# Cancer Prevention and Treatment on Chinese Social Media: Machine Learning-Based Content Analysis Study

Keyang Zhao, Xiaojing Li, Jingyang Li

Submitted to: Journal of Medical Internet Research
on: January 02, 2024

**Disclaimer: © The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

# Table of Contents

# Cancer Prevention and Treatment on Chinese Social Media: Machine Learning-Based Content Analysis Study

Keyang Zhao[1*] DPhil; Xiaojing Li[1, 2*] Prof Dr; Jingyang Li[3] DPhil

[1]School of Media & Communication Shanghai Jiao Tong University Shanghai CN
[2]Institute of Psychology and Behavioral Science Shanghai Jiao Tong University Shanghai CN
[3]School of Software Shanghai Jiao Tong University Shanghai CN
[*]these authors contributed equally

**Corresponding Author:**
Xiaojing Li Prof Dr
School of Media & Communication
Shanghai Jiao Tong University
800 Dongchuan Rd.
Minhang District
Shanghai
CN

## Abstract

**Background:** Nowadays, new media has played an important role in providing information about cancer prevention and treatment. A growing body of work has been devoted to examining the access and communication effects of cancer information on social media. However, there has been limited understanding of the overall presentation of cancer prevention and treatment on social media. Further, research on comparing the differences between medical social media and common social media remained limited.

**Objective:** Based on big data analytics, this study aimed to comprehensively map the characteristics of cancer treatment and prevention information on medical social media and common social media, which was promisingly helpful in cancer coverage and patients' treatment decision.

**Methods:** We collected all posts (N=60,843) from 4 medical WeChat official accounts (classified as medical social media in this paper), and 5 health and lifestyle WeChat official accounts (classified as common social media in this paper). By applying latent Dirichlet allocation topic model, we extracted cancer-related posts (N=8,427) and obtained 6 cancer themes in common social media and medical social media separately. After manually labeling posts according to our codebook, we adopted a neural-based method to label different articles automatically. To be more specific, we defined our task as a multi-label task and chose different pre-trained models, say, Bert and Glove, to learn document level semantic representations for labelling.

**Results:** Themes in common social media were more related to lifestyle, while medical social media were more related to medical attributions. Early screening and testing, healthy diet, and physical exercise were the most frequently mentioned preventive measures. Compared with common social media, medical social media mentioned vaccinations to prevent cancer more frequently. Both types of media provided limited coverage of radiation prevention (including sun protection) and breastfeeding. Surgery, chemotherapy, and radiotherapy were the most mentioned treatment measures. Medical social media discussed treatment information more than common social media.

**Conclusions:** Cancer prevention and treatment information on social media revealed a lack of balance. The focus on cancer prevention and treatment information was mainly limited to a few aspects. The cancer coverage on preventive measures and treatments in social media required further improvement. Additionally, the study's findings underscored the potential of applying machine learning to content analysis as a promising research paradigm for mapping the key dimensions of cancer information on social media. The findings provided methodological and practical significance in future study and health promotion.

(JMIR Preprints 02/01/2024:55937)
DOI: https://doi.org/10.2196/preprints.55937

### Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

✓ **Only make the preprint title and abstract visible.**

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in  <a href="http

# Original Manuscript

# Cancer Prevention and Treatment on Chinese Social Media□Machine Learning-Based Content Analysis Study

Keyang Zhao[1], Xiaojing Li[1, 2*], Jingyang Li[3]

[1]*School of Media & Communication, Shanghai Jiao Tong University, Shanghai, China*

[2]*Institute of Psychology and Behavioral Science, Shanghai Jiao Tong University, Shanghai, China*

[3]*School of Software, Shanghai Jiao Tong University, Shanghai, China*

*The first two authors contributed equally.*

* **Corresponding author:** Xiaojing Li

**Correspondence address:** School of Media & Communication, Shanghai Jiao Tong University, 800 Dongchuan Rd., Shanghai 200240, P. R. China

**Email:** lixiaojing@sjtu.edu.cn

**Phone:** +86-21-34207088; +86-13918611103

**Biographical note:**

***Xiaojing Li*** is a professor and the Vice Dean of the School of Media & Communication at Shanghai Jiao Tong University in China. Her research focused on new media uses and effects, health and medical communication, especially interested in the role of new media played in public health and Chinese society.

***Keyang Zhao*** is a doctoral student of the School of Media & Communication at Shanghai Jiao Tong University. Her research focused on new media use and effects, health and medical communication.

***Jingyang Li*** is a doctoral student of the School of Software at Shanghai Jiao Tong University. His research focused on natural language understanding.

Original Paper

# Cancer Prevention and Treatment on Chinese Social Media□Machine Learning-Based Content Analysis Study

## Abstract

**Background:** Nowadays, new media has played an important role in providing information about cancer prevention and treatment. A growing body of work has been devoted to examining the access and communication effects of cancer information on social media. However, there has been limited understanding of the overall presentation of cancer prevention and treatment on social media. Further, research on comparing the differences between medical social media and common social media remained limited.

**Objective:** Based on big data analytics, this study aimed to comprehensively map the characteristics of cancer treatment and prevention information on medical social media and common social media, which was promisingly helpful in cancer coverage and patients' treatment decision.

**Methods:** We collected all posts (N=60,843) from 4 medical WeChat official accounts (accounts with professional medical backgrounds, classified as medical social media in this paper), and 5 health and lifestyle WeChat official accounts (accounts with non-professional medical backgrounds, classified as common social media in this paper). By applying latent Dirichlet allocation topic model, we extracted cancer-related posts (N=8,427) and obtained 6 cancer themes in common social media and medical social media separately. After manually labeling posts according to our codebook, we adopted a neural-based method to label different articles automatically. To be more specific, we defined our task as a multi-label task and chose different pre-trained models, say, Bert and Glove, to learn document level semantic representations for labelling.

**Results:** A total of 4479 articles from medical social media (MSM) and 3948 articles from common social media (CSM) related to cancer were analyzed. Among these, 35.52% (2993/8427) contained preventive information, and 44.43% (3744/8427) contained treatment information. Themes in common social media were more related to lifestyle, while medical social media were more related to medical attributions. Early screening and testing, healthy diet, and physical exercise were the most frequently mentioned preventive measures. Compared with common social media, medical social media mentioned vaccinations to prevent cancer more frequently. Both types of media provided limited coverage of radiation prevention (including sun protection) and breastfeeding. Surgery, chemotherapy, and radiotherapy were the most mentioned treatment measures. Compared to MSM (13.49%, 1137/8427), CSM (35.52%, 2993/8427) focused more on prevention.

**Conclusions:** Cancer prevention and treatment information on social media revealed a lack of balance. The focus on cancer prevention and treatment information was mainly limited to a few aspects. The cancer coverage on preventive measures and treatments in social media required further improvement. Additionally, the study's findings underscored the potential of applying machine learning to content analysis as a promising research paradigm for mapping the key dimensions of cancer information on social media. The findings provided methodological and practical significance in future study and health promotion.

**Keywords:** Social media; cancer information; text mining; supervised machine learning; content analysis

## Introduction

In 2020, 4.57 million new cancer cases emerged in China, constituting 23.7% of the world's total [1]. Notably, many of these cancers can be prevented [2, 3]. According to the World Health Organization (WHO), 30-50% of cancers could be prevented by early detection and reducing exposure to known lifestyle and environmental risks [4]. This emphasized the imperative to advance education on cancer prevention and treatment.

Mass media was not only a primary channel for disseminating cancer information but also a potent force in setting health agenda for the public [5, 6]. Previous studies have underscored the necessity of understanding how certain cancer-related content was presented in the media. For example, the frequently mentioned specific cancer types in news reports had the potential to influence the public's perception of the actual incidence of cancer [7].

Nowadays, social media has played an essential role in disseminating health information, coordinating resources, and promoting health campaigns aimed at educating individuals about preventative measures [8]. Additionally, it influenced patients' decision-making processes regarding treatment [9]. A study revealed that social media use correlated with enhancing awareness of cancer screening in the general population [10]. In recent years, there has been a notable surge in studies evaluating cancer-related content on social media. Nevertheless, previous studies usually concentrated on specific cancer types [11] and limited aspects of cancer-related issues [12]. The most recent comprehensive systematic content analysis of cancer coverage, conducted in 2013, indicated that cancer news coverage has heavily focused on cancer treatment and devoted very little attention to prevention, detection, or coping [13].

Evaluating cancer prevention information on social media was crucial for future endeavors by health educators and cancer-controlling organizations. Moreover, providing more reliable medical information to individuals helped alleviate feelings of fear and uncertainty [14]. And in particular, patients tended to seek information online when faced with risky treatment decisions, such as chemotherapy [15]. Therefore, it was significant to evaluate what kinds of treatment information on social media comprehensively.

Although many studies have explored cancer-related posts from the perspectives of cancer patients [16] and caregivers [17], analysis of posts from medical professionals was found to be inadequate [18]. The paradox arose from the expectation that medical professionals, given their professional advantages, should have taken the lead in providing cancer education on social media. Nevertheless, a substantial number of studies have highlighted the prevalence of unreliable medical information on social media [19]. A Japanese study highlighted a concerning phenomenon: despite efforts by medical professionals to promote cancer screening online, a significant number of anti-screening activists disseminated contradictory messages on the internet, potentially undermining the effectiveness of cancer education initiatives [20]. Hence, there was an urgent need for the accurate dissemination of health information on social media, with greater involvement from scientists or professional institutions, to combat the spread of misinformation [21]. Despite efforts to study professional medical websites [22] and applications [23], a comprehensive understanding of the content posted on medical social media was still lacking. Further study was needed to compare the differences between cancer information on social media from professional medical and non-professional sources to improve the cancer education.

