DATA NOTE

Open Access

A gender-specific geodatabase of five cancer types with the highest frequency of occurrence in Iran

Sharareh Faramarzi[1], Behzad Kiani[2] , Mohammedreza Hoseinkhani[1] and Neda Firouraghi[1*]

Abstract

Objectives Cancer is a global health challenge with complex characteristics. Despite progress in research and treatment, a universally effective prevention strategy is lacking. Access to reliable information, especially on occurrence rates, is vital for cancer management. This study aims to create a database containing individual and spatially integrated data on commonly diagnosed cancers in Iran from 2014 to 2017, serving as a valuable resource for spatial-epidemiological approaches.

Data description This database encompasses several files related to cancer data. The first file is an Excel spreadsheet, containing information on newly diagnosed cancer cases from 2014 to 2017. It provides demographic details and specific characteristics of 482,229 cancer patients. We categorized this data according to the International Agency for Research on Cancer (IARC) reporting rules to identify cancers with the highest incidence. To create a geodatabase, individual data was integrated at the county level and combined with population data. Files 2 and 3 contain gender-specific spatial data for the top cancer types and non-melanoma skin cancer. Each file includes county identifications, the number of cancer cases for each cancer type per year, and gender-specific population information. Lastly, there is a user's guide file to help navigate through the data files.

Keywords Cancer incidence, Cancer epidemiology, Spatial epidemiology, Public health, Geographic Information Systems (GIS), Iran

Objective

Cancer has emerged as a notable focal point within public health across various communities, standing as the third primary contributor to mortality [1]. Regrettably, there has been a surge in cancer occurrences in recent times, amplifying its importance as a foremost concern within the healthcare domain [2]. The impact of cancer extends beyond developed nations and encompasses low- and middle-income countries as well, where resources for prevention, early detection, and treatment are often inadequate. Roughly 70% of global cancer cases are concentrated in these lower- and middle-income nations [3, 4]. Annually, over 50,000 new cases are identified within the Iranian population [5]. The living conditions of individuals can have a significant impact on their overall health. Cancer trends can vary among different populations and regions, highlighting the influence of factors such as work-related and industrial settings [6], socioeconomic situation [7], healthcare accessibility

*Correspondence:
Neda Firouraghi
n.firooraghi@gmail.com
[1]Department of Medical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[2]UQ Centre for Clinical Research, Faculty of Medicine, The University of Queensland, Brisbane, Queensland, Australia

[8], and environmental exposures that play a crucial role on cancer incidence [9, 10]. This wide variety of factors influencing cancer incidence showcases the dynamic, multidisciplinary nature of cancer research.

The geographical variation and the impact of environmental factors underscore the importance of spatial analysis in understanding patterns across both space and time. Spatial analysis of diseases contributes to recognize high-risk areas and patterns of occurrence to provide evidence-based information in order to enact efficient screening and disease management strategies in areas with high priority [11–16]. Utilizing spatial techniques such as local Moran *I*, hot spot analysis, and spatiotemporal scanning can aid in identifying high-risk zones. This approach can help generate hypotheses for further analysis on the relationships between risk factors and cancer incidence, providing valuable insights into the connections between environmental hazards, individual lifestyles, and cancer incidence within communities [17, 18].

In this research, we utilized Geographic Information System (GIS) to create a geodatabase encompassing high prevalent cancer cases in Iran between 2014 and 2017. This database is a set of individual and spatial data which can serve as a valuable resource for identifying spatial and temporal patterns of high-risk and low-risk areas for cancer incidence. In forthcoming studies, the geodatabase can be combined with socio-environmental factors such as poverty rate, income and lifestyle factors to assess the risk factors [13]. By integrating these datasets, researchers can take proactive measures in resource allocation and tailor interventions to specific geographic locations with higher cancer risks. This geodatabase might be useful for both spatial epidemiological research and machine learning algorithms in terms of classification or clustering studies.

