Data in Brief 57 (2024) 111029

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A dataset of breast cancer risk factors in Cuban women: Epidemiological evidence from Havana

José Manuel Valencia-Moreno[a,b,*], Jose Angel Gonzalez-Fraga[a], Everardo Gutierrez-Lopez[a], Hugo Alexis Cantero-Ronquillo[c]

[a] *Universidad Autónoma de Baja California, Mexico*
[b] *Universidad de las Ciencias Informáticas, La Habana, Cuba*
[c] *Hospital Universitario Clínico-Quirúrgico Comandante Manuel Fajardo, La Habana, Cuba*

ARTICLE INFO

Article history:
Received 10 September 2024
Revised 2 October 2024
Accepted 7 October 2024
Available online 22 October 2024

Dataset link: Breast cancerrisk factors inCubanwomen (Original data)

Keywords:
Breast cancer
Risk factors
Cuba
Epidemiology
Machine learning
Public health
Data analysis

ABSTRACT

This dataset compiles breast cancer risk factors from 1697 Cuban women who attended consultations at the Hospital Universitario Clínico-Quirúrgico Comandante Manuel Fajardo in Havana, Cuba. The data were collected to develop a breast cancer risk estimation model specifically tailored to the Cuban population. The dataset includes 23 variables encompassing internationally recognized risk factors such as family history of breast cancer, lifestyle habits, demographic characteristics, and clinical outcomes. The data were extracted from electronic records and anonymized to protect patient privacy, in compliance with the principles of the Declaration of Helsinki and with the approval of the hospital's scientific and ethics committees. This dataset can be employed in the development of predictive models and in comparative studies of risk factors across different populations. It is important to note that the data originate from a single

[*] Corresponding author at: Universidad Autónoma de Baja California, Mexico.
E-mail address: jova@uabc.edu.mx (J.M. Valencia-Moreno).

https://doi.org/10.1016/j.dib.2024.111029

hospital, which may limit their representativeness at the national level.

Specifications Table

Subject	Health and Medical Sciences.
Specific subject area	Breast cancer risk factors and epidemiology in Cuban women.
Type of data	Table, Raw, csv.
Data collection	The retrospective data were stored and collected from electronic records of breast cancer patients who attended consultations at the Hospital Universitario Clínico-Quirúrgico Comandante Manuel Fajardo.
Data source location	Hospital Universitario Clínico-Quirúrgico Comandante Manuel Fajardo in Havana, Cuba.
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/7jhddnpz2p.*1* Direct URL to data: https://data.mendeley.com/datasets/7jhddnpz2p/1 None
Related research article	Jose Manuel Valencia-Moreno, Jose Angel Gonzalez-Fraga, Everardo Gutierrez-Lopez, Vivian Estrada-Senti, Hugo Alexis Cantero-Ronquillo, Vitaly Kober. Breast cancer risk estimation with intelligent algorithms and risk factors for Cuban women, Computers in Biology and Medicine, Volume 179, 2024, 108,818, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2024.108818

1. Value of the Data

- The data can guide public health policymakers in Cuba in formulating strategies for the prevention and control of breast cancer.
- This dataset serves as a valuable resource for understanding the epidemiology of breast cancer among Cuban women.
- It enables comparative analysis of breast cancer risk factors in Cuban women residing in Havana, helping to identify potential differences when compared to other populations.
- The collected data provide a solid foundation for developing new predictive models for breast cancer, tailored to populations with sociodemographic and genetic characteristics similar to those of Cuban women.
- The dataset can be utilized to improve the accuracy of existing predictive models by incorporating risk factors specific to the Cuban population, thereby enabling the calibration of global models to local contexts.
- The dataset is a valuable resource for students, academics, and researchers in epidemiology, public health, data science and machine learning.

2. Background

The motivation for compiling this dataset arose from the need to develop a breast cancer risk estimation model specifically adapted to Cuban women [1]. While models such as the Gail model in the United States and the Tyrer–Cuzick model in the United Kingdom exist, they have been designed and validated in different populations, limiting their accuracy and applicability to the Cuban population due to genetic, environmental, and socioeconomic differences [2–4]. Comparative studies have shown that these differences can significantly impact the effectiveness of predictive models when applied in diverse contexts [5,6]. For this reason, many countries have undertaken efforts to collect data and develop risk estimation models tailored to their specific populations [7–10]. Following this approach, this dataset was compiled with the aim of developing a predictive model based on intelligent algorithms that more accurately reflects the unique characteristics of Cuban women [1].

3. Data Description

The dataset is provided in comma-separated values (CSV) format and in its raw form. The data were extracted from digital databases and comprise a total of 1697 records of women who attended consultations and whose data were collected. This dataset includes 23 variables, including breast cancer diagnosis and a unique serial number for each case, which serves to ensure patient anonymity in accordance with the Declaration of Helsinki [11].

