BMC Cancer

RESEARCH Open Access

Detection of differences in physical symptoms between depressed and undepressed patients with breast cancer: a study using K-medoids clustering

Jianyao Tang[1,2], Bingqian Guo[1,2], Chuhan Zhong[1,2], Jing Chi[1,2], Jiaqi Fu[1,2], Jie Lai[1,2], Yujie Zhang[1,2], Zihan Guo[1,2], Shisi Deng[1,2] and Yanni Wu[1*]

Abstract

Background To detect the differences in physical symptoms between depressed and undepressed patients with breast cancer (BC), including common symptoms, co-occurring symptoms, and symptom clusters based on texts derived from social media and expressive writing.

Methods A total of 1830 texts from social media and expressive writing were collected. The Chi-square test was used to compare the frequency of physical symptoms between depressed and undepressed patients with BC. Symptom lexicon of BC and K-medoids Clustering were used for mining physical symptoms and cluster analysis.

Results The common physical symptoms reported by texts included general pains (59.38%), fatigue (26.60%), vomiting (24.82%), swelling of limbs (21.69%), difficulty sleeping (21.56%), nausea (16.78%), alopecia (15.14%), loss of appetite (13.78%), dizziness (11.60%), and concentration problems (11.19%). The frequency of difficulty sleeping (depressed 28.40%; undepressed 18.16%; $P=0.002$) in depressed patients was higher than undepressed patients with BC. High co-occurrence was observed in both commonly mentioned symptoms and those less commonly mentioned but frequently co-occurring with them. There were 5 symptom clusters identified in depressed patients and 6 symptom clusters in undepressed patients. Pain-related symptom cluster and gastrointestinal symptom cluster were both identified in the depressed and undepressed patients. The novel immune system impairment symptom cluster consisting of bleeding and fever was found in the undepressed patients.

Conclusions This study found that difficulty sleeping was reported more frequently, and identified difficulty sleeping-pain symptom cluster in depressed patients. The novel immune system impairment symptom cluster in undepressed patients was detected. Healthcare providers can provide targeted care to depressed and undepressed patients based on these differences. These findings demonstrate that social media can provide new perspectives on symptom experiences. The combination of digital tools and traditional clinical tools for symptom management in follow-up has great potential in the future.

*Correspondence:
Yanni Wu
yanniwuSMU@126.com

Full list of author information is available at the end of the article

Clinical trial number Not applicable.

Keywords Breast cancer, Depression, Social media, Physical symptoms, K-medoids clustering

Introduction

Breast cancer (BC) has been the most common cancer worldwide and resulted in more than 2.26 million cases and 685,000 deaths in 2020 [1]. Most of patients with BC require chemotherapy, radiotherapy, or endocrine therapy according to their pathology [2]. The long-term side effects of the treatment significantly decrease their quality of life and bring them great psychological distress [3]. The global prevalence of depression among patients with BC is around 30% [4]. And it was associated with a 30% increase in risk of all-cause mortality and a 29% increase in risk of specific mortality in BC [5].

The severity of depression in BC has attracted great attention from researchers. It was found that patients with BC who had more depressive symptoms reported more physical symptoms [6]. Patients with psychological distress reported higher levels of fatigue and pain, poorer levels of sleep quality and self-rated health [6–8]. And there was a significant indirect effect between low levels of depressive symptoms and low levels of fatigue and pain interference [9]. It indicated that the frequency and intensity of physical symptoms were different in depressed and undepressed patients with BC, and the level of depressive symptoms might be alleviated by lowering the level of physical symptoms.

With the advances in technology, digital health tools such as wearable devices and remote monitoring systems provide new pathways for symptom monitoring. Remote cancer symptom monitoring based on patient self-reported outcomes can improve clinical outcomes and reduce treatment-related toxic effects [10]. Data via the web and social media also can play an important role for the detection of cancer symptoms [11]. Large amounts of text are rapidly generated every moment, providing a huge data resource for symptom discovery. Lindvall et al. believed that analysis of patient self-reported texts allows to improve monitoring and proactively manage patients' symptoms [12]. Additionally, textual data provides more comprehensive and detailed information, and helps us better understand the patients' disease journey [13]. As reported by patients and survivors with BC, social media platforms have become an essential part of oncology self-management and addressing psychosocial needs [14]. Compared to older people, young people are the active group on the Internet. And physical symptoms of patients with BC obtained only through social media may be biased. Expressive writing, as a form of emotional disclosure in which participants are asked to write about their feelings and experiences of the traumatic event, is an effective method of exploring the patients' experience. Thus, a combination of social media and expressive writing can provide a better understanding of what physical symptoms patients with BC are experiencing.

To sum up, depression was highly reported in survivors with BC [4], and depressive symptoms might be indirectly lowered by lowering levels of physical symptoms [9]. Therefore, identifying differences in physical symptoms reported by depressed and undepressed patients with BC enables the provision of targeted interventions and more precise care to patients. However, the majority of studies that compare symptoms between depressed and undepressed patients have concentrated on single symptoms, while studies that address symptom clusters has mainly focused on BC patients [15, 16]. Moreover, there was a lack of other perspectives to discover the patient's experience of symptoms. For example, Grotmol et al. used scale to explore differences in symptom burden between depressed and undepressed patients with cancer, but the exploration did not go far to symptom clusters [17]. In a systematic evaluation of symptom clusters in BC patients at different stages of treatment, the assessment tool for all 32 included studies was a scale [18]. To address these research gaps, we proposed a novel approach that employs textual data to detect differences from common symptoms, co-occurring symptoms, and symptom clusters in depressed and undepressed patients with BC. The key contributions of this paper are as follows:

1. Develop a symptom lexicon of BC that can deal with the colloquialization of expression.
2. Conduct the first comparison of physical symptoms between depressed and undepressed patients with BC using texts from social media and expressive writing, which will reveal differences in multiple dimensions and provide a basis for targeted interventions.

Materials and methods

Data collection

Texts were derived from social media as well as expressive writing. The term "BC" was used to identify posts on Chinese social media platforms such as RED, and Tieba from the construction until January 1, 2024. Posts containing the labels "BC" and "I was diagnosed with depression" were associated with depressed patients with BC [19]. Moreover, we paid special attention to comment section under the depressive posts with BC to track more patients. Each post or comment with a user ID, timestamp and content was regarded as a separate

post. Ultimately, 637 posts about depression were found among the 1,416 target posts we gathered from patients with BC.

The texts of the expressive writing were derived from a large multicenter randomized controlled trial of expressive writing [20], and consent was obtained from participants for their use. Participants were asked to write for at least 20 min on four consecutive days, or for at least 20 min at least three times a week for four weeks. Depressed patients were assessed by the Hospital Anxiety Depression Scale. Ultimately, we obtained 76 texts from depressed patients with BC and 340 texts from undepressed patients from the trial.

Data preprocessing

To identify suitable texts for subsequent data analysis, we performed text cleansing and word segmentation, removing information such as pages, links, and images. Moreover, we manually went through the full range of obtained information and filtered out some posts that did not fit. Firstly, posts that mentioned a combination of other cancers or major illnesses besides depression were filtered out. If a post was made by a patient with BC comorbid with other diseases, other posts made by her were also excluded. Then, given that bystander feelings were not a substitute for the patient's own experience of symptoms, posts reported by bystanders were also removed. As for the expressively written texts, they were written by patients themselves. And the patients were excluded if they had other cancers or other major illnesses in the expressive writing trial [20]. All texts that did not report physical symptoms were filtered out. Finally, there were 733 texts consisting of 480 from social media and 253 from expressive writing. Among these texts, 490 texts were generated by undepressed patients with BC and 243 texts were generated by depressed patients with BC.

