Cancer Stage Prediction Based on Patient Online Discourse

Mukund Jha
Computer Science
Columbia University
New York, NY 10027
mj2472@columbia.edu

Noémie Elhadad
Biomedical Informatics
Columbia University
New York, NY 10032
noemie@dbmi.columbia.edu

Abstract

Forums and mailing lists dedicated to particular diseases are increasingly popular online. Automatically inferring the health status of a patient can be useful for both forum users and health researchers who study patients' online behaviors. In this paper, we focus on breast cancer forums and present a method to predict the stage of patients' cancers from their online discourse. We show that what the patients talk about (content-based features) and whom they interact with (social network-based features) provide complementary cues to predicting cancer stage and can be leveraged for better prediction. Our methods are extendable and can be applied to other tasks of acquiring contextual information about online health forum participants.

1 Introduction

In this paper we investigate an automated method of inferring the stage of a patient's breast cancer from discourse in an online forum. Such information can prove invaluable both for forum members, by enriching their use of this rapidly developing and increasingly popular medium, and for health researchers, by providing them with tools to quantify and better understand patient populations and how they behave online.

Patients with chronic diseases like diabetes or life-threatening conditions like breast cancer get a wealth of information from medical professionals about their diagnoses, test results, and treatment options, but such information is not always satisfactory or sufficient for patients. Much of that is essential to their everyday lives and the management of their condition escapes the clinical realm. Furthermore, patients feel informed and empowered by exchanging experiences and emotional support with others in the same circumstances. Thus, it is not surprising that patient communities have flourished on the Web over the past decade, through active disease-specific discussion forums and mailing lists.

For health professionals, this new medium presents exciting research avenues related to theories of psycho-social support and how patients manage their conditions. Qualitative analyses of forums and mailing list posts show that breast cancer patients and survivors provide and seek support to and from their peers and that support, while also emotional, is largely informational in nature (Civan and Pratt, 2007; Meier et al., 2007). Emotional support may include words of encouragement and prayers. Examples of informational support are providing personal experiences with a treatment, discussing new research, explaining a pathology report to a peer, as well as exchanging information pertinent to patients' daily lives, such as whether to shave one's head once chemotherapy starts.

Given the kinds of benefits that patients and survivors seek and provide in online forums, it seems likely that they would be inclined to gravitate toward others whose circumstances most closely resemble their own, beyond sharing the general diagnosis of breast cancer. In fact, focus groups and surveys conducted with breast cancer patients identified and emphasized the need for online cancer forum participants to identify other patients of a particular age, stage of illness, or having opted for similar treatment (Rozmovits and Ziebland, 2004; van Uden-Kraan et al., 2008).

The stage of a patient's cancer, in particular, can be a crucial proxy for finding those whose experiences are likely similar and relevant to one's own. For breast cancer, there are five high-level standard stages (0 to IV). While they do not give the whole picture about a particular cancer (the stages

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 64–71,
Uppsala, Sweden, 15 July 2010.

themselves can be described with finer granularity and they do no not encompass additional information like hormonal sensitivity), physicians have traditionally relied on them for prognosis and determining treatment options. For patients and survivors, they are a useful way to communicate to their peers their health status, as evidenced by the members' signatures on forums and mailing lists (Meier et al., 2007).

Although many forums provide pre-set profile fields for users to populate with important background information, such as the stage of their cancer (e.g., the popular forum on `breastcancer.org`), in practice, only a fraction of members have a complete profile. Thus, an automated way of inferring member profile information via the social network created by a forum's users would help fill in the blanks.

Beyond identifying other patients in a forum in similar circumstances, such a tool can have numerous practical benefits for both forum users and health researchers who study patients' online behavior. When a patient searches for a particular piece of information in a forum, incorporating contextual information about the user into the search mechanism can improve search results. For example, a search tool can rank higher the posts that were authored by patients with the same stage. For health researchers, questions which bring a better understanding of forum usage (i.e., "are patients with stage IV cancer more or less active in a forum than patients with early stage cancer") can be answered accurately only if all members of the forums are taken into account, not just the ones who filled out their member profiles. Furthermore, in the context of health communication, the more information is available about an individual, the more effective the message can be, from generic to personalized to targetted to tailored (Kreuter et al., 2000). Our research contributes an automated method to acquiring contextual information about forum participants. We focus on cancer stage as an exmple of context information.