Table 1 presents the profile of the variables included in the dataset. These variables represent significant risk factors and clinical characteristics associated with breast cancer in the studied

Table 1
Profile of the dataset of Cuban women.

Name	Description	Role	Type	Values	Missing
Id	Sequential registration number	id	Quantitative	1 - 1697	0
Age	Patient's age in completed years	regular	Quantitative	20 - 90	0
Menarche	Age at first menstruation	regular	Quantitative	8 - 17	0
menopause	Age at menopause	regular	Qualitative	No, 0, 30 - 60	0
Agefirst	Age at first successful birth	regular	Qualitative	No, 0, 9–46	0
Children	Number of live births the patient has had	regular	Qualitative	0 - 5, 5+	0
breastfeeding	Months of breastfeeding	regular	Qualitative	No, 0 - 72 (months)	0
Nrelbc	Family history of first-degree breast cancer	regular	Qualitative	Aunt, cousin, daughter, grandmother, mother, no, sister (and combinations)	0
Biopsies	Number of biopsies performed on the patient	regular	Quantitative	0 - 5	1
hyperplasia	Presence of atypical hyperplasia	regular	Qualitative	No, yes	0
Race	Patient's race	regular	Qualitative	Black, mixed, white	0
Year	Year of breast cancer diagnosis	regular	Quantitative	2001 - 2018	537
Imc	Body mass index	regular	Quantitative	5.0 - 88.8	7
Weight	Patient's weight in kilograms	regular	Quantitative	13 - 240	10
Exercise	Number of days of physical activity per week	regular	Qualitative	No, NO, diary, 0 - 7	0
Alcohol	Alcohol consumption	regular	Qualitative	No, yes	0
Tobacco	Tobacco consumption	regular	Qualitative	No, yes	0
Allergies	Type of allergies the patient has	regular	Qualitative	Dermatitis, laryngitis, medicines, no, none, other, rhinitis (and combination)	0
Emotional	Emotional predisposition	regular	Qualitative	Sad, joy	0
Depressive	Presence of depressive symptoms	regular	Qualitative	No, yes	0
histologicalclass	Histological classification of cancer	regular	Quantitative	1 - 11	537
Birads	Birads classification of mammography	regular	Qualitative	3A, 3B, 3C, 4B, 5B, 5C, 6	537
Cancer	Breast cancer diagnosis	label, target	Qualitative	No, yes	0

population of Cuban women. Each variable is described below, providing key information about its nature and its relevance in the context of the study.

In addition, Table 1 details the role of each variable, its type, the amount of missing data, and its domain. It should be noted that the variables “nrelbc” and “allergies” contain multivalued entries, i.e. combinations of different reported values. These details are critical for assessing data quality, performing necessary cleaning tasks, and identifying potential limitations in the analysis.

4. Experimental Design, Materials and Methods

International health organizations have identified a set of key risk factors for breast cancer [12–14]. Based on this information, we conducted a comparative review of these risk factors to determine which should be collected in the context of Cuban women. From this review, a list of internationally accepted risk factors was compiled and selected for data collection.

First, the project was presented to the Scientific Committee of the Hospital Universitario Clínico-Quirúrgico Comandante Manuel Fajardo in Havana, Cuba, where it received the necessary approval. Subsequently, the proposal was submitted for review by the Medical Ethics Committee of the same hospital, which also granted its approval.

With the necessary approvals, retrospective data were requested from the hospital’s electronic database. Due to the nature of the data, informed consent could not be obtained directly from patients; instead, the hospital anonymized the data to ensure privacy and confidentiality.

Most variables were based on patient self-reports, while others, such as histological classification, BIRADS, year of diagnosis, and breast cancer diagnosis, were obtained through standard medical procedures.

Data were collected from patients during mammography and from women with clinical suspicions of breast cancer. Thus, the study population included Cuban women who attended consultations for breast cancer evaluation and received a diagnosis, either positive or negative.

Limitations

There are limitations in the dataset due to variability in capturing and interpreting certain qualitative values, particularly the “nrelbc” variable. This variable includes multiple values such as “no”, “mother”, “sister”, “daughter”, and combinations thereof, adding complexity to its analysis. This heterogeneity may cause inconsistencies in categorizing family history, which could affect the accuracy of predictive models.

For the variables “menopause”, “agefirst”, and “exercise”, the values “No” and “0” (zero) were used to indicate the absence of these characteristics. Additionally, in the “breastfeeding” variable, some values include the word “months.” These inconsistencies may lead to misinterpretation during analysis.

Moreover, the primary limitation is that the data come from a single hospital in Havana, which may not represent the entire Cuban population, limiting the generalizability of the findings. These issues could introduce biases and reduce the precision of predictive models.

To mitigate these limitations, future studies should standardize coding and analysis of qualitative variables like “nrelbc,” re-code “No” and “0” (zero) values, and collect data from multiple health centers in different Cuban regions to enhance representativeness and improve model robustness.

