative aspects of only one of these phenomena. Taking into account the instances-concepts configuration, we computed the Absolute Contributions to inertia (7) of each term for the two factors, in order to name and interpret the principal axis. Visual inspection of the main contributes to inertia (Tables I and II) highlights some interesting evidences. In particular, the nouns 'medical_test' [Absolute Contributions to Inertia for the

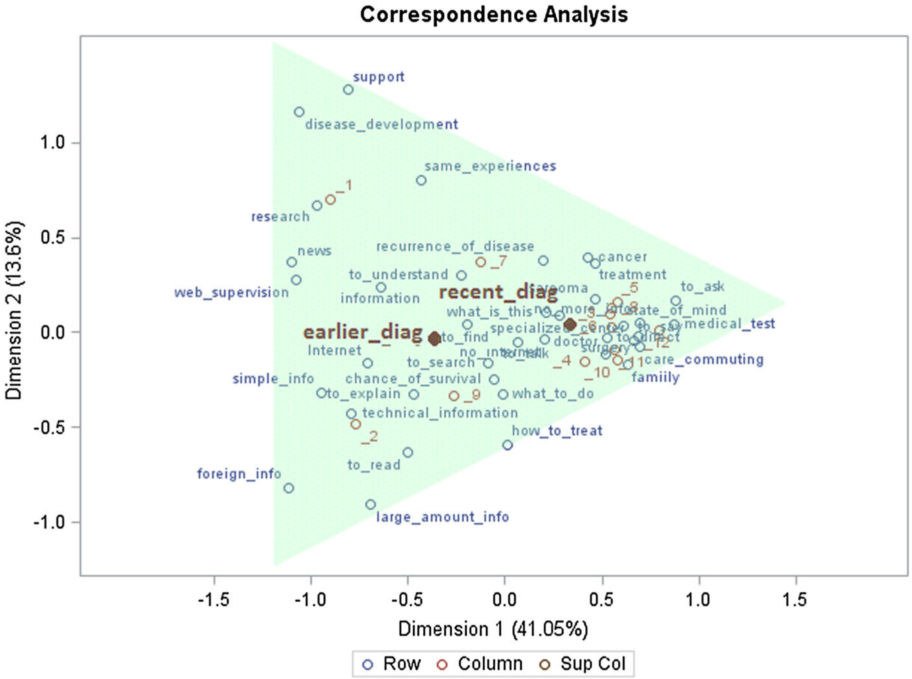

Figure 2. Two-dimensional correspondence map

Table I. The main absolute contributions to inertia and corresponding row coordinates of the first factorial axis

Term	Dim1	Contr1 (%)
medical_test	0.870	8.29
Doctor	0.521	7.25
to_say	0.662	4.73
Internet	−0.707	10.30
simple_information	−0.948	9.05
Research	−0.969	5.90
technical_information	−0.792	5.04

Table II. The main absolute contributions to inertia and corresponding row coordinates of the second factorial axis

Term	Dim2	Contr2 (%)
same_experiences	0.801	16.12
support	1.280	12.28
disease_development	1.161	11.29
research	0.669	8.49
large_amount_info	−0.904	11.90
foreign_info	−0.817	7.65
technical_information	−0.425	4.39

Row-Points = 8.29%] and ‘doctor’ [7.25] are those who mainly contribute to the characterization of the positive semi-axis of the first factor. The most significant verb for the same semi-axis is ‘to say’ [4.73%] and its synonyms ‘to indicate’ and ‘to recommend’.

Moreover, first factor’s negative semi-axis is described by the concepts ‘Internet’ [10.30%], ‘simple_information’ [9.05%], ‘research’ [5.90%], and ‘technical_information’ [5.40%]. According to that, we can conclude that the first component discriminates between the two principal sources of information:

- The positive semi-axis represents the respondents that prefer to receive official information and recommendations from their doctor;
- The negative semi-axis represents the respondents that integrate the medical information with active search on Internet.

For what concerns the second axis, we found that the key concepts for the positive semi-axis are ‘same_experiences’ [16.12%], ‘support’ [12.2%], ‘disease_development’ [11.29%], and medical ‘research’ [8.49%]. On the negative semi-axis, the most relevant terms are ‘large_amount_info [11.9%]’, ‘foreign_info’ [7.65%], and ‘technical_information’ [4.39%].

We, then, assume that the second factor distinguishes between the following:

- Social searchers (positive semi-axis): respondents that need information about their disease’s development and look for support from patients with similar experiences.
- Intensive searchers (negative semi-axis): respondents that use the Internet to collect the maximum amount of information on specific topic, even from not Italian sources.