For this study, we defined social media as internet-based platforms that were "characterized by social interactive functions such as reading, commenting, retweeting, and timely interaction"[24]. Based on the above definition, we further classified two types of media from the aspects of owners, content, and writers: common social media and medical social media. Medical social media (MSM) was a form of social media owned by professional medical institutions or organizations. It mainly provided medical and health information by medical professions, including medical-focused accounts on social media and mobile health applications. Common social media (CSM) referred to social media owned or managed by people not from medical backgrounds. It mainly provided health and lifestyle contents.

Similar to Facebook, WeChat was the most popular social media in China, that more than 90% of smartphones have installed it. A study has shown that 63.26% of people prefer to obtain health information from WeChat [25]. Unlike other Chinese social media platforms, WeChat has a broader user base covering a wide range of age groups [26]. WeChat Public Accounts (WPA) operated within the WeChat platform, providing services and information to the public. Numerous hospitals and primary care institutions in China have increasingly registered WPAs to provide healthcare and medical services, health education information, etc. [27]. Therefore, this study selected WPA as the research object.

Based on big data analytics, this study aimed to comprehensively map the characteristics of cancer treatment and prevention information on MSM and CSM, which was promisingly helpful in cancer coverage and patients' treatment decision. To address the previous mentioned research gaps, two research questions were generated.

*RQ 1: What were the characteristics of cancer prevention information discussed on social media? What were the differences between MSM and CSM?*

*RQ 2□What were the characteristics of cancer treatment information discussed on social media? What were the differences between MSM and CSM?*

## Methods

### Data Collection and Processing

We selected representative WPAs with reference to the reports of "Ranking of influential health WeChat public accounts" [28] and "2021 National rankings of best hospitals by specialty" [29]. We chose 4 medical WPAs as the object of MSM in this paper: Doctor Dingxiang (□□□□), 91Huayi (□□□), The Cancer Hospital of Chinese Academy of Medical Sciences (□□□□□□□□□□□), Fudan University Shanghai Cancer Center (□□□□□□□□□□). We also chose 5 health and lifestyle WOAs classified as CSM in this paper: Health Times (□□□□), Family Doctor (□□□□), CCTV lifestyle (CCTV □□□), Road to Health (□□□□), Life Times (□□□□).

Until now, we implemented a python crawler for the retrieval of posts from the aforementioned WPAs. Subsequently, a filtration process was implemented to eliminate data deemed noisy and unreliable. Note that our attention should be directed towards WPAs that offer substantial information, defined as containing no fewer than a certain number of characters, we deleted documents that contain less than 100 Chinese characters. Furthermore, we deleted the figures and the videos inside the remaining documents. Eventually, we conducted an analysis at the paragraph level. According to our analysis through random sampling, the noise in articles in WPAs mostly come from the advertisement, which appears in form of a specific paragraph. Therefore, we only kept the paragraphs without the advertising key words. We finally collected 60,843 posts from these WPAs, including 20654 articles from MSM and 40189 articles from CSM.

The workflow chart (Figure1) showed all procedures after completing data collection and preprocessing. After obtaining the meaningful raw documents, we performed word-level segmentation on the texts. We then eliminated the insignificant stopwords and substituted specific types of cancers with a general term in order to conduct a coarse-grained LDA-based filtering. Subsequently, we conducted fine-grained LDA topic modeling on the filtered documents without replacing keywords to visualize the topics extracted from the WPAs (presumably referring to the raw documents). Furthermore, we utilized a manually labeled codebook to train an LSTM network for document classification into various categories. Finally, the data analysis was performed using both the topic distribution derived from the fine-grained LDA and the classified documents.

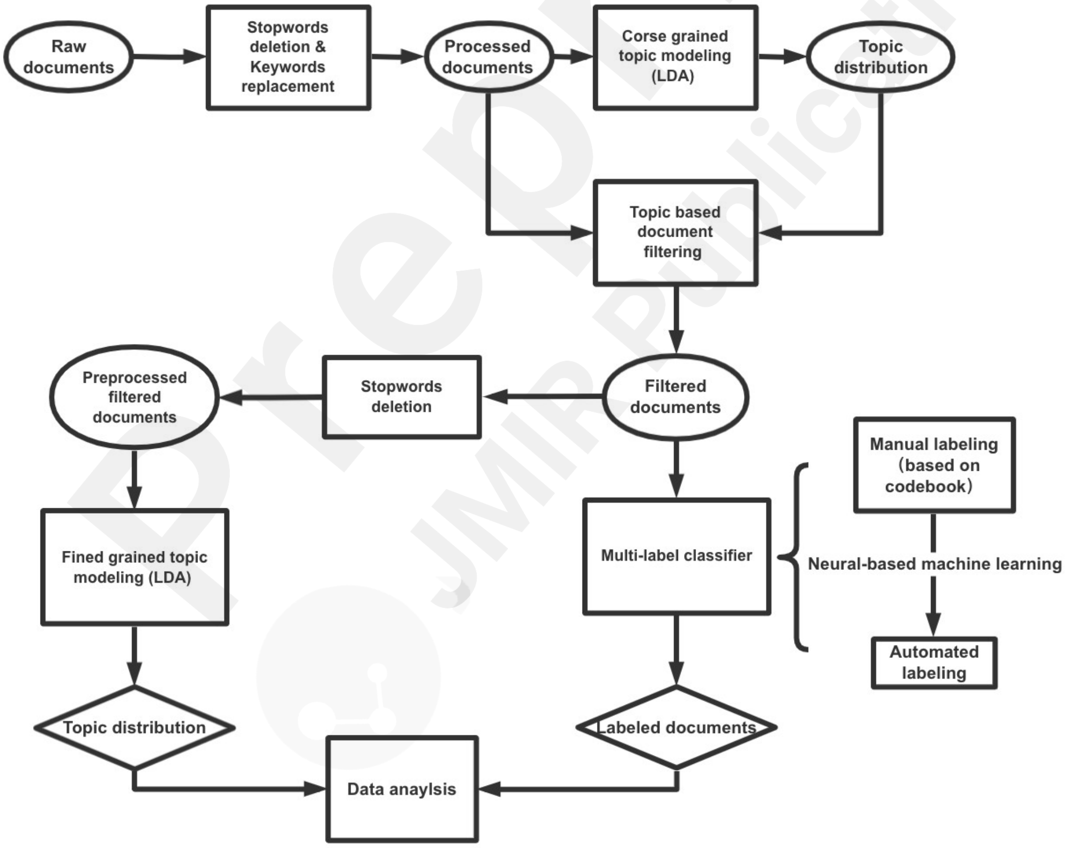

Raw documents
Stopwords deletion & Keywords replacement
Processed documents
Corse grained topic modeling (LDA)
Topic distribution
Topic based document filtering
Filtered documents
Stopwords deletion
Preprocessed filtered documents
Fined grained topic modeling (LDA)
Topic distribution
Multi-label classifier
Neural-based machine learning
Manual labeling (based on codebook)
Automated labeling
Labeled documents
Data anaylsis

**Figure 1.** Workflow chart.

## Latent Dirichlet Allocation Topic Modeling

The Latent Dirichlet allocation (LDA) was a generative statistic model that allowed sets of observations to be explained by unobserved groups that explained why some parts of the data were similar [30]. LDA algorithm could be used to speculate on the topic distribution of the document.

When comparing LDA with other NLP methods like LSTM-based deep learning, it was worth noting that LDA stood out as an unsupervised learning algorithm. Unlike its counterparts, LDA had the ability to uncover hidden topics without relying on labeled training data. Its strength lied in its capability to automatically identify latent topics within documents by analyzing statistical patterns of word co-occurrences. In addition, LDA offered interpretable outcomes. It assigned a probability distribution to each document, representing its association with various topics. Similarly, it assigned a probability distribution to each topic, indicating the prevalence of specific words within that topic. This feature enabled researchers to comprehend the principal themes present in their corpus, as well as the extent to which these themes were manifested in individual documents.

The foundational principle of Latent Dirichlet Allocation (LDA) involved utilizing probabilistic inference to estimate the distribution of topics and word allocations. Specifically, when considering words as observations within documents, LDA assumed that each document was composed of a mixture of a small number of topics, and each word's presence can be attributed to one of these topics. This approach allowed for overlapping content among documents, rather than strict categorization into separate groups. For a deeper understanding of the technical and theoretical aspects of the LDA algorithm, readers were encouraged to refer to the research conducted by Blei, D. M., Ng, A. Y., and Jordan, M. I., titled "Latent Dirichlet Allocation," published in the Journal of Machine Learning Research [30]. In this context, our primary focus was on the application of the algorithm to our corpus, and the procedure can be outlined as follows:

***Document Selection:*** Initial document selection involves employing a methodological approach to sample documents from the corpus, which may entail random selection or be guided by predetermined criteria such as document relevance or popularity within the social media context.

***Topic Inference:*** Utilizing Latent Dirichlet Allocation (LDA) or a similar topic modeling technique, we infer the underlying topical structure within each document. This involves modeling documents as mixtures of latent topics represented by a Dirichlet distribution, from which topic proportions are sampled.

***Topic Assignment to Words:*** Following the determination of topic proportions, we proceed to assign topics to individual words in the document. Employing a multinomial distribution, each word is probabilistically associated with one of the inferred topics based on the topic proportions previously derived.

***Word Distribution Estimation:*** Each topic is characterized by a distinct distribution over the vocabulary, encapsulating the likelihood of observing specific words within that topic. Through a Dirichlet distribution, we estimate the word distribution for each inferred topic.

***Word Generation:*** Finally, employing the multinomial distribution once more, we generate words for the document by sampling from the estimated word distribution corresponding to the topic assigned to each word. This iterative process yields a synthetic text that mirrors the statistical properties of the original corpus.

In order to filter the non-cancer-topic documents in our case, we replaced cancer-related words with "□□"(cancer or tumor in Chinese) in all documents and conducted an LDA analysis to compute the topic distribution of each document and reserve the documents related with the topics that have "□□" in top 10 words.

In our study, we used the Python packages including jieba and gensim to segment the document and extract per-topic-per-word probabilities from the model. During the segmentation, we filtered the meaningless words using a stopword dictionary and transform each document into a clean version with only meaningful words.

