Published in final edited form as:

Qual Life Res. 2016 March ; 25(3): 547–557. doi:10.1007/s11136-015-1156-7.

Symptom clusters in women with breast cancer: an analysis of data from social media and a research study

Sarah A. Marshall[1], **Christopher C. Yang**[2], **Qing Ping**[2], **Mengnan Zhao**[2], **Nancy E. Avis**[3], and **Edward H. Ip**[1,3]

Edward H. Ip: eip@wakehealth.edu

[1] Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA

[2] College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA

[3] Department of Social Sciences and Health Policy, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA

Abstract

Purpose—User-generated content on social media sites, such as health-related online forums, offers researchers a tantalizing amount of information, but concerns regarding scientific application of such data remain. This paper compares and contrasts symptom cluster patterns derived from messages on a breast cancer forum with those from a symptom checklist completed by breast cancer survivors participating in a research study.

Methods—Over 50,000 messages generated by 12,991 users of the breast cancer forum on MedHelp.org were transformed into a standard form and examined for the co-occurrence of 25 symptoms. The k-medoid clustering method was used to determine appropriate placement of symptoms within clusters. Findings were compared with a similar analysis of a symptom checklist administered to 653 breast cancer survivors participating in a research study.

Results—The following clusters were identified using forum data: menopausal/psychological, pain/fatigue, gastrointestinal, and miscellaneous. Study data generated the clusters: menopausal, pain, fatigue/sleep/gastrointestinal, psychological, and increased weight/appetite. Although the clusters are somewhat different, many symptoms that clustered together in the social media analysis remained together in the analysis of the study participants. Density of connections between symptoms, as reflected by rates of co-occurrence and similarity, was higher in the study data.

Correspondence to: Edward H. Ip, eip@wakehealth.edu.

Compliance with ethical standards

Conflict of interest All authors declare they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the research study of breast cancer survivors. Permission to extract data from the forum MedHelp.com was granted by site managers, and all data from this source were collected anonymously.

Conclusions—The copious amount of data generated by social media outlets can augment findings from traditional data sources. When different sources of information are combined, areas of overlap and discrepancy can be detected, perhaps giving researchers a more accurate picture of reality. However, data derived from social media must be used carefully and with understanding of its limitations.

Keywords

Social media; Online forum; MedHelp; Text mining; Symptom cluster; Breast cancer

Introduction

In recent years social media has profoundly transformed the way millions of people communicate and interact on a daily basis. Social media consists of a variety of Internet-based platforms and networks, such as blogs, forums, and virtual communities, which allow for the exchange of user-generated content in a variety of formats [1, 2]. An increasing percentage of Americans, upwards of 74.4 %, reported household Internet use in 2013 [3], and 71 % of those persons used social media [4]. The most popular social media sites are Facebook and Twitter, which generate 2.7 billion “likes” per day and 500 million tweets per day, respectively [2]. A variety of social media outlets cater specifically to users with health-related concerns. One such forum, MedHelp.org (previously MedHelp.com), hosts a plethora of interactive discussion boards centered on various diseases and illnesses, including breast cancer. MedHelp.org offers user support in the form of information, commiseration, and camaraderie [5]. In 2014 the site attracted over twelve million unique visitors [5].

The use of social media as a data source for health care research has burgeoned in recent years [6]. Data derived from social media has been used to track the spread of infectious diseases such as influenza or foodborne illness [7, 8], detect adverse reactions to medication as part of broader pharmacovigilance strategies [9–11], and better understand “hard-to-reach” populations such as the homeless [2, 6, 12]. Postings on social media have been analyzed to identify individuals at risk for depression [13], and patterns of usage have been studied in relationship to mental health [14]. In the future an even broader set of applications for using social media data is likely to be developed and refined.

Using data derived from social media offers researchers certain advantages over traditional sources of information such as clinical trials. Vast quantities of data can be extracted with relative ease and low cost in a short period of time [15–17], and as the percentage of Internet users approaches 100, a certain degree of ecological validity is assured. Using data from social media may bypass the need for expensive and time-consuming recruitment of subjects and augment study of persons who might otherwise be difficult to identify (i.e., members of an affinity group linked by an uncommon disease) [17, 18]. The anonymity of the Internet may facilitate discussion of “taboo” subjects such as mental illness and drug use [6]. Such data may also be free from bias introduced by formal studies and afford insights that would otherwise be hard to capture.

Despite these benefits, concerns regarding social media data remain [15, 17, 19, 20]. Internet usage is increasing in broad segments of the population [21], but generalizability is

hampered by the fact that users still tend to be younger and better educated than non-users [22, 23]. Information presented online may not comport to the quality or format desired by researchers. Use of abbreviations and typographical errors appears to be common [20]. Information may be inaccurate and impossible to verify, and important details may be missing. Ethical concerns related to privacy of users, confidentiality, and informed consent also persist [1, 24].

Given the large volume of unstructured and text-based data inherent in social media, computer algorithms are typically required for the efficient extraction of information. Text mining is defined as “the process of extracting meaningful information from large amounts of unstructured texts using computational methods” (p. 777) [19]. Simply using keywords to classify data has been used widely in the past, but such an approach may fail to recognize words that lack formal definitions or have multiple meanings [20, 23]. Machine learning algorithms with adaptive natural language processing may help circumvent such limitations. An overview of text mining techniques and strategies is presented by Harpaz et al. [19] and Taurob et al. [23].