Symptom lexicon

The symptom lexicon was developed from a pre-constructed emotional lexicon of patients with BC, which contained emotion words, symptom words, and BC terminology [21]. With references to physical symptoms mentioned in the Distress Thermometer [22], the MD Anderson Symptom Inventory [23], and the BC Chemotherapy Symptom Assessment Scale [24], symptom words were extracted from the colloquial expressions in the text corpus. In order to better map to the specialized expressions, researchers with different cultural backgrounds and geographical characteristics manually checked each symptom. For example, the term "腹泻 (diarrhea)" usually can be described "腹泄 (diarrhea)", "拉肚子 (diarrhea)", "拉稀 (diarrhea)" in China. Particularly, it can also be described as "肚疴 (diarrhea)" in the Southern China. We have checked some relevant studies on the development of Chinese consumers' health words to verify the rationality of our mapping [25–27]. For instance, in a study focusing on consumer health words for skin, gastroenterology, endocrinology, and cardiovascular, the term "腹泻 (diarrhea)" also mapped to "拉肚子 (diarrhea)", "拉稀 (diarrhea)" [27]. It proved that our terminological mapping is reasonable. The word "肝郁 (liver depression)" is a Traditional Chinese Medicine term. Clinically, it is mainly characterized by depression and moodiness [28]. And it was removed as we primarily focused on physical symptoms. In addition, there were some words that described pain symptoms without identifying the location of the pain, such as "疼 (ache)", "痛 (pain)", "刺痛 (tingling)", "隐隐作痛 (vague pain)" .These words that did not specify the location of the pain but describe its occurrence were categorized as general pain. Then we ended up with 28 physical symptoms. Moreover, terms were put into the Biomedical information ontology system [29] to enrich its synonyms and colloquial words, which better suit users' daily writing characteristics. Ultimately, we classified the symptoms according to the systems or organs that were impacted: the digestive, musculoskeletal, neurological, cardiovascular, respiratory, sensory, integumentary, and systemic [30]. There were 8 impacted organs or systems, 28 physical symptoms and 308 phrases in the final symptom lexicon of BC. (see supplementary material 1)

Data analysis

Common physical symptoms

We counted the frequency of each physical symptom in their texts to explore whether physical symptoms differed in depressed and undepressed patients with BC. The Chi-square test was used to compare the frequency of physical symptoms between the two groups. *P* values were two-sided, and a significant difference was indicated when *P* value was less than 0.05. Given that the small number of texts we collected, physical symptoms with a frequency of 10% or more were considered common physical symptoms [31].

Co-occurring physical symptoms

Rather than just one symptom, patients with cancer usually had several [32]. The relationship between the symptoms could be explored further at the level of symptom co-occurrence or even symptom clusters. Co-occurring physical symptoms were regarded as being mentioned in a text at the same time in this study. The similarity of the symptoms was obtained by calculating the number of times each symptom occurred and the number of times the symptoms co-occurred [33]. The formula is as follows:

$$similarity[symptom\,(i)\,,symptom\,(j)] = \frac{Co-occurrence[symptom\,(i)\,,symptom\,(j)]}{sqrt(Occurrence\,[symptom\,(i)] * Occurrence[symptom\,(j)]}$$

Physical symptom clusters

Clustering methods that rely on differences in similarity or distance, such as K-means Clustering, K-medoids Clustering, and hierarchical clustering, are among the most typical clustering methods [34]. The K-means Clustering uses the cluster mean as the virtual centroid of the cluster for the subsequent iteration of partitioning, making it vulnerable to the impact of sample outliers. Similarly, based on the selection of distance matrices, hierarchical clustering is sensitive to noise and outliers, and has high computational complexity [33]. Unlike K-means Clustering, K-medoids Clustering uses actual points in the dataset as clustering centers, which makes it more stable when it comes to outliers [35]. And this is especially important when dealing with texts, which can contain a lot of noise and outliers. Additionally, the actual points serve as clustering centers, making the clustering results easier to understand and interpret. Moreover, it has been used in a wide range of fields, demonstrating their effectiveness in processing complex data types, including textual data [36]. Therefore, we choose K-medoids Clustering as our clustering method. For each iteration, the clustering effect is recalculated by replacing one of the original centroids with a randomly selected non-centroid [37]. The original centroid is kept unless the clustering results show an improvement. If not, this replacement is kept. The iteration ends when the replacement no longer enhances the clustering effect.

Silhouette Width (SW) and Average Silhouette Width (ASW) are the metric for evaluating the effectiveness of clustering. The formula for the SW and ASW is as follows:

$$\mathrm{S_i(C,d)} = \frac{\mathrm{b\,(i)} - \mathrm{a\,(i)}}{\max\{\mathrm{a\,(i)\,,b\,(i)}\}}$$

$$\bar{S}\,(C,d) = \frac{1}{n}\sum\nolimits_{i=1}^{n} S_i(C,d)$$

a(i) is the average distance of symptom *i* to the points in the cluster to which it was. The smaller the value, the better the clustering.

b(i) is the average distance of symptom *i* to the points in the nearest cluster to which it was not assigned. The large the value, the better the clustering.

Given that clusters should exhibit homogeneity and clear separation, higher values of SW and ASW signify a better quality of clustering. Therefore, they can also be used to determine the number of clusters. The optimal number of clusters is the value of K corresponding to the maximum value of ASW [38]. The ASW ranges from −1 to 1. It indicates a more appropriate categorization of the symptom as its value approaches 1. And as the value of ASW gets closer to -1, it indicates that the partitioning of this symptom is not appropriate. The symptom has the same effect in both clusters when the ASW value is 0 [39]. We created a waiting list for placing symptoms with negative ASW. For these symptoms, we attempt to reassign to different clusters and calculate the new SW and ASW. If the SW value is better than the previous best, update the best SW, ASW and cluster assignment. If no better clustering is found, try to reassign again. Finally, if the symptom cannot be reassigned, it will be removed permanently. (see supplementary materials 2)

Results

Distribution of physical symptoms and affected organs or systems

The distribution of affected organs or systems was shown in Fig. 1. In total, 28 symptoms were mentioned 2,117 times in the 733 self-reported symptom texts; 25 symptoms (89%) were mentioned more than 10 times. It could be clearly seen that musculoskeletal symptoms were the most frequently reported of all symptoms, accounting for 30% (623/2,117). And of these, general pain accounted for more than 60%. The digestive symptoms ranked second (24%, 515/2,117), followed by the systemic symptoms (24%, 508/2,117), nervous symptoms (12%, 259/2,117), integumentary symptoms (7%, 151/2,117), respiratory symptoms (1%, 27/2,117), cardiovascular symptoms (1%, 26/2,117) and sensory symptoms.

Occurrence and frequency of physical symptoms

The occurrence and frequency of the 28 symptoms were shown in Table 1. The eight symptoms—general pain, fatigue, vomiting, swelling of the limbs, difficulty sleeping, nausea, alopecia, and lack of appetite—were collectively noted more than 100 times out of all the self-reported symptom texts. Additionally, taste sense altered and bleeding were mentioned less than 10 times, with drop in libido being not mentioned whether in the undepressed self-reported texts or in the depressed self-reported texts. The most often mentioned physical symptoms of undepressed patients with BC were general pain, fatigue, vomiting, swelling of the limbs, and nausea. Nevertheless, among the self-reported symptoms of depression, the most common symptoms were no longer nausea, but rather difficulty sleeping. There was a statistically significant difference in general pain (P=0.001), swelling of the limbs (P=0.016), difficulty sleeping (P=0.002), nausea (P=0.013), concentration problems (P=0.022), skin changes (P=0.031), oral ulcers (P=0.010), and back pain (P=0.024) reported in the depressed and undepressed texts.

