CLUSTER ANALYSIS IDENTIFIES VARIABLES RELATED TO PROGNOSIS OF BREAST CANCER DISEASE

Neyva Maria Lopes ROMEIRO[1]
Mara Caroline Torres dos SANTOS[2]
Carolina PANIS[3]
Tiago Viana Flor de SANTANA[2]
Paulo Laerte NATTI[1]
Daniel RECH[3]
Eliandro Rodrigues CIRILO[1]

- ABSTRACT: This work presents a cluster analysis approach aiming to determine distinct groups based on clinicopathological data from patients with breast cancer (BC). For this purpose, the clinical variables were considered: age at diagnosis, weight, height, lymph nodal invasion (LN), tumor-node-metastasis (TNM) staging and body mass index (BMI). Ward's hierarchical clustering algorithm was used to form specific groups. Based on this, BC patients were separated into four groups. The Kruskal-Wallis test was performed to assess the differences among the clusters. The intensity of the influence of variables on the prognosis of BC was also evaluated by calculating the Spearman's correlation. Positive correlations were obtained between weight and BMI, TNM and LN invasion in all analyzes. Negative correlations between BMI and height were obtained in some of the analyzes. Finally, a new correlation was obtained, based on this approach, between weight and TNM, demonstrating that the trophic-adipose status of BC patients can be directly related to disease staging.

- KEYWORDS: Cluster; hierarchical; Kruskal-Wallis test; Sperman's correlation; breast cancer.

[1] Universidade Estadual de Londrina - UEL, Departamento de Matemática, CEP: 86051990, Londrina, PR, Brasil. E-mail: *nromeiro@uel.br, plnatti@uel.br, ercirilo@uel.br*

[2] Universidade Estadual de Londrina - UEL, Departamento de Estatística, CEP: 86057970, Londrina, PR, Brasil. E-mail: *mara_torres14@hotmail.com, tiagodesantana@uel.br*

[3] Universidade Estadual do Oeste do Paraná - UNIOESTE, Laboratório de Tumor Biológico e Hospital do Câncer de Francisco Beltrão, CEP: 85605010, Francisco Beltrão, PR, Brasil. E-mail: *carolpanis@hotmail.com, dr.rech@gmail.com*

1 Introduction

The growing number of existing diseases, and its several clinicopathological features, have led us to the need to provide tools that help clinicians to solve how to proceed in specific situations. In many cases, it is necessary to join efforts from different areas of knowledge to generate results that can serve as a basis for some type of improvement in the diagnosis, treatment or remission of the disease (CARELS *et al.*, 2016; FRANCESCHINI *et al.*, 2013).

Advances in the last decade have improved disease diagnosis for several chronic diseases, especially cancer. Historically, there has been a significant growth in diagnosis for the most common types of tumors, and breast cancer is a good example. After the 90s, its increase stabilized, occurring reduction in the frequency of the disease in some countries (JATOI; MILLER, 2003; PINTO *et al.*, 1991; SANTOS, 2018). Although breast cancer mortality rate has decreased worldwide, many women are still dying, in spite of treatment advances. This fact indicates that the current information used by clinicians to stratify patients and make decisions regarding their treatment are not enough. Thus, it is important to develop new strategies that help to address the putative different outcomes of cancer.

It is known that strategies based on prevention and treatment are adapted according to individual characteristics of the patients. These features are obtained by using information from data banks containing, for example, medical history, lab results, demographic data and daily graphs. Aiming to process and efficiently extract this information, several approaches are being used.

From multivariate statistics it is possible to build and classify groups based on specific characteristics from each observation (CHATFIELD *et al.*, 1980; JOHNSON *et al.*, 2007; SCHMID *et al.*, 2007).

Therefore, data clustering can be a pivotal tool in Medicine, aiming to discover subgroups of patients that can have distinct diagnosis and treatment overcomes. Further investigation concerning specific subgroups altogether with clinical guidelines can help to design strategies that help clinicians to take care of their patients.