Studies have shown that cancer patients, including women with breast cancer, may experience eight or more symptoms simultaneously [25, 26], and some of these symptoms may persist for years after diagnosis [27]. Moreover, the total symptom burden experienced by individuals has been shown to be negatively correlated with quality of life [28, 29]. For example, a study of 240 Veterans Affairs patients with mixed cancer types showed that each additional symptom predicted a significant decline in quality of life as measured by the Fact-G Sum Quality of Life scale ($p < 0.001$) [25].

Strategies to identify and manage symptoms should consider how symptoms are connected and influence each other. One approach to further this aim has been the use of symptom clusters. Symptom clusters in oncology were initially defined as “three or more concurrent symptoms [that] are related to each other” (p. 465) [30]. In 2005 this definition was revised to allow for the presence of only two symptoms in a cluster, with the authors further stating that symptom clusters should be stable and independent, and symptoms within a cluster should be more strongly related to each other than with symptoms outside the cluster [31].

A variety of clinical and statistical approaches have been used to identify symptom clusters in the literature [32–34]. Examples of such methods include hierarchical cluster analysis, principal component analysis, and explorative factor analysis among others [35–39]. Some of the clusters postulated to exist in breast cancer include fatigue and sleep difficulties [40, 41], fatigue, pain, and disturbed sleep [42], pain, depression, and fatigue [43], depression, fatigue, and disturbed sleep [44], fatigue, perceived cognitive impairment, and mood problems [45], menopausal [46], psychoneurological and gastrointestinal (GI) [47], and chemotherapy-related and hormonal [48].

In this article we identify and examine symptom patterns generated by data extracted from a social media platform (MedHelp.org) intended for use by women with breast cancer. These women are known to have a strong presence on social media, and investigation into the symptom experience of women with breast cancer is an area of active research. We then

compare these findings to an analysis of symptoms reported by breast cancer survivors enrolled in a research study responding to a symptom checklist. We aim to compare symptom clusters derived from MedHelp.org and the research study in order to judge the reliability and usefulness of social media data.

Few studies to date have compared patient-reported outcomes collected from a social media platform with findings reported in a more traditional research study. One such investigation used social media to distribute a survey instrument to persons with a history of cerebral aneurysm, but did not examine the unstructured postings of users on a social media site [17]. Our study contributes to a growing literature regarding the use of user-reported data on social media platforms in scientific contexts as well as the field of symptom cluster research.

Methods

Data sources

Two sources of data were used in this investigation: a research study of breast cancer survivors and an online health forum (MedHelp.org). The research study, described in detail elsewhere [49], included 653 women ≥ age 25 diagnosed with stage I, II, and III breast cancer in the past eight months at two academic medical centers in the USA. All participants were required to be over the age of 18 and read and understand English. The eight-month time frame was chosen to better understand the early period of cancer treatment and management. Additional characteristics of study participants are shown in Table 1. Informed consent was obtained from all subjects.

Study participants completed self-administered questionnaires by mail that contained a variety of measures including a 39-item symptom checklist derived from the Women's Health Initiative [50]. For each symptom, women were asked to rate whether it occurred in the past month and if so, how bothersome it was. Response categories included: 1 = symptom did not occur, 2 = symptom occurred and was mild, 3 = symptom occurred and was moderate, and 4 = symptom occurred and was severe. Instructions stated that mild referred to a symptom that did not interfere with usual activities, moderate to a symptom that interfered somewhat with usual activities, and severe to a symptom that was so bothersome that usual activities could not be performed.

Out of these 39 symptoms listed on the initial survey, 25 symptoms were selected for inclusion in the present analyses. The rationale for choosing 25 symptoms out of the 39 was the enhancement of interpretability of the derived symptom clusters. The basis for choosing these 25 symptoms was a combination of factors such as wanting to include symptoms that are known to be especially common among cancer patients in general and breast cancer patients specifically, and to correspond better with symptom clusters that have already been reported in the literature. We elected to compare symptoms rated moderate or severe (T1) or severe only (T2) rather than include mild symptoms that did not impair daily functioning.

The other source of data used in this study was MedHelp.org, a Web site that hosts forums for various health-related concerns, including breast cancer. This forum allows users to ask questions, share information, and provide support to one another, which may lead to an

iterative discussion between users. Akin to other social media sites, users have the ability to create profiles, share personal information, and "friend" others. However, provision of basic information such as age, gender, and location is entirely optional, and clinical characteristics such as stage of cancer or time since diagnosis cannot be reliably ascertained. The site maintains records of threads, providing a copious amount of unstructured data thematically organized around a variety of topics. Forum data include rich contextual detail that can be explored at the level of individual posts and a voluminous amount of information when data from all users are combined.

With permission from MedHelp.org, 50,426 publicly available messages posted by 12,991 users in the breast cancer sub-forum from October 1, 2006 to September 21, 2014, were crawled. Each post or comment was transformed into a structured record with a user ID, timestamp, and message content. The presence or absence of the 25 pre-selected symptoms was noted for each post without consideration of severity.

In order to identify these symptoms, each message was parsed using a standardized consumer health vocabulary (CHV). The CHV groups each symptom with a variety of terms or phrases with similar meaning. For example, when searching the text for the standard symptom "mood changes," a variety of related terms such as "altered moods" will also be detected. Similarly, variations of the standard symptom such as "changes mood" or "change moods" are counted as well. This reduces the possibility of missing symptoms that are not described using a standard nomenclature [51, 52].

