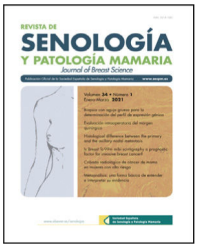
Revista de Senología y Patología Mamaria 35 (2022) 175–183

Sociedad Española de Senología y Patología Mamaria

REVISTA DE SENOLOGÍA Y PATOLOGÍA MAMARIA
Journal of Breast Science

www.elsevier.es/senologia



ORIGINAL ARTICLE

Covariate clustering: Women with breast cancer in southwestern Paraná, Brazil

Neyva Maria Lopes Romeiro[a,*], Carolina Panis[b], Mara Caroline Torres dos Santos[a], Daniel Rech[b], Paulo Laerte Natti[a], Eliandro Rodrigues Cirilo[a]

[a] *Mathematics Department, Universidade Estadual de Londrina, Campus Universitário, Londrina, Brazil*
[b] *Laboratory of tumor biology, Universidade Estadual do Oeste do Paraná, UNIOESTE and Cancer Hospital, Francisco Beltrão, Brazil*

Received 24 November 2021; accepted 14 December 2021
Available online 20 January 2022

KEYWORDS
Breast cancer;
Clusters;
Body mass index;
TNM staging;
Menopause

Abstract
Introduction: Due to the high incidence and aggressiveness of breast cancer, understanding specific factors associated with the profile of the disease is necessary. Thus, the study aimed to analyze data from 155 patients with breast cancer, grouping them according to their clinicopathological characteristics, attended at a reference hospital for Oncology, in 2015–2020, in the southwest region of Paraná, Brazil.
Material and Methods: Using multivariate statistical analysis, sample data were divided into three clusters. The heterogeneity between clusters was obtained by Ward's method. The clinicopathological variables obtained from the patients' medical records were: the presence of intratumoral emboli and lymph nodes, menopausal status, molecular subtype, histological grade, TNM staging of the disease, tumor size, age at diagnóstico, weight, height, and body mass index.
Results: It is observed that 70% of the patients were in menopause at diagnóstico, 31.5% had tumors containing emboli, and 41% had positive lymph nodes. The prevalence of Luminal B subtype, intermediate histological grade, and TNM staging II was verified. The prevalence of the disease was higher in women aged over 50 years, representing 66% of cases. The BMI of the patients ranged from 17.63 kg/m^2 to 51.26 kg/m^2, with 73.55% above 25 kg/m^2. Using the spatial distribution of patients, cluster analysis identified the regions with the worst averages of clinicopathological variables and the highest number of cancer cases.
Conclusion: Through the statistical analysis, it was possible to determine the heterogeneity of the data, so the patients were separated into three clusters. When analyzing the obtained clusters, each one of them had specific characteristics.
© 2021 SESPM. Published by Elsevier España, S.L.U. All rights reserved.

* Corresponding author.
E-mail address: nromeiro@uel.br (N.M.L. Romeiro).

https://doi.org/10.1016/j.senol.2021.12.005
0214-1582/© 2021 SESPM. Published by Elsevier España, S.L.U. All rights reserved.

PALABRAS CLAVE
Cáncer de mama;
Clústeres;
Índice de masa corporal;
Estadificación de TNM;
Menopausia

Agrupamiento covariado: mujeres con cáncer de mama en el suroeste de paraná, Brasil

Resumen
Introducción: Debido a la alta incidencia y agresividad del cáncer de mama, es necesario el conocimiento de factores específicos asociados al perfil de la enfermedad. Así, el objetivo del estudio fue analizar datos de 155 pacientes con cáncer de mama, agrupándolas según sus características clínico-patológicas, atendidas en un hospital de referencia en Oncología, en el período 2015–2020, en la región suroeste de Paraná, Brasil.
Material y métodos: A partir de la utilización de un análisis estadístico multivariado, los datos de la muestra se dividieron en tres grupos. La heterogeneidad entre clústeres se obtuvo mediante el método de Ward. Las variables clínico-patológicas obtenidas de la historia clínica de las pacientes fueron: presencia de émbolos y ganglios linfáticos intratumorales, estado menopáusico, subtipo molecular, grado histológico, estadificación TNM de la enfermedad, tamaño tumoral, edad al momento del diagnóstico, peso, talla, e índice de masa corporal.
Resultados: Se observa que el 70% de las pacientes se encontraba en menopausia al momento del diagnóstico, el 31,5% tenía tumores con émbolos y el 41% tenía ganglios positivos. Se verificó la prevalencia de subtipo luminal B, grado histológico intermedio y estadificación TNM II. La prevalencia de la enfermedad fue mayor en mujeres mayores de 50 años, lo que representa el 66% de los casos. El IMC de los pacientes osciló entre 17,63 kg/m^2 y 51,26 kg/m^2, con un 73,55% encima de 25 kg/m^2. El análisis de clúster, utilizando la distribución espacial de pacientes, identificó las regiones con los peores promedios de variables clínico-patológicas y el mayor número de casos de cáncer.
Conclusión: A través del análisis estadístico fue posible determinar la heterogeneidad de los datos, por lo que las pacientes fueron separadas en tres grupos. Al analizar los clústeres obtenidos, pudo verificarse que cada uno de ellos presentaba características específicas.
© 2021 SESPM. Publicado por Elsevier España, S.L.U. Todos los derechos reservados.