Our research question is whether it is possible to predict the stage of individuals' cancer based on their online discourse. By discourse we mean both the information she conveys and whom she talks to in a forum. Following ethical guidelines in processing of patient data online, we focus on a popular breast cancer forum with a large number of participants (Eysenbach and Till, 2001). We show that the content of members' posts and the stage of their interlocutors can provide complementary clues to identifying cancer stages.

2 Related Work

Researchers have begun to explore the possibility of diagnosing patients based on their speech productions. Content analysis methods, which rely on patient speech transcripts or texts authored by patients, have been leveraged for understanding cancer coping mechanisms (Graves et al., 2005; Bantum and Owen, 2009), psychiatric diagnoses (Oxman et al., 1988; Elvevaag et al., 2010), and the analysis of suicide notes (Pestian et al., 2008). In all cases, results, while not fully accurate, are promising and show that patient-generated content is a valuable clue to diagnosis in an automated framework.

Our work departs from these experiments in that we do not attempt to predict the psychological state of a patient, but rather the status of a clinical condition. Staging breast cancer provides a way to summarize the status of the cancer based on clinical characteristics (the size of the tumor, whether the cancer is invasive or not, whether cancer cells are present in the lymph nodes, and whether the cancer has spread beyond the breast). There are five high-level stages for breast cancer. Stage 0 describes a non-invasive cancer. Stage I represents early stage of an invasive cancer, where the tumor size is less than 2 centimeters and no lymph nodes are involved (that is, the cancer has not spread outside of the breast). Stages II and III describe a cancer with larger tumor size and/or the cancer has spread outside of the breast. Stage IV describes a cancer that have metastasized to distant parts of the body, such as lungs and bones.

In our work, we analyze naturally occurring content, generated by patients talking to each other online. As such, our sample population is much larger than in earlier works (typically less than 100 subjects). Like the researchers who focus on content analysis, we rely on the content generated by patients, but we also hypothesize that whom the patients interact with can help the prediction of cancer stage.

In particular, we build a social network based on patients' interactions to boost text-based predictions. Graph-based methods are becoming increasingly popular in the NLP community, and similar approaches have been employed and

shown to perform well in other areas like question answering (Jurczyk, 2007) (Harabagiu et al., 2006), word-sense disambiguation (Niu et al., 2005), and textual entailment (Haghighi, 2005).

3 Methods

Our methods to predict cancer stage operate in a supervised framework. We cast the task of stage prediction as a 4-way classification (Stage I to IV). We hypothesize that the discourse of patients online, as defined by the content of their posts in a forum, can be leveraged to predict cancer stage. Furthermore, we hypothesize that the social network derived by whom patients interact with can provide an additional clue for stage detection.

We experimented with three methods of predicting cancer stage:

Text-based stage prediction A classifier is trained given the post history of a patient.

Network-based stage prediction A social network representing the interactions among forum members is built, and a label propagation algorithm is applied to infer the stage of individual patients.

Combined prediction A classifier which combines text-based and network-based features.

Next we describe each method in detail, along with our dataset and our experimental setup.

3.1 Data Collection and Preprocessing

We collected posts from the publicly available discussion board from `breastcancer.org`. It is a popular forum, with more than 60,000 registered members, and more than 50,000 threads discussed in 60 subforums. To collect our dataset, we crawled the content of the most popular subforums.[1]

Collected posts were translated from HTML into an XML format, keeping track of author id, thread id, position of the post in the thread, body of the post, and signature of the author (which is kept separated from the body of the post). The content of the posts was tokenized, lower-cased and stemmed. Images, URLs, and stop words were removed.

[1]There were 17 such subforums: "Just Diagnosed," "Help Me Get Through Treatment," "Surgery - Before, During, and After," "Chemotherapy - Before, During and After," "Radiation Therapy - Before, During and After," "Hormonal Therapy - Before, During and After," "Alternative, Complementary and Holistic Treatment," "Stage I and II Breast Cancer," "Just Diagnosed with a Recurrence or Metastasis," "Stage III Breast Cancer," "Stage IV Breast Cancer Survivors," "HER2/neu Positive Breast Cancer," "Deperession, Anxiety and Post Traumatic Stress Disorder," "Fitness and Getting Back in Shape," "Healthy Recipes for Everyday Living," "Recommend Your Resources," "Clinical Trials, Research, News, and Study Results."