Drawing on these results, we propose to name the first factor ‘source of information’ and the second factor ‘typology of internet searcher’, according to the ‘triangular pattern’.

Finally, we matched the attitude toward oncological information and the patient characteristics collected through the questionnaire. To this purpose, we projected patient characteristics as illustrative variables on the correspondence map.

We found that the time from diagnosis better discriminates among the sources of information (one dimension) of interviewed patients. In Figure 2, the positive semi-axis of the first component contains respondents with a recent diagnosis, who only rely on health professionals’ communication and require general information on their disease and on specialized center for an appropriated treatment.

The negative semi-axis is mainly associated with patients with an early diagnosis. They need to integrate the medical information with all the information available, primarily from web media. They probably reached a deeper understanding of their pathology during the years of their illness, so, at the time of the interview, they use not trivial source (foreign and/or specialized sites) to satisfy their informational requirements.

5. Discussion and conclusions

Extant healthcare literature has widely stated that significant advancements in oncological research have led to the emergence and formalization of the concept of cancer care continuum. One of the main consequences of the cancer care continuum has been patient empowerment. Promoting patient self-care and autonomy from health professionals, patient empowerment has increased the importance of communication to patients and of the process of patients' information seeking. Web media progressively complements traditional media with respect to satisfying these emerging informational needs. In fact, the Internet is the second most used source of health advice after physicians.

To better target web communication, health policy makers need to extend their understanding of patients' requirements. Few and not recent studies, however, have examined patient communication requirements in depth. In this paper, we propose a new methodology to investigate the communication needs of oncological patients, and we provide an application for a rare oncological pathology. We have designed and conducted an empirical study on a sample of patients affected by sarcomas and treated in a highly specialized research center in Northern Italy. We have collected the detailed descriptions of patients' communication requirements by means of semi-structured interviews. Then, we have applied text mining and CA techniques on the transcriptions of the interviews in order to detect the main concepts expressed by patients and to associate specific concepts to Internet user patients.

Our findings on rare cancers are consistent with existing literature on common oncological diseases. Interviewed patients, mainly patients with a recent diagnosis, show a substantial preference for information coming from health professional. Our work contributes to show some specific behaviors of respondents, in particular, with respect to long-term patients. They seem to show a greater aptitude toward web search. The e-patients using the Internet for integrating medical information have two primary requirements: (1) find on the Internet the larger amount of quality and reliable information, specifically on medical issues and (2) find on the internet support and share their experience to better understand the possible development of their disease.

This preliminary evidence can suggest the main directions for the development of a website patient oriented. An effective web communication must provide a supervised and complete information on the primary medical topic. Also, this website can facilitate the use of social networks in order to support the creation of a community of cancer patients, for example, by means of a supervised forum.

In other words, there are yet similar experiences for common cancer (e.g., http://ecancer.org), that patients using non-Italian web sources perhaps already exploit. The Italian policy makers can take inspiration from these projects to implement in Italy an affective web communication for rare oncological patients.

Because of the small sample size, our empirical findings have to be considered preliminary and interpreted cautiously. Notwithstanding, it is worth pointing to the validity of the approach adopted, that is, quantitative text analysis applied to qualitative interviews. Moreover, this work represents a first attempt to construct the basic vocabulary for the textual analysis on the topic of rare cancer communication needs, and our findings could be a useful starting point for future researches.

Even examining such a small number of interviews has allows (1) detecting completely different behaviors, in terms of propensity toward information seeking; (2) linking these behaviors to different types of sources of information and of tools used; and (3) identifying different combinations of behaviors, sources, and tools in different phases of the disease.

We intend to extend this introductory study following two main directions. First of all, we are planning to assess the evidence coming from this work and to administer a structured questionnaire to a wider sample of sarcoma patients.

In the future, we would like extend our sample of textual data by taking advantage of the increasing availability of this kind of data on the Internet. In particular, we are interested in monitoring social media specifically devoted to supporting oncological patients. We believe that informal communication analyzed using text mining techniques can become a rich and up-to-date source of information on patients' communication requirements.

Acknowledgements

The authors would like to thank Prof. Paolo Casali and his research team for the insightful comments. Special thanks also go to Dr. Laura Gangeri who conducted the interviews. This research has received financial support from the Italian Health Ministry under the "Ricerca Finalizzata 2009" scheme (RF- 2009-1551083). The usual disclaimer applies.