During the LDA analysis, to obtain the optimal number of topics, the main target was to compute the topic coherence for different numbers of topics and choose the model that gave the highest topic coherence. Coherence provided the probabilistic coherence of each topic. The coherence score was a score indicating whether the words in the same topic make sense when they were extracted by those topics. The higher the score for a specific number k, the more closely related the words. In this part, we used the Python package pyLDAvis to compare the coherence with different topic numbers. After that, we only reserved the documents that were related to cancer topics and obtain 4479 MSM articles and 3948 CSM articles.

Among filtered articles, we performed another LDA analysis to extract topics for original articles, without the replacement of cancer-related words. We calculated the coherence score with pyLDAvis and obtained both 6 topics for MSM and CSM articles, respectively.

We further visualized the topic modeling results in bar graphs with the Y-axis, indicating the top 10 keywords associated with that topic, and the X-axis, which showed the weight of each keyword (to reveal the extent to which a certain keyword contributes to that topic). Based on the top 10 most relevant keywords to each topic, we generalized and presented the name of each topic at the bottom of each graph (Figure 2 and Figure 3).

## Manual Content Analysis

### *Coding Procedure*

Based on codebook, two independent coders engaged in discussion regarding the coding rules, ensuring a shared understanding of the conceptual and operational distinctions among the coding items. To ensure the reliability of the coding process, both coders independently coded 100 randomly selected articles. Upon completion of the pilot coding, any disagreement was resolved through discussion between the two coders.

For the subsequent coding phase, each coder was assigned an equitable proportion of articles, with 10% of the cancer-related articles randomly sampled from both MSM samples (n=450) and CSM samples (n=394). Manual coding was performed on a total of 844 articles, served as the training dataset for the machine learning model. The operational definitions of each coding variable were shown in Multimedia Appendix 1.

### Coding Measures

Cancer Preventive Measures

Coders identified whether an article mentioned any of the following cancer preventive measures [31-35]: (1) avoid tobacco use, (2) maintain a healthy weight, (3) healthy diet, (4) exercise regularly, (5) limit alcohol use, (6) get vaccinated , (7) reduce exposure to ultraviolet radiation and ionizing radiation, (8)avoid urban air pollution and indoor smoke from household use of solid fuels, (9) early screening and detection, (10) breastfeeding, (11) controlling chronic infections , (12) other preventive measures.

Cancer Treatment Measures

Coders identified whether an article mentioned any of the following treatment [36]: (1) surgery (including cryotherapy, lasers, hyperthermia, photodynamic therapy, cuts with scalpels), (2) radiotherapy, (3) chemotherapy, (4) immunotherapy, (5) targeted therapy, (6) hormone therapy, (7) stem cell transplant, (8) precision medicine, (9) cancer biomarker testing, (10) other treatment measures.

### Neural-based Machine Learning

In this part, we attempted to label each article using neural network. As mentioned above, we manually labeled 450 MSM articles and 394 CSM articles. We divided the labeled data into the training set and tested set according to the ratio of 4:1. We adopted Bert pretrained model and since Bert can only accept inputs that have less than 512 tokens [37], we cut each document into pieces with 510 tokens (because Bert automatically adds [CLS] and [SEP]) and adjacent pieces had 384 overlap tokens. We first used a Bert-based encoder to encode each piece and predict its labels using a multi-output decoder. After predicting labels for each piece, we collected pooling outputs for all pieces in the same document and predicted final labels for each document using an LSTM network.

## Results

### Cancer Topics on Social Media

By applying LDA, we obtained 6 topics for MSM and CSM articles respectively. The distribution of topics on MSM and CSM was shown in Table 1. And the weight of keywords in each topic was shown in Figure 2 and Figure 3.

**Table 1.** Distribution of topics on MSM and CSM (n=8427) [a-b]

| Media Type | Topic Number | Topic description | Number of articles | n (%) | Top 10 Keywords |
|---|---|---|---|---|---|
| **MSM** | | | | | |
| | **Topic 1** | Liver cancer and stomach caner | 1519 | 18.03 | Cancer (□□), Liver cancer (□□), Stomach cancer (□□), Factors (□□), Food (□□), Disease (□□ ), Helicobacter pylori (□□), Exercise (□□), Patient ( □□), Diet (□□) |
| | **Topic 2** | Female and cancer | 1611 | 19.12 | Breast cancer (□□□), Female (□□), Patient (□□), Lung cancer (□□), Surgery (□□), Tumor (□□), Mammary gland (□□), Expert (□□), Ovarian cancer (□□□), Lump (□□) |
| | **Topic 3** | Breast cancer | 1093 | 12.97 | Breast cancer (□□□), Surgery (□□), Thyroid (□□□), Lump (□□), Breast (□□), Patient (□□), Female (□□), Screening and testing (□□), Mammary gland (□□), Tumor (□□) |

| | | | | | |
|---|---|---|---|---|---|
| | **Topic 4** | Cervical cancer | 1019 | 12.09 | Vaccine (□□), Cervical cancer (□□□), Virus (□□), Cervix (□□), Patient (□□), Nation (□□), Female (□□), Nasopharynx cancer (□□□), Medicine (□□), Hospital (□□) |
| | **Topic 5** | Clinical cancer treatment | 2548 | 30.24 | Tumor (□□), Patient (□□), Screening (□□), Chemotherapy (□□), Clinic (□□), Symptom (□□), Hospital (□□), Surgery (□□), Medicine (□□), Disease (□□) |
| | **Topic 6** | Diet and cancer risk | 1741 | 20.66 | Patient (□□),Tumor (□□), Food (□□), Polyp (□□), Professor (□□), Nutrition (□□), Expert (□□), Surgery (□□), Cancer (□□), Disease (□□) |
| **CSM** | | | | | |
| | **Topic 1** | Cancer causing substances | 1136 | 13.48 | Foods (□□), Nutrition (□□), Carcinogen (□□□), Food (□□) , Ingredient (□□), Vegetable (□□), Cancer (□□), Body (□□), Lump (□□), Formaldehyde (□□) |
| | **Topic 2** | Cancer treatment | 1319 | 15.65 | Patient (□□), Cancer (□□), Hospital (□□), Lung Cancer (□□), Tumor (□□), Medicine (□□), Disease (□□), Professor (□□), Surgery (□□), Clinic (□□) |
| | **Topic 3** | Female and cancer risk | 1599 | 18.97 | Screening and Testing (□□), Female (□□), Disease (□□), Breast Cancer (□□□), Cancer (□□), Lung Cancer (□□), Patient (□□), Body (□□), Tumor (□□), Risk (□□) |
| | **Topic 4** | Exercise, diet and cancer risk | 1947 | 23.10 | Cancer (□□), Exercise (□□), Food (□□), Risk (□□), Body (□□), Disease (□□), Suggestion (□□), Patient (□□), Fat (□□), Hospital (□□) |
| | **Topic 5** | Screening and diagnosis of cancer | 1790 | 21.24 | Screening and Testing (□□), Disease (□□), Hospital (□□), Stomach Cancer (□□), Symptom (□□), Patient (□□), Cancer (□□), Liver Cancer (□□), Female (□□), Suggestion (□□) |
| | **Topic 6** | Disease and body parts | 869 | 10.31 | Disease (□□), Intestine (□□), Food (□□), Hospital (□□), Oral Cavity (□□), Patient (□□), Teeth (□□), Cancer (□□), Ovary (□□), Garlic (□□) |

NOTE.

[a] In each article, different topics may appear at the same time. Therefore, the total frequency of each topic did not equate to the total number of 8427 articles.

[b] To ensure the accuracy of the results, directly translating sampling texts from Chinese into English was not convenient due to differences in semantic elements. Cancer screening mainly referred to detect the possibility of getting cancer before cancer symptom occurs. Diagnosis tests referred to confirm whether cancer occurs after finding suspect symptoms. But in Chinese, "□□" included two meanings simultaneously. Therefore, we translated it into screening and testing.

On MSM, Topic 5 was the most frequent one (30.24%, 2548/8427), followed by Topic 6 (20.66%, 1741/8427) and Topic 2 (19.12%, 1611/8427). Topic 5 and Topic 6 were both related to clinical treatments. Topic 5 was more concerned with the cancer diagnosis. The keywords in Topic 6, such as "Polyp," "Tumor," and "Surgery," emphasized the risk and diagnose of precancerous lesions. Topic 2 was more concerned with cancer surgery with breast cancer, lung cancer, and ovarian cancer. As observed from the results, MSM focused more on specific cancers with higher incidence in China: stomach cancer, liver cancer, lung cacner, breast cancer and cervical cancer [10].

On CSM, Topic 4 (23.10%, 1947/8427) took the highest proportion, followed by Topic 5 (21.24%, 1790/8427) and Topic 3 (18.97%, 1599/8427). Topic 6 took the least proportion. Both Topic1 and Topic 4 were related to lifestyle. Specifically, Topic 1 was identified as cancer-causing substances with keywords like "food", "nutrition", "carcinogen" used more frequently. Topic 4 was identified as exercise, diet, and cancer risk. Topic 3 and Topic 5 focused on cancer screening and diagnosis. Topic 3 was female and cancer. The screening and testing of breast cancer were more frequently discussed in articles on this topic. Topic 5 focused more on early detection and diagnosis of stomach cancer and lung cancer, with keywords like "screening" and "symptom".