Nb. of threads	26,160
Nb. of posts	524,247
Nb. of threads with < 20 posts	22,334
Nb. of users with profile Stage I	2,226
Nb. of users with profile Stage II	2,406
Nb. of users with profile Stage III	1,031
Nb. of users with profile Stage IV	749
Total Nb. of users with profile	6,412
Nb. of active users profiled Stage I	1,317
Nb. of active users profiled Stage II	1,400
Nb. of active users profiled Stage III	580
Nb. of active users profiled Stage IV	448
Total Nb. of active users with profile	3,745

Table 1: General statistics of the dataset.

To post in `breastcancer.org`, users must register. They have the option to enter a profile with pre-set fields related to their breast cancer diagnosis; in particular cancer stage between stage I and IV. We collected the list of members who entered their stage information, thereby providing us with an annotated set of patients with their corresponding cancer stage. Table 1 shows various statistics for our dataset. Active users are defined as members who have posted more than 50 words overall in the forums. Note the low number of user with profile information (approximately 10% of the overall number of registered participants in the forum).

3.2 Text-Based Stage Prediction

We trained a text-based classifier relying on the full post history of each patient. The full post history was concatenated. Signature information, which is derived automatically from the patient's profile (and thus contains stage information) was removed from the posts. The classifier relied on unigrams and bigrams only. Table 2 shows statistics about post history length, measured as number of words authored by a forum member.

3.3 Network-Based Stage Prediction

We hypothesize that patients tend to interact in a forum with patients with similar stage. To test this

Stages	Min	Max	Average	Median
I	4	609,608	8,429	3,123
II	2	353,731	8,142	3,112
III	8	211,655	9,297	3,189
IV	10	893,326	17,083	326

Table 2: Statistics about number of words in post history.

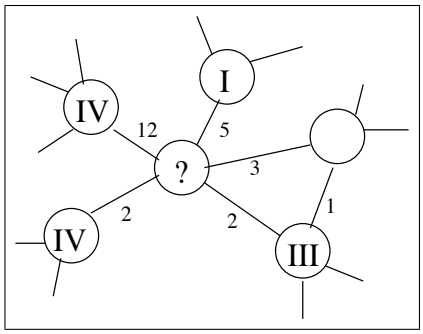

Figure 1: Nodes in the social network of forum member interaction.

hypothesis, we represent the interactions of the patients as a social network. The nodes in the network represent patients, and an edge is present between two nodes if the patients interact with each other, that is they are part of the same threads often. Weights on edges represent the degree of interaction. Higher weight on an edge between two forum members indicates they interact more often. More precisely, we build an undirected, weighted network, where the nodes representing training instances are labeled with their provided stage information and their labels are fixed. Figure 1 shows an example of node and its immediate neighbors in the network. Of his five neighbors, four represent training instances and have a fixed stage, and one represents a user with an unknown stage.

A label propagation algorithm is applied to the network, so that every node in the network is assigned a stage between I and IV (Raghavan et al., 2007). Given a node and its immediate neighbors, it looks for the most frequent labels, taking into account the edge weights. In our example, the propagated label for the central node will be stage IV. This label, in turn, will be used to assign a label to the other nodes. When building the social network of interactions, we experimented with the following parameters.

Nodes in the network. We experimented with including all the forum members who participated in a conversation thread. Thus, it includes all the members, even the ones without a known cancer stage. This resulted in a network of 15,035 forum participants. This way, the network covers more interactions among more users, but is very sparse in its initial labeling (only the training instances in the dataset of active members with a known label are labeled). The label propagation algorithm assigns labels to all the nodes, but we test its accuracy only on the test instances. We also experimented with including only the patients in the training and testing sets, thereby reducing the size of the network but also decreasing the sparsity of the labeling. This resulted in a network of 3,305 nodes.[2]

Drawing edges in the network. An edge between two users indicate they are frequently interacting. One crude way is to draw an edge between every user participating in the same thread, this however does not provide an accurate picture and hence does not yield good results. In our approach we draw an edge in two steps. First, since threads are often long and can span over multiple topics, we only draw an edge if the two individuals' posts are within five posts of each other in the thread. Second, we then look for any direct references made by a user to another user in their post. In forum threads, users usually make a direct reference by either by explicitly referring to each other using their real name or internet aliases or by quoting each other, i.e., repeating or stating what the other user has mentioned in her post. For example in *"Hey Dana, I went through the same thing the first time I went to my doctor..."*, the author of the post is referring to another user with name *'Dana'*. We rely on such explicit references to build accurate graph.[3] To find direct explicit references, we search in every post of a thread for any mention of names (real or aliases) of users participating in the thread and if one is found we draw an edge between them.