In this context, many researchers have been developing studies with the objective of interpreting data available on breast cancer, considering different clinical conditions, such as height, weight, age, body mass index (BMI), lymph node invasion and TNM staging (AZRAD *et al.*, 2019; CHEN *et al.*, 2017; COX *et al.*, 2002; GAJDOS *et al.*, 2000; KUROZUMI *et al.*, 2019; KYUWAN *et al.*, 2019; LAUBY-SECRETAN *et al.*, 2016; MAEHLE *et al.*, 2004; MARTEL *et al.*, 2018; NEWMAN *et al.*, 1986; PAPA *et al.*, 2013; SMETANA *et al.*, 2016; SUN *et al.*, 2017; TRETLI, 1989; XIA *et al.*, 2018), and other parameters (AL-AMMAR *et al.*, 2018; BURFORD *et al.*, 2017; CASTILLO-OTINIANO *et al.*, 2019; CHRISTIAN *et al.*, 2015; HORIGOME *et al.*, 2019; HUANG *et al.*, 2019; KRUGER *et al.*, 2018; KULLDORFF *et al.*, 1997; LAMINO *et al.*, 2011; NATAL *et al.*, 2019; NEVO *et al.*, 2016; PINEDA-HIGUITA *et al.*, 2019; RENDA *et al.*, 2019; ZAPATA *et al.*, 2010).

Excessive body fat is an emerging risk factor for breast cancer development (CHEN *et al.*, 2017). Young obese patients are often diagnosed in advanced stages of disease, demonstrating that the combination of both risk factors, overweight and young age at diagnosis, decrease the disease-free survival and overall survival of these women (GAJDOS *et al.*, 2000). Furthermore, the weight gain in adulthood seems to accelerate the development of breast cancer (AZRAD *et al.*, 2019). Obesity is also a factor that has been linked to poor prognosis in breast cancer patients. It has been demonstrated that high BMI is directly correlated with the failure to identify positive lymph nodes in biopsies (COX *et al.*, 2002).

Obese patients can have a high risk of lymph nodal metastasis (LAUBY-SECRETAN *et al.* 2016) than eutrophic patients. It is also known that the BMI did not show a statistically significant relationship with disease prognosis, when only the status of the hormone receptors were considered. However, if the status of the lymph nodes and the hormone receptor were considered together with BMI, this association was strong and reversed in the lymph node-positive group with ER-negative tumours. These findings point out that the conjoint analysis of clinicopathological data from breast cancer patients can completely change the perspective of their prognosis.

Therefore, statistics may be a powerful tool to be considered in patients' data analysis, to provide reliable associations among parameters that can not be understood in isolation. Considering that clinicians are not used to this type of analysis, interdisciplinary study in this field is essential.

Thus, here is presented an analysis of data correlations for clinicopathological variables used to evaluate breast cancer prognosis. Several works have looked for correlations between variables as age and body mass index, TNM and age, obesity and lymph node invasion, triple-negative and obesity, height and weight (COX *et al.*, 2002; GAJDOS *et al.*, 2000; MAEHLE *et al.*, 2004; SUN *et al.*, 2017; TRETLI, 1989), respectively.

However, it is necessary to analyze it in a larger set of variables, aiming a better understanding of their behavior in relation to disease prognosis.

This work presents an exploratory study of data treatment when considering the set of variables age, weight, height, BMI, lymph nodal invasion (LN) and tumor-node-metastasis (TNM) staging by using a design based on a cluster analysis tool. The intensity of the influence of each variable on the prognosis of cancer was evaluated by calculating the Spearman's correlation (BEST *et al.*, 1975; BEWICK *et al.*, 2004; HECKE, 2012; ORNSTEIN *et al.*, 2016; SCHMID *et al.*, 2007). For data statistical analysis was used R software (R CORE TEAM, 2020).

2 DESCRIPTION OF THE PROBLEM

2.1 Data from the patients

The data used in this study were obtained from the patients medical records. Initially, 361 women were considered, diagnosed with breast cancer between April

2015 and May 2018. However, due to missing information, it was decided to evaluate a smaller group of 129 patients. Confidentiality of the data was maintained according to the clinical research guidelines.