Data analysis

Symptoms were noted to co-occur when found in the same post on the social media forum or when both were endorsed as present on the study checklist. The rates of symptom occurrence and co-occurrence were used to calculate a similarity index as follows:

$$\text{Similarity}\left[\text{sympotom}\,(i)\,,\quad \text{symptom}\quad (j)\right]=\frac{\text{Co\_ occurrence}\left[\text{symptom}\,(i)\,,\quad \text{symptom}\,(j)\right]}{\sqrt{\text{Occurrence}\left[\text{symptom}\,(i)\right]\quad\times\quad\text{Occurrence}\left[\text{symptom}\,(j)\right]}}$$

(1)

Using this equation, a 25 × 25 matrix inclusive of the previously selected symptoms was derived for all possible symptom pairings among the group of research participants and forum members.

K-medoid clustering, a method of partitioning data similar to K-means clustering, was conducted using each matrix to determine the placement of individual symptoms within clusters. Compared to K-means clustering, the K-medoid clustering method assigns a member in a cluster with minimal overall cost as the new "centroid," or medoid, for the partitioning of the next iteration [53]. There are two advantages of using the K-medoid clustering method for the current application. First, it allows for the most cost efficient grouping of symptoms together in terms of distance from a specific "anchoring" symptom—

the medoid. Second, it improves the interpretability of the cluster solution. The maximum number of clusters allowed was set as five for both analyses. This number was based on a review of the literature on symptom clusters of breast cancer patients [40–47], and chosen to ensure symptoms would combine with other symptoms rather than form many small clusters with very few symptoms in each one. Limiting the number of clusters also helps prevent symptoms from failing to cluster altogether.

A random sample of postings crawled by the data mining algorithm was also inspected by hand to determine whether the algorithm was detecting symptoms accurately and that reported instances of symptom co-occurrence were genuine. In order to examine the symptom of restless sleep in more detail, an additional random sample of 100 posts containing the keyword “sleep” was read to determine whether sleep difficulties were present. Findings were compared to the results of the automated extraction of information to determine whether true instances of restless sleep were captured or whether postings were being misclassified as containing restless sleep when in fact it was not indicated.

Results

Rates of symptom co-occurrence and similarity scores

The most commonly reported symptom on the MedHelp forum was “general aches,” which was noted to occur 1091 times (2.2 % of posts). Fatigue was noted in 689 posts (1.4 %), followed by depressed mood in 508 posts (1.0 %). In contrast, the most common symptom of at least moderate severity in the research group (T1, moderate and severe symptoms) was fatigue, reported by 366 out of 653 persons (56.0 %). Fatigue was followed by restless sleep at 307 (47.0 %) and muscle pains at 171 (26.2 %) instances. The most common symptoms rated by study members as “severe” (T2, severe symptoms only) were fatigue (101, 15.5 %), restless sleep (75, 11.5 %), and muscle pain (52, 8.0 %).

The three most commonly co-occurring symptoms on MedHelp were general aches-headaches, fatigue-general aches, and fatigue-nausea, occurring together in 0.3, 0.2, and 0.1 % of posts respectively (Table 2). In T1 the most common symptom pairs were fatigue-restless sleep, hot flashes-night sweats, and fatigue-muscle pains, found in 37.1, 22.4, and 21.4 % of subjects, respectively. The same three symptom pairs were reported most commonly in T2. Using the similarity index defined in Eq. (1), which adjusts co-occurrences for the prevalence of the respective symptoms, the most similar symptoms on MedHelp were determined to be general aches-headaches, hot flashesmood changes, and constipation-diarrhea (Table 3). The study data showed that the most similar symptoms in T1 were hot flashes-night sweats, depression-mood changes, and fatigue-restless sleep, and the most similar symptoms in T2 were loss of interest in work-lowered work performance, hot flashes-night sweats, and general aches-muscle pains. Similarity scores were higher using study data compared to forum data.

Cluster analysis

Results of k-medoid clustering are shown graphically for social media (Fig. 1) and the two research study groups (Figs. 2, 3). Social media data were noted to include only 20 of the

possible 25 symptoms used in the analysis of the research study. Absent symptoms included avoidance of social affairs, decreased efficiency, loss of interest in work, lowered work performance, and restless sleep. Four clusters were generated using social media, and five were generated using symptoms rated as either moderate/severe or severe only in the research group. In the social media data, besides the four identified clusters, a single symptom—increased appetite—formed its own group and did not meet the definition of a symptom cluster. The number of symptoms per cluster ranged from 2 to 12.

The four clusters identified in the social media group were menopausal/psychological, pain/fatigue, gastrointestinal, and miscellaneous. The five clusters identified in T1 were menopausal, pain, fatigue/sleep/gastrointestinal, psychological, and increased appetite/weight. For T2 the five clusters were menopausal, pain, fatigue/psychological/gastrointestinal, gastrointestinal, and increased appetite/weight. A side-by-side comparison of clusters found in the three analyses is shown in Table 4.

The density of connections between symptoms was lower for the social media group than the study groups. However, density decreased in the research analysis when considering only symptoms rated as severe. In this study, density reflects the overall rates of co-occurrence and similarity between symptoms and is depicted visually by the number and thickness of lines in the three figures.

Review of the data mining algorithm

An additional review of findings reported by the data algorithm was undertaken to better understand the relationships between symptoms that were noted to occur together. Some co-occurrences seem very accurate (post 1), while other instances of co-occurrences that were detected may in fact be spurious (post 2). In post 2, an ulcerating sore on the breast is misidentified as a mouth ulcer (neither symptom was among the 25 used to formulate the analysis).