We observed that users refer to each other very

[2]This number of nodes is less than the numbers of overall active members in our gold standard because some active members have either posted in threads with only one post or with more than 20 posts.

[3]An alternative approach is to identify quotes in posts. In our particular dataset, quotes did not occur often, and thus were ignored when assessing the degree of interaction between two forum members.

frequently using their real names instead of internet names (which are long and often arbitrary). These are often hard to detect because no data is present which link users' forum aliases to their real name. We use following approach to extract real names of the users.

Extracting real names. For every user, we extract the last ten words (signature) from every post posted by the user and concatenate them after removing all stop words and other common signature terms (like thanks, all the best, love, good luck etc.) using a pre-compiled list. We then mine for the most frequent name occurring in the concatenated text using standard list of names and extracting capitalized words. We also experimented with using Named Entity Recognizers, but our simple rule based name extractor gave us better results with higher precision. Finally, we map the extracted real name with the user's alias and utilize them to find direct references between posts.

Weights Computation. The weight of an edge between two nodes represents the degree of interaction between two corresponding users (the more often they communicate, the higher the weight). Since the label propagation algorithm takes into account the weighted frequency of neighboring nodes, these weights are crucial. We compute the weights in following manner: for each pair of users with an existing edge (as determined above), we iterate through their posts in common threads, and add the cosine similarity score between the two posts to the weight of the edge. For edges made through direct references we add the highest cosine similarity score between any two pair of posts in that particular thread. This way we weigh higher the edges made through direct reference as we are more confident about them.

The full network of all users (15,035 nodes) had 480,051 edges, and the restricted network of dataset users (3,305 nodes) had 28,152 edges.

3.4 Combining Text-Based and Network-Based Predictions

To test the hypothesis that text-based and network-based predictions model different aspects of patients and thus provide complementary cues to stage prediction, we trained a classifier which incorporates text-based and network-based features.

The combined classifier contained the following features: text-based predicted label, confidence score of the text-based prediction, network-based predicted label, percentage of immediate neighbors in the network with a stage I label, stage II, III and IV labels (neighbors in the network with no labels do not contribute to the counts). For instance, the central node in Figure 1 is assigned the feature values 1/4, 0, 1/4 and 1/2 for the ratio of stage I, II, III and IV neighbors.

3.5 Experimental Setup

Our dataset for the three models consisted of the 3,745 active members. For all the models, we follow a five-fold stratified cross validation scheme. The text-based classification was carried out with BoosTexter (Schapire and Singer, 2000), trained with 800 rounds of boosting. The label propagation on the social network was carried out in R.[4] The final decision-tree classification was carried out in Weka, relying on an SVM classifier with default parameters (Hall et al., 2009).

4 Results

Table 3 shows the results of the text-based prediction, the network-based prediction and the combined prediction for each stage measured by Precision, Recall and F-measure. For comparison, we report on the results of a baseline text-based prediction. The baseline prediction assigns a stage based on the explicit mention of stage in the post history of a patient. In practice, it is a rule-based prediction with matching against the pattern "`stage [IV|four|4]`" for stage IV prediction, and similarly for other stages. The text-based prediction yields better results than the baseline, with a marked improvement for each stage.

The network-based prediction performs only slightly worse than the text-based predictions. The hypothesis that whom the patient interacts with in the forums helps predict stage holds. To verify this point further, we computed for each stage the average ratio of neighbors per stage based on the social network of interactions, as shown in Figure 2. For instance, stage IV patients interact mostly with their peers (49% of their posts are shared with other stage IV users), and to some extent with other patients (18% of their posts with stage I patients, 20% with stage II patients, and 13% with stage III patients). Except for stage III patients, all other patients are mostly interacting with similarly staged patients.

[4] `www.r-project.org`

Baseline				Text Based			
Stage	Precision	Recall	F	Stage	Precision	Recall	F
I	76.2	26.4	39.3	I	54.9	63.9	59.1
II	79.4	18.7	30.3	II	51.6	55.0	53.2
III	76.6	35.0	48.0	III	52.7	30.3	38.5
IV	76.4	50.7	60.9	IV	82.5	71.2	76.4
Network Based				Combined			
Stage	Precision	Recall	F	Stage	Precision	Recall	F
I	50.4	56.7	53.4	I	57.1	65.4	61.0
II	49.6	49.1	49.3	II	56.6	53.5	55.0
III	65.7	27.7	39.0	III	56.1	48.3	51.9
IV	59.3	83.7	69.4	IV	84.7	81.3	83.0

Table 3: Stage prediction results (Precision, Recall, and F-measure).

