Original Article

Healthc Inform Res. 2025 April;31(2):175-188.
https://doi.org/10.4258/hir.2025.31.2.175
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Symptom and Sentiment Analysis of Older People with Cancer and Caregivers: A Text Mining Approach Using Korean Social Media Data

Kyunghwa Lee[1], **Soomin Hong**[2]

[1]College of Nursing, Konyang University, Daejeon, Korea
[2]Red Cross College of Nursing, Chung-Ang University, Seoul, Korea

Objectives: This study examined the symptoms and emotions expressed by older adults with cancer and their caregivers in South Korean online cancer communities. It aimed to identify narrative patterns and provide insights to inform personalized care strategies. **Methods:** We analyzed 6,908 user-generated posts collected from major online cancer communities in South Korea. Keyword frequency analysis, term frequency-inverse document frequency, 2-gram analysis, and latent Dirichlet allocation-based topic modeling were applied to explore language patterns. Sentiment analysis identified 12 emotional categories, and Pearson correlation coefficients were calculated to examine associations between symptoms and emotional expressions. All data were cleaned and standardized prior to analysis. **Results:** Many users expressed anxiety (20.63%) and depression (19.59%), frequently associated with chemotherapy and sleep disturbances. Among reported symptoms, sleep problems carried the highest negative sentiment (79.81%), underscoring their profound impact on well-being. Topic modeling consistently revealed seven recurring themes, including treatment decision-making, symptom management, and concerns about family, demonstrating the layered and personalized experiences of older cancer patients and their caregivers. **Conclusions:** This study explored treatment-related and symptom-related difficulties faced by older adults with cancer. Many reported significant emotional strain, especially anxiety, depression, and sleep disturbances. These findings highlight the necessity for supportive strategies addressing both psychological and physical aspects of care. Future research could investigate the utility of large language models in analyzing these narratives, provided the data is ethically managed and appropriate for such use.

Keywords: Natural Language Processing, Neoplasms, Aged, Signs and Symptoms, Caregivers

Submitted: February 3, 2025
Revised: 1st, March 28, 2025; 2nd, April 10, 2025
Accepted: April 15, 2025

Corresponding Author
Soomin Hong
Red Cross College of Nursing, Chung-Ang University. 84, Heukseok-ro, Building 102, Room 712, Dongjak-gu, Seoul 06974, Korea. Tel: +82-2-600-5645, E-mail: smhong@cau.ac.kr; soominsnow@naver.com (https://orcid.org/0000-0002-5884-1799)

I. Introduction

Age is a well-established risk factor for cancer, and cancer prevalence is increasing with population aging [1,2]. In 2020, the cancer incidence rate among South Koreans aged 65 and older was 1,552 per 100,000 people, significantly higher compared to younger age groups [3]. This trend is expected to impose considerable social and economic burdens on healthcare systems.

Advances in medical technology have significantly expanded therapeutic options and improved life expectancy for older adults with cancer. However, many still experience

substantial distress throughout their cancer journey, contributing to a high prevalence of frailty in this population [4]. Additionally, older adults often report negative emotions following a cancer diagnosis [5], and these emotional responses, together with frailty, are strongly associated with diminished quality of life [6].

Caregivers support older adults with cancer throughout diagnosis, treatment, and survivorship stages, frequently experiencing fear of death and significant caregiving burdens [7,8]. Thus, it is essential to consider both patient and caregiver well-being when planning supportive care. Nevertheless, large-scale studies focusing specifically on this population, especially those employing unstructured, real-world data such as online narratives, remain limited. Existing Korean research has primarily concentrated on digital literacy, treatment decision-making, and end-of-life planning [9,10], with relatively little attention to analyzing physical and psychological symptoms using natural language processing (NLP) techniques.

As unstructured digital text data become increasingly available, NLP and text mining techniques have emerged as valuable tools for health research [11,12]. By employing sentiment analysis and topic modeling, these approaches can uncover symptoms, emotional patterns, and thematic concerns. Advances in NLP could enhance the efficiency and scope of oncology research, potentially transforming clinical practice [13].

Therefore, this study aimed to explore the symptoms and emotions expressed by older adults with cancer and their caregivers by applying NLP and text mining techniques to posts from online cancer communities in South Korea. The insights gained are intended to guide the development of person-centered nursing interventions for this population.

II. Methods

1. Study Design

The current study employed an NLP and text mining approach to analyze symptom expressions and emotional states in Korean-language online posts authored by older cancer patients and their caregivers. The data were collected from major social media platforms in South Korea.

2. Data Collection

Data were collected from online cancer communities on prominent South Korean platforms such as Naver and Daum, covering the period from January 2010 to October 2024. These communities included forums dedicated to specific cancer types (e.g., breast, lung, colorectal) and general platforms where patients and caregivers discussed their treatment experiences and feelings.

To develop a corpus suitable for NLP, we implemented a text mining strategy focusing on posts by or concerning older adults with cancer, utilizing search terms reflective of this group's experiences. Examples included "seniors diagnosed with colorectal cancer," "cancer treatment in the elderly," and "experiences following cancer treatment." Python libraries Selenium and BeautifulSoup4 were used for data scraping, adhering to ethical guidelines and minimizing server load. Personal identifiers, such as usernames, were removed from the dataset.

From an initial set of 8,789 posts, we curated a refined corpus of 6,908 posts by removing duplicates, promotional content, and overly brief entries. Posts were categorized into 11 topic-based Excel sheets (e.g., "elderly cancer patients," "treatment experiences"), informed by previous research highlighting functional and comorbid considerations among older cancer patients [14].