Introduction

Breast cancer is the most common malignant neoplasm in women.[1] Age, over 50 years old, is the most critical risk factor.[2] Other determining factors for the development of the disease are genetic, hereditary, late menopause, obesity, sedentary lifestyle, and frequent exposure to ionizing radiation.[3,4] Such factors are mainly responsible for the clinicopathological differences found in the literature on breast cancer.[5–11]

Specific studies involving the Brazilian population point to classic risk factors, such as aging and menopausal status.[12,13] Other studies show more complex associations, also observed worldwide, such as the development of tumors with a worse prognosis, such as triple-negative, in obese and overweight women.[11,14] Factors such as social vulnerability[15] and a history of psychological stress[16] have also been reported as possible risks associated with the presence of breast cancer in women living in southern Brazil. However, studies referring to regional risk factors are rare and inconclusive.

In this context, it is intended to categorize, through statistical analysis, possible risk factors for breast cancer, targeting patients in the southwest region of Paraná, Brazil. It is known that mathematical analysis can be a powerful tool to assess patient data, providing reliable associations between variables that often cannot be understood in isolation. Considering that physicians may not be familiar with statistical analysis, such interdisciplinary studies become essential.

Data from breast cancer patients can be analyzed using various mathematical tools. We highlight the multivariate analysis that studies the correlation of two or more variables with different information.[17,18] The analysis of these groups can provide relevant information about part of the total sample. In this way, clustering performs a more specific descriptive analysis of the groups within the sample.

It is observed that many studies perform statistical analysis considering correlations of a few variables.[5,8–11,19] In this line of study, to categorize possible risk factors identified in women diagnosed with breast cancer, an exploratory data study is presented considering 11 clinico-pathological variables.

Materials and methods

Sample

The sequential data used contain information from biopsy samples taken serially from women who had lesions suggestive of breast cancer, visualized by imaging tests and physical examinations, in the period from May 2015 to March 2020. Data confidentiality was maintained following clinical research guidelines. The Institutional Ethics Board approved the study under the number CAAE 35524814.4.0000.0107, including 155 patients with a confirmed breast cancer diagnóstico through biopsy. These patients from the 8th Health Regional of the State of Paraná, which covers 25 municipalities, divided into three regions, were treated at the Oncology hospital in Francisco Beltrão, Paraná, Brazil.

Medical records were consulted to obtain data. All patients signed consent, and each protocol followed the principles of medical research involving humans described in the Declaration of Helsinki.

Variables

In this study, variables with different characteristics and applications are considered. The clinical and pathological variables were used to characterize patients and tumors (breast cancer). The spatial distribution of breast cancer cases in the three regions and municipalities of the 8th Health Region of the State of Paraná was also analyzed. Numbers labeled patients.

Furthermore, 11 clinicopathological variables are used to describe characteristics of the disease, such as the presence of intratumoral emboli, the presence of lymph nodes, the menopausal status, the molecular subtype, the histological grade, the TNM staging of the disease, tumor size (cm), age at diagnóstico (years), weight (kg), height (m) and body mass index (BMI) in (kg/m^2).

Histopathological evaluation is essential for the diagnóstico of neoplasia. In this context, for the histological grade variable, the criteria were adopted as being well, moderately, and little differentiated. For molecular subtype, variables were determined as recommended by the St Gallen Consensus.[20] The TNM staging variable was classified concerning the stages of the disease as described by the American Joint Committee on Cancer, Breast Cancer Staging System.[21]

For a better understanding of the work, the variables are classified as:

Binary: intratumoral emboli, lymph node invasion, and menopausal status;

Categorical: molecular subtype, histological grade, and TNM staging;

Quantitative: tumor size, age at diagnóstico, weight, height, and BMI.

Statistical methods

Several methodologies analyze characteristics that differentiate the data from a sample, dividing it into clusters.[17,18,22] In this article, we used the software R.[23] The packages Hmisc[24] and Agricolae[25] were used to facilitate the interpretation of the analysis performed on the data. Cluster analysis was performed by calculating the Euclidean distance between clinicopathological variables.

Initially, to determine which data are more homogeneous with each other, the Euclidean distance method is used. Next, Ward's hierarchical agglomerative method is used to generate the heterogeneous groups among themselves. The result of the analysis, presented in a dendrogram, helps identify the division of groups, thus generating clusters.

Once the clusters were obtained, the calculation of Spearman's lineal correlation between the variables allowed us to understand the influence that one variable exerts over another, enabling the identification of possible risk factors associated with the groups.[26] The Spearman's correlation coefficient varies between −1 and 1, where, to determine the significance of the correlations, the *p*-value was calculated.

To extract characteristics that distinguish the data from different clusters and calculate the correlation, the test of means was used. This procedure allows us to calculate, for each cluster variable, those that present different significant means and those that are just sample variations.

Results

Approximately 31% of the patients had tumors containing intratumoral emboli. It is also noted that the presence of positive lymph nodes was observed in 41% of patients and that 70% are classified as menopausal women at diagnóstico.

On average, the patients in this study have a higher frequency of tumors of the Luminal B molecular subtype, intermediate histological grade, moderately differentiated, and a median TNM stage II, with variations between 0 and IV.