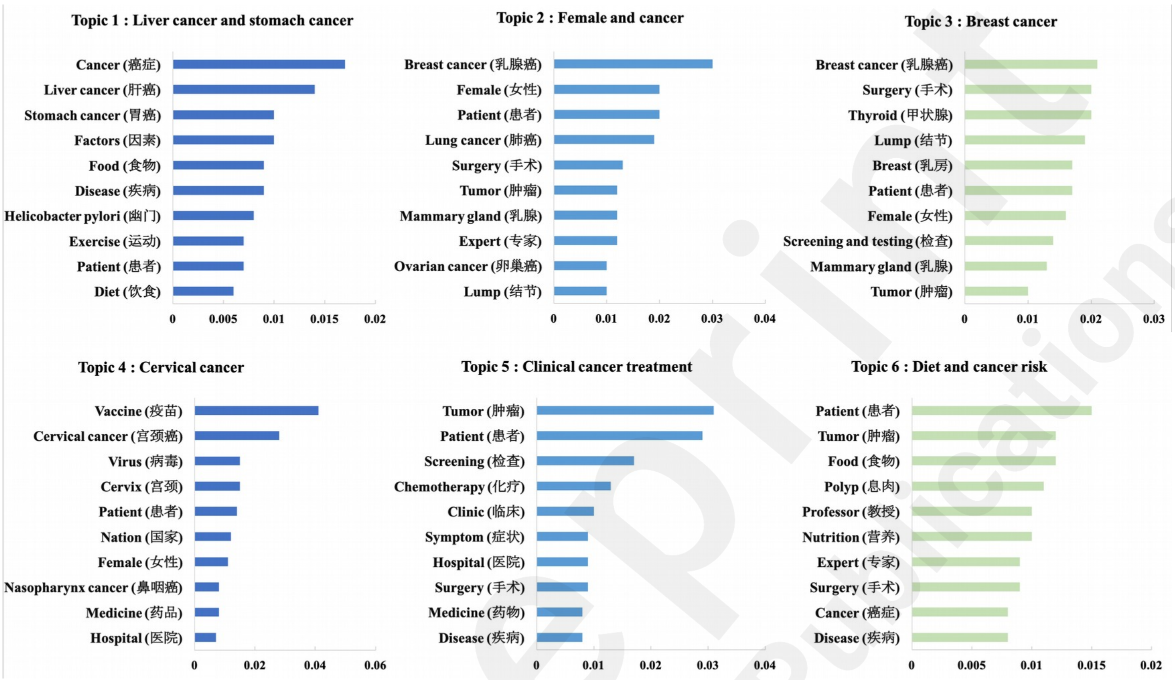


**Figure 2.** Cancer topics on MSM.

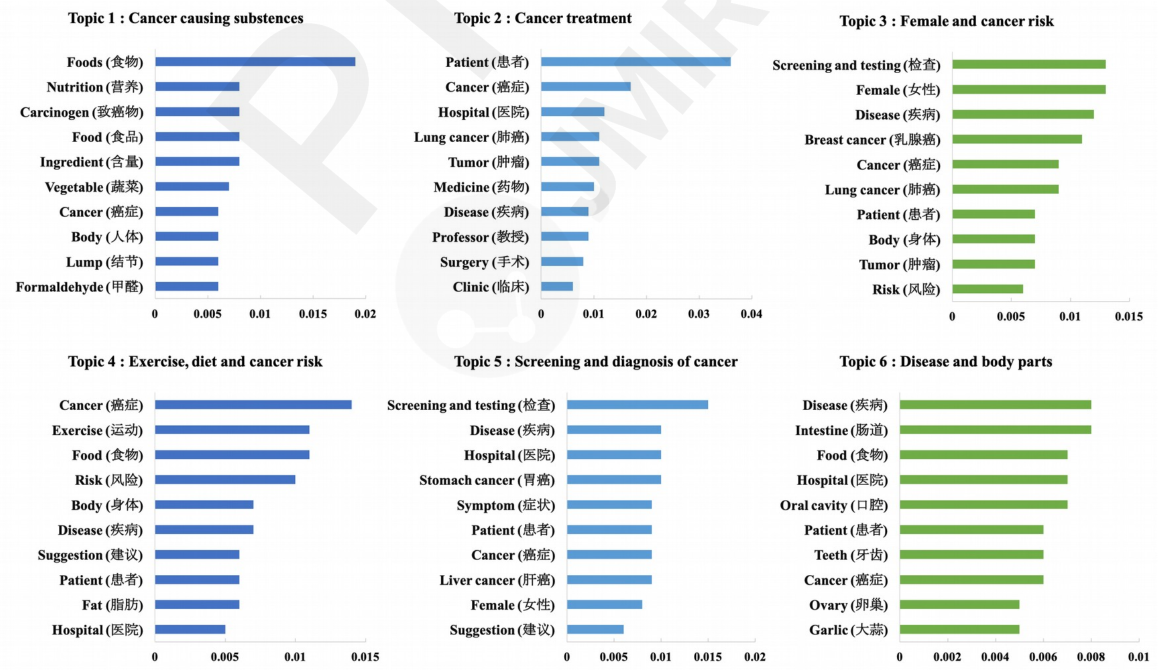


**Figure 3.** Cancer topics on CSM.

## Cancer Prevention Information

Our experiment on the test set shown that the machine learning model achieves F1 scores over 85 on both prevention and treatment for both MSM and CSM. As for the sub-classes of prevention and treatment, due to the sparsity of the labels, we achieved F1 scores no less than 70 on the dense ones (occurrence rate larger than 10%) and no less than 50 on sparse ones (occurrence rate less than 10%). Then, we deleted items of "other prevention measures" and "other treatment measures" for semantic ambiguity.

Table 2 presented the distribution of cancer preventive information on MSM(n=4479) and CSM(n=3948).

**Table 2**□Distribution of cancer prevention information on MSM and CSM.

| Type of cancer prevention measures | Number of articles on MSM, n(%),n=4479 | Number of articles on CSM, n(%),n=3948 |
|---|---|---|
| Articles containing preventive information | 25.39% (1137) | 47.01% (1856) |
| Early Screening and Testing | 16.45% (737) | 27.48% (1085) |
| Healthy Diet | 6.21% (278) | 15.15% (598) |
| Get Vaccinated | 5.83% (261) | 2.86% (113) |
| Avoid Tobacco Use | 4.51% (186) | 9.32% (368) |
| Exercise Regularly | 3.01% (135) | 16.74% (661) |
| Limit Alcohol Use | 2.86% (128) | 7.12% (281) |
| Avoid Urban Air Pollution and Indoor Smoke from Household Use of Solid Fuels | 0.42% (19) | 1.62% (64) |
| Maintain a Healthy Weight | 0.40% (18) | 4.89% (193) |
| Practice Safe Sex | 0.27% (12) | 0.10% (4) |
| Controlling Chronic Infections | 0.07% (3) | 0.81% (32) |
| Reduce Exposure to Radiation | 0.04% (2) | 0.03% (1) |
| Breastfeeding | 0.02% (1) | 0.03% (1) |

### *MSM: Cancer Preventive Information*

The descriptive results of the distribution of cancer preventive information on MSM were as follows (n=4479). Articles containing preventive information accounted for 25.39% (1137/4479) of the total number of MSM cancer-related articles. The most frequently mentioned measure was "early screening and testing" (16.45%, 737/4479). "Healthy diet" (6.21%, 278/4479) and "get vaccinated" (5.83%, 261/4479) were the second and third frequently mentioned preventive measures. The least mentioned preventive measures were "controlling chronic infections" (0.07%, 3/4479), "reduce exposure to radiation" (0.04%, 2/4479), and "breastfeeding" (0.02%, 1/4479), each appearing in only 1-3 articles.

### *CSM: Cancer Preventive Information*

A total of 47.01% (1856/3948) of articles on CSM (n=3948) referred to cancer preventive information. Among these, "early screening and testing" (27.48%, 1085/3948) was the most commonly mentioned preventive measure. "Exercise regularly"(16.74%, 661/3948), "healthy diet"(15.15%, 598/3948%), "avoid tobacco use" (9.32%, 368/3948), were the three most frequently mentioned lifestyle-related preventive measures. Other lifestyle-related preventive measures included "limit alcohol use"(7.12%, 281/3948) and "maintain a healthy weight"(4.89%, 193/3948). The least mentioned preventive measures were "practice safe sex" (0.10%, 4/3948), "reduce exposure to radiation" (0.03%, 1/3948), and "breastfeeding" (0.03%, 1/3948), each mentioned in only 1-4 articles.

### *Cancer Preventive Information on Social Media*

Table 3 listed the overall distribution of cancer prevention information on social media (n=8427). Notably, CSM exhibited a greater focus on cancer prevention (22.02%, 1856/3948) compared to MSM (13.49%, 1137/8427). Both of them addressed the significance of early screening and testing. However, MSM placed more emphasis on vaccination as a preventive measure. In addition to lifestyle related preventive measure, both CSM and MSM showed a relatively lower emphasis on measures related to avoiding exposure to environmental carcinogens, such as "avoiding air pollution and indoor smoke" and "reducing exposure to radiation." "Breastfeeding" was the least mentioned preventive measure (0.02%, 2/3948) in both social media types.

**Table 3**□Distribution of cancer prevention information on social media.

| **Type of cancer prevention measures** | **Number of articles on MSM, n(%)** | **Number of articles on CSM, n(%)** | **Number of articles overall, n(%), n=8427** |
|---|---|---|---|
| **Articles containing preventive information** | 1137 (13.49%) | 1856 (22.02%) | 2993 (35.52%) |
| **Early screening and testing** | 737 (8.75%) | 1085 (12.88%) | 1822 (21.62%) |
| **Healthy diet** | 278 (3.30%) | 598 (7.10%) | 876 (10.40%) |
| **Get vaccinated** | 261 (3.10%) | 113 (1.34%) | 374 (4.44 %) |
| **Avoid tobacco use** | 186 (2.21%) | 368 (4.37%) | 554 (6.57%) |
| **Exercise regularly** | 135 (1.60%) | 661 (7.84%) | 796 (9.45 %) |
| **Limit alcohol use** | 128 (1.52%) | 281 (3.33%) | 409 (4.85%) |
| **Avoid urban air pollution and indoor smoke from household use of solid fuels** | 19 (0.23%) | 64 (0.76%) | 83 (0.98%) |
| **Maintain a healthy weight** | 18 (0.21%) | 193 (2.29%) | 211 (2.50%) |
| **Practice safe sex** | 12 (0.14%) | 4 (0.05%) | 16 (0.19%) |
| **Controlling chronic infections** | 3 (0.04%) | 32 (0.38%) | 35 (0.42%) |
| **Reduce exposure to radiation** | 2 (0.02%) | 1 (0.01%) | 3 (0.04%) |
| **Breastfeeding** | 1 (0.01%) | 1 (0.01%) | 2 (0.02%) |

## Cancer Treatment Information

Table 4 presented the distribution of cancer treatment information on MSM(n=4479) and CSM (n=3948).