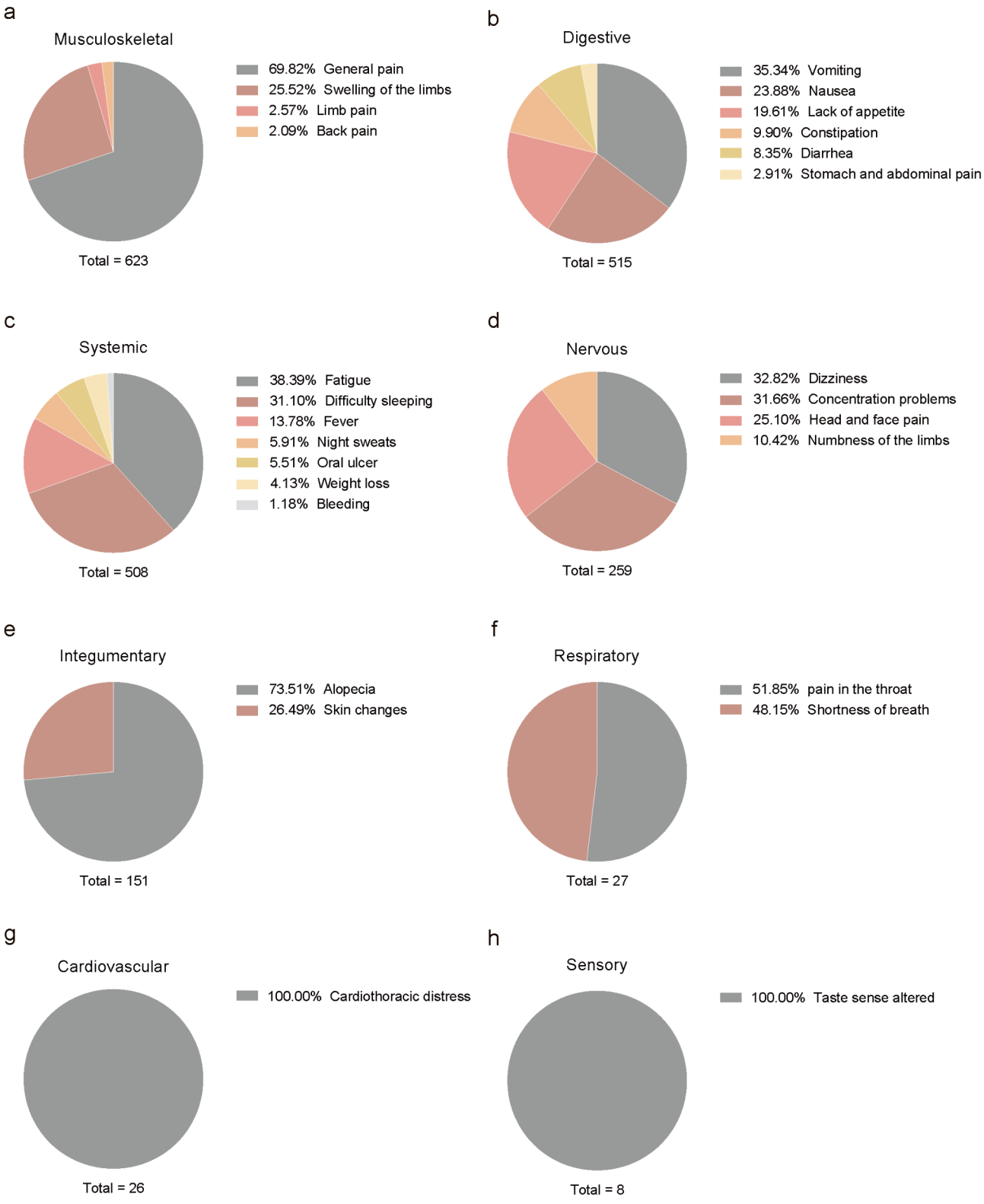

Fig. 1 Distribution of physical symptoms and affected organs or systems

Table 1 Occurrence and frequency of physical symptoms

Symptom	System	Self-reported symptoms (all; *N* = 733), *n* (%)	Self-reported symptoms (undepressed; *N* = 490), *n* (%)	Self-reported symptoms (depressed; *N* = 243), *n* (%)	*P*
General Pain	Musculoskeletal	435 (59.38)	312 (63.67)	123 (50.62)	0.001*
Fatigue	Systemic	195 (26.60)	132 (26.94)	63 (25.93)	0.770
Vomiting	Digestive	182 (24.82)	132 (26.94)	50 (20.58)	0.061
Swelling of the limbs	Musculoskeletal	159 (21.69)	119 (24.29)	40 (16.46)	0.016*
Difficulty sleeping	**Systemic**	**158 (21.56)**	**89 (18.16)**	**69 (28.40)**	**0.002***
Nausea	Digestive	123 (16.78)	94 (19.18)	29 (11.93)	0.013*
Alopecia	Integumentary	111 (15.14)	78 (15.92)	33 (13.58)	0.406
Lack of appetite	Digestive	101 (13.78)	69 (14.08)	32 (13.17)	0.736
Dizziness	Nervous	85 (11.60)	61 (12.45)	24 (9.88)	0.306
Concentration problems	Nervous	82 (11.19)	64 (13.06)	18 (7.41)	0.022*
Fever	Systemic	70 (9.55)	46 (9.39)	24 (9.88)	0.832
Head and face pain	Nervous	65 (8.87)	43 (8.78)	22 (9.05)	0.901
Constipation	Digestive	51 (6.96)	39 (7.96)	12 (4.94)	0.130
Diarrhea	Digestive	43 (5.87)	34 (6.94)	9 (3.70)	0.079
Skin changes	Integumentary	40 (5.46)	33 (6.73)	7 (2.88)	0.031*
Night sweats	Systemic	30 (4.09)	20 (4.08)	10 (4.12)	0.983
Oral ulcer	Systemic	28 (3.82)	25 (5.10)	3 (1.23)	0.010*
Numbness of the limbs	Nervous	27 (3.68)	22 (4.49)	5 (2.06)	0.100
Cardiothoracic distress	Cardiovascular	26 (3.55)	17 (3.47)	9 (3.70)	0.872
Weight loss	Systemic	21 (2.86)	13 (2.65)	8 (3.29)	0.625
Limb pain	Musculoskeletal	16 (2.18)	11 (2.24)	5 (2.06)	0.870
Stomach and abdominal pain	Digestive	15 (2.05)	12 (2.45)	3 (1.23)	0.414
Pain in the throat	Respiratory	14 (1.91)	12 (2.45)	2 (0.82)	0.220
Back pain	Musculoskeletal	13 (1.77)	13 (2.65)	0 (0.00)	0.024*
Shortness of breath	Respiratory	13 (1.77)	9 (1.84)	4 (1.65)	1.000
Taste sense altered	Sensory	8 (1.09)	8 (1.63)	0 (0.00)	0.104
Bleeding	Systemic	6 (0.82)	5 (1.02)	1 (0.41)	0.670
Drop in libido	Systemic	0 (0.00)	0 (0.00)	0 (0.00)	/

*: $P < 0.05$

Bold: emphasize that symptom was reported more frequently in depressed patients with BC and was statistically different.

Difference of the common physical symptoms between depressed and undepressed patients with BC

The frequency of common physical symptoms among the undepressed and depressed patients with BC was shown in Fig. 2. A total of 10 symptoms were mentioned with a frequency of more than 10%. Notably, compared to the self-reported symptoms in the depressive patients texts, there was a higher frequency for most of the common physical symptoms in undepressed patients texts, especially general pain (undepressed 63.67%, depressed 50.62%, $P=0.001$), and swelling of limbs (undepressed 24.29%, depressed 16.46%, $P=0.016$). Specifically, difficulty sleeping (undepressed 18.16%, depressed 28.40%; $P=0.002$) was the only common symptom mentioned more frequently than undepressed patients with BC.