The dispersion of tumor size was observed, ranging from 0.9 cm to 15 cm. The average age of patients is 56.6 years, and the prevalence of the disease was higher in women aged over 45 years, representing 75% of cases. The average weight, when diagnosed, was close to 72.5 kg, but one of the patients weighed 120 kg. Furthermore, only 25% of patients had a BMI of less than 24.8 kg/m^2.

Clinicopathological correlations

Checking the influence that one variable exerts on the other allows a better understanding of the data from a sample, making it possible to identify possible risk factors. Thus, Spearman's lineal correlation is used to estimate the correlation between each pair of variables, evaluating possible connections between clinicopathological variables.

Statistical analysis reveals the existence of significant associations between some of the variables in the sample. These results can be seen through the color map in Table 1, where stronger colors reveal the existence of significant associations ($p < 0.01$) between the variables.

We should highlight the positive and significant correlations of intratumoral emboli with the presence of lymph node invasion and TNM staging. The formation of intratumoral emboli occurs due to tumor-induced coagulation changes. This event facilitates the spread of the disease, explaining its correlation with lymph node invasion.[27]

TNM staging variable presents a positive and significant correlation with the lymph node invasion variable. This correlation was also expected, as the TNM staging calculation uses lymph node invasion as one of its parameters. These results show that the mathematical model follows the clinical classification used to establish the TNM staging.

It is noted that the correlations involving the variables presence of intratumoral emboli, lymph node invasion, and TNM staging did not show significant connections with the variables age and BMI, so that the correlations presented describe risk factors independent of age and the patients' body weight at diagnóstico. However, it is known that both age and obesity are considered determinant risk factors. Table 1 shows a positive and robust correlation between menopause and the patient's age. It is an expected association, as women experience hormonal failure with aging. On the other hand, the data do not show a significant correlation between menopause and being overweight.

Table 1 Spearman's correlations and *p*-value (in parentheses) for the 155 patients diagnosed with breast cancer.

Variables	Lymph node invasion	Menopausal status	Molecular subtype	Histological grade	TNM staging	Tumor size	Age	Weight	Height	BMI
Intratumoral emboli	0.46 (<0.001)	-0.04 (0.588)	-0.01 (0.947)	-0.03 (0.721)	0.36 (<0.001)	0.16 (0.044)	-0.06 (0.478)	-0.02 (0.828)	0.04 (0.646)	-0.04 (0.607)
Lymph node invasion		-0.10 (0.205)	0.08 (0.342)	0.16 (0.044)	0.60 (<0.001)	0.09 (0.283)	-0.08 (0.303)	-0.03 (0.737)	-0.06 (0.423)	0.00 (0.964)
Menopausal status			-0.07 (0.391)	-0.07 (0.415)	-0.04 (0.662)	-0.14 (0.087)	0.71 (<0.001)	-0.03 (0.742)	-0.14 (0.082)	0.05 (0.573)
Molecular subtype				0.23 (0.005)	0.19 (0.017)	0.17 (0.039)	-0.07 (0.373)	-0.08 (0.348)	-0.05 (0.572)	-0.05 (0.555)
Histological grade					0.18 (0.026)	0.16 (0.053)	-0.03 (0.743)	0.03 (0.666)	-0.02 (0.772)	0.04 (0.610)
TNM staging						0.39 (<0.001)	-0.05 (0.564)	0.00 (0.952)	0.01 (0.925)	0.00 (0.992)
Tumor size							-0.06 (0.462)	0.08 (0.344)	0.13 (0.113)	0.04 (0.620)
Age								-0.07 (0.366)	-0.24 (0.003)	0.02 (0.830)
Weight									0.38 (<0.001)	0.91 (<0.001)
Height										0.00 (0.958)

Stronger colors reveal the existence of significant associations ($p < 0.01$) between the variables.

Due to the heterogeneity of the behavior patterns of the clinical parameters evaluated in breast cancer, it is necessary to analyze the sample data in smaller groups. Thus, probably more specific correlations can be evidenced from the data.

Considering the data from the complete sample, Ward's method resulted in the formation of hierarchical groups by similarities, suggesting a marked reduction in similarity when 2 or 3 clusters are obtained, as shown in Fig. 1. The dendrogram graphically indicates the evolutionary history of the distance matrix; that is, the observations of each group are similar to each other, and each group is heterogeneous compared to the other, so three groups were the ideal choice. In addition to the dendrogram, the fusion test was performed, indicating that the first most significant distance between the groups is found in the third group. Thus, it was decided to keep 3 clusters, denoted by C_1, C_2, and C_3, respectively, even because, clinically, there was a better representation of information about the patients.

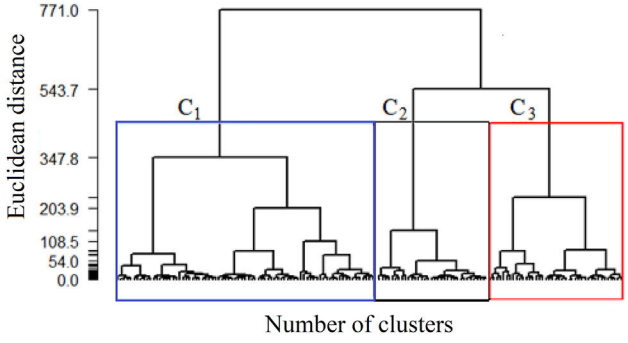


Fig. 1 Hierarchical formation of groups by similarity, using Euclidean distance.