References

1. Verdecchia A, Francisci S, Brenner H, Gatta G, Micheli A, Mangone L, Kunkler I, EUROCARE-4 Working Group. Recent cancer survival in *Europe*: *a 2000–02 period analysis of EUROCARE-4 data*. *The Lancet Oncology* 2007; **8**(9):784–796. doi:10.1016/S1470-2045(07)70246-2.
2. Kuijpers W, Groen WG, Aaronson NK, van Harten WH. A systematic review of web-based interventions for patient empowerment and physical activity in chronic diseases: relevance for cancer survivors. *Journal of Medical Internet Research* 2013; **15**(2). doi:10.2196/jmir.2281.
3. McCorkle R, Ercolano E, Lazenby M, Schulman-Green D, Schilling LS, Lorig K, Wagner EH. Self-management: enabling and empowering patients living with cancer as a chronic illness. *CA: A Cancer Journal for Clinicians* 2011; **61**:50–62. doi:10.3322/caac.20093.
4. Anderson RM, Funnell MM. Patient empowerment: myths and misconceptions. *Patient Education and Counseling* 2010; **79**(3):277–282. doi:10.1016/j.pec.2009.07.025.
5. Samoocha D, Bruinvels DJ, Elbers NA, Anema JR, van der Beek AJ. Effectiveness of web-based interventions on patient empowerment: a systematic review and meta-analysis. *Journal of Medical Internet Research* 2010; **12**(2). doi:10.2196/jmir.1286.
6. Zebrack BJ. Cancer survivor identity and quality of life. *Cancer Practice* 2000; **8**(5):238–242.
7. Warren E *et al.* Do cancer-specific websites meet patient's information needs? *Patient Education and Counseling* 2014; **95**(1):126–136. doi:10.1016/j.pec.2013.12.013.
8. Zachariae R, Pedersen CG, Jensen AB, Ehrnrooth E, Rossen PB, Von Der Maase H. Association of perceived physician communication style with patient satisfaction, distress, cancer-related self-efficacy, and perceived control over the disease. *British Journal of Cancer* 2003; **88**(5):658–665. doi:10.1038/sj.bjc.6600798.
9. Murthy D, Gross A, Oliveira D. Understanding cancer-based networks in Twitter using social network analysis. In Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on, 2011, September; 559–566, IEEE.