Co-occurrence physical symptoms of depressed and undepressed patients with BC

The physical symptoms co-occurrence heatmaps of depressed and undepressed patients with BC were shown in Figs. 3 and 4. The shade of the colour block represented the proportion of times a symptom (on the x-axis) co-occurred with a different symptom (on the y-axis). The higher the proportion, the darker the colour block.

In the heatmap of depressed patients with BC, general pain, fatigue, vomiting, swelling of the limbs, difficulty sleeping, and nausea (on the y-axis) basically were dark colour blocks, since there were commonly mentioned physical symptoms. It explained the darker hues in the top half of the heatmap and the lighter hues in the bottom half. The colour block in the lower right corner became lighter, as the symptoms became less frequent. Rarely mentioned symptoms, such as shortness of breath, taste sense altered, bleeding, and drop in libido, were often mentioned in conjunction with common symptoms and rarely alone.

In the heatmap of undepressed patients with BC, the squares of general pain, fatigue, and vomiting (on the y-axis) were dark in colour. It exceeded the heatmap of depression in terms of the range of dark colour blocks.

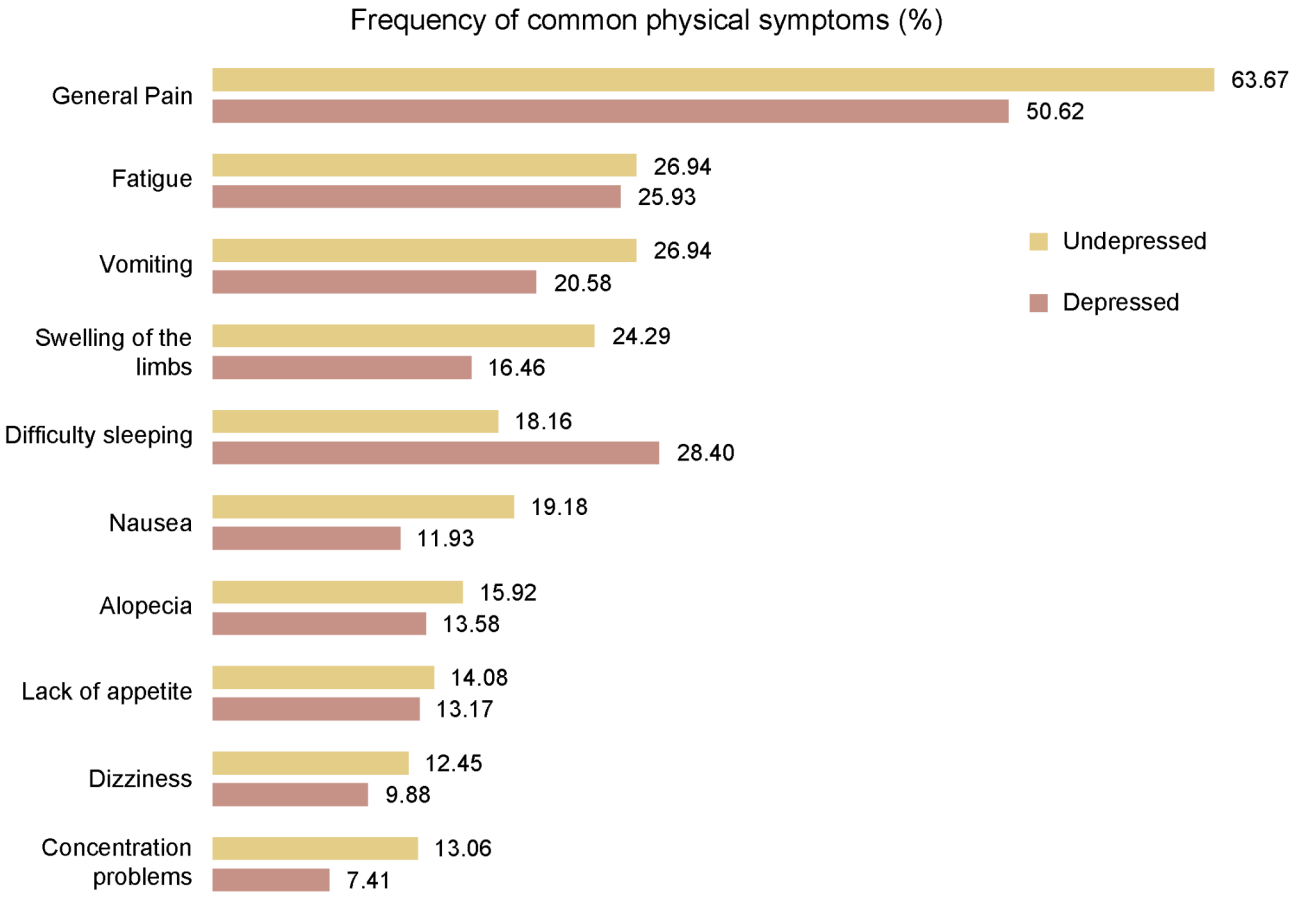

Fig. 2 Difference of physical common symptoms between depressed and undepressed patients with BC

There was no great difference in common symptoms (on the x-axis) between the depressed and undepressed heatmap. For the less commonly cited symptoms (on the x-axis), Fig. 3 appeared to show a strong correlation with a specific symptom, and Fig. 4 showed a generalized correlation.

Physical symptom clusters of depressed and undepressed patients with BC

The clustering results of the depressed and undepressed patients with BC were shown in the Figs. 5 and 6. Each circle represented a symptom, and the size of the circle represented the number of times each symptom occurred. Symptoms were connected by edges, and the thickness of the edges represented the similarity of the two symptoms. Different clusters were distinguished by different colours. Symptoms of the same cluster were filled with the same colour and enclosed in a rectangle. The medoid symptom of each cluster was marked with a solid circle, such as vomiting and general pain.

The best K value of clustering in depressed patients with BC was 5, and the symptoms fell into 5 clusters. Cluster 1 covered symptoms related to pain and difficulty sleeping, such as general pain, pain in the throat, head and face pain, limb pain, and difficulty sleeping. Cluster 2 primarily consisted of digestive symptoms, such as nausea, vomiting, loss of appetite, and stomach and abdominal pain. Cluster 3 contained the symptoms of skin changes and concentration problems, which may be associated with receiving anthracycline-based chemotherapeutic agents [40]. Cluster 4 seemed to indicate diarrhea, which would result in weight loss. Cluster 5 contained the symptoms of dizziness, fatigue, and cardiothoracic distress, which may be the result of poor circulation due to cardiotoxicity.

The best K value of clustering in the undepressed patients with BC was 6, and the symptoms were grouped into 6 clusters. Cluster 1 covered most of the symptoms associated with pain, such as general pain, back pain, and head and face pain, although swelling of the limbs and fatigue were included. The reason why fatigue is included is that pain is one of the factors that causes fatigue [41]. Cluster 2 comprised some of gastrointestinal symptoms, such as nausea and vomiting. Cluster 3 contained the symptoms of lack of appetite, difficulty sleeping and cardiothoracic distress. Some symptoms of cardiothoracic distress such as palpitations are associated with insomnia. Cluster 4 covered symptoms of fever and bleeding,