To prevent overrepresentation of particular cancer types, we assessed the distribution of cancer discussions across forums. Although breast, lung, and prostate cancers were frequently discussed, stratification reviews confirmed balanced representation. NLP techniques, including topic modeling, were applied to the entire dataset to extract prevalent symptoms and emotional patterns.

Data processing was independently reviewed by a nursing professor and a data scientist, with discrepancies resolved through consensus. Demographic and clinical details varied due to the public nature of the posts. Although some authors explicitly identified themselves as patients or caregivers, many did not, precluding systematic classification. This limi-

Table 1. Descriptive statistics of the dataset

Category	Frequency
Total data count	6,907
Unique texts in the dataset	6,907
Sentence length (word count)	
Minimum	13
Maximum	383
Average	194.8
Median	205
Total word count	306,914
Total vocabulary size	87,409

https://doi.org/10.4258/hir.2025.31.2.175

Table 2. Frequency and TF-IDF (term frequency-inverse document frequency) analysis of keywords in the current dataset

Rank	Keywords	Frequency	Percentage (%)	TF-IDF
1	Older adults	4,770	2.40	0.023
2	Surgery	3,018	1.52	0.023
3	Chemotherapy	1,855	0.93	0.021
4	Premature ejaculation	1,505	0.76	0.016
5	Hepatocellular carcinoma	1,284	0.65	0.011
6	(in) Case	1,281	0.65	0.011
7	Abnormalities	1,216	0.61	0.011
8	Lung cancer	1,205	0.61	0.011
9	Colon cancer	1,202	0.61	0.011
10	Breast cancer	1,165	0.59	0.010
11	People	1,151	0.58	0.012
12	Aging	1,086	0.55	0.011
13	Occurrence	943	0.48	0.009
14	Immunity	904	0.46	0.009
15	Risk	889	0.45	0.008
16	Effect	854	0.43	0.008
17	(medical) Checkup	820	0.41	0.008
18	Health	809	0.41	0.008
19	Prevention	797	0.40	0.008
20	Because	760	0.38	0.008
21	Thoughts	727	0.37	0.009
22	Results	725	0.37	0.008
23	Reviews	711	0.36	0.008
24	Exercise	682	0.34	0.007
25	Degree	679	0.34	0.008
26	Function	670	0.34	0.006
27	Condition	669	0.34	0.008
28	Metastasis	646	0.33	0.008
29	Cause	640	0.32	0.006
30	Admission	615	0.31	0.007
31	Radiotherapy	567	0.29	0.007
32	Increasing	561	0.28	0.006
33	Progress	548	0.28	0.006
34	Method	543	0.27	0.006
35	Father	539	0.27	0.007
36	Mother	528	0.27	0.007
37	Gastric cancer	521	0.26	Not in the top 50
38	Pain	503	0.25	0.006
39	Now	493	0.25	0.007
40	Age	487	0.25	0.006
41	Research	486	0.24	Not in the top 50

Continued on the next page.

Table 2. Continued

Rank	Keywords	Frequency	Percentage (%)	TF-IDF
42	Treatment	484	0.24	Not in the top 50
43	Female	483	0.24	Not in the top 50
44	Cell	469	0.24	Not in the top 50
45	Dementia	467	0.23	0.006
46	Help	467	0.23	Not in the top 50
47	Other	460	0.23	Not in the top 50
48	Management	459	0.23	Not in the top 50
49	Screening	457	0.23	Not in the top 50
50	Symptoms	Not in the top 50	N/A	0.007
51	Diagnosis	Not in the top 50	N/A	0.009
52	Disease	Not in the top 50	N/A	0.012
53	Hospital	Not in the top 50	N/A	0.03
54	Patients	Not in the top 50	N/A	0.022
55	Examination	Not in the top 50	N/A	0.016
56	Physician	Not in the top 50	N/A	0.006

N/A: not applicable.

This table displays the top 50 keywords based on frequency and TF-IDF values. Some high-frequency function words, such as "because," "now," and "other," were retained due to their contextual importance in expressing causal reasoning, temporal status, or references to additional symptoms and treatment effects.

tation, along with potential engagement bias, was recognized in the analysis.

3. Data Analysis

The data analysis comprised several systematic steps designed to comprehensively examine the text data.

1) Text preprocessing

For processing Korean text, we employed the Okt tokenizer from the KoNLPy library [15], optimized for Korean language analysis. Stopwords—such as particles, verb endings, and general terms—were removed to improve the analytical accuracy. Additionally, spacing inconsistencies in synonymous terms (e.g., variations of "chemotherapy") were standardized using regular expressions.

2) Extraction and categorization of treatment and symptoms

We developed a custom symptom-treatment dictionary encompassing 22 categories, including "surgery," "chemotherapy," "radiotherapy," and "sleep disorders." To extract relevant mentions, substring matching was conducted using the str.contains() function from the pandas library (version 2.2.3) [15]. All applicable categories were recorded when multiple symptoms or treatments appeared in a single post, enabling comprehensive classification and frequency analysis.

Additionally, keyword frequency and bigram (2-gram) analyses identified commonly used terms and co-occurring word pairs within the corpus. These analyses provided insights extending beyond symptoms and treatments, highlighting recurring language patterns and frequent word associations. In some cases, high-frequency function words such as "because," "now," and "other" were retained due to their contextual significance in conveying causal explanations, temporal states, or referencing additional symptoms or treatments. Despite their generality, these words carried significant emotional and narrative implications in user expressions.