From the descriptive means and test of means, Table 2, it appears that clusters C_1, C_2, and C_3 do not show significant differences in the variables intratumoral emboli, presence of lymph node invasion, molecular subtype, grade, TNM

Table 2 Descriptive means and test of the means of variables for each cluster.

	Full sample	C_1	C_2	C_3
Number of patients	155	79	35	41
Binary variables				
Intratumoral emboli	0.31	0.27	0.31	0.39
Lymph node invasion	0.41	0.33	0.51	0.49
Menopausal status	0.70	**0.99**^{a}	**0.26**^{c}	**0.51**^{b}
Categorical variables				
Molecular subtype	2.43	2.34	2.66	2.39
Grade	1.93	1.95	1.91	1.90
TNM staging	2.06	1.96	2.09	2.22
Quantitative variables				
Tumor size (cm)	3.18	3.11	3.51	3.03
Age (years)	56.64	**67.06**^{a}	**42.57**^{c}	**48.56**^{b}
Weight (kg)	72.54	**68.66**^{b}	**62.59**^{c}	**88.54**^{a}
Height (m)	1.60	1.59	1.61	1.62
BMI (kg/m^2)	28.25	**27.14**^{b}	**24.11**^{c}	**33.94**^{a}

Different letters indicate a significant difference between the means of a variable in the different clusters ($p < 0.01$). This table aims to verify, using a hypothesis test, whether there is a difference between groups depending on the evaluated covariate. In this way, the means of each group are compared, in which equal letters mean that statistically, there is no difference. Otherwise, different letters mean that there are differences between the groups. It is observed that this comparison occurs between groups two by two for each variable analyzed.

staging, tumor size, and height. On the other hand, all clusters show significant differences in the menopausal status, age, weight, and BMI variables, in bold in Table 2. Thus, in our mathematical modeling, the variable menopause, age, and BMI at diagnóstico were selected based on their contribution to characterizing the influence of each variable in breast cancer prognosis.

Discussion

In this study, we showed that cluster analysis is an interesting approach to categorizing breast cancer patients according to combined clinicopathological features. This analysis resulted in 3 utterly distinct clusters. C_1 was formed by 99% of menopausal women, older, with a mean age of 67 years, and overweight, with a mean BMI of 27.14 kg/m^2. C_2 included younger women with a mean of 42.57 years; most were not in menopause, had lymph node invasion, and had average weight. C_3 was composed of women of middle ages, obese, in menopause, and with lymph node invasion.

Information regarding the categorical and quantitative variables in each cluster is shown in Fig. 2. Note that Fig. 2 (a) shows the prevalence of tumors of the Luminal A and B subtypes in C_1 and C_3 and Luminal B in C_2. It means that considering the molecular subtype of breast cancer, C_2 has patients with the worst clinical prognosis compared to the others since Luminal B tumors are very aggressive.[28]

It was observed that with the division of clusters, it was possible to characterize the heterogeneity of behavior between the clinicopathological variables. Quantifying the intensity of the statistical dependence of the set of variables in each cluster will allow us to understand the influence that one variable exerts over another, making it possible to identify possible risk factors associated with the groups.

Thus, from the analysis of each cluster, where stronger colors reveal the existence of significant associations ($p < 0.01$) between the variables, as shown in Table 3, the data confirm some common characteristics, highlighting the strong correlation between the variables weight and BMI in all clusters, and the correlation between menopausal status and age at diagnóstico, in C_3. Similarly, correlations between intratumoral emboli, the presence of lymph node invasion, and TNM staging are present in almost all clusters. It is worth mentioning that, despite these characteristics having already been observed in the data of the complete sample, Table 1, it is now possible to analyze these correlations in the context of the particularities of each cluster. Next, the analysis of each cluster was carried out.

In C_1, menopausal, older, and overweight patients, significant correlations between the variables intratumoral emboli, lymph node invasion, and TNM staging are observed. This statement is justified by the analysis of the correlations obtained in Table 3, that is, the correlation of:

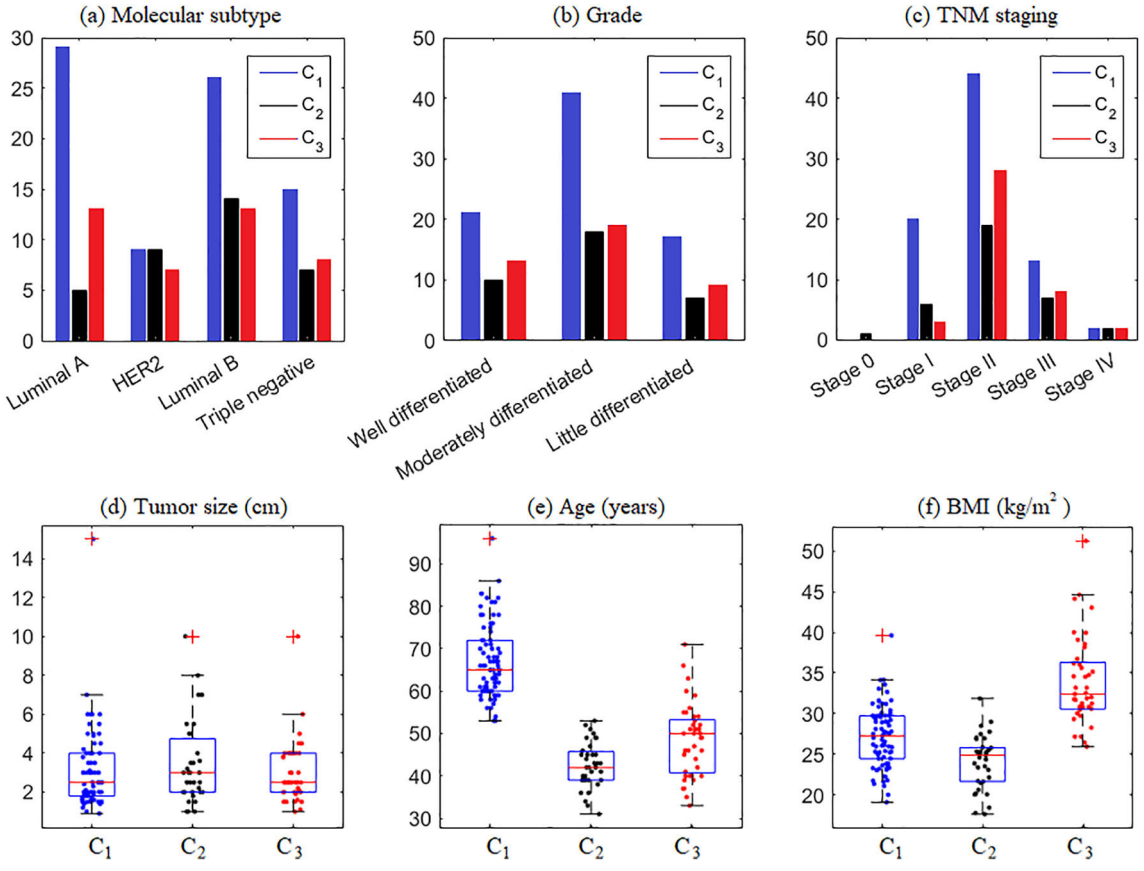


Fig. 2 Distribution of categorical and quantitative variables of clusters C_1 (blue), C_2 (black) and C_3 (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3 Spearman's correlations and *p*-value (in parentheses) of variables for clusters C_1, C_2 and C_3.

Variables	Lymph node invasion	Menopausal status	Molecular subtype	Histological grade	TNM staging	Tumor size	Age	Weight	Height	BMI
Characteristics of cluster C_1: 79 patients (99% menopause, mean age 67 years, overweight)										
Intratumoral emboli	0.43 (<0.001)	-0.19 (0.097)	-0.08 (0.506)	-0.04 (0.754)	0.35 (0.002)	0.13 (0.254)	0.12 (0.310)	-0.02 (0.852)	0.00 (0.991)	-0.06 (0.571)
Lymph node invasion		-0.16 (0.155)	0.12 (0.293)	0.21 (0.060)	0.59 (<0.001)	0.06 (0.600)	0.24 (0.036)	0.01 (0.938)	-0.21 (0.059)	0.06 (0.620)
Menopausal status			0.13 (0.252)	-0.01 (0.924)	-0.17 (0.131)	0.06 (0.616)	0.08 (0.486)	-0.09 (0.421)	-0.03 (0.777)	-0.09 (0.433)
Molecular subtype				0.24 (0.031)	0.22 (0.052)	0.19 (0.101)	0.05 (0.654)	-0.02 (0.844)	-0.10 (0.359)	0.05 (0.647)
Histological grade					0.30 (0.007)	0.28 (0.014)	-0.03 (0.774)	0.06 (0.627)	-0.04 (0.694)	0.07 (0.522)
TNM staging						0.38 (0.001)	0.23 (0.042)	0.00 (0.968)	-0.17 (0.130)	0.04 (0.749)
Tumor size							0.19 (0.100)	0.15 (0.191)	-0.02 (0.876)	0.19 (0.097)
Age								0.06 (0.614)	0.01 (0.953)	0.04 (0.735)
Weight									0.59 (<0.001)	0.88 (<0.001)
Height										0.19 (0.101)
Characteristics of cluster C_2: 35 patients (25% menopause, mean 42 years, eutrophic BMI)										
Intratumoral emboli	0.53 <0.001	0.31 (0.074)	0.09 (0.596)	0.00 (0.985)	0.31 (0.072)	0.09 (0.625)	0.13 (0.473)	-0.06 (0.728)	-0.10 (0.552)	-0.13 (0.442)
Lymph node invasion		0.18 (0.303)	-0.01 (0.933)	0.05 (0.764)	0.59 (<0.001)	-0.01 (0.948)	-0.10 (0.581)	-0.25 (0.153)	0.07 (0.685)	-0.25 (0.149)
Menopausal status			-0.18 (0.310)	-0.30 (0.079)	0.16 (0.369)	-0.13 (0.445)	0.19 (0.287)	0.14 (0.435)	0.09 (0.603)	0.04 (0.825)
Molecular subtype				0.18 (0.311)	-0.11 (0.527)	-0.08 (0.632)	0.14 (0.413)	-0.23 (0.177)	-0.36 (0.034)	0.04 (0.800)
Histological grade					-0.10 (0.562)	-0.15 (0.399)	-0.19 (0.271)	-0.08 (0.648)	-0.16 (0.351)	0.05 (0.759)
TNM staging						0.43 (0.010)	0.00 (0.987)	-0.24 (0.169)	0.20 (0.256)	-0.31 (0.071)
Tumor size							-0.16 (0.345)	0.37 (0.029)	0.16 (0.363)	0.22 (0.203)
Age								-0.08 (0.663)	-0.23 (0.193)	0.05 (0.775)
Weight									0.28 (0.106)	0.76 (<0.001)
Height										-0.27 (0.113)
Characteristics of cluster C_3: 41 patients (50% menopause, mean 49 years, obese)										
Intratumoral emboli	0.42 (0.006)	-0.12 (0.457)	0.07 (0.642)	-0.03 (0.865)	0.41 (0.007)	0.27 (0.086)	-0.12 (0.437)	-0.34 (0.032)	0.15 (0.345)	-0.34 (0.027)
Lymph node invasion		-0.02 (0.883)	0.05 (0.759)	0.19 (0.242)	0.60 (<0.001)	0.12 (0.444)	-0.01 (0.929)	-0.15 (0.347)	0.02 (0.918)	-0.16 (0.309)
Menopausal status			0.04 (0.821)	-0.06 (0.699)	0.13 (0.407)	-0.30 (0.055)	0.76 (<0.001)	0.13 (0.411)	-0.12 (0.453)	0.18 (0.273)
Molecular subtype				0.25 (0.112)	0.41 (0.007)	0.32 (0.038)	-0.16 (0.325)	-0.16 (0.331)	0.25 (0.111)	-0.31 (0.049)
Histological grade					0.24 (0.133)	0.18 (0.251)	-0.14 (0.378)	0.26 (0.098)	0.15 (0.337)	0.12 (0.458)
TNM staging						0.36 (0.019)	-0.07 (0.660)	-0.20 (0.211)	0.14 (0.377)	-0.28 (0.074)
Tumor size							-0.33 (0.037)	-0.21 (0.189)	0.39 (0.011)	-0.40 (0.009)
Age								0.18 (0.257)	-0.37 (0.019)	0.32 (0.042)
Weight									0.04 (0.803)	0.84 (<0.001)
Height										-0.46 (0.003)