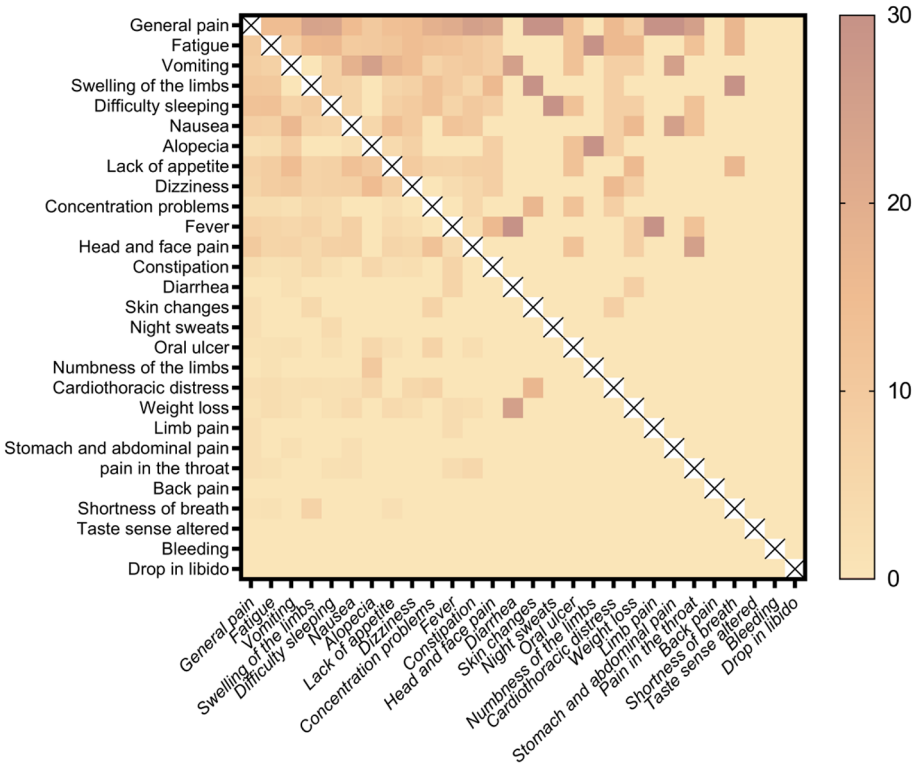

Fig. 3 The heatmap of depressed patients with BC. The heatmap is a representation of the percentage of occurrences of a symptom (on the y-axis) in co-occurrence with another (on the x-axis) out of all occurrences of the symptom (on the x-axis)

which appeared to be related to immune system impairment. Patients with BC undergoing chemotherapy experienced a drop in white blood cells and platelets, making them more susceptible to viral and bacterial invasion as well as unintended injuries. Cluster 5 contained symptoms of diarrhea, stomach and abdominal pain, and constipation, clearly suggesting disorder of digestive function. Cluster 6 seemed to defy interpretation, including the symptoms of shortness of breath, weight loss, limb pain, and numbness of the limbs.

Discussion

This study, based on social media and expressive written texts evaluated common physical symptoms, co-occurring physical symptoms, and physical symptom clusters between depressed and undepressed patients with BC to detect the difference in symptoms between these two groups.

In our study, pain was mentioned more frequently than other symptoms. On the one hand, pain, as a subjective feeling of patient, amplifies the patient's discomfort and their negative emotions [42]. On the other hand, this was due to the double categorization of pain, such as "头痛 (headache)" should be categorized only as head and face pain, but was actually categorized as general pain and head and face pain. Few reports addressed sexual health issues. Actually, patients with BC had dramatically worsened sexual health, especially those who had undergone mastectomy alone or were on hormone therapy [43, 44]. However, no patients reported this problem. It may be a result of differences in data sources and cultural background, with East Asian people reluctant to talk about sexuality in public [45]. Most symptoms were reported less frequently by depressed patients than by undepressed patients with BC. Leis et al. [46] found that depressed users wrote fewer characters in tweets than undepressed users, due to their reduced interest and language skills. Disfluencies in verbal expression were associated with depressive symptoms [47]. It suggests that depressed patients may experience a variety of symptoms, but few are mentioned in the text. Therefore, it was difficult to obtain the full information about symptoms through text alone. In the future, we need to combine traditional assessment tools with texts to better tap into the patient's symptom experience. For example, we can develop an effective system with features such as peer support and peer communities to help nurses better follow-up. Patients can post health-seeking texts. Healthcare providers can send health questionnaires to the patient and view the texts to understand the dynamic changes in the patient's symptoms for personalized interventions.

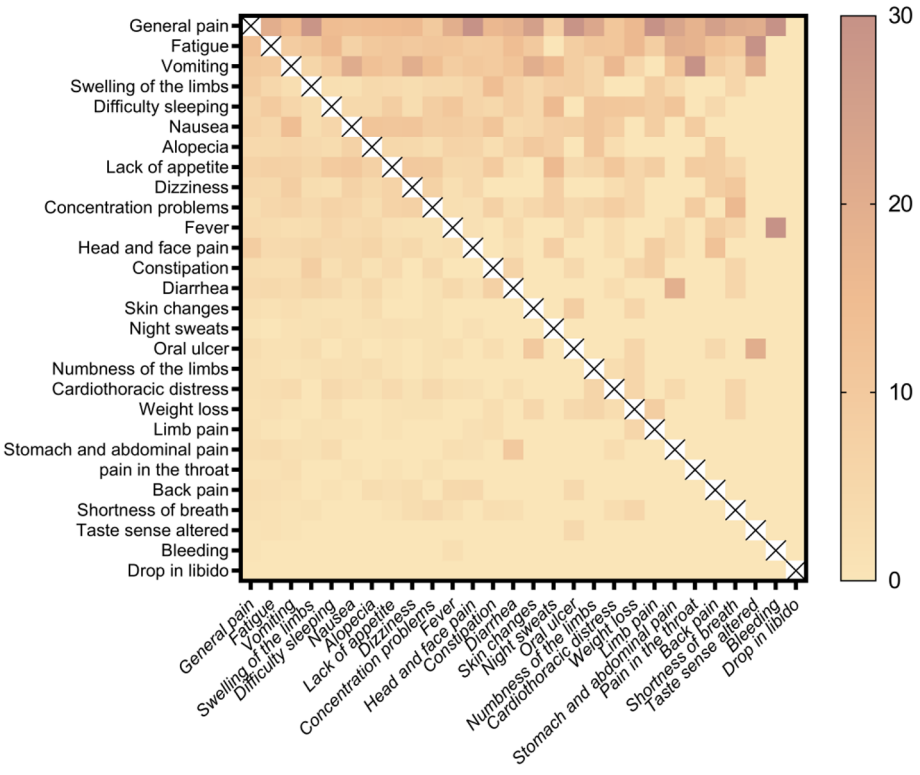

Fig. 4 The heatmap of undepressed patients with BC. The heatmap is a representation of the percentage of occurrences of a symptom (on the y-axis) in co-occurrence with another (on the x-axis) out of all occurrences of the symptom (on the x-axis)

High co-occurrence rates were found between general pain, fatigue, vomiting, difficulty sleeping, and nausea both in the heatmaps of undepressed and depressed patients with BC. This finding was in line with other studies [48]. Moreover, we found that rare symptoms are typically co-occurring with common symptoms. It could be explained by the fact that common physical symptoms persist during a BC patient's therapy while unusual symptoms are side effects of certain medicines. For instance, treatments that increase white blood cell counts, such as granulocyte colony-stimulating factor, often result in back pain [49].

The symptom of difficulty sleeping was reported more frequently in the depressed patients with BC, and was significant different from those reported by undepressed patients. In the clustering of depressed patients, sleep difficulties were clustered with pain-related symptoms. Previous studies have shown that pain and sleep disorders are highly co-morbid, with a recognized bidirectional relationship [50]. The hyperreactivity of the hypothalamic-pituitary-adrenal axis to stressors mediates the relationship between sleep deprivation and higher pain sensitivity [51]. Therefore, co-management of pain and sleep difficulties is of great clinical implication. Progressive muscle relaxation, and mindfulness-based stress reduction have been demonstrated to be effective measures in alleviating the difficulty sleeping-pain symptom cluster [52]. It suggests that we can provide targeted interventions for depressed patients undergoing such symptom cluster.