Ethics Statement

Ethical approval for the use of these data was granted by the Scientific Committee and the Medical Ethics Committee of the Hospital Universitario Clínico-Quirúrgico Comandante Manuel

Fajardo in September 2023. Individual informed consent was not required for the analysis of depersonalized health records provided by the hospital, as all data were anonymized, and patients were given the option to opt out of clinical data sharing.

Data Availability

Breast cancerrisk factors inCubanwomen (Original data) (Mendeley Data).

CRediT Author Statement

José Manuel Valencia-Moreno: Conceptualization, Methodology, Investigation, Writing – original draft; **Jose Angel Gonzalez-Fraga:** Methodology, Writing – review & editing; **Everardo Gutierrez-Lopez:** Methodology, Writing – original draft; **Hugo Alexis Cantero-Ronquillo:** Resources, Data curation.

Acknowledgments

We would like to express our gratitude to Dr. Maria Vidal Ledo of the National School of Public Health, Havana, Cuba. We also thank the following physicians for their invaluable assistance: Dr. Dalsy Torres Avila, Dr. Manuel Ortega Soto, Dr. Heydi Bustamante Abreu, and Dr. Monica Gomez Ramos. We also acknowledge the support of the Universidad Autónoma de Baja California in the realization of this work.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.M. Valencia-Moreno, J.A. Gonzalez-Fraga, E. Gutierrez-Lopez, V. Estrada-Senti, H.A. Cantero-Ronquillo, V. Kober, Breast cancer risk estimation with intelligent algorithms and risk factors for Cuban women, Comput. Biol. Med. 179 (2024) 108818, doi:10.1016/j.compbiomed.2024.108818.

[2] M.H. Gail, L.A. Brinton, D.P. Byar, et al., Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, J. Natl. Cancer Inst. 81 (24) (1989) 1879–1886, doi:10.1093/jnci/81.24.1879.

[3] J. Tyrer, S.W. Duffy, J. Cuzick, A breast cancer prediction model incorporating familial and personal risk factors, Stat. Med. 23 (7) (2004) 1111–1130, doi:10.1002/sim.1668.

[4] J. Cuzick, A.R. Brentnall, C. Segal, et al., A new model for breast cancer risk prediction and its validation in a prospective study of 2 Million Women in the UK, Br. J. Cancer 117 (6) (2017) 877–883, doi:10.1038/bjc.2017.229.

[5] J.P. Costantino, M.H. Gail, D. Pee, et al., Validation studies for models projecting the risk of invasive and total breast cancer incidence, J. Natl. Cancer Inst. 91 (18) (1999) 1541–1548, doi:10.1093/jnci/91.18.1541.

[6] A.M. McCarthy, K. Armstrong, Risk prediction models for breast cancer: challenges to implementation, Cancer Epidemiol., Biomark. Prev. 23 (10) (2014) 2324–2334, doi:10.1158/1055-9965.EPI-14-0590.

[7] R.D. Nindrea, E. Usman, Y. Katar, I.Y. Darma, H.H. Warsiti, N.P. Sari, Dataset of Indonesian women's reproductive, high-fat diet and body mass index risk factors for breast cancer, Data Br. 36 (2021) 107107, doi:10.1016/j.dib.2021.107107.

[8] P.E. Oguntunde, A.O. Adejumo, H.I. Okagbue, Breast cancer patients in Nigeria: data exploration approach, Data Br. 15 (2017) 47–57, doi:10.1016/j.dib.2017.08.038.

[9] P.O. Awodutire, O.A. Kolawole, O.R. Ilori, Data on the survival times of breast cancer patients in a Teaching Hospital, Osogbo, Data Br. 32 (2020) 106109, doi:10.1016/j.dib.2020.106109.

[10] R.K. Matsuno, J.P. Costantino, R.G. Ziegler, et al., Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American Women, J. Natl. Cancer Inst. 103 (12) (2011) 951–961, doi:10.1093/jnci/djr154.

[11] World Medical Association, World medical association declaration of Helsinki: ethical principles for medical research involving human subjects, JAMA 310 (20) (2013) 2191–2194, doi:10.1001/jama.2013.281053.

[12] World Health Organization (WHO) (2024). Global breast cancer initiative implementation framework: assessing, strengthening and scaling up of services for the early detection and management of breast cancer: executive summary. wawww.who.int [Internet]. Available from: https://www.who.int/publications/i/item/9789240067134.

[13] Pan American Health OrganizationPrevention: Breast Cancer Risk Factors and Prevention, Pan American Health Organization, Geneva, Switzerland, 2016 Available from https://www.paho.org/en/documents/prevention-breast-cancer-risk-factors-and-prevention .

[14] Centers for Disease Control and Prevention. (2024). Breast Cancer Risk Factors. Available from: https://www.cdc.gov/breast-cancer/risk-factors/index.html.