> Post 1: "… I am still in a lot of pain not so much in my elbow area but it feels as if it is in the bone and is slowly making its way up my arm and is now into my **shoulders and neck (Neck-Skull Aches)**…it was doing well for a while but now the **headaches (Headaches)** are slowly coming back…"
>
> Post 2: "I about a month ago I have what started out like a pimple on my breast. I popped the pimple and a **ulcerating (not Mouth Ulcer)** sore formed. I saw a doctor who prescribed antibiotics, stated no lump under the sore, just thought it was a skin infection. It took quite a while but the sore has pretty much healed but is still quite **tender on this same breast (Breast sensitivity)**…"

Five of the 25 symptoms chosen to conduct the analysis, including restless sleep, were not detected in any forum postings. After reviewing a random sample of 100 posts containing the keyword "sleep," we found that approximately one-fifth of these posts described sleep-related difficulties. However, none of these posts contained the specific phrase "restless sleep." This may suggest a limitation of the expanded CHV used by the algorithm to detect variations in standard terminology. Given that no instances of restless sleep were detected, there were no "false positives" in which restless sleep was erroneously counted.

Discussion

Comparison of symptom cluster patterns collected in a non-standard setting with those derived by symptom checklists administered as part of a research study generated multiple interesting findings. There was significant overlap in terms of the same symptoms clustering together in the analyses of both social media content and study subjects (T1). For example, hot flashes, night sweats, and vaginal dryness clustered together in both groups, as did abdominal pain, constipation, and nausea, and general aches, muscle pains, and neck-skull aches. Significant discrepancies were also seen. Psychological and menopausal symptoms formed separate clusters using research study data, but clustered together in the social media group. The social media dataset yielded a cluster containing fatigue and pain, while fatigue clustered with restless sleep rather than pain using study data. This finding is not especially surprising given that different relationships between these symptoms have been reported in the literature, including pain grouping with fatigue [42, 43] and fatigue with sleep difficulties [40, 41, 44]. These discrepant patterns may be attributable to the use of different study populations or statistical methods.

Several symptom clusters made intuitive sense in terms of having a common etiology or pathophysiology, such as the menopausal cluster and pain cluster in the research study group, and are also consistent with previously reported clusters [46, 47]. Other clusters were harder to rationalize, such as the grouping of fatigue and sleep problems with GI symptoms in the research study population (T1), or the combination of bloating, difficulty concentrating, and sleeping too much in the social media group. Further investigation could reveal connections between symptoms that clustered together that may not be readily apparent. It is possible had additional clusters beyond the initial five been allowed, menopausal and psychological symptoms in the forum group and GI and sleep symptoms in the research group would not have stayed together in the same cluster. We determined five clusters to be the optimal limit in terms of balancing statistical criteria and interpretability.

Density of connections between symptoms was highest in the T1 research study group, which included moderate and severe symptoms. Density was lower in the T2 group, which only included severe symptoms. Such a result is not surprising when considering that respondents to the checklist are likely to note more symptoms being present when the range of severity is extended to moderate and severe versus only severe. When a larger number of symptoms are reported, there will be more instances of symptoms occurring together.

The social media analysis had lower density compared to both study groups, likely because spontaneous reporting of symptoms on the forum generates fewer symptoms than a survey instrument that specifically elicits the presence of symptoms. Checklists likely trigger patients to name symptoms that are mild in severity, previously unnoted, and perhaps otherwise difficult to describe and articulate. For example, presumably some fraction of posters who described sleep difficulties would have endorsed restless sleep if asked specifically. Forum users may tend to describe only symptoms that are most frequent and bothersome in nature, and they may be reluctant to reveal certain symptoms that may carry stigma, such as psychological symptoms. This matter requires further investigation.

As noted, data derived from social media often carry significant limitations, and the data collected in our study reflect some of these concerns. Symptoms described by users often lack desired details such as elaboration of the nature or temporality of symptoms, especially compared to validated and reliable survey instruments. Some symptoms appeared difficult to detect algorithmically given users' unique lexicon and use of non-standard terminology, suggesting a need for refinement of the data mining engine. There was significant uncertainty regarding users themselves, some of whom apparently did not belong to the target population. We noted instances of family members of breast cancer patients or persons being evaluated for possible cancer using the forum. The number of posts that were misinterpreted is unknown; however, the fact that so many posts were crawled may minimize the overall impact of such errors [17]. Quality control efforts must continue to ensure that big data sources are scrutinized with a high degree of accuracy. Despite concerns regarding the source of data, the analysis of social media data generated results that appear reasonable in light of what is currently known about symptom clusters.

This study has several limitations worth noting. The configuration of symptom clusters in the research study was sensitive to dichotomization of symptom severity. Using moderate and severe symptoms as opposed to mild symptoms appeared to generate a set of clusters most comparable with results from the social media analysis. Also, the symptom clusters would have differed had the number of clusters allowed been changed or had a different set of symptoms from the 25 we selected been used. Data from MedHelp were generated spontaneously based on the needs of the users, and users likely did not report every symptom experienced. Study data likely reflect the true frequency of symptoms, whereas forum data reflect the concerns of users.

Use of "big data" such as that derived from online social networking sites offers researchers the titillating prospect of large volumes of easily accessible information. More work is needed to compare the results of research using social media with those garnered from traditional research studies. These efforts could further elucidate the pitfalls and potential of online approaches to research. Whether the data overlap or disagree, taking both perspectives into account will likely yield a more accurate and nuanced understanding of reality. As refinements in the extraction and use of data from social media continue to accrue, confidence in the validity and reliability of findings will increase. In its current form, social media data should be seen as a supplement, rather than a replacement, for carefully controlled research studies.

Acknowledgments

Funding This study was funded by the following Grants—National Science Foundation SES-1424875, National Institutes of Health R21AG042761, and Department of Defense #DAMD 17-01-1-0446.

