BMC Research Notes

DATA NOTE

Open Access

WLCD: a dataset of lifestyle in relation with women's cancer

Alireza Ardalani[1] and Mojtaba Daneshvar[2*]

Abstract

Objectives Social media text mining has been widely used to extract information about the experiences and needs of patients regarding various diseases, especially cancer. Understanding these issues is necessary for further management in primary care. Researchers have identified that lifestyle factors such as diet, exercise, alcohol, and Smoking are associated with cancer risks, particularly women's cancer. Considering the growing trend in the global burden of women's cancer, it is essential to monitor up-to-date data sources using text mining.

Data description We have prepared six independent datasets regarding lifestyle components and women's cancer: (1) a dataset of nutrition containing 10,161 tweets; (2) a dataset of exercise containing 9412 tweets; (3) a dataset of alcohol containing 2132 tweets; (4) a dataset of Smoking containing 4316 tweets; and (5) a dataset of lifestyle (term) containing 1861 tweets. We also construct an additional dataset: (6) a dataset by summing other components containing 27,882 tweets. These data are provided to discover people's perspectives, knowledge, and experiences regarding lifestyle and women's cancer. Hence, it should be valuable for healthcare providers to develop more efficient patient management approaches.

Keywords Twitter, Text-mining, Cancer, Women, Lifestyle

Objective

Cancer is one of the leading causes of mortality and morbidity worldwide. Growing trends in cancer burden, especially among women, have become a significant global health issue [1]. Lifestyle factors, including unhealthy diet, physical inactivity [2], smoking, and alcohol use [3], are among the risk factors of cancer targeted for primary control. On the other hand, cancer progression and treatments might affect different aspects of lifestyle in cancer patients [4].

Nowadays, the data mining of social media platforms has become an important emerging tool for understanding the experiences and needs of cancer patients. There is a wealth of information available that can be used to gain insight into the patient experience relating to lifestyle patterns [5]. In a previous study, assisted with Twitter data related to breast cancer, researchers identified that physical activity and healthy eating are important factors in symptom management in cases [6]. Another study by analyzing tweets related to site-specific cancers found that physical activity and alcohol consumption are among lifestyle habits that might be associated with liver and breast cancer [7].

By analyzing social media conversations, researchers can identify patterns and trends related to these factors, which can be used to develop targeted public health policies to prevent or manage cancer risk [5, 8].

*Correspondence:
Mojtaba Daneshvar
aref.daneshvar@gmail.com
[1]Iran University of Science and Technology, Tehran, Iran
[2]Tehran University of medical sciences, Tehran, Iran

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data set 1	*Alcohol*	*Text file (.txt), Excel file (.xlsx)*	*OSF (https://doi.org/10.17605/OSF.IO/WC89Z)*
Data set 2	*Diet*	*Text file (.txt), Excel file (.xlsx)*	*OSF (https://doi.org/10.17605/OSF.IO/WC89Z*
Data set 3	*Exercise*	*Text file (.txt), Excel file (.xlsx)*	*OSF (https://doi.org/10.17605/OSF.IO/WC89Z*
Data set 4	*Smoking*	*Text file (.txt), Excel file (.xlsx)*	*OSF (https://doi.org/10.17605/OSF.IO/WC89Z*
Data set 5	*Lifestyle (term)*	*Text file (.txt), Excel file (.xlsx)*	*OSF (https://doi.org/10.17605/OSF.IO/WC89Z*
Data set 6	*Total*	*Text file (.txt), Excel file (.xlsx)*	*OSF (https://doi.org/10.17605/OSF.IO/WC89Z*

This approach can also help healthcare providers better address cancer patients' psychological and emotional needs [9]. By analyzing online discussions, investigators can gain insights into patients' awareness and identify opportunities for providing proper support and resources associated with lifestyle modification approaches [8, 10].

Social media data mining provides a unique opportunity for public health strategists to understand better people's attitudes toward the association between lifestyle and women's cancer and healthcare delivery [11]. By leveraging this information, researchers and healthcare providers can develop targeted interventions that promote healthy lifestyles and improve treatment outcomes, especially among cancer patients [11, 12]. The main objective of this research is to provide Twitter-based datasets containing tweets related to lifestyle and women's cancer.

Data description

This study collected tweets related to Women, Lifestyle, and Cancer as a Dataset (WLCD). We have used the following keywords for each section: (1) Lifestyle components including diet (*"diet", "nutrition", "eating", "food", and "feed"*), physical activity (*"exercise", "training", "workout", "gym", "fitness", "yoga", "aerobic", "athlete", "sedentary", "jogging", "running", "physical activity"*), alcohol (*"alcohol", "drink", "ethanol", "liquor", "drunk"*), Smoking (*"smoke", "smoking", "cigar", "cigarette", "tobacco", "smoker", "shisha", "vape"*), and lifestyle (*"lifestyle", "lifestyle"*); (2) Women (*"mother", "women", "woman", "female", "wife", "wives", "gynecologic", "ovarian", "ovary", "cervix", "cervical", "breast", "endometrium", "endometrial"*); (3) Cancer (*"cancer", "carcinoma", "tumor", "chemotherapy", "radiotherapy", "chemo", "cancerous"*). Combinations of keywords were used to reach pertinent search queries and obtain tweets related to lifestyle and women's cancer. Two independent researchers applied refining search queries and manual review of extracted tweets to ensure the quality and relevance of the queries. Data were gathered from January 1st to December 31st, 2022. Data collection was not limited to location, user, and originality (retweets and quotes included). Tweets were extracted via the Twitter API and presented as multiple datasets, stored at the "Open Science Framework" (OSF: https://osf.io/wc89z/). We have prepared five datasets according to predefined lifestyle components and a cumulative dataset of components. The name of the datasets and the number of tweets are as follows: (1) Diet and women's cancer (10,161 Tweets, see Table 1, dataset 1); (2) Exercise and women's cancer (9412 Tweets, see Table 1, dataset 2); (3) Alcohol and women's cancer (2132 Tweets, see Table 1, dataset 3); (4) Smoking and women's cancer (4316 Tweets, see Table 1, dataset 4); (5) Lifestyle and women's cancer (1861 Tweets, see Table 1, dataset 5); (6) The final dataset of lifestyle components and women's cancer (27,882 Tweets, see Table 1, dataset 6). Datasets contain the tweet's text, and are prepared in Excel (.xlsx) and text (.txt) formats (see Table 1).