Stronger colors reveal the existence of significant associations ($p < 0.01$) between the variables.

(a) intratumoral emboli with lymph node invasion and TNM staging,
(b) TNM staging with lymph node invasion, tumor grade, and size.

It is known that, clinically, the formation of intratumoral emboli occurs due to coagulation alterations induced by tumors, facilitating the process of spreading the disease. Furthermore, the larger the tumor size, the more advanced the TNM stage of the disease is.

In C_2, composed of patients aged between 31 and 52 years, most of them not menopausal and with an average BMI of 24.11 kg/m^2. There are significant correlations between intratumoral emboli and lymph node invasion, without association with the obesity variable. This statement is justified by the analysis of the correlations obtained in Table 3, that is, the correlation of:

(a) intratumoral emboli with lymph node invasion,
(b) TNM staging with lymph node invasion and tumor size.

Clinically, these correlations act in favor of the same biological event, which in this case would be favoring tumor spread. In addition, this association has a significant clinical meaning since this cluster is characterized by the incidence of the disease in young women, which gives them a risk of highly aggressive tumors.[29] Thus, these women are not in menopause at diagnóstico is another factor of worse prognosis because estrogen acts as fuel for breast cancer.[30]

In C_3, composed of patients considered young, obese, and with a prevalence of TNM staging in stages, II and III, presents a strong correlation between the menopausal status variables and age at diagnóstico, in addition to other correlations previously observed in the preceding clusters. This statement is justified by the analysis of the correlations obtained in Table 3, that is, the correlation of:

(a) intratumoral emboli with lymph node invasion and TNM staging,
(b) TNM staging with intratumoral emboli, lymph node invasion, and molecular subtype,
(c) menopausal status with age at diagnóstico.

Clinically, the strong correlation between the variables menopausal status and age at diagnóstico configures a worse disease prognosis for non-menopausal women. There is no correlation between obesity and variables associated with breast cancer in these data. On the other hand, in the literature, it is observed that obesity is a risk factor for breast cancer and is associated with the occurrence of highly aggressive tumors.[31,32]

Spatial distribution of clinicopathological variables

Table 4 presents the descriptive means of the variables considered in each cluster for the three regions of the 8th Health Regional of the State of Paraná, that are: Fronteira, Vale do Iguaçu, and Vale do Marrecas. Note that the means of the variables in each region that stand out the most are cluster C_2, in bold in Table 4. Although the Vale do Iguaçu region contains the smallest number of patients, those in C_2 of this region showed that the youngest patients are in a very advanced stage of the disease, TNM III, with the worst prognosis.

Regarding the high frequency of patients who developed the disease in the Vale do Marrecas region, this area is

Table 4 Means of the variables in each cluster for the regions of the 8th Health Regional of the State of Paraná.