Abbreviations

CHV	Consumer health vocabulary
GI	Gastrointestinal

References

1. Grajales FJ III, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: A review and tutorial of applications in medicine and health care. Journal of Medical Internet Research. 2014; 11(16):e13.
2. Young SD. Behavioral insights on big data: Using social media for predicting biomedical outcomes. Trends in Microbiology. 2014; 22(11):601–602. [PubMed: 25438614]
3. File, T.; Ryan, C. Computer and internet use in the United States: 2013. Resource document, US Census. 2014. http://www.census.gov/content/dam/Census/library/publications/2014/acs/acs-28.pdf. Accessed February 26, 2015
4. Duggan, M.; Ellison, NB.; Lampe, C.; Lenhart, A.; Madden, M. Social media update 2014. Resource document, Pew Research Center. 2015. http://www.pewinternet.org/2015/01/09/social-media-update-2014/. Accessed February 26, 2015
5. Chaung KY, Yang CC. Interaction patterns of nurturant support exchanged in online health social networking. Journal of Medical Internet Research. 2012; 14(3):e54. [PubMed: 22555303]
6. Capurro D, Cole K, Echavarria MI, Joe J, Neogi T, Turner AM. The use of social networking sites for public health practice and research: A systematic review. Journal of Medical Internet Research. 2014; 16(3):e79. [PubMed: 24642014]
7. Corley CD, Cook DJ, Mikler AR, Singh KP. Using web and social media for influenza surveillance. Advances in Experimental Medicine and Biology. 2010; 680:559–564. [PubMed: 20865540]
8. Kuehn BM. Agencies use social media to track food-borne illness. Journal of the American Medical Association. 2014; 312(2):117–118. [PubMed: 24963655]
9. Liu M, Hu Y, Tang B. Role of text mining in early identification of potential drug safety issues. Methods in Molecular Biology. 2014; 1159:227–251. [PubMed: 24788270]
10. Vaughan Sarrazin MS, Cram P, Mazur A, Ward M, Reisinger HS. Patient perspectives of dabigatran: Analysis of online discussion forums. Patient. 2014; 7(1):47–54. [PubMed: 24030706]
11. Abou Taam M, Rossard C, Cantaloube L, Bouscaren N, Roche G, Pochard L, et al. Analysis of patients' narratives posted on social media websites on benfluorex's (mediator) withdrawal in France. Journal of Clinical Pharmacy and Therapeutics. 2014; 39(1):53–55. [PubMed: 24304185]
12. Alshaikh F, Ramzan F, Rawaf S, Majeed A. Social network sites as a mode to collect health data: A systematic review. Journal of Medical Internet Research. 2014; 16(7):e171. [PubMed: 25048247]
13. Karmen C, Hsiung RC, Wetter T. Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. Computer Methods and Programs in Biomedicine. 2015; 120(1):27–36. [PubMed: 25891366]
14. Lloyd A. Social media, help or hindrance: What role does social media play in young people's mental health? Psychiatria Danubia. 2014; 26(1):340–346.
15. Leng HK. Methodological issues in using data from social networking sites. Cyberpsychology, Behavior, and Social Networking. 2013; 16(9):686–689.
16. Bainbridge WS. The scientific research potential of virtual worlds. Science. 2007; 317(5837):472–476. [PubMed: 17656715]
17. Chen M, Mangubat E, Ouyang B. Patient-reported outcome measures for patients with cerebral aneurysms acquired via social media: Data from a large nationwide sample. Journal of Neurointerventional Surgery. 2014 doi:10.1136/neurintsurg-2014-011492.
18. Park K, Harris M, Khavari N, Khosla C. Rationale for using social media to collect patient-reported outcomes in patients with celiac disease. Journal of Gastrointestinal and Digestive System. 2014; 4(1):166. [PubMed: 25392743]
19. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: The promise, challenges, and state of the art. Drug Safety. 2014; 37(10):777–790. [PubMed: 25151493]
20. Peek N, Holmes JH, Sun J. Technical challenges for big data in biomedicine and health: Data sources, infrastructure, and analytics. Yearbook of Medical Informatics. 2014; 9(1):42–47. [PubMed: 25123720]