**Table 4**□Distribution of cancer treatment information on MSM and CSM.

| Type of cancer treatment measures | Number of articles on MSM, n(%),n=4479 | Number of articles on CSM, n(%),n=3948 |
|---|---|---|
| **Articles containing treatment information** | 66.22% (2966) | 19.72% (778) |
| **Surgery** | 45.66% (2045) | 10.61% (419) |
| **Chemotherapy** | 25.5% (1122) | 7.22% (285) |
| **Radiation therapy** | 24.74% (1108) | 5.88% (232) |
| **Cancer biomarker testing** | 8.48% (380) | 1.39% (55) |
| **Targeted therapy** | 8.46% (379) | 4.58% (181) |
| **Immunotherapy** | 7.08% (317) | 0.56% (22) |
| **Hormone therapy** | 1.05% (47) | 0.35% (14) |
| **Stem cell transplantation therapy** | 0.11% (5) | 0 |

### *MSM: Cancer Treatment Information*

Cancer treatment information appeared in 66.22% (2966/4479) on MSM (n=4479). "Surgery" (45.66%, 2045/4479) was the most frequently mentioned treatment measure, followed by "Chemotherapy" (25.5%, 1122/4479) and "radiation therapy" (24.74%, 1108/4479). The proportions of "cancer biomarker testing" (8.48%, 380/4479), "targeted therapy" (8.46%, 379/4479), and "immunotherapy" (7.08%, 317/4479) were comparable. Only a minimal percentage of articles, 1.05% (47/4479), addressed "hormone therapy." Furthermore, "stem cell transplantation therapy" was mentioned in just 5 articles (0.11%, 5/4479).

### *CSM: Cancer Treatment Information on CSM*

Cancer treatment information accounted for only 19.72% (778/3948) on CSM (n=3948). Surgery (10.61%, 419/3948) was the most frequently mentioned treatment measure, followed by "chemotherapy" (7.22%, 285/3948) and "radiation therapy" (5.88%, 232/3948). Relatively, the frequency of "targeted therapy" (4.58%, 181/3948) was similar to the first three types. However, "cancer biomarker testing" (1.39%, 55/3948), "immunotherapy" (0.56%, 22/3948), and "hormone therapy" (0.35%, 14/3948) appeared rarely on CSM. Notably, there was no article on the CSM mentioning stem cell transplantation.

### *Cancer Treatment Information on Social Media*

The overall distribution of cancer treatment information on social media (n=8427) was shown in Table 5. A total of 44.43% (3744/8427) of articles contained treatment information. MSM (35.20%, 2966/8427) discussed treatment information much more frequently than CSM (9.23%, 778/8427). Furthermore, the frequency of all types of treatment measures mentioned on MSM was higher than CSM. The three most frequently mentioned types of treatment measures were surgery (29.24%, 2464/8427), chemotherapy (16.70%, 1407/8427), and radiation therapy (15.90%, 1340/8427). Relatively, MSM (4.51%, 380/8427) showed a higher focus on cancer biomarker testing compared to CSM (0.65%, 55/8427).

**Table 5**□Distribution of cancer treatment information on social media.

| Type of cancer treatment measures | Number of articles on MSM, n(%) | Number of articles on CSM, n(%), | Number of articles overall, n(%), n=8427 |
|---|---|---|---|
| **Articles containing treatment information** | 2966 (35.20%) | 778 (9.23%) | 3744 (44.43%) |
| **Surgery** | 2045 (24.27%) | 419 (4.97%) | 2464 (29.24%) |
| **Radiation therapy** | 1108 (13.15%) | 232 (2.75%) | 1340 (15.90%) |
| **Chemotherapy** | 1122 (13.31%) | 285 (3.38%) | 1407 (16.70%) |
| **Immunotherapy** | 317 (3.76%) | 22 (0.26%) | 339 (4.02%) |
| **Targeted Therapy** | 379 (4.50%) | 181 (2.15%) | 560 (6.65%) |
| **Hormone Therapy** | 47 (0.56%) | 14 (0.17%) | 61 (0.72%) |
| **Stem cell transplant** | 5 (0.06%) | 0 (0.00%) | 5 (0.06%) |
| **Cancer biomarker testing** | 380 (4.51%) | 55 (0.65%) | 435 (5.16%) |

## Discussion

### Cancer Topics on MSM and CSM

In the distribution of topics within MSM, treatment-related topics constituted the largest proportion, featuring keywords related to medical examinations. Conversely, in CSM, the distribution of topics appeared more balanced, with keywords frequently associated with cancer risk and screening. Overall, the distribution of topics on MSM and CSM revealed that CSM placed greater emphasis on lifestyle factors and early screening and testing. Specifically, topics on CSM concerned more about the early screening of cancer and focused on cancer types with high incidence rates. In contrast, topics of MSM centered more on clinical treatment, medical testing, and cervical cancer vaccine in cancer prevention. Additionally, MSM focused more on the types of cancers that are easier to screen and prevent, including liver cancer, stomach cancer, breast cancer, cervical cancer, and colon cancer.

### Cancer Preventive Information on MSM and CSM

Through content analysis, 35.52% (2993/8427) of articles on social media contained preventive information, and 44.43% (3744/8427) contained treatment information. Compared to MSM (13.49%, 1137/8427), CSM (35.52%, 2993/8427) focused more on prevention.

Primary prevention mainly involved choosing healthy behaviors to lower the risk of getting cancer, proven to have long-term effects on cancer prevention. Secondary prevention was mainly concerned with inhibiting or reversing carcinogenesis, including early screening and detection, treatment, or removal of precancerous lesions [38]. In comparison to cancer screening and treatment, primary prevention was the most cost-effective way to reduce the cancer burden.

From our results, "early screening and testing (21.62%,1822/8427)" was the most frequently mentioned preventive measure on both MSM and CSM. According to a cancer study from China, behavioral risk factors were the main cause of cancer [39]. However, measures related to primary prevention were not mentioned frequently. Additionally, lifestyle-related measures on MSM, including "healthy diet", "exercise regularly", "avoid tobacco use", and "limit alcohol use" were mentioned much less frequently than CSM.

In addition, "avoiding tobacco use" (6.57%, 554/8427) and "limit alcohol use" (4.85%, 409/8427) were rarely mentioned, despite tobacco and alcohol being the leading causes of cancer. In China, public policies on the production, sale, and consumption of alcohol were weak compared with the Western countries. Notably, traditional Chinese customs often held the belief that drinking moderately was beneficial for people’s health [40]. Moreover, studies indicated that the smoking rate among adult men exceeded 50% in 2015. In 2018, 25.6% of Chinese adults aged 18 and above were smokers, totaling approximately 282 million smokers in China (271 million males and 11 million females) [41]. These statistics aligned with the consistently high incidence of lung cancer among Chinese men [42]. Simultaneously, the incidence and mortality of lung cancer in Chinese women were more likely associated with second-hand smoke exposure or occupation-related risk factors.

Although MSM (3.01%, 261/8427) mentioned vaccination more frequently than CSM (1.34%, 113/8427), vaccination was not frequently presented on social media overall (4.44%, 374/8427). The introduction of HPV vaccination in China has lagged for more than 10 years compared with the Western countries. One bivalent vaccine was approved by China Food and Drug Administration in 2017, it hasn't been included in the national immunization schedules until now [43].

According to the “European Code Against Cancer” [44], breastfeeding was recommended as a measure to prevent breast cancer. However, there were no articles mentioned the role of breastfeeding in preventing breast cancer on social media.

One of the least frequently mentioned measures was "radiation protection", which also included sun protection. Although skin cancer was not as common in China as in Western countries, China has the largest population in the world. A study showed that only 55.2% of Chinese people knew that UV radiation causes skin cancer [33]. Additional efforts should have been made to enhance public awareness of skin cancer prevention through media campaigns.

Overall, our results indicated that social media was more focused on secondary prevention, epically MSM. Since the outcomes of primary prevention were challenging to identify in individuals, studies on cancer education may partly explain why primary prevention was often overlooked [45].

## Cancer Treatment Information on MSM and CSM

Compared to a related content analysis study in the United States, we obtained similar findings indicating that media placed greater emphasis on treatment [46]. Treatment information on MSM was more diverse than on CSM. The proportion of the three most common cancer treatments mentioned in MSM (surgery, chemotherapy, and radiation therapy) far exceeded that on CSM. Notably, CSM (2.75%, 232/8427) mentioned radiation therapy less frequently compared to MSM (13.15%, 1108/8427), which was one of the most common cancer treatment measures in clinical practice.

In addition to common treatment methods, other treatment methods such as targeted therapy (6.65%, 560/8427) and immunotherapy (4.02%, 339/8427) were rarely discussed. This could be attributed to the high cost associated with these treatment methods. A study revealed that each newly diagnosed cancer patient in China faced out-of-pocket expenses of US$4,947, constituting 57.5% of the family's annual income with an unaffordable 77.6% economic burden [47]. In 2017, the Chinese government released the National Health Insurance Coverage (NHIC) policy to enhance the accessibility and affordability of innovative anti-cancer medicines, resulting in reduced prices and increased availability and utilization of 15 negotiated innovative anti-cancer drugs. However, a study indicated that the availability of these innovative anti-cancer drugs remained low. As of 2019, the NHIC policy on anti-cancer medicines had benefited 44,600 people, while the number of new cases in China in 2020 was 4.57 million [48]. The promotion of innovative therapy information assisted patients in gaining a better understanding of cancer treatment options [49].