	Fronteira			Vale do Iguaçu			Vale do Marrecas		
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
Number of patients	30	10	15	17	5	9	32	20	17
Binary variables									
Intratumoral emboli	0.30	0.40	**0.47**	0.18	**0.40**	0.33	0.28	0.25	**0.35**
Lymph node invasion	0.30	**0.60**	0.47	0.29	**0.80**	0.67	0.38	0.40	0.41
Menopausal status	0.97	0.10	0.33	1.00	0.40	0.67	1.00	0.30	0.59
Categorical variables									
Molecular subtype	2.27	2.70	2.27	2.29	2.40	**2.78**	2.44	2.70	2.29
Grade	1.87	**2.30**	1.93	1.76	1.60	**1.89**	**2.13**	1.80	1.88
TNM staging	1.93	2.00	2.27	1.76	**3.00**	2.22	2.09	1.90	2.18
Quantitative variables									
Tumor size (cm)	**3.22**	2.38	2.99	2.84	**3.80**	3.50	3.15	**4.00**	2.81
Age (years)	65.60	**43.00**	46.87	69.59	**42.40**	50.44	67.09	**42.40**	49.06
Weight (kg)	71.07	60.62	84.91	66.48	56.88	83.90	67.55	65.00	94.19
Height (m)	1.59	1.59	1.62	1.58	1.62	1.61	1.60	1.62	1.63
BMI (kg/m^2)	28.24	23.92	**32.76**	26.63	21.58	**32.58**	26.37	24.84	**35.70**

The means of the variables that stood out among the clusters in the respective regions are shown in bold. Higher mean frequencies of patients who presented the presence of intratumoral emboli and lymph node; higher means for categorical variables Molecular subtype, grad, and TNM staging, configuring worse prognosis. Largest tumor size averages. Lower and higher mean age and BMI, respectively.

reported as the highest pesticide trade in in Paraná state, and it includes all municipalities in the 8th Health Regional.[33] Therefore, knowledge about the spatial distribution of pesticide use can be used as a variable for the prognosis of breast cancer and give a possible interpretation of the data presented in Table 4. Thus, new analyzes are suggested to assess the impacts of extrinsic factors on breast cancer patients, especially environmental factors, health habits, and diet.

Conclusion

The cluster analysis allowed us to identify essential factors in breast cancer and their relationship with characteristics of worse prognosis, such as BMI, TNM staging, the presence of intratumoral emboli, and lymph nodal invasion. Due to the heterogeneity of the clinical parameters evaluated, the sample data was analyzed in smaller groups, which resulted in three hierarchical groups categorized by similarities based on cluster analysis. All groups showed significant differences in menopausal status, age and BMI; therefore, these variables were selected.

Through the statistical analysis, it was possible to determine the heterogeneity of the data, so the patients were separated into three clusters. This analysis identified that the group composed of older, postmenopausal and obese patients with intratumoral emboli, and lymph node invasion, configured characteristics of worse prognosis. As young, non-menopausal, and eutrophic patients who presented intratumoral emboli and lymph node invasion had characteristics associated with the development of tumors with poor clinical prognosis, regardless of obesity. Another group was formed by patients considered young, in menopause, and obese who presented a prevalence of TNM staging in stages II and III, reflecting the failures in the late search for health services for screening the disease at earlier stages. When analyzing the obtained clusters, each one of them had singular characteristics. Thus, with the division of the groups, it was possible to characterize the heterogeneity of characteristics of the clinicopathological variables, especially when considering that BMI is not a classical risk factor used to predict patients' prognosis in breast cancer. Quantifying the intensity of the statistical dependence of the set of variables in each cluster allowed us to understand the influence that one variable exerts over another, enabling the identification of possible risk factors associated with the groups.

Funding

The authors declare that they have not received funding for this study.

Ethical disclosures

All patients signed consent forms. The study was approved by the Institutional Ethical Committee. The authors have no ethical issues to declare.