3) Sentiment analysis

Sentiment analysis was performed using a predefined system of 12 emotional categories: fear, anger, anxiety, loss, depression, frustration, gratitude, determination, acceptance, relief, calmness, and hope. Two researchers independently annotated emotional content, resolving disagreements by consensus. Sentiment scores were normalized to a 0–1 scale, where 0 indicated "not present at all" and 1 represented "strongly felt," ensuring consistent and interpretable emotion ratings.

https://doi.org/10.4258/hir.2025.31.2.175

Table 3. Frequent 2-gram phrases in the dataset

Category	Word 1	Word 2	Frequency
Treatment/Surgery	Treatment	Review	268
	Treatment	Method	224
	Surgery	Cost	186
	Treatment	Effect	117
	Surgery	Recovery	86
	Treatment	Process	66
Disease/Symptoms	Erectile dysfunction	Treatment	352
	Premature ejaculation	Treatment	343
	Chronic	Disease	183
	Premature ejaculation	Solution	149
	Premature ejaculation	Overcome	103
	Early	Symptoms	52
Examination/Test	Examination	Results	95
	Regular	Checkup	65
	Blood	Test	64
	Endoscopy	examination	61
Cost/Insurance	Surgery	Cost	186
	Health	Insurance	70
	Care	Insurance	68
Management	Lifestyle	Habits	124
	Health	Status	76
	Health	Management	48

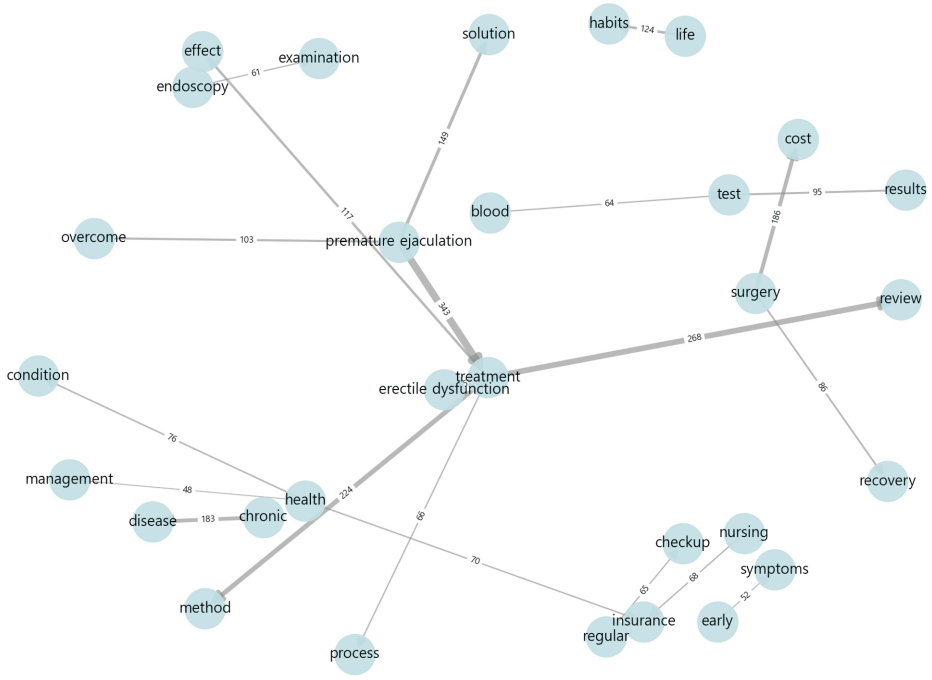

Figure 1. Keyword co-occurrence network based on 2-gram analysis from online cancer community posts.

4) Topic modeling

Topic modeling employed the latent Dirichlet allocation (LDA) algorithm from the gensim library. Optimal topic numbers were identified by calculating coherence and perplexity scores across a range of topics (2–10). Preprocessing and modeling were conducted using Python (version 3.11) with libraries including KoNLPy, gensim, scikit-learn, and re.

Korean-language texts were preprocessed using the Okt analyzer from KoNLPy. Stopwords, including topic/subject markers, verb endings (e.g., "do," "be"), and general functional terms (e.g., "about," "subject," "content"), were removed. Regular expressions standardized spacing inconsistencies in semantically equivalent terms (e.g., "anticancer treatment" as one or two words).

Key topic terms were extracted using term frequency-inverse document frequency (TF-IDF) values computed via the TfidfVectorizer function from the scikit-learn library (version 1.5.2) [16]. Topics lacking a clearly dominant term distribution were labeled qualitatively. Two researchers reviewed the top 10 keywords per topic and assigned thematic labels by consensus.

5) Sentiment correlation analysis of symptoms and treatments

Pearson correlation analysis explored associations between symptoms and sentiment. Correlation coefficients and corresponding p-values were calculated for each symptom or treatment category to determine statistical significance ($p < 0.05$).

4. Ethical Considerations

We adhered strictly to ethical guidelines while utilizing publicly shared data. Posts originated from open cancer forums on South Korean platforms (e.g., Naver, Daum), where users voluntarily shared experiences. All data were anonymized, excluding personally identifiable information. The study complied with legal requirements and each platform's policies, including those outlined in the Personal Information Protection Act.

To mitigate the risk of misinterpreting context-dependent social media content, two independent researchers reviewed sentiment and symptom categorizations, resolving discrepancies through consensus. Careful attention was given to preserving the intended meaning of each post.

Although the data were publicly accessible, ethical oversight was maintained throughout the study. Informed consent was not required, as the research involved no direct human subjects and met exemption criteria. We acknowledge ongoing ethical discussions concerning social media research and support continued efforts to refine best practices for digital health studies.