We identified novel symptom cluster consisting of fever and bleeding that are associated with immune system problems. In the analysis of symptom clusters experienced during chemotherapy for BC [18, 53], many studies reported pain-fatigue-sleep disorder symptom cluster, menopausal symptom cluster, gastrointestinal symptom cluster, and psychological cluster, with limited reports of bleeding and fever. It is noteworthy that the hazard ratio for myelosuppression following chemotherapy among BC patients is estimated to be 19% [54]. Therefore, it is necessary to pay attention to this cluster of symptoms in patients. Based on our clustering outcomes, health care providers can deliver a basket of targeted health education materials that encompass infection and bleeding prevention, regular blood monitoring. The authoritative and accessible nature of health information can promote patient self-management [55].

There were several limitations in our study. While text can provide detailed information about the disease journey, it is also a product of the patient's freely expressed thoughts and experiences. Thus, determining the treatment stage of the text is often challenging, which is why

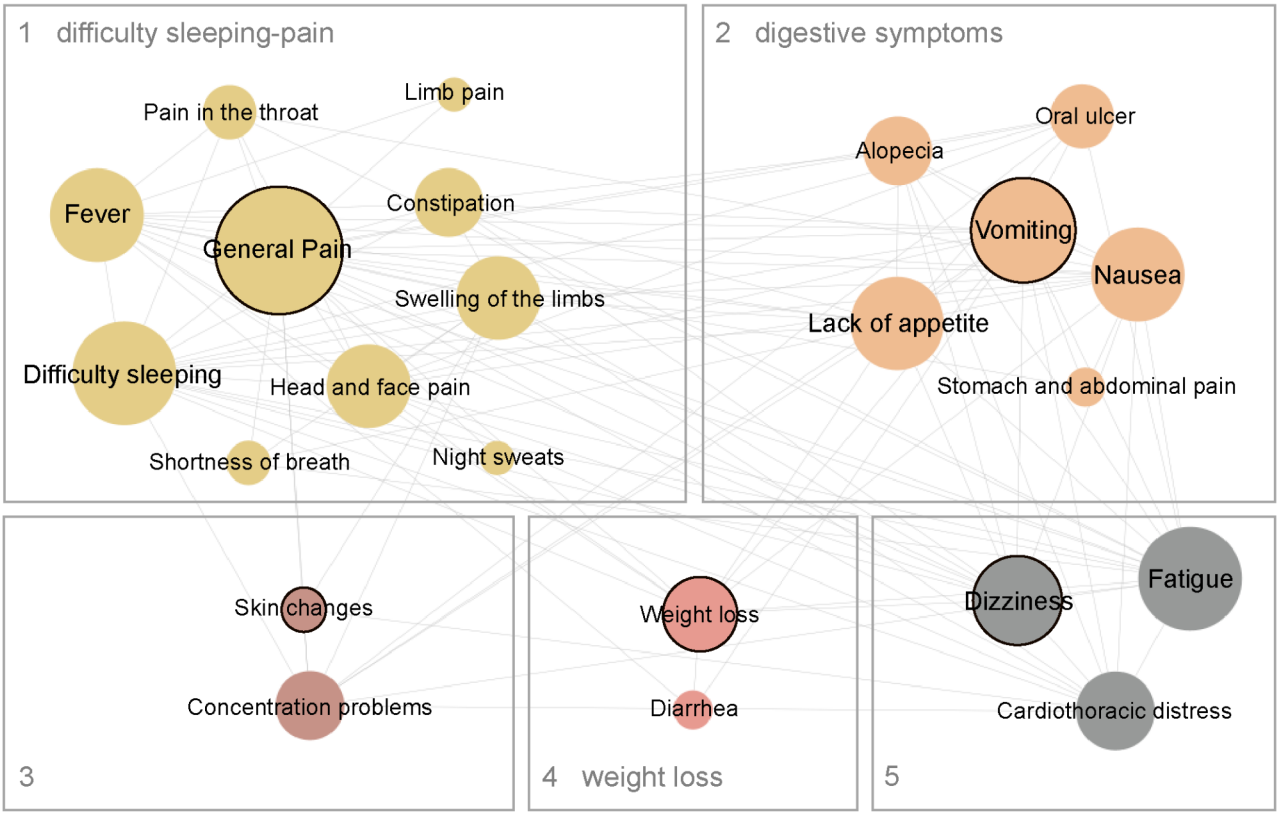

Fig. 5 Physical symptom clusters of depressed patients with BC. For example, Cluster 4 contained the symptoms of weight loss and diarrhea. And weight loss was at the center of the clustering

we could not track the trajectory of the patient's symptoms. Traditional clinical assessment allows us to visualize the severity of the symptoms for patients, but in text, we can not know the severity level of a symptom at the time of posting or recording. Second, the lack of a systematic Chinese consumer health lexicon meant that, despite our best efforts to enrich the symptom lexicon in BC, gaps may still exist. This could result in an underestimation of the physical symptom counts. Additionally, the Hospital Anxiety Depression Scale was used to assess the psychological state of the participants in the expressive writing texts. Although this scale has been widely used to screen potentially depressed patients, a clinical diagnosis of depression requires specialized doctors. Therefore, there was no guarantee that a patient exhibiting depressive symptoms was clinically depressed. There is a cultural bias in reporting symptoms, such as breast symptoms and sexual health. People with different cultural backgrounds differ in reporting such symptoms. For example, individuals from East Asia are often reluctant to disclose them in public [45].

Differences in physical symptoms are clinically significant in depressed and undepressed patients with BC. Sleep difficulties are particularly prominent in depressed patients, and progressive muscle relaxation can be offered to alleviate the symptom cluster of pain and sleep difficulties. Comprehensive health information can also be provided based on the clustering results when creating educational checklists. Future research should encompass longitudinal studies to track symptom trajectories and validate findings. Additionally, it is recommended that oncologists and digital health experts collaborate to further refine a systematic lexicon of cancer symptoms and clustering models, as they have a wide range of applications. For symptom monitoring on social media and identification of patient self-reported symptoms, the assistance of a symptom lexicon may be necessary.

Conclusions

We explore the differences in physical symptoms between depressed and undepressed patients with BC based on the texts derived from social media and expressive writing. Difficulty sleeping was reported more frequently in depressed patients, and it was clustered together with pain. The incidence of myelosuppression was high, and a symptom cluster of immune system impairment, consisting of fever and bleeding, was recognized in undepressed patients with BC. These findings may provide clinical