Data description

This data was collected across the entire country of Iran, located in western Asia. Cancer has become the third most prominent cause of death in the country, with increasing in occurrence in recent years. This rise can be attributed to the country's rapid advancements in industrialization and modernization, as well as significant changes in people's lifestyles and environment. These transformations have the potential to impact the occurrence and distribution of different types of cancer [19–21]. Incidence rates of cancer may vary across different geographic locations, potentially due to differing environmental factors [22]. Lifestyle [23] and environmental factors play significant roles in contributing to this phenomenon [24]. Pesticides and industrial chemicals have been linked to an increased risk of cancer [25, 26]. Urbanization causes air pollution, and sedentary

lifestyles, and increased exposure to carcinogens [27, 28]. Climate variations and geographic characteristics can affect cancer rates, with regions experiencing increased ultraviolet radiation exposure having higher rates of skin cancer [29]. Moreover, lifestyle factors such as smoking, diet, physical activity, and alcohol consumption can all influence cancer risk and may vary by geographic location [30, 31] and areas with lower socioeconomic status may have higher rates of cancer due to factors such as limited access to healthcare, and unhealthy living conditions [28, 32].

In this study the data was obtained for entire country from three different sources. Population data was gathered from Iran's statistical center, through the most recent national census in 2016 [33]. According to this data, Iran had an estimated total population of 80 million. The country boundaries, in the county scale, were provided by the Ministry of the Interior in vector map shapefile format. Iran is composed of 417 counties and 31 provinces, encompassing a total land area of 1,648,195 km [33]. During the years 2014–2017, 482,229 cancer cases data were obtained from the Iranian National Population-Based Cancer Registry (INPCR) [34]. The INPCR records newly diagnosed cases of cancer with malignant primary tumors. In the case of metastatic cancers, the focus is on tracing back to the primary tumor, and only information about the primary tumor is recorded for that patient. Tumor topography, morphology, and grade are coded in this registry using the third edition of the International Classification of Diseases for Oncology (ICD-O) [35].

Population-based cancer data has been collected from various sources including death certificates, clinical investigations such as X-ray, endoscopy, imaging, ultrasound, exploratory surgery (such as laparotomy), cytology, and pathology. To ensure accuracy and completeness, strategies like training staff, following standardized procedures and guidelines (based on ICD-O, the International Agency for Research on Cancer (IARC), and WHO guidelines), conducting audits, utilizing validation checks, and comparing data with other sources are implemented [21]. The university cancer registry secretariat uses the Sima-ye-Saratan system to process and control the quality of data. They check for duplicate records and ensure internal consistency before submitting the data to the national registry. Patient information is entered and checked for duplicates before tumor information is added [34].

This database includes 3 data files and a help file (Table 1). Data file 1 includes individual data of cancer cases diagnosed between 2014 and 2017, over the whole country. This data contains sex, age, diagnosis year, code of tumor topography, code of tumor morphology and behavior, code of tumor grade, source of diagnosis report,

Table 1 Dataset overview

Label	Name of data file	File types (File extension)	Data repository and identifier
Data file 1	Individual data of cancer cases	MS Excel file (*.xlsx)	Harvard Dataverse, (https://doi.org/10.7910/DVN/7ZK41X) [42]
Data file 2	Spatial data of top cancer types in females	Polygonal features shapefile format (*.shp)	Harvard Dataverse, (https://doi.org/10.7910/DVN/7ZK41X) [43]
Data file 3	Spatial data of top cancer types in males	Polygonal features shapefile format (*.shp)	Harvard Dataverse, (https://doi.org/10.7910/DVN/7ZK41X) [44]
Data file 4	Help file	MS Excel file (*.xlsx)	Harvard Dataverse, (https://doi.org/10.7910/DVN/7ZK41X) [45]

and county ID. We conducted an examination to prepare the gender-specific spatial data of top cancer types in the matter of incidence. The individual data has been carefully categorized according to IARC reporting rules, and we have identified the five cancer types with the highest incidence rates in Iran, in addition to non-melanoma skin cancer. It is noteworthy that non-melanoma skin cancer is often excluded from global cancer case counts. The exclusion is primarily due to its widespread occurrence and predominant treatment within primary healthcare facilities, contributing to potential under-reporting in national cancer registry data [36]. Furthermore, we have geocoded and linked this data to the county level boundaries as a geographical reference and incorporated population data as well.