21. Madden, M.; Zickurh, K. 65 % of online adults use social networking sites. Resource document, Pew Internet and American life project. 2011. http://www.pewinternet.org/Reports/2011/SocialNetworkingSites.aspx. Accessed September 24, 2015
22. Cavallo DN, Chou WY, McQueen A, Ramirez A, Riley WT. Cancer prevention and control interventions using social media: User-generated approaches. Cancer Epidemiology, Biomarkers and Prevention. 2014; 23(9):1953–1956.
23. Taurob S, Tucker CS, Salathe M, Ram N. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. Journal of Biomedical Informatics. 2014; 49:255–268. [PubMed: 24642081]
24. Gustafson DL, Woodworth CF. Methodological and ethical issues in research using social media: A metamethod of Human Papillomavirus vaccine studies. BMC Medical Research Methodology. 2014; 14:127. [PubMed: 25468265]
25. Chang VT, Hwang SS, Feuerman M, Kasimis BS. Symptom and quality of life survey of medical oncology patients at a veterans affairs medical center: A role for symptom assessment. Cancer. 2000; 88(5):1175–1183. [PubMed: 10699909]
26. Deshields TL, Potter P, Olsen S, Liu J. The persistence of symptom burden: Symptom experience and quality of life of cancer patients across one year. Supportive Care in Cancer. 2014; 22(4): 1089–1096. [PubMed: 24292095]
27. Naughton MJ, Weaver KE. Physical and mental health among cancer survivors: Considerations for long-term care and quality of life. North Carolina Medical Journal. 2014; 75(4):283–286. [PubMed: 25046097]
28. Deshields TL, Potter P, Olsen S, Liu J, Dye L. Documenting the symptom experience of cancer patients. The Journal of Supportive Oncology. 2011; 9(6):216–223. [PubMed: 22055891]
29. Portenoy RK, Thaler HT, Kornblith AB, Lepore JM, Friedlander-Klar H, Coyle N, et al. Symptom prevalence, characteristics and distress in a cancer population. Quality of Life Research. 1994; 3(3):183–189. [PubMed: 7920492]
30. Dodd MJ, Miaskowski C, Paul SM. Symptom clusters and their effect on the functional status of patients with cancer. Oncology Nursing Forum. 2001; 28(3):465–470. [PubMed: 11338755]
31. Kim HJ, McGuire DB, Tulman L, Barsevick AM. Symptom clusters: Concept analysis and clinical implications for cancer nursing. Cancer Nursing. 2005; 28(4):270–282. [PubMed: 16046888]
32. Fan G, Filipczak L, Chow E. Symptom clusters in cancer patients: A review of the literature. Current Oncology. 2007; 14(5):173–179. [PubMed: 17938700]
33. Kirkova J, Aktas A, Walsh D, Davis MP. Cancer symptoms clusters: Clinical and research methodology. Journal of Palliative Medicine. 2011; 14(10):1149–1166. [PubMed: 21861613]
34. Xiao C. The state of science in the study of cancer symptom clusters. European Journal of Oncology Nursing. 2010; 14(5):417–434. [PubMed: 20599421]
35. Denieffe S, Cowman S, Gooney M. Symptoms, clusters, and quality of life prior to surgery for breast cancer. Journal of Clinical Nursing. 2014; 23(17–18):2491–2502. [PubMed: 24329603]
36. Walsh D, Rybicki L. Symptom clustering in advanced cancer. Supportive Care in Cancer. 2006; 14(8):831–836. [PubMed: 16482450]
37. Fan G, Hadi S, Chow E. Symptom clusters in patients with advanced-stage cancer referred for palliative radiation therapy in an outpatient setting. Supportive Cancer Therapy. 2007; 4(3):157–162. [PubMed: 18632482]
38. Tsai JS, Wu CH, Chiu TY, Chen CY. Significance of symptom clustering in palliative care of advanced cancer patients. Journal of Pain and Symptom Management. 2010; 39(4):655–662. [PubMed: 20226623]
39. Gleason JF, Case D, Rapp SR, Ip E, Naughton M, Butler JM, et al. Symptom clusters in patients with newly-diagnosed brain tumors. Journal of Supportive Oncology. 2007; 5(9):427–433. [PubMed: 18019850]
40. Broeckel JA, Jacobsen PB, Horton J, Balducci L, Lyman GH. Characteristics and correlates of fatigue after adjuvant chemotherapy for breast cancer. Journal of Clinical Oncology. 1998; 16(5): 1689–1696. [PubMed: 9586880]
41. Berger AM, Farr L. The influence of daytime inactivity and nighttime restlessness on cancer-related fatigue. Oncology Nursing Forum. 1999; 26(10):1663–1671. [PubMed: 10573683]

42. Byar KL, Berger AM, Bakken SL, Cetak MA. Impact of adjuvant breast cancer chemotherapy on fatigue, other symptoms, and quality of life. Oncology Nursing Forum. 2006; 33(1):E18–E26. [PubMed: 16470230]
43. Gaston-Johansson F, Fall-Dickson JM, Bakos AB, Kennedy MJ. Fatigue, pain, and depression in pre-autotransplant breast cancer patients. Cancer Practice. 1999; 7(5):240–247. [PubMed: 10687593]
44. Ho SY, Rohan KJ, Parent J, Tager FA, McKinley PS. A longitudinal study of depression, fatigue, and sleep disturbances as a symptom cluster in women with breast cancer. Journal of Pain and Symptom Management. 2015; 49(4):707–715. [PubMed: 25461671]
45. Bender CM, Ergyn FS, Rosenzweig MQ, Cohen SM, Sereika SM. Symptom clusters in breast cancer across 3 phases of the disease. Cancer Nursing. 2005; 28(3):219–225. [PubMed: 15915067]
46. Glaus A, Boehme C, Thurlimann B, Ruhstaller T, Hsu Schmitz SF, Morant R, et al. Fatigue and menopausal symptoms in women with breast cancer undergoing hormonal cancer treatment. Annals of Oncology. 2006; 17(5):801–806. [PubMed: 16507565]
47. Kim HJ, Barsevick AM, Tulman L, McDermott PA. Treatment-related symptom clusters in breast cancer: A secondary analysis. Journal of Pain and Symptom Management. 2008; 36(5):468–479. [PubMed: 18718735]
48. Fu OS, Crew KD, Jacobson JS, Greenlee H, Yu G, Campbell J, et al. Ethnicity and persistent symptom burden in breast cancer survivors. Journal of Cancer Survivorship. 2009; 3(4):241–250. [PubMed: 19859813]
49. Avis N, Levine B, Naughton M, Case LD, Naftalis E, Van Zee KJ. Age related longitudinal changes in depressive symptoms following breast cancer diagnosis and treatment. Breast Cancer Research and Treatment. 2013; 139(10):199–206. [PubMed: 23588951]
50. Barnabei VM, Cochrane BB, Aragaki AK, et al. Menopausal symptoms and treatment-related effects of estrogen and progestin in the women's health initiative. Obstetrics and Gynecology. 2005; 105:1063–1073. [PubMed: 15863546]
51. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. Journal of the American Medical Informatics Association. 2006; 13(1):24–29. [PubMed: 16221948]
52. Jiang L, Yang CC. Using co-occurrence analysis to expand consumer health vocabularies from social media data. In Proceedings of IEEE international conference on healthcare informatics. 2013:74–81.
53. Kaufman, L.; Rousseeuw, PJ. Clustering by means of medoids. In: Dodge, Y., editor. Statistical data analysis based on the L1-norm and related methods. Birkhauser; North-Holland: 1987. p. 405-416.