When combining the text-based and the network-based predictions in an overall classifier the prediction yields the best results. These results confirm the potential in combining the two facets of patient discourse, content and social interaction.

The results presented in the table correspond to a network built with the full set of users, including those without any profile information. When restricting the network on the patients with stage labels only, we obtained similar results (F-measures of 56% for stage I, 52% for stage II, 43% for stage III, and 79% for stage IV). This shows that it is worth modeling the full set of interactions and the full network structure, even when a large number of nodes have missing labels.

Finally, we also experimented with building networks with no weights or with weights without the 5-post-apart restriction. In both cases, the results of the network-based and combined predictions are lower than those presented in Table 3. We interpret this fact as a confirmation that our edge weighting strategy models to a promising extent the degree of interaction among patients.

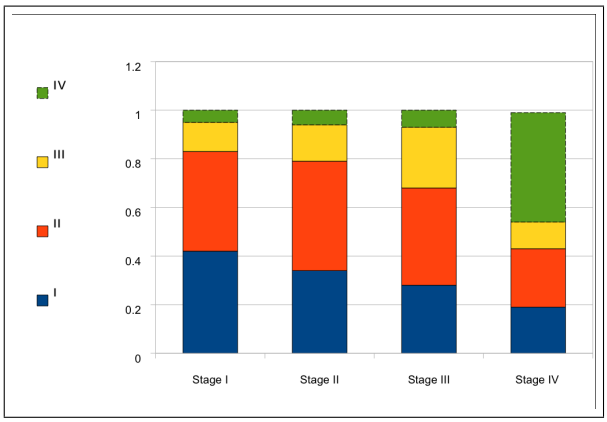

Figure 2: Distribution of stage-wise interactions.

5 Discussion

Text-based prediction. Results confirm that cancer stage can be predicted by a patient's online discourse. When examining the unigrams and bigrams picked up by the classifier as predictive of stage, we can get a sense of the frequent topics of discussion of patients. For instance, the phrases "tumor mm" (referring to tumor size in millimeters) and "breast radiation" were highly predictive of stage I patients. The words "hat" and "hair" were highly predictive of stages II and III, while stage IV patients were predicted by the presence of the phrases "bone met." (which stands for bone metastasis), "met lung" "liver," and "lymphedema" (which is a side effect of cancer treatment linked to the removal of lymph nodes and tumor).

Figure 3 shows the overall accuracy of the text-based classifier, when tested against the amount of text available for the classification. As expected, the longer the post history, the more accurate the classification.

Representing degree of interaction among patients. In our experiments, we observed that the weigthing scheme of edges had a strong impact on the overall accuracy of stage prediction. The more interaction was modeled (through distance in thread and identification of explicit references), the better the results. This confirms the hypothesis that dialogue is helpful in predicting cancer stage, and emphasizes the need for accurate techniques

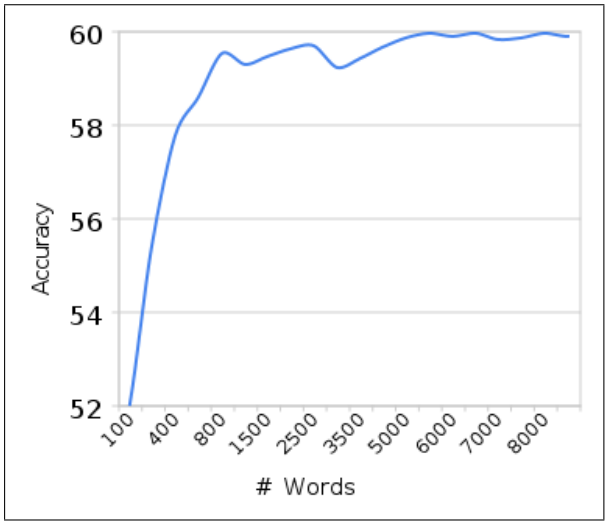

Figure 3: Overall text-based prediction accuracy against post history length.

to model interaction among forum participants in a social network.

Discourse of Stage IV patients. Both the text-based and the network-based predictions provide higher precision and recall for the stage IV patients. This is emphasized by Figure 2, where we see that, in our dataset, stage IV patients talk mostly to each other. These results suggest that stage IV patients have particular discourse, which separates them from other patients. This presents interesting avenues for future investigation.