10. Fox S, Duggan M. Health online 2013, 2013, Health.
11. Brombacher A, Hopma E, Ittoo A, Lu Y, Luyk I, Maruster L, …, Wortmann H. Improving product quality and reliability with customer experience data. *Quality and Reliability Engineering International* 2012; **28**(8):873–886. doi: 10.1002/qre.1277
12. Menon R, Tong LH, Sathiyakeerthi S, Brombacher A, Leong C. The needs and benefits of applying textual data mining within the product development process. *Quality and Reliability Engineering International* 2004; **20**(1):1–15. doi:10.1002/qre.536.
13. Zhang K, Narayanan R, Choudhary A. Voice of the customers: mining online customer reviews for product feature-based ranking. In Proceedings of the 3rd Conference on Online Social Networks, 2010. USENIX Association.
14. Jenkins V, Fallowfield L, Saul J. Information needs of patients with cancer: results from a large study in UK cancer centres. *British Journal of Cancer* 2001; **84**(1):48. doi:10.1054/bjoc.2000.1573.
15. Powell J, Inglis N, Ronnie J, Large S. The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study. *Journal of Medical Internet Research* 2011; **13**(1). doi:10.2196/jmir.1600.
16. Rutten LJF, Arora NK, Bakos AD, Aziz N, Rowland J. Information needs and sources of information among cancer patients: a systematic review of research (1980–2003). *Patient Education and Counseling* 2005; **57**(3):250–261.
17. Bianchet K, Salvador M, Michilin N, Ciolfi L, Giacomello E, De Paoli P, Truccolo I. Il parere dei pazienti circa la qualitá del materiale informativo di carattere divulgativo: un'indagine in campo oncologico. *AIDAInformazioni: rivista di Scienze dell'informazione* 2005; **23**(3); 13–26.
18. Cox J, Dale BG. Key quality factors in web site design and use: an examination. *International Journal of Quality & Reliability Management* 2002; **19**(7):862–888.
19. Belanche D, Casaló LV, Guinalíu M. Website usability, consumer satisfaction and the intention to use a website: the moderating effect of perceived risk. *Journal of Retailing and Consumer Services* 2012; **19**(1):124–132. doi:10.1016/j.jretconser.2011.11.001.
20. Alford P, Duan Y, Taylor J. An analysis of the key factors affecting the success of a re-launched destination marketing website in the UK. In In Information and Communication Technologies in Tourism 2014. Springer International Publishing: Berlin, 2013; 637–649.
21. Bosch-Sijtsema P, Bosch J. User involvement throughout the innovation process in high-tech industries. *Journal of Product Innovation Management* 2014. doi:10.1111/jpim.12233.
22. Van Kleef E, van Trijp H, Luning P. Consumer research in the early stages of new product development: a critical review of methods and techniques. *Food Quality and Preference* 2005; **16**(3):181–201. doi:10.1016/j.foodqual.2004.05.012.
23. Coussement K, Van den Poel D. Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems* 2008; **44**(4):870–882.
24. Liberati C, Camillo F. Subjective Business Polarization: Sentiment Analysis meets Predictive Modeling, in New Trends in Database and Information Systems, Studies in Advances in Intelligent Systems and Computing, Catania B. et al. (Eds.), Springer-Verlag, 2014; 329–338.
25. Liberati C, Camillo F. Discovering hidden concepts in predictive models for texts' polarization. *International Journal of Data Warehousing and Mining* 2015; **11**(4):29–48.
26. Lebart L, Salem A, Berry L. Exploring Textual Data. Kluwer Academic Publisher: Dordrecht, 1998.
27. Greenacre M. Theory and Applications of Correspondence Analysis. Academic Press: NewYork, 1984.
28. Berelson B, Lazarsfeld PF. The Analysis of Communications Content. University of Chicago and Columbia University: Chicago and New York, 1948.
29. Berelson B. Content Analysis in Communication Research. The Free Press: Glencoe, IL, 1952.
30. Weber RP. Basic Content Analysis. Sage: Beverly Hills, 1985.
31. McKevitt P, Partridge D, Wilks Y. Approaches to natural language discourse processing. *Artificial Intelligence Review* 1992; **6**:333–364.
32. Indurkhya N, Damerau FJ. Handbook of Natural Language Processing. Chapman and Hall: Boca Raton FL, 2010.
33. Benzécri J. Analyse des Donnèes. Dunod: Paris, 1973.
34. Benzécri JP. Pratique de l'analyse des donnees, T. 3. Linguistique & Lexicologie: Dunod, Paris, 1981.
35. Lebart L, Morineau A, Piron M. Statistique exploratoire multidimensionnelle. Dunod: Paris, 1998.
36. Escofier B, Pagès J. Analyses factorielles simples et multiples; objectifs, méthodes et interprétation. Dunod: Paris, 1988.
37. Bolasco S. Analisi multidimensionale dei dati: metodi, strategie e criteri d'interpretazione. Carocci: Roma, 1999.
38. Jambu M. Exploratory and Multivariate Data Analysis. Academic Press Inc: San Diego, 1991.

Appendix

Table A1. Inertia and chi-square decomposition

Singular value	Principal inertia	Chi-square	Percent	Cumulative percent
0.64	0.41	684.87	41.05	41.05
0.37	0.13	226.85	13.60	54.64
0.34	0.11	193.69	11.61	66.25
0.27	0.07	125.16	7.50	73.75
0.26	0.07	117.35	7.03	80.78
0.22	0.05	81.89	4.91	85.69
0.22	0.05	78.10	4.68	90.37
0.18	0.03	54.87	3.29	93.66
0.17	0.03	46.56	2.79	96.45
0.15	0.02	37.28	2.23	98.68
0.11	0.01	21.95	1.32	100.00
Totale	0.99	1668.56	100.00	
Degrees of Freedom = 407				