Data file 2 includes spatial data for five cancer types with the highest frequency of occurrence and non-melanoma skin cancer in females. The cancer types covered are breast, non-melanoma skin cancer, thyroid, stomach, colon and brain and nervous system. This file provides aggregated data based on county boundaries and includes county identification number and name, geographical coordinates (longitude and latitude), number of cancer cases for each cancer per year, total and females' population. Data file 3 is assigned for spatial data of top cancer types in males. These files include stomach, non-melanoma skin cancer, prostate, bladder, trachea, bronchial, lung cancer (TBL) and colon cancer. Similar to the previous file, each file comprises aggregated data on cancer cases and includes county identification number and name, geographical coordinates, number of cancer type per year, total and males' population. These data files are in shapefile format which is a digital format used for storing both the geometric location and relevant attribute information of vector-based data. It is commonly used for spatial data storage [37]. Data file 4 is a help file in Microsoft Excel format that provides a description of the fields used in the previous files. This file includes two sheets, first one has been designed to assist in understanding and utilizing the data from other files and the second sheet contains the list of counties ID and names. The data does not include any identification data of patients.

These data files can be used by researchers in different disciplines such as spatial epidemiology, health and cancer epidemiology, public health, and health service

research. These data can be used for spatial visualization (hotspot analysis [38, 39], and local Moran's clustering [40], spatio-temporal analysis such as purely temporal, purely spatial, spatial-temporal, and spatial variation to detect spatial patterns of different types of cancer [14, 41]. By incorporating sociodemographic and environmental variables, a regression model or artificial intelligence methods can be employed to explore and understand the correlation between cancer incidence rates and lifestyle or environmental factors. GIS is well-equipped to integrate diverse data from multiple sources, including spatial, temporal, and descriptive elements, within a unified framework. Table 1 presents the specifics of each dataset and offers access links to these data.

Limitation
The data recorded and reported by INPCR had a delay of 4 years, which means the most recent data is not available to us. Additionally, the prevalence of non-melanoma skin cancer, often treated in primary healthcare settings, may lead to under-reporting in national cancer registry data. Another limitation we encountered was the lack of an annual census in Iran. Instead, we had to rely on the latest census conducted in 2016.

Abbreviations
GIS   Geographic Information Systems
INPCR   Iranian National Population-Based Cancer Registry
ICD-O   International Classification of Diseases for Oncology
IARC   International Agency for Research on Cancer
TBL   Trachea, Bronchus, Lung

Data availability
The data described in this data note can be freely and openly accessed on Harvard Data-verse (https://doi.org/10.7910/DVN/7ZK41X) [46]; consisting of four datafiles:
Data file 1 (https://doi.org/10.7910/DVN/7ZK41X)[42],

Data file 2 (https://doi.org/10.7910/DVN/7ZK41X)[43],
Data file 3 (https://doi.org/10.7910/DVN/7ZK41X)[44],
Data file 4 (https://doi.org/10.7910/DVN/7ZK41X)[45].
Additional information can be found in Table 1.

Declarations

Ethics approval and consent to participate
The research protocol bearing the assigned protocol number 4001462 has undergone a comprehensive evaluation and scrutiny by the Ethics Committee of Mashhad University of Medical Sciences. Upon assessment, the committee has concluded that the study complies with the ethical guidelines and regulations pertaining to research involving human participants data. The research procedure did not entail any direct involvement or interaction with human participants. As a result, obtaining individual informed consent was considered unnecessary for this particular study as determined by the Ethics Committee of Mashhad University of Medical Sciences.

Consent for publication
Not applicable. Since the study used data without including any personally identifiable information, the requirement for informed consent was not applicable. The study was conducted with ethical considerations, aligning with the regulations and guidelines laid out by the Ethics Committee of Mashhad University of Medical Sciences.

Competing interests
The authors declare no competing interests.