### Practical Implications

This research highlighted that MSM did not take full advantage of its professional background in providing comprehensive cancer information for the public. In fact, MSM held substantial potential for contributing to cancer education. The findings from the content analysis also contributed to applied implications for practitioners. They offered valuable insights for experts to evaluate the performance of social media, monitor the type of information available to the public and cancer patients, and guide communication professionals and medical professionals in designing educational and persuasive messages based on the content that was widely covered or received less attention.

### Limitations and Future Directions

This study had some limitations. First, we only collected 60,843 articles from 9 WeChat public accounts in China. Future research could enhance the scope by collecting data from diverse countries and social media platforms. Second, our manual labeling only extracted 10% of the samples, the accuracy of the machine learning model can be improved by training the model with more labeled articles. Finally, our results only represented the media's presentation, and the impact of this information on individuals remained unclear. Further work could examine the influence on behavioral intentions or actions related to cancer prevention among the audience.

### Conclusions

The analysis of cancer-related information on social media unveiled an imbalance between prevention and treatment content. In general, there were more treatment information on social media. Compared to MSM, CSM mentioned more prevention information. On MSM, the proportion of treatment information was more than preventive information. But on CSM, the two were equal. The focus on cancer prevention and treatment information was mainly limited to a few aspects. Both of them concerned more on secondary cancer prevention rather than primary prevention. The cancer coverage on preventive measures and treatments in social media required further improvement. Additionally, the findings underscored the potential of applying machine learning to content analysis as a promising research paradigm for mapping the key dimensions of cancer information on social media. The findings provided methodological and practical significance in future study and health promotion.

## Acknowledgements

This study funded by The Major Program of Chinese National Foundation of Social Sciences under Grant "The Challenge and Governance of Smart Media on News Authenticity" [number 23&ZD213].

## Conflicts of Interest

The authors declare no conflict of interest.

## Abbreviations

**WPA**: WeChat public account
**MSM**: Medical social media
**CSM**: Common social media

## MULTIMEDIA APPENDIX 1

## References

1. IARC. The Global Cancer Observatory. 2020; URL: https://www.iarc.who.int/faq/latest-global-cancer-data-2020-qa/. [Accessed 2023-12-25]
2. Yu S, Yang CS, Li J, You W, Chen J, Cao Y, et al. Cancer prevention research in China. Cancer Prevention Research. 2015;8(8):662-674. doi: 10.1158/1940-6207.CAPR-14-0469.
3. Xia C, Dong X, Li H, Cao M, Sun D, He S, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chinese medical journal. 2022;135(05):584-590.
4. WHO. Cancer. 2023; URL: https://www.who.int/news-room/facts-in-pictures/detail/cancer. [Accessed 2023-12-27]
5. Pagoto S, Waring ME, Xu R. A call for a public health agenda for social media research. Journal of Medical Internet Research. 2019;21(12):e16661. doi: 10.2196/16661.
6. Tekeli-Yesil S, Tanner M. Understanding the Contribution of Conventional Media in Earthquake Risk Communication: A Qualitative Study. Journal of Emergency Management and Disaster Communications.0(0):1-23. doi: 10.1142/s2689980924500052.
7. Jensen JD, Scherr CL, Brown N, Jones C, Christy K, Hurley RJ. Public estimates of cancer frequency: Cancer incidence perceptions mirror distorted media depictions. Journal of Health Communication. 2014;19(5):609-624. doi: 10.1080/10810730.2013.837551.
8. Banaye Yazdipour A, Niakan Kalhori SR, Bostan H, Masoorian H, Ataee E, Sajjadi H. Effect of social media interventions on the education and communication among patients with cancer: a systematic review protocol. BMJ Open. 2022;12(11):e066550. doi: 10.1136/bmjopen-2022-066550.
9. Wallner LP, Martinez KA, Li Y, Jagsi R, Janz NK, Katz SJ, et al. Use of online communication by patients with newly diagnosed breast cancer during the treatment decision process. JAMA Oncology. 2016;2(12):1654-1656. doi: 10.1001/jamaoncol.2016.2070.
10. Sun D, Li H, Cao M, He S, Lei L, Peng J, et al. Cancer burden in China: trends, risk factors and prevention. Cancer Biol Med. 2020;17(4):879-895. doi: 10.20892/j.issn.2095-3941.2020.0387.
11. Basch CH, Menafro A, Mongiovi J, Hillyer GC, Basch CE. A Content Analysis of YouTube Videos Related to Prostate Cancer. Am J Mens Health. 2017;11(1):154-157. doi: 10.1177/1557988316671459.

12. Silva CV, Jayasinghe D, Janda M. What can Twitter tell us about skin cancer communication and prevention on social media? Dermatology. 2020;236(2):81-89. doi: 10.1159/000506458.

13. Hurley RJ, Riles JM, Sangalang A. Online cancer news: Trends regarding article types, specific cancers, and the cancer continuum. Health Communication. 2013;29(1):41-50. doi: 10.1080/10410236.2012.715538.

14. Mishel MH, Germino BB, Lin L, Pruthi RS, Wallen EM, Crandell J, et al. Managing uncertainty about treatment decision making in early stage prostate cancer: A randomized clinical trial. Patient Education and Counseling. 2009;77(3):349-359. doi: 10.1016/j.pec.2009.09.009.

15. Brown P, Kwan V, Vallerga M, Obhi HK, Woodhead EL. The Use of Anecdotal Information in a Hypothetical Lung Cancer Treatment Decision. Health Communication. 2019;34(7):713-719. doi: 10.1080/10410236.2018.1433415.

16. Crannell WC, Clark E, Jones C, James TA, Moore J. A pattern-matched Twitter analysis of US cancer-patient sentiments. J Surg Res. 2016;206(2):536-542. doi: 10.1016/j.jss.2016.06.050.

17. Gage-Bouchard EA, LaValley S, Mollica M, Beaupin LK. Cancer Communication on Social Media Examining How Cancer Caregivers Use Facebook for Cancer-Related Communication. Cancer Nursing. 2017;40(4):332-338. doi: 10.1097/Ncc.0000000000000418.

18. Reid BB, Rodriguez KN, Thompson MA, Matthews GD. Cancer-specific Twitter conversations among physicians in 2014. Journal of Clinical Oncology. 2015;33(15_suppl):e17500-e17500. doi: 10.1200/jco.2015.33.15_suppl.e17500.

19. Warner EL, Waters AR, Cloyes KG, Ellington L, Kirchhoff AC. Young adult cancer caregivers' exposure to cancer misinformation on social media. Cancer. 2021;127(8):1318-1324. doi: 10.1002/cncr.33380.

20. Okuhara T, Ishikawa H, Okada M, Kato M, Kiuchi T. Assertions of Japanese Websites for and Against Cancer Screening: a Text Mining Analysis. Asian Pacific Journal of Cancer Prevention. 2017;18(4):1069-1075. doi: 10.22034/apjcp.2017.18.4.1069.

21. Qin L, Zhang X, Wu A, Miser JS, Liu Y-L, Hsu JC, et al. Association Between Social Media Use and Cancer Screening Awareness and Behavior for People Without a Cancer Diagnosis: Matched Cohort Study. Journal of Medical Internet Research. 2021;23(8):e26395. doi: 10.2196/26395.

22. Denecke K, Nejdl W. How valuable is medical social media data? Content analysis of the medical web. Information Sciences. 2009;179(12):1870-1880. doi: 10.1016/j.ins.2009.01.025.

23. Bender JL, Yue RYK, To MJ, Deacken L, Jadad AR. A lot of action, but not in the right direction: Systematic review and content analysis of smartphone applications for the prevention, detection, and management of cancer. Journal of Medical Internet Research. 2013;15(12):e287. doi: 10.2196/jmir.2661.

24. Li X, Liu Q. Social media use, eHealth literacy, disease knowledge, and preventive behaviors in the COVID-19 pandemic: Cross-sectional study on Chinese netizens. Journal of Medical Internet Research. 2020;22(10):e19684. doi: 10.2196/19684.

25. Zhang X, Wen D, Liang J, Lei J. How the public uses social media wechat to obtain health information in china: a survey study. BMC Medical Informatics and Decision Making. 2017;17(2):71-79.

26. Elad B. WeChat Statistics By Device Allocation, Active Users, Country Wise Traffic, Demographics and Marketing Channels, Social Media Traffic. 2023; URL: https://www.enterpriseappstoday.com/stats/wechat-statistics.html.]