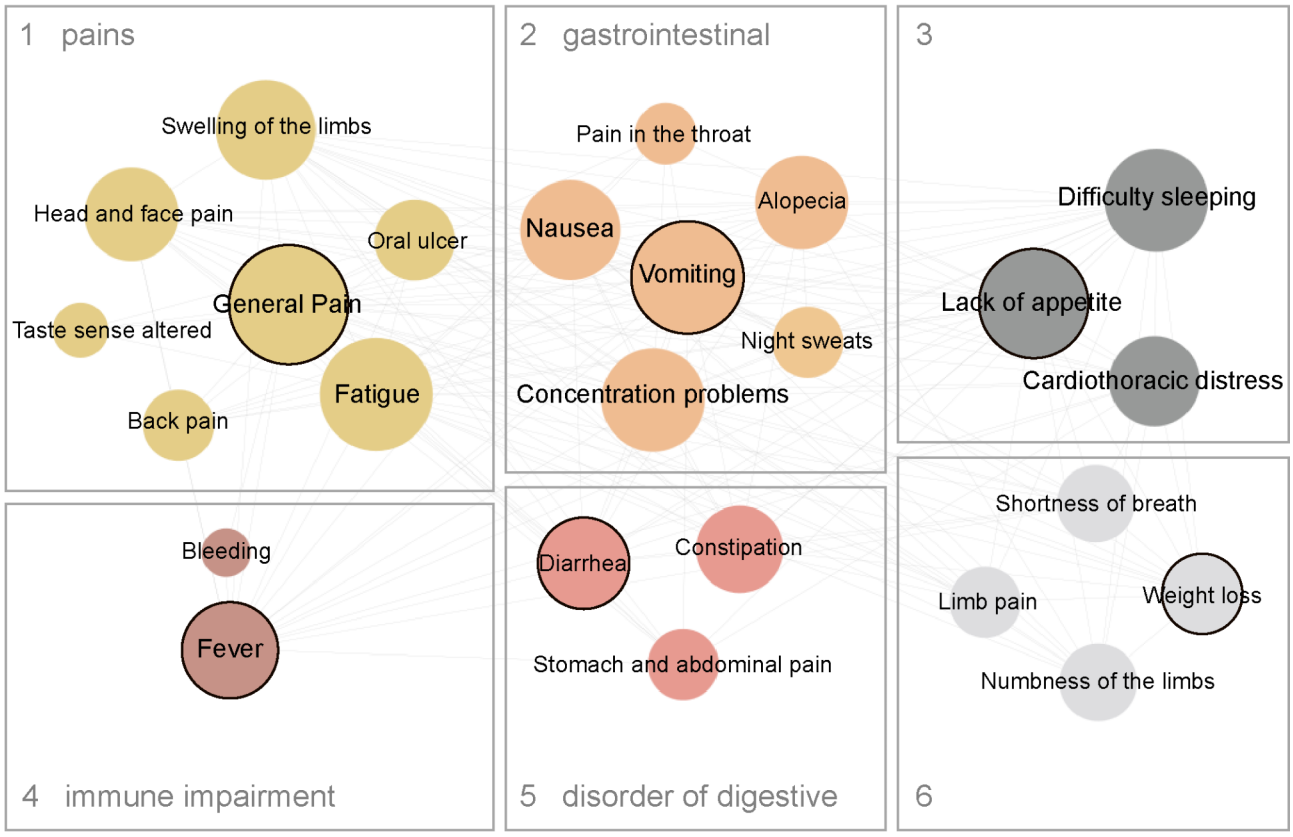

Fig. 6 Physical symptom clusters of undepressed patients with BC. For example, Cluster 4 contained the symptoms of fever and bleeding. And fever was at the center of the clustering

assistance in implementing intervention strategies for the management of diverse patients. Our findings demonstrate that textual data can provide a new perspective for understanding patients' symptom experiences. Due to the abundance and richness of social media texts, it will be more meaningful to combine them with traditional clinical data to understand patients' symptoms comprehensively.

List of abbreviations

ASW	Average Silhouette Width
BC	Breast Cancer
SW	Silhouette Width

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-024-13387-z.

Supplementary Material 1: Symptom Lexicon of BC.

Supplementary Material 2: Code of cluster analysis.

Acknowledgements
We sincerely thank all the participants for sharing their symptom experiences, and support from the 3 hospitals.

Author contributions
JT and YW were responsible for design of the research. JT, BG, CZ, and JC collected and analysed the data. JT wrote the manuscript. JF, JL, YZ, ZG, and SD raised the suggestion about the manuscript. All authors read and approved the final manuscript.

Funding
This study was funded by the National Natural Science Foundation of China (grant number 72304131) and the Outstanding Youths Development Scheme of Nanfang Hospital, Southern Medical University (grant number 2023J005).

Data availability
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Ethics approval
and consent participate.
This study followed the Declaration of Helsinki and ethical principles and guidelines. The Medical Ethics Committee of Nanfang Hospital of Southern Medical University approved the study (number NFEC-2022-484). Participants were informed and signed informed consent form prior to participating in the study.

Author details
[1]Nanfang Hospital, Southern Medical University, Guangzhou, China
[2]School of Nursing, Southern Medical University, Guangzhou, China

Received: 4 November 2024 / Accepted: 23 December 2024
Published online: 07 January 2025