Table A2. Row coordinates

Term	Dim1	Dim2	Dim3	Dim4
Doctor	0.521	−0.118	−0.107	−0.197
Internet	−0.707	−0.160	0.037	−0.071
state_of_mind	0.612	0.039	−0.718	0.188
to_search	−0.091	−0.164	−0.021	−0.125
medical_test	0.870	0.043	−0.600	−0.547
to_say	0.662	−0.040	0.154	−0.154
simple_info	−0.948	−0.314	−0.121	−0.064
same_experiences	−0.436	0.801	0.014	0.015
information	−0.638	0.238	0.044	−0.020
technical_information	−0.792	−0.425	−0.196	0.188
care_commuting	0.696	−0.072	1.044	−0.156
what_is_this	0.211	0.104	0.241	−0.058
sarcoma	0.468	0.174	−0.342	0.673
Surgery	0.684	−0.018	0.201	−0.114
specialized_center	0.277	0.093	0.509	0.251
Famiily	0.635	−0.165	0.307	−0.130
research	−0.969	0.669	−0.032	−0.032
Cancer	0.428	0.397	0.092	−0.014
large_amount_info	−0.692	−0.904	−0.122	0.079
to_find	−0.192	0.039	0.252	−0.023
to_talk	0.111	−0.041	−0.163	0.173
foreign_info	−1.116	−0.817	−0.095	−0.266
News	−1.102	0.371	−0.136	−0.104
to_read	−0.505	−0.628	−0.102	0.115
no_more_info	0.693	0.050	−0.071	0.827
no_internet	0.215	−0.035	0.523	0.558
to_direct	0.528	−0.025	0.218	0.165
how_to_treat	0.013	−0.589	−0.102	0.653
recurrence_of_disease	0.198	0.382	0.294	0.572
web_supervision	−1.079	0.276	−0.048	−0.128
to_understand	−0.223	0.302	0.462	0.176
disease_development	−1.060	1.161	−0.004	−0.083
to_explain	−0.469	−0.327	0.402	−0.013
treatment	0.467	0.360	0.168	−0.226
what_to_do	−0.010	−0.327	0.356	−0.237
chance_of_survival	−0.057	−0.250	0.034	0.566
Support	−0.811	1.280	−0.326	−0.274
to_ask	0.883	0.166	−0.559	0.590

Table A3. Absolute contributions to inertia for the row-points

Term	Contr1 (%)	Contr2 (%)	Contr3 (%)	Contr4 (%)
Doctor	7.25	1.12	1.09	5.68
Internet	10.30	1.59	0.10	0.57
state_of_mind	4.44	0.05	21.58	2.29
to_search	0.09	0.90	0.02	0.95
medical_test	8.29	0.06	13.95	17.95
to_say	4.73	0.05	0.90	1.41
simple_info	9.05	2.99	0.52	0.23
same_experiences	1.58	16.12	0.01	0.01
information	3.39	1.43	0.06	0.02
technical_information	5.04	4.39	1.09	1.56
care_commuting	3.89	0.13	30.92	1.07
what_is_this	0.35	0.26	1.62	0.14
sarcoma	1.57	0.66	2.96	17.75

(Continues)

Table A3. (Continued)

Term	Contr1 (%)	Contr2 (%)	Contr3 (%)	Contr4 (%)
Surgery	3.28	0.01	1.00	0.50
specialized_center	0.53	0.18	6.29	2.37
Famiily	2.53	0.52	2.09	0.58
research	5.90	8.49	0.02	0.03
Cancer	1.10	2.84	0.18	0.01
large_amount_info	2.30	11.90	0.25	0.17
to_find	0.18	0.02	1.08	0.01
to_talk	0.05	0.02	0.37	0.65
foreign_info	4.72	7.65	0.12	1.47
News	4.43	1.52	0.24	0.22
to_read	0.90	4.18	0.13	0.25
no_more_info	1.68	0.03	0.06	13.11
no_internet	0.15	0.01	3.24	5.73
to_direct	0.89	0.01	0.54	0.48
how_to_treat	0.00	3.21	0.11	7.15
recurrence_of_disease	0.12	1.35	0.94	5.49
web_supervision	3.40	0.67	0.02	0.26
to_understand	0.14	0.76	2.10	0.47
disease_development	3.12	11.29	0.00	0.11
to_explain	0.58	0.85	1.50	0.00
treatment	0.57	1.03	0.26	0.74
what_to_do	0.00	0.80	1.11	0.76
chance_of_survival	0.01	0.47	0.01	4.35
Support	1.63	12.28	0.93	1.02
to_ask	1.82	0.19	2.58	4.45

Authors' biographies

Rosa Falotico received her PhD in Statistics (2011) from Università degli Studi di Milano-Bicocca, Italy, where she is currently a postdoctoral researcher. Her research interests include textual analysis, health management, and pharmaceutical market.

Caterina Liberati received her PhD in Statistics in 2006 from the University of Bologna, Italy. Since 2010, she is assistant professor in Statistics for Economy at the University of Milano-Bicocca, where she collaborates with Bicocca Applied Statistics Center (B-ASC). Her current research interests include all the aspects related to business analytics and applied statistics, in particular, multidimensional modeling for business applications and dynamic patterns.

Paola Zappa is a postdoctoral fellow at the faculty of Economics and at the Social Network Analysis Research Center, University of Italian Switzerland (CH). Her research interests lie in the domains of Management and Organizational Theory, with a particular emphasis on collaboration and communication dynamics in healthcare.