References
1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359–86.
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. Cancer J Clin. 2023;73(1):17–48.
3. Akimana B, Abbo C, Balagadde-Kambugu J, Nakimuli-Mpungu E. Prevalence and factors associated with major depressive disorder in children and adolescents at the Uganda Cancer Institute. BMC Cancer. 2019;19(1):466.
4. Li Q, Lin Y, Xu Y, Zhou H. The impact of depression and anxiety on quality of life in Chinese cancer patient-family caregiver dyads, a cross-sectional study. Health Qual Life Outcomes. 2018;16(1):230.
5. Mohebbi M, Mahmoodi M, Wolfe R, Nourijelyani K, Mohammad K, Zeraati H, et al. Geographical spread of gastrointestinal tract cancer incidence in the Caspian Sea region of Iran: spatial analysis of cancer registry data. BMC Cancer. 2008;8(1):137.
6. Hopf NB, Bolognesi C, Danuser B, Wild P. Biological monitoring of workers exposed to carcinogens using the buccal micronucleus approach: a systematic review and meta-analysis. Mutat Research/Reviews Mutat Res. 2019;781:11–29.
7. González LV, Sotos FE, de Miguel Ibáñez R. Colorectal Cancer screening in Castilla La Mancha, Spain: the influence of Social, Economic, Demographic and Geographic factors. J Community Health. 2022;47(3):446–53.
8. Ambroggi M, Biasini C, Del Giovane C, Fornari F, Cavanna L. Distance as a barrier to Cancer diagnosis and treatment: review of the literature. Oncologist. 2015;20(12):1378–85.
9. Viñas Casasola MJ, Fernández Navarro P, Fajardo Rivas ML, Gurucelain Raposo JL, Alguacil Ojeda J. [Municipal distribution of the incidence of the most common tumours in an area with high cancer mortality]. Gac Sanit. 2017;31(2):100–7.
10. Syriopoulou E, Morris E, Finan PJ, Lambert PC, Rutherford MJ. Understanding the impact of socioeconomic differences in colorectal cancer survival: potential gain in life-years. Br J Cancer. 2019;120(11):1052–8.
11. Sahar L, Foster SL, Sherman RL, Henry KA, Goldberg DW, Stinchcomb DG, et al. GIScience and cancer: state of the art and trends for cancer surveillance and epidemiology. Cancer. 2019;125(15):2544–60.
12. Graves BA. Integrative literature review: a review of literature related to geographical information systems, healthcare access, and health outcomes. Perspect Health Inf Manag. 2008;5:11.
13. Goshayeshi L, Pourahmadi A, Ghayour-Mobarhan M, Hashtarkhani S, Karimian S, Shahhosein Dastjerdi R et al. Colorectal cancer risk factors in north-eastern Iran: a retrospective cross-sectional study based on geographical information systems, spatial autocorrelation and regression analysis. Geospat Health. 2019;14(2).
14. Kiani B, Raouf Rahmati A, Bergquist R, Hashtarkhani S, Firouraghi N, Bagheri N, et al. Spatio-temporal epidemiology of the tuberculosis incidence rate in Iran 2008 to 2018. BMC Public Health. 2021;21(1):1093.
15. Firouraghi N, Mohammadi A, Hamer DH, Bergquist R, Mostafavi SM, Shamsoddini A, et al. Spatio-temporal visualisation of cutaneous leishmaniasis in an endemic, urban area in Iran. Acta Trop. 2022;225:106181.
16. Montazeri M, Hoseini B, Firouraghi N, Kiani F, Raouf-Mobini H, Biabangard A, et al. Spatio-temporal mapping of breast and prostate cancers in South Iran from 2014 to 2017. BMC Cancer. 2020;20(1):1170.
17. Faramarzi S, Kiani B, Faramarzi S, Firouraghi N. Cancer patterns in Iran: a gender-specific spatial modelling of cancer incidence during 2014–2017. BMC Cancer. 2024;24(1):191.
18. Firouraghi N, Bergquist R, Fatima M, Mohammadi A, Hamer DH, Shirzadi MR, et al. High-risk spatiotemporal patterns of cutaneous leishmaniasis: a nationwide study in Iran from 2011 to 2020. Infect Dis Poverty. 2023;12(1):49.
19. Almasi Z, Rafiemanesh H, Salehiniya H. Epidemiology characteristics and trends of incidence and morphology of stomach cancer in Iran. Asian Pac J Cancer Prev. 2015;16(7):2757–61.
20. Rohani-Rasaf M, Abdollahi M, Jazayeri S, Kalantari N, Asadi-Lari M. Correlation of cancer incidence with diet, smoking and socio- economic position across 22 districts of Tehran in 2008. Asian Pac J Cancer Prev. 2013;14(3):1669–76.
21. Roshandel G, Ferlay J, Ghanbari-Motlagh A, Partovipour E, Salavati F, Aryan K, et al. Cancer in Iran 2008 to 2025: recent incidence trends and short-term predictions of the future burden. Int J Cancer. 2021;149(3):594–605.
22. Khazaei S, Ayubi E, Soheylizad M, Manosri K. Incidence rate and distribution of common cancers among Iranian children. Middle East J Cancer. 2017;8(1):39–42.
23. Simonian M, Khosravi S, Mortazavi D, Bagheri H, Salehi R, Hassanzadeh A, et al. Environmental risk factors Associated with sporadic colorectal Cancer in Isfahan, Iran. Middle East J Cancer. 2018;9(4):318–22.
24. Khorrami Z, Pourkhosravani M, Rezapour M, Etemad K, Taghavi Shahri Seyed M, Künzli N et al. Multiple Air pollutant exposure and lung cancer in Tehran, Iran. ISEE Conference Abstracts.2021(1).
25. Xie PP, Zong ZQ, Qiao JC, Li ZY, Hu CY. Exposure to pesticides and risk of colorectal cancer: a systematic review and meta-analysis. Environ Pollut. 2024:123530.
26. Ayuso-Álvarez A, García-Pérez J, Triviño-Juárez J-M, Larrinaga-Torrontegui U, González-Sánchez M, Ramis R, et al. Association between proximity to industrial chemical installations and cancer mortality in Spain. Environ Pollut. 2020;260:113869.
27. White AJ, Keller JP, Zhao S, Carroll R, Kaufman JD, Sandler DP. Air Pollution, Clustering of Particulate Matter Components, and breast Cancer in the Sister Study: a U.S.-Wide cohort. Environ Health Perspect. 2019;127(10):107002.
28. Rana N, Gosain R, Lemini R, Wang C, Gabriel E, Mohammed T et al. Sociodemographic disparities in gastric adenocarcinoma: a Population-based study. Cancers (Basel). 2020;12(1).
29. Symanzik C, John SM. [Skin cancer from solar ultraviolet radiation exposure at work]. Dermatologie (Heidelb). 2024;75(2):104–11.
30. Babaei M, Pirnejad H, Rezaie J, Roshandel G, Hoseini R. Association between socioeconomic factors and the risk of gastric Cancer incidence: results from an ecological study. Iran J Public Health. 2023;52(8):1739–48.
31. Salamat F, Semnani S, Honarvar MR, Fazel A, Roshandel G. 10-Year trends in Dietary intakes in the high- and low-risk areas for esophageal Cancer: a Population-based ecological study in Northern Iran. Middle East J Dig Dis. 2020;12(2):89–98.
32. Golzari SE, Ghabili K, Khanli HM, Tizro P, Rikhtegar R. Access to cancer medicine in Iran. Lancet Oncol. 2013;14(3):e87.
33. Statistical Centre of Iran (SCI),Population statistics of Iran 2016. Tehran, Iran. 2020 [Available from: https://www.amar.org.ir/.
34. Roshandel G, Ghanbari-Motlagh A, Partovipour E, Salavati F, Hasanpour-Heidari S, Mohammadi G, et al. Cancer incidence in Iran in 2014: results of the Iranian National Population-based Cancer Registry. Cancer Epidemiol. 2019;61:50–8.