Limitations

- This study assessed only Twitter users that may not represent the general population. Data from other social media, such as Facebook, Instagram, and Reddit, might be needed to have more comprehensive results.
- People may report inaccurate or incomplete information about their lifestyle and health status due to social undesirability.
- Habits and experiences reported by users may be timely, leading to potential misinterpretation.

Abbreviations

WLCD Women Lifestyle and Cancer Dataset

Acknowledgements

None.

Authors' contributions

AA and MD contributed to the study designation, data extraction, and writing the manuscript. The final manuscript was read and approved by the authors.

Funding
None.

Data Availability
The data described in this Data note can be freely and openly accessed on the OSF repository under reference number wc89z https://osf.io/wc89z/. Please see Table 1 for details and links to the data.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Received: 4 May 2023 / Accepted: 11 August 2023
Published online: 22 August 2023

References

1. Ginsburg O, et al. The global burden of women's cancers: a grand challenge in global health. The Lancet. 2017;389(10071):847–60. https://doi.org/10.1016/S0140-6736(16)31392-7. 2017/2//.
2. McTiernan A, Irwin M, VonGruenigen V. Weight, physical activity, Diet, and prognosis in breast and gynecologic cancers. J Clin Oncol. 2010;28(26):4074–80. https://doi.org/10.1200/JCO.2010.27.9752. 2010/9//.
3. Keyvani V, Kheradmand N, Navaei ZN, Mollazadeh S, Esmaeili S-A. Epidemiological trends and risk factors of gynecological cancers: an update. Med Oncol. 2023;40(3):93–3. https://doi.org/10.1007/s12032-023-01957-3. 2023/2//.
4. van Broekhoven MECL, et al. Illness perceptions and changes in lifestyle following a gynecological cancer diagnosis: a longitudinal analysis. Gynecol Oncol. 2017;145(2):310–8. https://doi.org/10.1016/j.ygyno.2017.02.037. 2017/5//.
5. Sugawara Y, Narimatsu H, Hozawa A, Shao L, Otani K, Fukao A. Cancer patients on Twitter: a novel patient community on social media. BMC Res Notes. 2012;5(1):699–9. https://doi.org/10.1186/1756-0500-5-699. 2012/12//.
6. Attai DJ, Cowher MS, Al-Hamadani M, Schoger JM, Staley AC, Landercasper J. Twitter Social Media is an effective Tool for breast Cancer patient education and support: patient-reported outcomes by Survey. J Med Internet Res. 2015;17. https://doi.org/10.2196/jmir.4721. no. 7, pp. e188-e188, 2015/7//.
7. Khandelwal S, Routray A. "Coverage and Evolution of Cancer and Its Risk Factors - A Quantitative Study with Social Signals and Web-Data," 2020, pp. 108–23.
8. Xu S, Markson C, Costello KL, Xing CY, Demissie K, Llanos AAM. Leveraging Social Media to Promote Public Health knowledge: Example of Cancer Awareness via Twitter. JMIR Public Health and Surveillance. 2016;2(1). https://doi.org/10.2196/publichealth.5205. e17-e17, 2016/4//.
9. Falisi AL, Wiseman KP, Gaysynsky A, Scheideler JK, Ramin DA, Chou W-yS. Social media for breast cancer survivors: a literature review. J Cancer Surviv. 2017;11(6):808–21. https://doi.org/10.1007/s11764-017-0620-5. 2017/12//.
10. Shaw G Jr, Sharma T, Ramakrishnan S et al. "Exploring Diabetes and Users' lifestyle choices in Twitter to improve health outcomes," in *Shaw. Exploring Diabetes and Users' Lifestyle Choices Proceedings of the Southern Association for Information Systems Conferen, Simons Island, Georgia, USA, 2019, pp. 15–17*
11. Singh T, et al. Social media as a Research Tool (SMaaRT) for Risky Behavior Analytics: Methodological Review. JMIR Public Health and Surveillance. 2020;6(4). https://doi.org/10.2196/21660. e21660-e21660, 2020/11//.
12. Tapi Nzali MD, Bringay S, Lavergne C, Mollevi C, Opitz T. What patients can tell us: topic analysis for social media on breast Cancer. JMIR Med Inf. 2017;5(3). https://doi.org/10.2196/medinform.7779. e23-e23, 2017/7//.

Publisher's Note