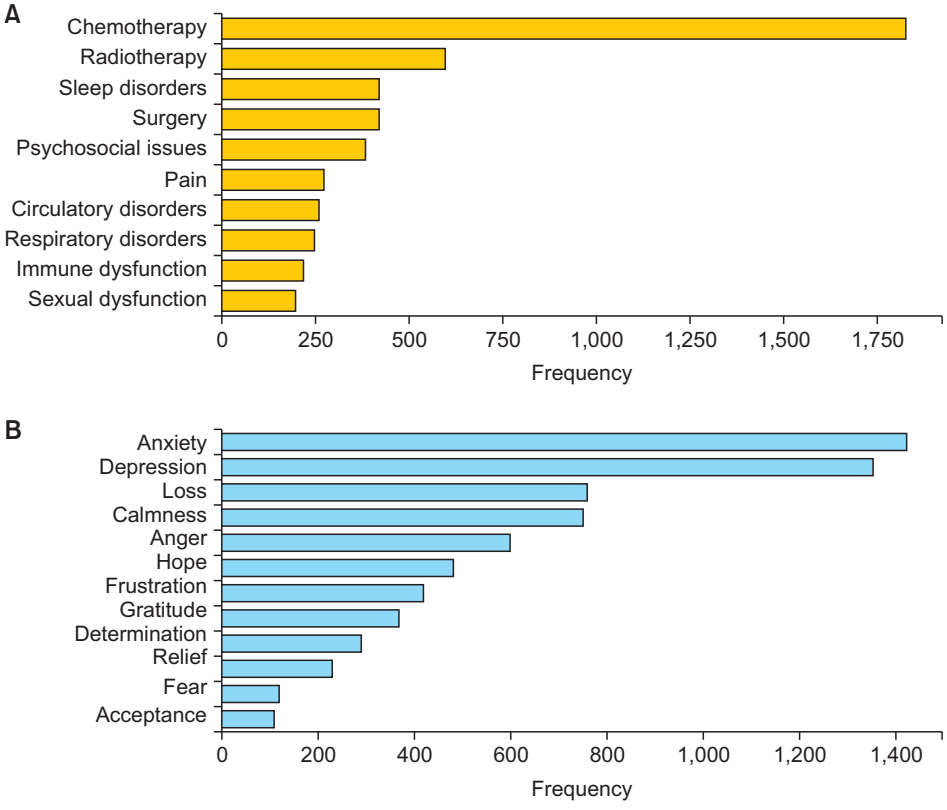

Figure 2. Distribution of symptom, treatment, and sentiment categories in online cancer community posts: (A) frequency of symptoms and treatments, (B) frequency of emotional sentiments.

https://doi.org/10.4258/hir.2025.31.2.175

Table 4. Integrated frequency and sentiment analysis of categorized symptoms and treatments

Rank	Category	Frequency (%)	List of frequently keywords related category (occurrence)	Positive (%)	Negative (%)	Neutral (%)	Total occurrences
1	Chemotherapy	1,828 (26.46)	Chemotherapy (1,783), Anticancer (503), Medication or drug (181)	15.81	65.81	18.38	1,828
2	Radiotherapy	597 (8.64)	Radiotherapy (399), Heavy ion therapy (40)	14.91	54.61	30.49	597
3	Sleep disorders	421 (6.09)	Worries (359), Insomnia (100)	11.4	79.81	8.79	421
4	Surgery	420 (6.08)	Surgery (174), Procedure (130), Anesthesia (51)	11.9	69.52	18.57	420
5	Psychosocial issues	386 (5.59)	Depression (155), Distress (143), Anxiety (99)	20.73	49.74	29.53	386
6	Pain	273 (3.95)	Painkiller (76), Abdominal pain (53), Headache (50)	8.79	63.37	27.84	273
7	Circulatory disorders	261 (3.78)	Bleeding (83), Anemia (60), Edema (49)	4.44	68.15	27.42	248
8	Respiratory disorders	248 (3.59)	Pneumonia (224), Dyspnea (20), Shortness of breath (11)	19.09	49.09	31.82	220
9	Immune dysfunction	220 (3.08)	Inflammation (175), Antibiotics (49)	4.02	41.21	54.77	199
10	Sexual dysfunction	199 (2.88)	Erectile dysfunction (187), Vaginitis (9), Decreasing libido (4)	15.81	65.81	18.38	1,828

III. Results

1. Overview of the Dataset

A descriptive analysis was conducted to examine the characteristics of the dataset, which comprised a total of 6,907 unique texts (Table 1). Sentence lengths ranged from 13 to 383 words, with a mean of 194.8 words and a median of 205 words, indicating relatively consistent text lengths. The dataset contained 306,914 total words and 87,409 unique terms, reflecting a combination of medical terminology and everyday language typically found in healthcare-related discourse.

2. Keyword Analysis Using Frequency and TF–IDF

Keyword analysis was performed by assessing term frequency and contextual relevance using TF-IDF (Table 2). Frequently occurring terms included "older adults," "surgery," and "chemotherapy," while TF-IDF analysis emphasized contextually significant yet less frequent terms, such as "premature ejaculation" and "hepatocellular carcinoma."

General terms like "research" and "treatment" demonstrated lower TF-IDF scores due to their broader usage, whereas terms such as "symptoms" and "diagnosis" exhibited higher TF-IDF scores despite their lower overall frequency. To confirm the relevance of specific sexual health-related terms, sample posts were manually reviewed. Most were related to prostate cancer treatments, and unrelated content was eliminated during preprocessing.

3. 2-Gram Analysis of Frequent Word Pairs

A 2-gram analysis was conducted to identify commonly co-occurring word pairs within the dataset, grouped into five thematic categories (Table 3, Figure 1). Particularly frequent were pairs such as "erectile dysfunction treatment" (352 occurrences), "premature ejaculation treatment" (343 occurrences), and "treatment method" (224 occurrences). The phrase "surgery cost" appeared in both the "treatment/surgery" and "cost/insurance" thematic categories.