27. Liang X, Yan M, Li H, Deng Z, Lu Y, Lu P, et al. WeChat official accounts' posts on medication use of 251 community healthcare centers in Shanghai, China: content analysis and quality assessment. Frontiers in Medicine. 2023;10. doi: 10.3389/fmed.2023.1155428.
28. NewRank. Ranking of influential health WeChat public accounts □□□□□□□□□□□□□□. 2018; URL: https://newrank.cn/public/info/rank_detail.html?name=health. [Accessed 2021-4-30]
29. Institute HM. 20210National rankings of best hospitals by specialty 2021 □□□□□□□□□□□□□. Hospital Management Institute of Fudan University; 2021; URL: https://rank.cn-healthcare.com/fudan/specialty-reputation/year/2021/sid/2. [Accessed 2021-05-01]
30. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3:993-1022.
31. WHO. Healrh topic: Cancer. URL: https://www.who.int/health-topics/cancer#tab=tab_2. [Accessed 2023-12-27]
32. Moore SC, Lee I-M, Weiderpass E, Campbell PT, Sampson JN, Kitahara CM, et al. Association of Leisure-Time Physical Activity With Risk of 26 Types of Cancer in 1.44 Million Adults. JAMA Internal Medicine. 2016;176(6):816-825. doi: 10.1001/jamainternmed.2016.1548.
33. Stephens PM, Martin B, Ghafari G, Luong J, Nahar VK, Pham L, et al. Skin Cancer Knowledge, Attitudes, and Practices among Chinese Population: A Narrative Review. Dermatology Research and Practice. 2018;2018:1965674. doi: 10.1155/2018/1965674.
34. IARC. Agents Classified by the IARC Monographs, Volumes 1–130. 2021; URL: https://monographs.iarc.who.int/agents-classified-by-the-iarc/. [Accessed 2023-12-25]
35. Han CJ, Lee YJ, Demiris G. Interventions Using Social Media for Cancer Prevention and Management: A Systematic Review. Cancer Nursing. 2018;41(6):E19-E31. doi: 10.1097/Ncc.0000000000000534.
36. NIH. Types of Cancer Treatment. URL: https://www.cancer.gov/about-cancer/understanding. [Accessed 2021-03-15]
37. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-Training With Whole Word Masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021;29:3504-3514. doi: 10.1109/TASLP.2021.3124365.
38. Loomans-Kropp HA, Umar A. Cancer prevention and screening: the next step in the era of precision medicine. npj Precision Oncology. 2019;3(1):3. doi: 10.1038/s41698-018-0075-9.
39. Sun D, Li H, Cao M, He S, Lei L, Peng J, et al. Cancer burden in China: trends, risk factors and prevention. Cancer Biol Med. 2020;17(4):879-895. doi: 10.20892/j.issn.2095-3941.2020.0387.
40. Tang Y-l, Xiang X-j, Wang X-y, Cubells JF, Babor TF, Hao W. Alcohol and alcohol-related harm in China: Policy changes needed. Bulletin of the World Health Organization. 2013;91(4):270-276. doi: 10.2471/BLT.12.107318.
41. Zhang M, Yang L, Wang L, Jiang Y, Huang Z, Zhao Z, et al. Trends in smoking prevalence in urban and rural China, 2007 to 2018: Findings from 5 consecutive nationally representative cross-sectional surveys. PLOS Medicine. 2022;19(8):e1004064. doi: 10.1371/journal.pmed.1004064.
42. Li J, Wu B, Selbæk G, Krokstad S, Helvik A-S. Factors associated with consumption of alcohol in older adults-a comparison between two cultures, China and Norway: The CLHLS and the HUNT-study. BMC Geriatrics. 2017;17(1):1-10. doi: 10.1186/s12877-017-0562-9.
43. Feng RM, Zong YN, Cao SM, Xu RH. Current cancer situation in China: good or bad news from the 2018 Global Cancer Statistics? Cancer Commun (Lond). 2019;39(1):22. doi: 10.1186/s40880-019-0368-6.

44. Scoccianti C, Key TJ, Anderson AS, Armaroli P, Berrino F, Cecchini M, et al. European code against cancer 4th edition: Breastfeeding and cancer. Cancer Epidemiology. 2015;39:S101-S106. doi: 10.1016/j.canep.2014.12.007.

45. Espina C, Porta M, Schüz J, Aguado Ildefonso H, Percival Robert V, Dora C, et al. Environmental and Occupational Interventions for Primary Prevention of Cancer: A Cross-Sectorial Policy Framework. Environmental Health Perspectives. 2013;121(4):420-426. doi: 10.1289/ehp.1205897.

46. Jensen JD, Moriarty CM, Hurley RJ, Stryker JE. Making sense of cancer news coverage trends: A comparison of three comprehensive content analyses. Journal of Health Communication. 2010;15(2):136-151. doi: 10.1080/10810730903528025.

47. Huang HY, Shi JF, Guo LW, Zhu XY, Wang L, Liao XZ, et al. Expenditure and financial burden for common cancers in China: a hospital-based multicentre cross-sectional study. Lancet. 2016;388:10-10. doi: 10.1016/S0140-6736(16)31937-7.

48. Daily Ps. 17 kinds of anticancer drugs were added to the national drug reimbursement list to reduce the cost burden of patients by more than 75% 17 □□□□□□□□□□□□□□□□ 75%. 2019; URL: http://www.gov.cn/xinwen/2019-02/13/content_5365211.htm.]

49. Fang W, Xu X, Zhu Y, Dai H, Shang L, Li X. Impact of the National Health Insurance Coverage Policy on the Utilisation and Accessibility of Innovative Anti-cancer Medicines in China: An Interrupted Time-Series Study. Front Public Health. 2021;9(1097):714127. doi: 10.3389/fpubh.2021.714127.

## MULTIMEDIA APPENDIX 1

***Definitions and Descriptions of Coding Items***

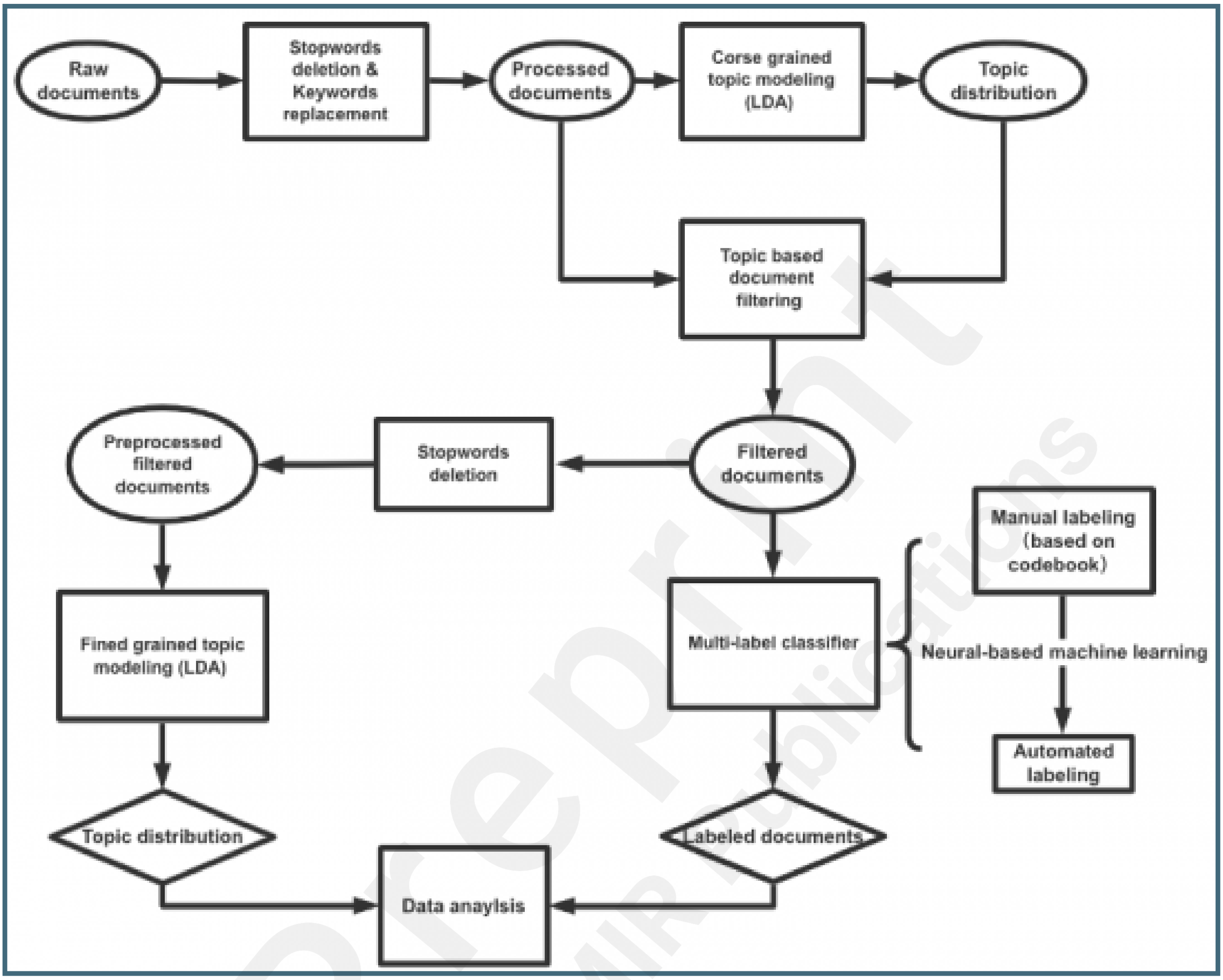
| Coding Item | Definition and descriptions |
|---|---|
| ***Cancer Prevention Measures*** | |
| Avoid tobacco use | Suggestions related to avoiding tobacco use to prevent cancer (including cigarettes and smokeless tobacco). |
| Maintain a healthy weight | Suggestions related to maintaining or keeping a healthy weight to prevent cancer. |
| Healthy diet | Suggestions related to a healthy diet to prevent cancer: eat whole grains, fruit, vegetables, and beans; limit red meat and processed foods; limit fast foods high in fat, starches or sugars. |
| Exercise regularly | Suggestions related to exercising regularly to prevent cancer, including all types of physical activities. |
| Limit alcohol use | Suggestions related to limiting alcohol use to prevent cancer. |
| Practice safe sex | Suggestions related to practicing safe sex to prevent cancer, like using condoms when having sex. |
| Get vaccinated | Suggestions related to getting vaccinated against hepatitis B and human papillomavirus (HPV) to prevent cancer. |
| Reduce exposure to ultraviolet radiation and ionizing radiation | Suggestions related to reducing exposure to ultraviolet radiation and ionizing radiation.<br>1. Ultraviolet radiation from sunlight.<br>2. Ionizing radiation includes medical radiation from tests to diagnose cancer such as x-rays, CT scans, fluoroscopy, and nuclear medicine scans. Radon gas in our homes. |
| Avoid urban air pollution and indoor smoke from household use of solid fuels | Suggestions related to avoiding urban air pollution and indoor smoke from household use of solid fuels to prevent cancer. |
| Early screening and testing | Suggestions related to early screening and testing to prevent cancer. |
| Breastfeeding | Suggestions related to breastfeeding to prevent breast cancer (female only). |
| Controlling chronic infections | Suggestions related to controlling chronic infections to prevent cancer. |
| Other | Preventive measures do not belong to any of the categories above. |
| ***Cancer Treatment*** | |