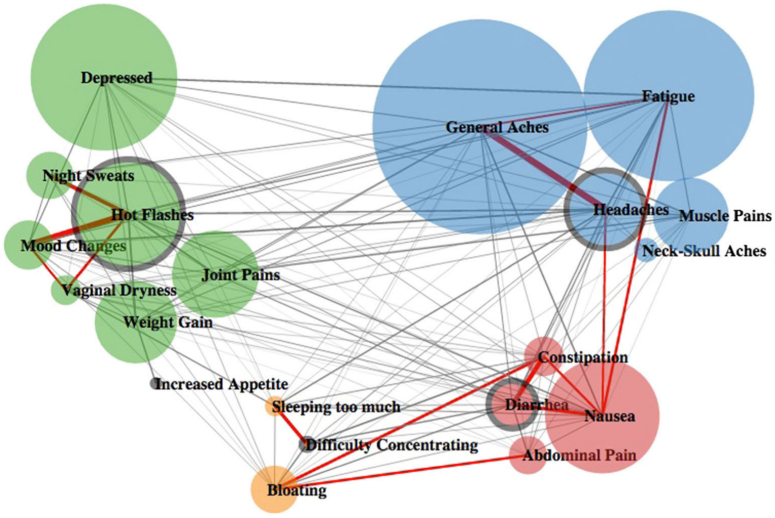

Fig. 1.

Clustering results for social media group. The size of the *circle* reflects the frequency of the symptom. The *thickness of the lines* connecting individual symptoms reflects the degree of similarity between the two. The medoid of each cluster is highlighted using a thicker edge. *Red linkages* represent the highest 10 % similarity between nodes

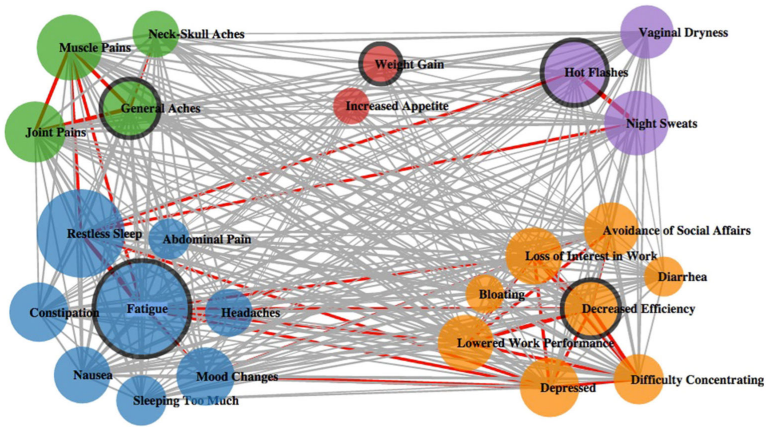

Fig. 2.

Clustering results for research study group T1 (moderate and severe symptoms). The size of the circle reflects the frequency of the symptom. The *thickness of the lines* connecting individual symptoms reflects the degree of similarity between the two. The medoid of each cluster is highlighted using a *thicker edge. Red linkages* represent the highest 10 % similarity between nodes

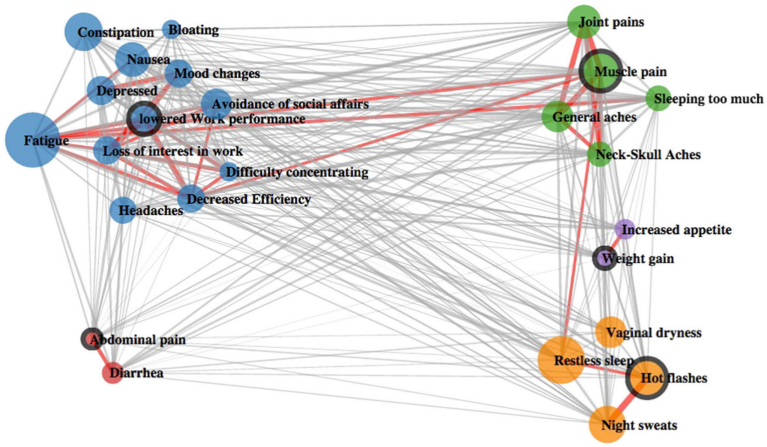

Fig. 3.