4. Categorization of Symptoms and Treatment in the Dataset

Of the 6,907 texts analyzed, 3,836 (55.54%) explicitly mentioned symptoms and treatments, indicating that over half of the posts referenced these topics. An average of 0.85 symptoms or treatment categories was identified per post, with up to five distinct categories appearing in a single entry.

The most frequent category was "chemotherapy" (1,828 instances, 26.46%), followed by "radiotherapy" (8.64%), "sleep disorders" (6.09%), and "surgery" (6.08%) (Figure 2A, Table 4). Psychosocial issues (5.59%) were also commonly noted, with "depression" (155 occurrences), "distress" (143 occurrences), and "anxiety" (99 occurrences) standing out. Other symptoms appearing frequently included pain, cardiopulmonary symptoms, immune issues, and sexual dysfunction, in descending order of frequency.

5. Seven Key Topics Identified from Topic Modeling

To identify the optimal number of topics for LDA, coherence and perplexity scores were evaluated for 2 to 10 potential topics [17]. Coherence peaked at seven topics (0.62), and perplexity was lowest at four topics (-6.86). A seven-topic model was ultimately selected, balancing semantic clarity and thematic coverage. Coherence scores declined beyond seven topics, suggesting semantic redundancy (Figure 3).

Topic modeling (Table 5) identified seven themes: (1) common questions about cancer among older adults; (2) decision-making related to cancer diagnosis and treatment; (3) cancer diagnosis and treatment considerations; (4) new cancer therapies and associated risks; (5) immune-function strategies for older adults with cancer; (6) symptom management among older male cancer patients; and (7) symptom management among older female cancer patients.

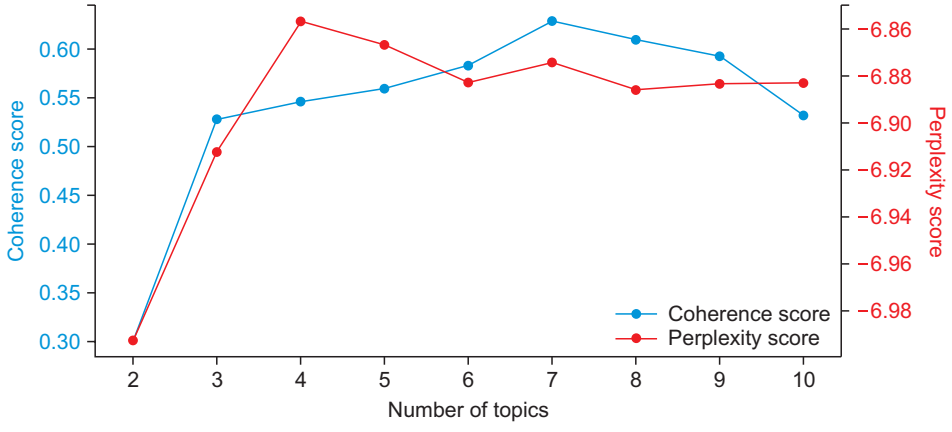

Figure 3. Optimal number of latent Dirichlet allocation topics based on coherence and perplexity scores.

https://doi.org/10.4258/hir.2025.31.2.175

Table 5. Key terms and weights from topic modeling in content from older adults with cancer

	Words	Weight
1. Common questions about cancer among older adults	Examination	0.0732
	Checkup	0.0398
	Abnormality	0.0256
	Breast cancer	0.0242
	Hepatocellular carcinoma	0.0222
	Colon cancer	0.0217
	Cancer screening	0.0204
	Function	0.0174
	Gastric cancer	0.0171
	Early	0.015
2. Decision-making in cancer diagnosis and treatment	Treatment	0.0751
	Surgery	0.0387
	Patients	0.0236
	Chemotherapy	0.0198
	Radiotherapy	0.0168
	Cancer care	0.013
	Metastasis	0.0126
	Review	0.0115
	In case	0.0115
	Anticancer	0.0109
3. Cancer diagnosis and consideration for treatments	Hospital	0.0403
	Surgery	0.0177
	Treatment	0.0152
	Chemotherapy	0.0143
	Consideration	0.0128
	Admission	0.0115
	Patients	0.011
	Mother	0.0097
	Biopsy	0.0096
	Father	0.0085
4. New cancer therapies and the risk	Effect	0.0211
	Prevention	0.0156
	Health	0.0132
	Occur	0.0131
	Research	0.0128
	Patients	0.0121
	Increasing	0.0121
	Colon cancer	0.0119
	Breast cancer	0.0118
	Risk	0.0111

Continued on the next page.

Table 5. Continued

	Words	Weight
5. Immune function strategies for older adults with cancer	Disease	0.0445
	Patients	0.034
	Pneumonia	0.0328
	Immunity	0.0279
	People	0.0233
	Distress	0.0144
	In case	0.0133
	Risk	0.0127
	Occur	0.0125
	Symptoms	0.0119
6. Symptom management for male older adults with cancer	Erectile dysfunction	0.2031
	PE	0.1343
	Treatment	0.1129
	Surgery	0.0457
	Review	0.0393
	Strategies	0.0322
	Workout	0.028
	PE symptom	0.0203
	Cost	0.0183
	Food	0.0176
7. Symptom management for female older adults with cancer	Constipation	0.0164
	Urinary incontinence	0.016
	Community	0.0158
	Treatment	0.014
	Uterus	0.0135
	Diabetes Mellitus	0.0131
	Pregnancy	0.0122
	Hypertension	0.0122
	Breast cancer	0.0118
	Disease	0.0117

PE: premature ejaculation.