Declaration of competing interest

The authors declare that they have no conflict to interest.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics. 2020. CA Cancer J Clin. 2020;70(1):7–30. https://doi.org/10.3322/caac.21590.
2. INCA. Instituto Nacional de Câncer José Alencar Gomes da Silva. https://www.inca.gov.br/institucional 2020.
3. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* Hoboken. 2018;68(6):394–424. https://doi.org/10.3322/caac.21492.
4. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer* New York. 2018;144(8):1941–53. https://doi.org/10.1002/ijc.31937.
5. Gajdos C, Tartter PI, Bleiweiss IJ, et al. Stage 0 to stage III breast cancer in young women. J Am Coll Surgeons. 2000;190(5):523–9. https://doi.org/10.1016/S1072-7515(00)00257-X.
6. Martel S, Poletto E, Ferreira AR, et al. Impact of body mass index on the clinical outcomes of patients with HER2-positive metastatic breast cancer. Breast. 2018;37:142–7. https://doi.org/10.1016/j.breast.2017.11.004.
7. Papa AM, Pirfo CBL, Murad AM, et al. Impact of obesity on prognosis of breast cancer. Rev Bras Oncologia Clin. 2013;9(31):25–30.
8. Azrad M, Blair CK, Rock CL, et al. Adult weight gain accelerates the onset of breast cancer. Breast Cancer Res Treat. 2019;176(3):649–56.
9. Cox CE, Dupont E, Whitehead GF, et al. Age and body mass index may increase the chance of failure in sentinel lymph node biopsy for women with breast cancer. The Breast J. 2002;8(2):88–91. https://doi.org/10.1046/j.1524-4741.2002.08203.x.
10. Maehle BO, Tretli S, Thorsen T. The associations of obesity lymph node status and prognosis in breast cancer patients: Dependence on estrogen and progesterone receptor status. APMIS. 2004;112:349–57. https://doi.org/10.1111/j.1600-0463.2004.apm1120605.x.
11. Sun H, Zou J, Chen L, Zu X, Wen G, Zhong J. Triple-negative breast cancer and its association with obesity (Review). Mol Clin Oncol. 2017;7(6):935–42. https://doi.org/10.3892/mco.2017.1429.
12. Borghesan DH, Agnolo CM, Gravena AA, et al. Risk factors for breast cancer in postmenopausal women in Brazil. Asian Pac J Cancer Prev. 2016;17(7):3587–93.
13. Gravena AAF, Romeiro Lopes TC, Demitto MO, et al. The obesity and the risk of breast cancer among pre and postmenopausal women. Asian Pac J Cancer Prev. 2018;19(9):2429–36. https://doi.org/10.22034/APJCP.2018.19.9.2429.
14. Jerônimo AFA, Weller M. Differential association of the lifestyle-related risk factors smoking and obesity with triple negative breast cancer in a Brazilian population. Asian Pac J Cancer Prev. 2017;18(6):1585–93. https://doi.org/10.22034/APJCP.2017.18.6.1585.
15. Kops NL, Bessel M, Caleffi M, et al. Body weight and breast cancer: Nested case-control study in southern Brazil. Clin Breast Cancer. 2018;18(5):797–803. https://doi.org/10.1016/j.clbc.2018.04.014.
16. Cormanique TF, Almeida LEDF, Rech CA, et al. Chronic psychological stress and its impact on the development of aggressive breast cancer. *Einstein* (São Paulo). 2015;13(3):352–6. https://doi.org/10.1590/S1679-45082015AO3344.
17. Chatfield C, Collins AJ. Introduction to Multivariate Analysis.US: Springer; 1980.
18. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis.Prentice Hall Upper Saddle River NJ: PEARSON; 2007.
19. Chen Y, Liu L, Zhou Q, Imam MU, et al. Body mass index had different effects on premenopausal and postmenopausal breast cancer risks: a dose-response meta-analysis with 3.318.796

subjects from 31 cohort studies. BMC Public Health. 2017;17. https://doi.org/10.1186/s12889-017-4953-9 936 pages.
20. Thomssen C, Balic M, Harbeck N, et al. St. Gallen/Vienna 2021: A brief summary of the consensus discussion on customizing therapies for women with early breast cancer. Breast Care. 2021(16):135–43. https://doi.org/10.1159/000516114.
21. Amin MB, Edge SB, Greene FL, et al. AJCC Cancer Staging Manual.8th ed. New York: Springer International Publishing; 2017;563–5854.
22. Villardón JLV. Introducción al análisis de clúster.Departamento de Estadística Universidad de Salamanca; 2007;1–22.
23. R Core Team. R: A language and environment for statistical computing.Vienna, Austria: R Foundation for Statistical Computing; 2020. R version 4.0.3 (2020-10-10) – "Bunny-Wunnies Freak Out": https://www.R-project.org/. Access in July 2021.
24. Hmisc: Harrell M, Frank E, Harrell Jr and with contributions from Dupont C and many others. *R package* version 4.4–0. https://CRAN.R-project.org/package=Hmisc 2020 Access in July 2021.
25. Agricolae: Statistical Procedures for Agricultural Research, Felipe de Mendiburu. *R package* version 1.3-3.https://CRAN.R-project.org/package=agricolae 2020 Access in July 2021.
26. Bolboacă SD, Jäntschi L, Pearson versus Spearman. Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. Leonardo J Sci. 2006;5(9):179–200.
27. Tsoi DT, Rowsell C, Mcgregor C, et al. Disseminated tumor embolism from breast cancer leading to multiorgan failure. J Clin Oncol. 2010;28(12):e180–3. https://doi.org/10.1200/JCO.2009.25.1009.
28. Hashmi AA, Aijaz S, Khan SM, et al. Prognostic parameters of luminal A and luminal B intrinsic breast cancer subtypes of Pakistani patients. World J Surg Oncol. 2018;16(1):1–6. https://doi.org/10.1186/s12957-017-1299-9.
29. Sundquist M, Thorstenson S, Brudin L, et al. Incidence and prognosis in early onset breast cancer. The Breast. 2002;11(1): 30–5. https://doi.org/10.1054/brst.2001.0358.
30. Rana A, Rangasamy V, Mishra R. How estrogen fuels breast cancer. Future Oncol. 2010;6(9):1369–71. https://doi.org/10.2217/fon.10.112.
31. Engin A. Obesity-associated Breast Cancer: Analysis of risk factors. Adv Exp Med Biol. 2017;960:571–606. https://doi.org/10.1007/978-3-319-48382-5_25.
32. Romeiro NML, dos Santos MCT, Panis C, et al. Cluster analysis evidences body mass index as an independent variable related to disease prognosis in breast cancer. Rev Bras Biom Lavras. 2021;39(4):536–55. https://doi.org/10.28951/rbb.v39i4.596.
33. Gaboardi SC, Zanetti L, Candiotto P, et al. Profile of pesticides use in the southwest of Paraná (2011-2016). Rev NERA. 2019;22 (46):13–40. https://doi.org/10.47946/rnera.v0i46.5566.