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and Mortality Worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209–49.
2. Early. and Locally advanced breast cancer: diagnosis and management. London: National Institute for Health and Care Excellence (NICE); 2023.
3. Prieto-Callejero B, Rivera F, Fagundo-Rivera J, Romero A, Romero-Martín M, Gómez-Salgado J, et al. Relationship between chemotherapy-induced adverse reactions and health-related quality of life in patients with breast cancer. Med (Baltim). 2020;99:e21695.
4. Pilevarzadeh M, Amirshahi M, Afsargharehbagh R, Rafiemanesh H, Hashemi S-M, Balouchi A. Global prevalence of depression among breast cancer patients: a systematic review and meta-analysis. Breast Cancer Res Treat. 2019;176:519–33.
5. Wang X, Wang N, Zhong L, Wang S, Zheng Y, Yang B, et al. Prognostic value of depression and anxiety on breast cancer recurrence and mortality: a systematic review and meta-analysis of 282,203 patients. Mol Psychiatry. 2020;25:3186–97.
6. McFarland DC, Shaffer KM, Tiersten A, Holland J. Physical Symptom Burden and its Association with distress, anxiety, and depression in breast Cancer. Psychosomatics. 2018;59:464–71.
7. Renna ME, Shrout MR, Madison AA, Lustberg M, Povoski SP, Agnese DM, et al. Distress disorder histories relate to Greater physical symptoms among breast Cancer patients and survivors: findings across the Cancer Trajectory. Int J Behav Med. 2023;30:463–72.
8. Sheridan D, Foo I, O'Shea H, Gillanders D, Williams L, Fallon M, et al. Long-term follow-up of pain and emotional characteristics of women after surgery for breast cancer. J Pain Symptom Manage. 2012;44:608–14.
9. Fisher HM, Winger JG, Miller SN, Wright AN, Plumb Vilardaga JC, Majestic C, et al. Relationship between social support, physical symptoms, and depression in women with breast cancer and pain. Support Care Cancer off J Multinatl Assoc Support Care Cancer. 2021;29:5513–21.
10. Haemmerle R, Paludo J, Haddad TC, Pritchett JC. The growing role of Digital Health Tools in the care of patients with Cancer: current use, Future opportunities, and barriers to effective implementation. Curr Oncol Rep. 2024;26:593–600.
11. Aiello AE, Renson A, Zivich PN. Social media- and Internet-Based Disease Surveillance for Public Health. Annu Rev Public Health. 2020;41:101–18.
12. Lindvall C, Deng C-Y, Agaronnik ND, Kwok A, Samineni S, Umeton R, et al. Deep learning for Cancer symptoms monitoring on the basis of Electronic Health Record Unstructured Clinical Notes. JCO Clin Cancer Inf. 2022;6:e2100136.
13. Mazza M, Piperis M, Aasaithambi S, Chauhan J, Sagkriotis A, Vieira C. Social media listening to Understand the Lived experience of individuals in Europe with metastatic breast Cancer: a systematic search and content analysis study. Front Oncol. 2022;12:863641.
14. Ure C, Cooper-Ryan AM, Condie J, Galpin A. Exploring strategies for Using Social Media to Self-Manage Health Care when living with and beyond breast Cancer: In-Depth qualitative study. J Med Internet Res. 2020;22:e16902.
15. Avis NE, Levine BJ, Case LD, Naftalis EZ, Van Zee KJ. Trajectories of depressive symptoms following breast cancer diagnosis. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2015;24:1789–95.
16. Galiano-Castillo N, Ariza-García A, Cantarero-Villanueva I, Fernández-Lao C, Díaz-Rodríguez L, Arroyo-Morales M. Depressed mood in breast cancer survivors: associations with physical activity, cancer-related fatigue, quality of life, and fitness level. Eur J Oncol Nurs off J Eur Oncol Nurs Soc. 2014;18:206–10.
17. Grotmol KS, Lie HC, Loge JH, Aass N, Haugen DF, Stone PC, et al. Patients with advanced cancer and depression report a significantly higher symptom burden than non-depressed patients. Palliat Support Care. 2019;17:143–9.
18. So WKW, Law BMH, Ng MSN, He X, Chan DNS, Chan CWH, et al. Symptom clusters experienced by breast cancer patients at various treatment stages: a systematic review. Cancer Med. 2021;10:2531–65.
19. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T et al. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17). 2017. pp. 3838–44.
20. Wu Y, Liu L, Zheng W, Zheng C, Xu M, Chen X, et al. Effect of prolonged expressive writing on health outcomes in breast cancer patients receiving chemotherapy: a multicenter randomized controlled trial. Support Care Cancer off J Multinatl Assoc Support Care Cancer. 2021;29:1091–101.
21. Li C, Fu J, Lai J, Sun L, Zhou C, Li W, et al. Construction of an emotional lexicon of patients with breast Cancer: development and sentiment analysis. J Med Internet Res. 2023;25:e44897.
22. Tuinman MA, Gazendam-Donofrio SM, Hoekstra-Weebers JE. Screening and referral for psychosocial distress in oncologic practice: use of the Distress Thermometer. Cancer. 2008;113:870–8.
23. Cleeland CS, Mendoza TR, Wang XS, Chou C, Harle MT, Morrissey M, et al. Assessing symptom distress in cancer patients: the M.D. Anderson Symptom Inventory. Cancer. 2000;89:1634–46.
24. Wen C, Lu Q, Ding Y, Pang D, Yu X, Yang P, et al. Development and psychometric testing of the symptoms assessment inventory for breast cancer patients receiving chemotherapy. Chin J Nurs. 2012;47:451–3. [in Chinese].
25. Tuo B. Construction of Chinese Consumer Health Vocabulary Using Diabetes Mellitus As an Example. M.S. thesis. Huazhong University of Science and Technology; 2019.[in Chinese].
26. Yao R. Construction of Consumer Health Vocabulary for Chronic Disease and Evaluation of Vocabulary Familiarity in the Example of Hypertension. M.S. thesis. Huazhong University of Science and Technology; 2024.[in Chinese].
27. Hou L, Kang H, Liu Y, Li L, Li J. Mining and standardizing Chinese consumer health terms. BMC Med Inf Decis Mak. 2018;18:120.
28. China Medical information Platform. Liver Depression. https://www.dayi.org.cn/disease/1161477?from=sg. Accessed 04 June 2024.
29. International Digital Economy Academy. Biomedical Informatics Ontology System. https://bios.idea.edu.cn/. Assessed 04 June 2024.
30. U. S. National Institutes of Health, National Cancer Institute. SEER Training Modules, Review: Introduction to the Human Body. https://training.seer.cancer.gov/anatomy/body/review.html. Assessed 19 May 2024.
31. Wu J, Wang L, Hua Y, Li M, Zhou L, Bates DW, et al. Trend and Co-occurrence Network of COVID-19 symptoms from large-scale Social Media Data: Infoveillance Study. J Med Internet Res. 2023;25:e45419.
32. Esper P. Symptom clusters in individuals living with advanced cancer. Semin Oncol Nurs. 2010;26:168–74.
33. Xu D, Tian Y. A Comprehensive Survey of Clustering algorithms. Ann Data Sci. 2015;2:165–93.
34. Pan D, Han Y, Jin Q, Wu H, Huang H. Study of typical electric two-wheelers pre-crash scenarios using K-medoids clustering methodology based on video recordings in China. Accid Anal Prev. 2021;160:106320.
35. Arora P, Deepali, Varshney S. Analysis of K-Means and K-Medoids Algorithm for Big Data. Procedia Comput Sci. 2016;78:507–12.
36. Ushakov AV, Vasilyev I. Near-optimal large-scale k-medoids clustering. Inf Sci. 2021;545:344–62.
37. Kaufmann L, Rousseeuw P. Clustering by Means of Medoids. Data Anal Based L1-Norm Relat Methods. 1987;:405–16.
38. Batool F, Hennig C. Clustering with the average Silhouette Width. Comput Stat Data Anal. 2021;158:107190.
39. Rousseeuw PJ, Silhouettes. A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
40. Kesler SR, Blayney DW. Neurotoxic effects of Anthracycline- vs nonanthracycline-based chemotherapy on cognition in breast Cancer survivors. JAMA Oncol. 2016;2:185–92.
41. Fabi A, Bhargava R, Fatigoni S, Guglielmo M, Horneber M, Roila F, et al. Cancer-related fatigue: ESMO Clinical Practice guidelines for diagnosis and treatment. Ann Oncol off J Eur Soc Med Oncol. 2020;31:713–23.
42. Tan MY, Goh CE, Tan HH. Contemporary English Pain descriptors as detected on Social Media using Artificial intelligence and emotion analytics Algorithms: cross-sectional study. JMIR Form Res. 2021;5:e31366.
43. Huberts AS, Clarijs ME, Pastoor H, van Rosmalen M, Koppert LB. Sexual well-being in patients with early-stage breast cancer at 1- and 2-year follow-up. J Sex Med. 2023;20:507–14.
44. Nimbi FM, Magno S, Agostini L, Di Micco A, Maggiore C, De Cesaris BM, et al. Sexuality in breast cancer survivors: sexual experiences, emotions, and

cognitions in a group of women under hormonal therapy. Breast Cancer Tokyo Jpn. 2022;29:419–28.
45. Nicolosi A, Glasser DB, Kim SC, Marumo K, Laumann EO, GSSAB Investigators' Group. Sexual behaviour and dysfunction and help-seeking patterns in adults aged 40–80 years in the urban population of Asian countries. BJU Int. 2005;95:609–14.
46. Leis A, Ronzano F, Mayer MA, Furlong LI, Sanz F. Detecting signs of Depression in Tweets in Spanish: behavioral and linguistic analysis. J Med Internet Res. 2019;21:e14199.
47. O'Dea B, Boonstra TW, Larsen ME, Nguyen T, Venkatesh S, Christensen H. The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: a longitudinal study. PLoS ONE. 2021;16:e0251787.
48. Whisenant M, Wong B, Mitchell SA, Beck SL, Mooney K. Trajectories of depressed mood and anxiety during chemotherapy for breast cancer. Cancer Nurs. 2020;43:22–31.
49. Huang W, Liu J, Zeng Y, Wu F, Li N, Chen K, et al. Randomized controlled clinical trial of polyethylene glycol recombinant human granulocyte colony-stimulating factor in the treatment of neutropenia after chemotherapy for breast cancer. Cancer Chemother Pharmacol. 2018;82:607–13.
50. Haack M, Simpson N, Sethna N, Kaur S, Mullington J. Sleep deficiency and chronic pain: potential underlying mechanisms and clinical implications. Neuropsychopharmacol off Publ Am Coll Neuropsychopharmacol. 2020;45:205–16.
51. Goodin BR, Smith MT, Quinn NB, King CD, McGuire L. Poor sleep quality and exaggerated salivary cortisol reactivity to the cold pressor task predict greater acute pain severity in a non-clinical sample. Biol Psychol. 2012;91:36–41.
52. He C-C, Lin D-M, Liu H-Z, Wang F-F, Guo X-F, Zhang X-B, et al. Nonpharmacological Interventions for Management of the Pain-fatigue-sleep disturbance symptom cluster in breast Cancer patients: a systematic review and network Meta-analysis of Randomized controlled trials. J Pain Res. 2023;16:2713–28.
53. Qi Y, Li H, Guo Y, Cao Y, Wong CL. Symptom clusters in breast Cancer patients receiving adjuvant chemotherapy: a systematic review. J Clin Nurs. 2024;33:4554–67.
54. Nurgalieva Z, Liu C-C, Du XL. Chemotherapy use and risk of bone marrow suppression in a large population-based cohort of older women with breast and ovarian cancer. Med Oncol Northwood Lond Engl. 2011;28:716–25.
55. Thapa DK, Visentin DC, Kornhaber R, West S, Cleary M. The influence of online health information on health decisions: a systematic review. Patient Educ Couns. 2021;104:770–84.

Publisher's note