6. Sentiment Analysis of Older Adults with Cancer

Sentiment analysis of 6,908 texts revealed that anxiety (1,425 occurrences, 20.63%) and depression (1,353 occurrences, 19.59%) were the most frequently expressed emotions (Figure 2B, Table 6). Anxiety was exemplified by statements such as, "My mother says she's terrified of chemotherapy and fears she won't endure it. I'm also scared—what if my desire only makes her suffer more?" Similarly, depression was evident in statements like, "This overwhelming situation makes me feel unprepared. I suppose I'll have to say goodbye soon." Loss (761 occurrences, 11.02%) and anger (599 occurrences, 8.67%) were also prevalent. Loss was reflected in statements such as, "She'll turn 80 next year, and with late-stage pancreatic cancer, what meaning does chemotherapy even have?" Anger was conveyed through statements including, "My mom quietly goes to the hospital for excessive treatments to avoid bothering her children. It's frustrating; though I love her, I feel irritated and upset." In contrast, positive emotions such as calmness (751 occurrences, 10.87%), hope (480 occurrences, 6.09%), and gratitude (369 occurrences, 5.34%)

 https://doi.org/10.4258/hir.2025.31.2.175

were relatively infrequent.

To ensure the reliability of sentiment annotations, inter-rater agreement was calculated using Cohen's kappa coefficient, yielding a value of 0.75, indicative of substantial agreement [18].

7. Sentiment Analysis of Treatment and Symptoms among Older Adults with Cancer

Pearson correlation analysis revealed weak but statistically significant relationships between certain symptoms and sentiment (Table 7). Sleep disorders exhibited a negative correlation with sentiment, while radiotherapy, respiratory, circulatory, and gastrointestinal issues were associated primarily with neutral sentiments ($p < 0.05$). Most symptoms lacked strong correlations, reflecting a predominantly neutral tone, possibly indicative of informational or inquiry-based content.

Chemotherapy (1,828 instances) had a high negative sentiment ratio (65.81%) (Table 4), followed closely by surgery (69.52%) and radiotherapy (54.61%). Sleep disorders showed the highest negative sentiment overall (79.81%).

Psychosocial issues (386 instances) demonstrated emotional variability, with 49.74% negative and 20.73% positive sentiments, indicating simultaneous distress and support. Pain (273 instances) and respiratory symptoms (248 instances) also showed predominantly negative sentiment. In contrast, immunity issues and sexual dysfunction were predominantly associated with neutral sentiments (31.82% and 54.77%, respectively), with sexual dysfunction showing the highest overall neutrality.

Table 6. Results of sentiment analysis of content from older adults with cancer

Rank	Sentiment	Frequency (%)
1	Anxiety	1,425 (20.63)
2	Depression	1,353 (19.59)
3	Loss	761 (11.02)
4	Calmness	751 (10.87)
5	Anger	599 (8.67)
6	Hope	480 (6.95)
7	Frustration	421 (6.09)
8	Gratitude	369 (5.34)
9	Determination	290 (4.20)
10	Relief	229 (3.31)
11	Fear	120 (1.74)
12	Acceptance	110 (1.59)

IV. Discussion

Despite advances in medical treatments, the emotional strain on older adults with cancer and their caregivers remains substantial. This study explored symptom expressions and emotional experiences among older adults with cancer and their caregivers through text mining and sentiment analysis of numerous online community posts.

The study used NLP techniques, including keyword frequency, TF-IDF, and 2-gram analyses, to identify recurring concerns and emotions within user narratives. Frequently mentioned terms such as "older adults," "surgery," and disease-specific terminology reflected clinical priorities, while references to "immunity" and "exercise" indicated interests related to health maintenance. These results emphasize the importance of patient-centered approaches addressing emotional, informational, and physical needs.

Our findings highlighted frequent mentions of male sexual

Table 7. Correlations between treatment, symptoms, and sentiments

Category	Negative correlation (p-value)	Neutral correlation (p-value)	Positive correlation (p-value)
Chemotherapy	-0.00002 (0.99)	0.015010 (0.21)	-0.01611 (0.18)
Radiotherapy	-0.01728 (0.15)	0.029857 (0.01)*	-0.00756 (0.53)
Respiratory issues	-0.01795 (0.14)	0.029803 (0.01)*	-0.00656 (0.56)
Circulatory issues	-0.01566 (0.20)	0.026421 (0.03)*	-0.00616 (0.61)
Sleep disorders	-0.02467 (0.04)*	0.017622 (0.14)	0.015997 (0.18)
Gastrointestinal issues	-0.01362 (0.26)	0.024026 (0.04)*	-0.00648 (0.59)
Pain	-0.01293 (0.28)	0.022284 (0.06)	-0.0056 (0.64)
Activity decrease	-0.01166 (0.33)	0.019669 (0.10)	-0.00458 (0.70)

*Significant correlation at $p < 0.05$.

dysfunction issues, notably treatments for premature ejaculation and erectile dysfunction. These concerns likely relate to other commonly cited terms such as surgery and colon cancer, as postoperative sexual dysfunction—including erectile dysfunction and ejaculatory disorders—is prevalent among male colorectal cancer patients [19]. In South Korea, older patients and their families often find discussing sexual health difficult, which might prompt them to seek related information online, possibly explaining the high frequency of these terms in our dataset.

The TF-IDF and 2-gram analyses identified terms that commonly co-occurred and carried significant contextual meaning. Expressions such as "erectile dysfunction treatment," "surgery cost," and "treatment reviews" indicated patient concerns about health outcomes and suggested reliance on shared patient experiences for treatment decisions [20]. These findings underline the necessity for comprehensive, patient-centered care that addresses both medical and psychosocial dimensions.