Clustering results for research study group T2 (severe symptoms only). The size of the *circle* reflects the frequency of the symptom. The *thickness of the lines* connecting individual symptoms reflects the degree of similarity between the two. The medoid of each cluster is highlighted using a *thicker edge. Red linkages* represent the highest 10 % similarity between nodes

Table 1

Characteristics of participants1—research study data

Age range	**Number of participants**
25–44	132
45–54	209
55–64	167
65–74	102
75+	43
Stage	Percentage of participants (%)
I	34.1
II	55.3
III	10.6
Treatment	Percentage of participants (%)
Chemotherapy	81
Radiation	86

Table 2

Ten most common symptom pairs for social media and research study groups based on number of co-occurrences (social media $n = 50{,}426$; research study $n = 653$)

Social media group		Research study group T1 moderate and severe symptoms		Research study group T2 severe symptoms only	
Symptom pair	**Co-occurrence (%)**	**Symptom pair**	**Co-occurrence (%)**	**Symptom pair**	**Co-occurrence (%)**
General aches–headaches	141 (0.28)	Fatigue–restless sleep	242 (37.1)	Hot flashes–night sweats	36 (5.5)
Fatigue–general aches	89 (0.18)	Hot flashes–night sweats	146 (22.4)	Fatigue–muscle pains	30 (4.6)
Fatigue–nausea	64 (0.13)	Fatigue–muscle pains	140 (21.4)	Fatigue–restless sleep	29 (4.4)
Depressed–fatigue	59 (0.12)	Hot flashes–restless sleep	126 (19.3)	Joint pains–muscle pains	29 (4.4)
General aches–nausea	49 (0.10)	Muscle pains–restless sleep	125 (19.1)	General aches–muscle pains	27 (4.1)
Hot flashes–mood changes	42 (0.08)	Decreased efficiency–fatigue	123 (18.8)	Fatigue–lowered work performance	26 (4.0)
Depressed–general aches	39 (0.08)	Fatigue–night sweats	121 (18.5)	Hot flashes–restless sleep	23 (3.5)
Fatigue–hot flashes	31 (0.06)	Night sweats–restless sleep	120 (18.4)	Decreased efficiency–fatigue[a]	22 (3.4)
General aches–muscle pains	31 (0.06)	Depressed–fatigue	118 (18.1)	Fatigue–nausea	22 (3.4)
Fatigue–headaches	29 (0.06)	Fatigue–hot flashes	117 (17.9)	General aches–joint pains	22 (3.4)
				Loss of interest in work–lowered work performance	22 (3.4)

[a] Four-way tie for eighth place

Table 3

Similarity index for ten most similar symptom pairs for social media and research study groups (social media $n = 50{,}426$; research study $n = 653$)

Social media group		Research study group T1 moderate and severe symptoms		Research study group T2 severe symptoms only	
Symptom pair	**Similarity[a]**	**Symptom pair**	**Similarity**	**Symptom pair**	**Similarity**
General aches–headaches	0.355	Hot flashes–night sweats	0.872	Loss of interest in work–lowered work performance	0.763
Hot flashes–mood changes	0.331	Depressed–mood changes	0.737	Hot flashes–night sweats	0.751
Constipation–diarrhea	0.246	Fatigue–restless sleep	0.722	General aches–muscle pains	0.652
Hot flashes–night sweats	0.204	General aches–joint pains	0.710	Joint pains–muscle pains	0.652
Difficulty concentrating–sleeping too much	0.183	Decreased efficiency–lowered work performance	0.709	Decreased efficiency–lowered work performance	0.624
Diarrhea–nausea	0.177	Loss of interest in work–lowered work performance	0.704	General aches–joint pains	0.621
Bloating–constipation	0.156	Decreased efficiency–loss of interest in work	0.702	Depressed–mood changes	0.556
Hot flashes–vaginal dryness	0.143	Decreased efficiency–difficulty concentrating	0.659	Decreased efficiency–loss of interest in work	0.462
Fatigue–Nausea	0.138	General aches–muscle pains	0.652	Joint pains–neck-skull aches	0.460
Bloating–Abdominal Pain	0.138	Joint pains–muscle pains	0.637	Fatigue–lowered work performance	0.457

[a] Similarity between symptoms i and j is defined as co-occurrence (i, j)/square root (occurrence (i) × occurrence (j)); higher value represents greater similarity

Table 4

Symptom clusters for social media and research study groups

	Social media group		Research study group Tl moderate and severe symptoms		Research study group T2 severe symptoms only	
	Symptom cluster	**Symptoms**	**Symptom cluster**	**Symptoms**	**Symptom cluster**	**Symptoms**
1	Menopausal/ psychological	Depressed, *hot flashes*[a], joint pains, mood changes, night sweats, vaginal dryness, weight gain	Menopausal	*Hot flashes,* night sweats, vaginal dryness	Menopausal	*Hot flashes,* night sweats, restless sleep, vaginal dryness
2	Pain/fatigue	Fatigue, general aches, *headaches,* muscle pains, neck-skull aches	Pain	*General aches,* joint pains, muscle pains, neck-skull aches	Pain	General aches, joint pains, *muscle pains,* neck-skull aches, sleeping too much
3	Gastrointestinal	Abdominal pain, constipation, *diarrhea,* nausea	Fatigue/ sleep/gastrointestinal	Abdominal pain, constipation, *fatigue,* headaches, mood changes, nausea, restless sleep, sleeping too much	Fatigue/ psychological/gastrointestinal	Avoidance of social affairs, bloating, constipation, decreased efficiency, depressed, difficulty concentrating, fatigue, headaches, loss of interest in work, *lowered work performance,* nausea, restless sleep
4	Miscellaneous	Bloating, *difficulty concentrating,* sleeping too much	Psychological	Avoidance of social affairs, bloating, *decreased efficiency,* depressed, diarrhea, difficulty concentrating, loss of interest in work, lowered work performance	Gastrointestinal	*Abdominal pain,* diarrhea
5		*Increased appetite*[b]	Increased weight/ appetite	Increased appetite, *weight gain*	Increased weight/appetite	Increased appetite, *weight gain*

[a]The medoid (anchor) of a symptom cluster is italicized

[b]Single-symptom group