6 Future Work and Conclusion

In this paper, we investigated breast cancer stage prediction based on the online discourse of patients participating in a breast cancer-specific forum. We show that relying on lexical features derived from the content of the posts of a patient provides promising classification results. Furthermore, even a simple social network representing patient interactions on a forum, yields predictions with comparable results. Combining the two approaches boosts results, as content and interaction seem to model complementary aspects of patient discourse.

Our experiments show that stage IV patients appear to exhibit specific textual and social patterns in forums. This point can prove useful to health researchers who want to quantify patient behaviors online.

The strategy of combining two facets of discourse (content and interactions) introduces several interesting research questions. In the future, we plan to investigate some of them. In a first step, we plan to better model the interactions of patients online. For instance, we would like to analyze the content of the posts to determine further if two patients are in direct communication, and the domain of their exchange (e.g., clinical vs. day-to-day vs. emotional). As we have observed that the way edges in the network are weighted has an impact on overall performance, we could then investigate whether the domain(s) of interaction among users (clinical matters vs. emotional and instrumental matters for instance) has an impact on predicting cancer stage by taking the different domains of interaction in account in the weight computation.

Finally, this work relies on a single, yet highly active and popular, forum. We would like to test our results on different breast cancer forums, but also on other disease-specific forums, where patients can be separated in clinically relevant groups.

Acknowledgments

We thank Phani Nivarthi for his help on data collection. This work is supported in part by a Google Research Award. Any opinions, findings, or conclusions are those of the authors, and do not necessarily reflect the views of the funding organization.

References

Erin Bantum and Jason Owen. 2009. Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment*, 21(1):79–88.

Andrea Civan and Wanda Pratt. 2007. Threading together patient expertise. In *Proceedings of the AMIA Annual Symposium*, pages 140–144.

Brita Elvevaag, Peter Foltz, Mark Rosenstein, and Lynn DeLisi. 2010. An automated method to analyze language use in patients with schizophrenia and their first degree-relatives. *Journal of Neurolinguistics*, 23:270–284.

Gunther Eysenbach and James Till. 2001. Ethical issues in qualitative research on internet communities. *BMJ*, 323:1103–1105.

Kristi Graves, John Schmidt, Julie Bollmer, Michele Fejfar, Shelby Langer, Lee Blonder, and Michael Andrykowski. 2005. Emotional expression and emotional recognition in breast cancer survivors: A controlled comparison. *Psychology and Health*, 20(5):579–595.

Aria Haghighi. 2005. Robust textual inference via graph matching. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'05*, pages 387–394.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Sanda Harabagiu, Finley Lacatusu, and Andrew Hickl. 2006. Answering complex questions with random walk models. In *Proceedings of SIGIR Conference (SIGIR'06)*, pages 220–227.

Pawel Jurczyk. 2007. Discovering authorities in question answer communities using link analysis. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM'07)*.

Matthew Kreuter, David Farrell, Laura Olevitch, and Laura Brennan. 2000. *Tailoring health messages: customizing communication using computer technology*. Lawrence Erlbaum Associates.

Andrea Meier, Elizabeth Lyons, Gilles Frydman, Michael Forlenza, and Barbara Rimer. 2007. How cancer survivors provide support on cancer-related internet mailing lists. *Journal of Medical Internet Research*, 9(2):e12.

Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. 2005. Word sense disambiguation using label propagation based semi-supervised learning. In *Proceedings of the ACL Conference (ACL'05)*, pages 395–402.

Thomas Oxman, Stanley Rosenberg, Paula Schnurr, and Gary Tucker. 1988. Diagnostic classification through content analysis of patient speech. *American Joural of Psychatry*, 145:464–468.

John Pestian, Pawel Matykiewicz, Jacqueline Grupp-Phelan, Sarah Arszman Lavanier, Jennifer Combs, and Robert Kowatch. 2008. Using natural language processing to classify suicide notes. In *Proceedings of BioNLP'08*, pages 96–97.

Usha Raghavan, Reka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physics Review*, page E 76 036106.

Linda Rozmovits and Sue Ziebland. 2004. What do patients with prostate or breast cancer want from an Internet site? a qualitative study of information needs. *Patient Education and Counseling*, 53:57–64.

Robert Schapire and Yoram Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

Cornelia van Uden-Kraan, Constance Drossaert, Erik Tall, Bret Shaw, Erwin Seydel, and Mart van de Laar. 2008. Empowering processes and outcomes of participation in online support groups for patients with breast cancer, arthritis, or fibromyalgia. *Qualitative Health Research*, 18(3):405–417.