Seven major topics related to cancer diagnosis, treatment decisions, symptom experiences, and exploration of new therapies emerged from topic modeling. The results suggest that older adults with cancer and their caregivers frequently turn to online communities for practical information and emotional support, which may not be fully provided during brief clinical interactions. Given the rising digital health engagement among older populations [21], expanding online support interventions may help fulfill their care needs and bridge gaps within current healthcare services.

Sentiment analysis indicated significant psychological distress among older adults with cancer, underscoring the critical need for emotional support. Anxiety and depression emerged as the most prevalent emotions, aligning with previous studies that identified the psychological impacts of cancer treatment, including concerns about side effects and uncertainty regarding outcomes [22,23]. Furthermore, our findings highlighted the substantial negative impact of sleep disturbances, which are known to exacerbate discomfort and reduce overall quality of life among cancer patients [24,25]. Thus, tailored interventions targeting improvements in sleep quality and emotional support are urgently needed.

Previous NLP research has examined emotional narratives related to cancer. For example, an analysis of lung cancer records found strong negative emotions linked to physical decline and concerns about treatment outcomes, consistent with our findings on anxiety and distress [26]. Another complementary study analyzing Reddit posts identified diverse emotional responses, including fear, sadness, and even unexpected moments of hope and joy [27]. In contrast, our study employed clearly defined NLP tools such as TF-IDF, 2-gram, and LDA analyses, enhancing our ability to systematically capture and understand emotional patterns in online communities.

Two complementary analyses were conduced to clarify emotional patterns associated with symptoms. While most symptoms exhibited minimal correlation with sentiment, sleep disorders showed a weak but statistically significant association with neutral sentiment. Psychosocial issues were predominantly associated with negative emotions, yet the considerable presence of neutral and positive responses suggested a more complex and dynamic emotional landscape [28]. Healthcare providers should recognize this variability, fostering emotional validation and supporting positive coping strategies.

Although several correlations between symptoms and sentiment were statistically significant ($p < 0.05$), their overall associations were generally weak. This might reflect the inherent limitations of user-generated content, which tends to be brief, contextually variable, and structurally inconsistent. These results should thus be interpreted cautiously, as they indicate general patterns rather than clear causal relationships. As noted in previous studies, individuals often describe the connection between physical symptoms and emotions as ambiguous or unique to each person [29]. Even individuals experiencing identical symptoms can differ markedly in how they are emotionally affected [30]. Future research could incorporate structured data collection methods or qualitative techniques to capture more detailed personal experiences.

This study has several limitations. Given the nature of social media, much of the content analyzed originated from family members caring for older adults. Consequently, we might not fully capture older adults' direct perspectives, potentially omitting essential aspects of their personal experiences.

While online sources offer valuable insights into people's experiences, they also have inherent limitations. For example, perspectives of individuals without internet access are not represented, restricting the diversity of viewpoints captured. This limitation highlights the need for comprehensive research incorporating varied perspectives from older adults concerning their health and care experiences.

Despite these limitations, the study provides meaningful insights into the experiences of older adults with cancer and their caregivers regarding symptom burden and emotional challenges. Future research could more clearly distinguish

 https://doi.org/10.4258/hir.2025.31.2.175

between patient and caregiver experiences by integrating structured data from electronic health records or through questionnaire-based methodologies. Additionally, advancements in NLP may facilitate automated speaker identification, enabling detailed subgroup analyses. Such methodological improvements could substantially enhance our understanding of this population's unique needs. Finally, our findings establish a foundational basis for developing psychological support interventions, and future research could further guide customized symptom management strategies suitable for clinical application.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. RS-2024-00346310).

ORCID

Kyunghwa Lee (https://orcid.org/0000-0003-0210-9176)
Soomin Hong (https://orcid.org/0000-0002-5884-1799)

References

1. National Cancer Institute. Age and cancer risk [Internet]. Bethesda (MD): National Cancer Institute; 2021 [cited at 2025 Apr 22]. Available from: https://www.cancer.gov/about-cancer/causes-prevention/risk/age.
2. Pilleron S, Sarfati D, Janssen-Heijnen M, Vignat J, Ferlay J, Bray F, et al. Global cancer incidence in older adults, 2012 and 2035: a population-based study. Int J Cancer 2019;144(1):49-58. https://doi.org/10.1002/ijc.31664
3. National Cancer Information Center. Cancer in statistics: age-specific cancer incidence rate [Internet]. Goyang, Korea; National Cancer Information Center; c2025 [cited at 2025 Apr 22]. Available from: https://www.cancer.go.kr/lay1/S1T639C642/contents.do.
4. Ethun CG, Bilen MA, Jani AB, Maithel SK, Ogan K, Master VA. Frailty and cancer: implications for oncology surgery, medical oncology, and radiation oncology. CA Cancer J Clin 2017;67(5):362-77. https://doi.org/10.3322/caac.21406
5. Abdallah M, Kadambi S, Parsi M, Rai M, Mendler JH, Wittink M, et al. Older patients' experiences following initial diagnosis of acute myeloid leukemia: a qualitative study. J Geriatr Oncol 2022;13(8):1230-5. https://doi.org/10.1016/j.jgo.2022.08.017
6. Xiao M, Chen X, Ji L, Qian X, Xiu M, Li Z, et al. Prevalence of Frailty and Its Impact on Quality of Life in Older Patients With Breast Cancer: A Prospective Cross-Sectional Study. J Clin Nurs 2024 Dec 9 [Epub]. https://doi.org/10.1111/jocn.17599.
7. Webb K, Sharpe L, Butow P, Dhillon H, Zachariae R, Tauber NM, et al. Toward the development of a model of caregiver-specific fear of cancer recurrence: a systematic review. J Psychosoc Oncol Res Pract 2022;4(3):1-10. https://doi.org/10.1097/OR9.0000000000000082
8. Choi YS, Bae JH, Kim NH, Tae YS. Factors influencing burden among family caregivers of elderly cancer patients. Asian Oncol Nurs 2016;16(1):20-9. https://doi.org/10.5388/aon.2016.16.1.20
9. Kim EY, Hong SJ. Decision-making experience of older patients with cancer in choosing treatment: a qualitative meta-synthesis study. J Korean Gerontol Nurs 2021;23(4):418-30. https://doi.org/10.17079/jkgn.2021.23.4.418
10. Su KH, Hyun SJ. The influence of digital informatization level, self-efficacy, and social support on digital health literacy in the elderly with cancer. Asian Oncol Nurs 2022;22(4):255-63. http://doi.org/10.5388/aon.2022.22.4.255
11. Movva R, Balachandar S, Peng K, Agostini G, Garg N, Pierson E. Topics, authors, and institutions in Large Language Model research: trends from 17K arXiv papers. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); 2024 Jun 16-21; Mexico City, Mexico. p. 1223-43. https://doi.org/10.18653/v1/2024.naacl-long.67
12. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3(1):141. https://doi.org/10.1038/s43856-023-00370-1
13. Webster P. Six ways large language models are changing healthcare. Nat Med 2023;29(12):2969-71. https://doi.org/10.1038/s41591-023-02700-1
14. Given B, Given CW. Older adults and cancer treat-