35. Fritz APC, Jack A, et al. International classification of diseases for Oncology. 3rd ed. Geneva, Switzerland: World Health Organization; 2000.
36. Ferlay J, Whelan CBS, Parkin DM. Chech and conversion programs for cancer registeries (IARC/IACR tools for Cancer Registries) 2005 [Available from: https://cri.tums.ac.ir/2/pbcr/References/Check%20and%20Conversion%20Programs%20for%20Cancer%20Registries.pdf.
37. ESRI, ESRI_shape. Last significant FDD update: 2020-05-29 [Available from: https://www.loc.gov/preservation/digital/formats/fdd/fdd000280.shtml.
38. Dadashi A, Mohammadi A, MohammadEbrahimi S, Bergquist R, Shamsoddini A, Hesami A et al. Spatial analysis of the 10 most prevalent cancers in north-eastern Iran, 2017–2018. J Spat Sci. 2021:1–21.
39. Cheng W, Washington S. Experimental evaluation of hotspot identification methods. Accid Anal Prev. 2005;37:870–81.
40. Anselin L. Local Indicators of Spatial Association—ISA. Geographical Anal. 2010;27:93–115.
41. Kulldorff M. A spatial scan statistic. Commun Stat Theory Methods 26:1481–961997.
42. 1 Df. [Available from: https://dataverse.harvard.edu/api/access/datafile/8139858.
43. 2 Df. [Available from: https://dataverse.harvard.edu/api/access/datafile/8094015.
44. 3 Df. [Available from: https://dataverse.harvard.edu/api/access/datafile/8094012.
45. 4 Df. [Available from: https://dataverse.harvard.edu/api/access/datafile/8139859.
46. Firouraghi N. A gender-specific geodatabase of five cancer types with the highest frequency of occurrence in Iran, 2024 [https://doi.org/10.7910/DVN/7ZK41X.

Publisher's Note