| | |
|---|---|
| ***Measures*** | |
| Surgery | Surgery, when used to treat cancer, is a procedure in which a surgeon removes cancer from the patient's body, including cuts with scalpels, cryotherapy, lasers, hyperthermia, photodynamic therapy. |
| Radiation therapy | Radiation therapy (also called radiotherapy) is a cancer treatment that uses high doses of radiation to kill cancer cells and shrink tumors. |
| Chemotherapy | Chemotherapy (also called chemo) is a type of cancer treatment that uses drugs to kill cancer cells, including external beam radiation therapy, brachytherapy, radioactive iodine, I-131. |
| Immunotherapy | Immunotherapy is a type of cancer treatment that helps your immune system fight cancer. Immunotherapy is a type of biological therapy, including immune checkpoint inhibitors, T-cell transfer therapy, monoclonal antibodies, immune system modulators, etc. |
| Targeted Therapy | Most targeted therapies are either small-molecule drugs or monoclonal antibodies. |
| Hormone Therapy | Hormone therapy is a cancer treatment that slows or stops the growth of cancer that uses hormones to grow. Hormone therapy is also called hormonal therapy, hormone treatment, or endocrine therapy. |
| Stem cell transplant | Stem cell transplants are procedures that restore blood-forming stem cells in people who have had theirs destroyed by the very high doses of chemotherapy or radiation therapy that are used to treat certain cancers. |
| Cancer biomarker testing | Biomarker testing is a way to look for genes, proteins, and other substances (called biomarkers or tumor markers) that can provide information about cancer. Biomarker testing for cancer treatment may also be called:<br>1. Tumor testing<br>2. Tumor genetic testing<br>3. Genomic testing or genomic profiling<br>4. Molecular testing or molecular profiling<br>5. Somatic testing<br>6. Tumor subtyping. |
| Other | Treatment measures do not belong to any of the categories above. |

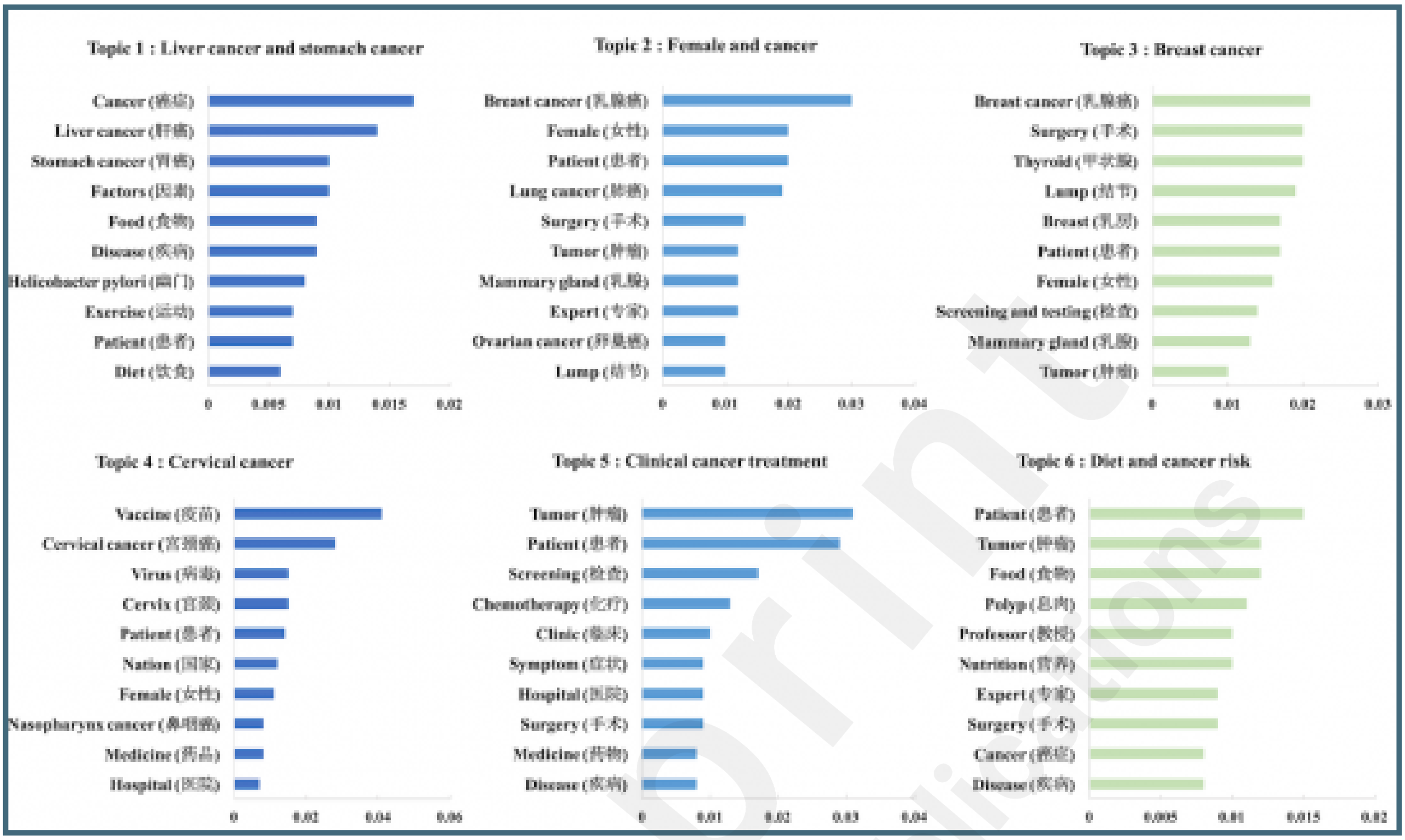
# Supplementary Files

# Figures

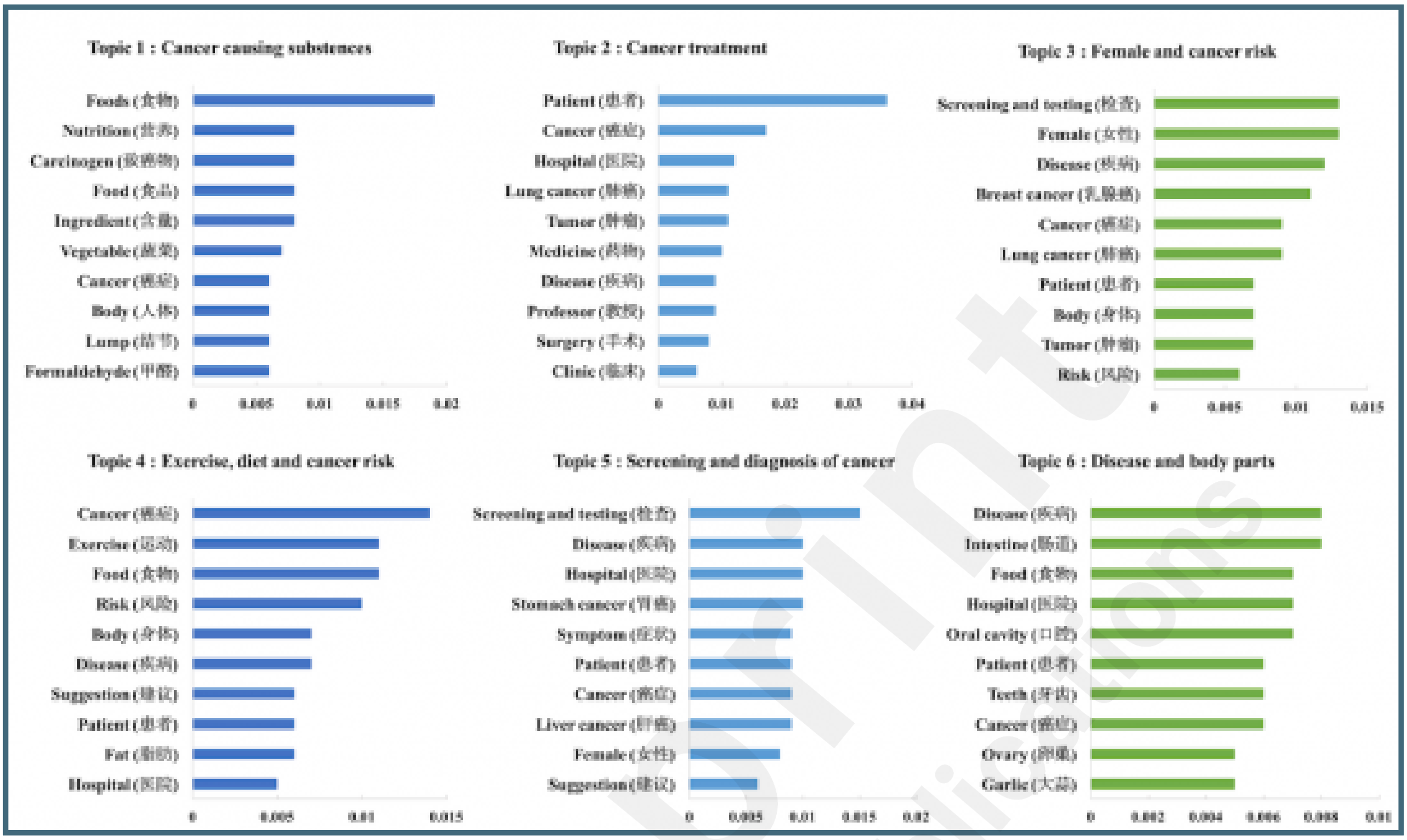
Workflow chart.


Raw documents
Stopwords deletion & Keywords replacement
Processed documents
Corse grained topic modeling (LDA)
Topic distribution
Topic based document filtering
Filtered documents
Stopwords deletion
Preprocessed filtered documents
Fined grained topic modeling (LDA)
Multi-label classifier
Manual labeling (based on codebook)
Neural-based machine learning
Automated labeling
Topic distribution
Labeled documents
Data anaylsis

Cancer topics on MSM.



Cancer topics on CSM.



# Multimedia Appendixes

Definitions and descriptions of coding items.
URL: http://asset.jmir.pub/assets/398f25f2e9ddb24df8424544b27aaaf5.pdf