ment. Cancer 2008;113(12 Suppl):3505-11. https://doi.org/10.1002/cncr.23939
15. Pandas Development Team. Pandas: Python data analysis library [Internet]. London, UK: Pandas; 2024 [cited at 2025 Apr 22]. Available from: https://pandas.pydata.org/docs/reference/index.html.
16. Scikit-learn Development Team. Scikit-learn: machine learning in Python [Internet]. Paris, France: Scikit-learn Developers: 2023 [cited at 2025 Apr 22]. Available from: https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.
17. Hasan M, Rahman A, Karim MR, Khan MS, Islam MJ. Normalized approach to find optimal number of topics in latent Dirichlet allocation (LDA). In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering. Singapore: Springer; 2021. p. 341-54. https://doi.org/10.1007/978-981-33-4673-4_27
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-74. https://doi.org/10.2307/2529310
19. Towe M, Huynh LM, El-Khatib F, Gonzalez J, Jenkins LC, Yafi FA. A review of male and female sexual function following colorectal surgery. Sex Med Rev 2019;7(3):422-9. https://doi.org/10.1016/j.sxmr.2019.04.001
20. DuMontier C, Loh KP, Soto-Perez-de-Celis E, Dale W. Decision making in older adults with cancer. J Clin Oncol 2021;39(19):2164-74. https://doi.org/10.1200/JCO.21.00165
21. Zhou W, Cho Y, Shang S, Jiang Y. Use of digital health technology among older adults with cancer in the United States: findings from a national longitudinal cohort study (2015-2021). J Med Internet Res 2023;25:e46721. https://doi.org/10.2196/46721
22. Lee AR, Leong I, Lau G, Tan AW, Ho RC, Ho CS, et al. Depression and anxiety in older adults with cancer: Systematic review and meta-summary of risk, protective and exacerbating factors. Gen Hosp Psychiatry 2023;81:32-42. https://doi.org/10.1016/j.genhosppsych.2023.01.008
23. Jayani RV, Hamparsumian A, Sun C, Li D, Cabrera Chien L, Moreno J, et al. The relationship of mental health symptoms to chemotherapy toxicity risk in older adults with cancer: results from the geriatric assessment-driven intervention study. Cancer 2024;130(22):3894-901. https://doi.org/10.1002/cncr.35482
24. Buttner-Teleaga A, Kim YT, Osel T, Richter K. Sleep disorders in cancer: a systematic review. Int J Environ Res Public Health 2021;18(21):11696. https://doi.org/10.3390/ijerph182111696
25. Di Nardo P, Lisanti C, Garutti M, Buriolla S, Alberti M, Mazzeo R, et al. Chemotherapy in patients with early breast cancer: clinical overview and management of long-term side effects. Expert Opin Drug Saf 2022;21(11):1341-55. https://doi.org/10.1080/14740338.2022.2151584
26. Elbers DC, La J, Minot JR, Gramling R, Brophy MT, Do NV, et al. Sentiment analysis of medical record notes for lung cancer patients at the Department of Veterans Affairs. PLoS One 2023;18(1):e0280931. https://doi.org/10.1371/journal.pone.0280931
27. Lal DM, Rayson P, Payne SA, Liu Y. Analysing emotions in cancer narratives: a corpus-driven approach. Proceedings of the 1st Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024; 2024 May 20; Torino, Italia. p. 73-83.
28. Silva S, Bartolo A, Santos IM, Pereira A, Monteiro S. Towards a better understanding of the factors associated with distress in elderly cancer patients: a systematic review. Int J Environ Res Public Health 2022;19(6):3424. https://doi.org/10.3390/ijerph19063424
29. Bekhuis E, Gol J, Burton C, Rosmalen J. Patients' descriptions of the relation between physical symptoms and negative emotions: a qualitative analysis of primary care consultations. Br J Gen Pract 2020;70(691):e78-e85. https://doi.org/10.3399/bjgp19X707369
30. Ebrahimi OV, Borsboom D, Hoekstra RH, Epskamp S, Ostinelli EG, Bastiaansen JA, et al. Towards precision in the diagnostic profiling of patients: leveraging symptom dynamics as a clinical characterisation dimension in the assessment of major depressive disorder. Br J Psychiatry 2024;224(5):157-63. https://doi.org/10.1192/bjp.2024.19