The study was approved by the Institutional Ethics Board under the number CAAE 35524814.4.0000.0107, and included patients diagnosed with breast cancer attended by the 8th Health Care Region of the State of Paraná at Francisco Beltrão Cancer Hospital, Paraná, Brazil, corresponding to a total of 27 municipalities. All patients signed consent forms and each protocol followed the principles for medical research involving human subjects described in the Declaration of Helsinki.

The measured variables were:

a) age (in years);

b) weight (in kg);

c) lymph nodal invasion (LN): This variable represents the existence, or non-existence, of metastasis in lymph nodes;

d) tumor-node-metastasis (TNM) staging: It represents the stage of the disease, established after the patient has been diagnosed. The international standard is used for classification, where T represents the existence of the tumor and its size; N refers to the existence of cancer in the lymph nodes and M describes the absence or existence of metastasis;

e) body mass index (BMI), is calculated by using the expression:

$$\text{BMI} = \frac{\text{Weight [kg]}}{\text{Height}^2\ [\text{m}^2]}. \tag{1}$$

All patients were categorized according to the classification of the World Health Organization (WHO), with low weight if BMI < 18.5 kg/m^2, normal weight if BMI is between 18.5 e 24.9 kg/m^2, overweight if BMI is between 25.0 e 29.9 kg/m^2, and obese if BMI > 30.0 kg/m^2 (WHO, 1995).

2.2 Statistical methods

Cluster analysis is a multivariate statistical tool used for the construction and classification of groups according to the characteristics of the variables, so that the groups are heterogeneous with each other, but the elements (patients) within each group have homogeneous characteristics (CHATFIELD *et al.* 1980; JOHNSON *et al.*, 2007; MADHULATHA, 2012; RODRIGUEZ *et al.*, 2019).

For the formation of groups, Ward's hierarchical agglomerative algorithm was considered (CHATFIELD *et al.*, 1980; JOHNSON *et al.*, 2007). In the first stage, each element (patient) is considered a cluster of unit size, totaling n groups. In the following stages, the clusters are combined until the end of the process, when a single cluster is obtained with all elements. At each step, Ward's algorithm combines two clusters that result in the lowest value of

$$SSR = \sum_{i=1}^{g_k} SS_i, \tag{2}$$

where g_k is the number of clusters in the step k,

$$SS_i = \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\cdot}\right)^T \left(X_{ij} - \bar{X}_{i\cdot}\right) \tag{3}$$

describes the sum of squares of the i-th cluster in step k, X_{ij} is the j-th element of the i-th cluster and $\bar{X}_{i\cdot}$ is the average of the i-th cluster.

Note that the number of desired clusters, described by g, represents a natural division of the elements, where $1 < g < n$. Thus, to determine the number of groups g, the behavior of the fusion level was studied in each step of Ward's algorithm, equation (2). The calculation of the fusion level is given by

$$d_{q,r} = \left[\frac{n_q n_r}{n_q + n_r}\right] \left(\bar{X}_{q\cdot} - \bar{X}_{r\cdot}\right)^T \left(\bar{X}_{q\cdot} - \bar{X}_{r\cdot}\right), \tag{4}$$

where $\bar{X}_{q\cdot}$ and $\bar{X}_{r\cdot}$ are averages and n_q and n_r are the sizes (number of elements) of the q-ths and r-ths clusters, respectively.

To compare the g groups formed, equation (4), according to clinicopathological variables, the Kruskal-Wallis non-parametric test (BEWICK *et al.*, 2004; HECKE, 2012) was adopted, whose only requirement is that the variables can be ordered. The Kruskal-Wallis technique tests the null hypothesis H_0, that the groups come from the same population, against the alternative hypothesis H_1, that the groups originate from different populations. Population are patients defined by the groups tested.

In order to perform the test, the complete sample (n elements) must be considered. For each variable, the values must be ordered and transformed into rank, assigning rank 1 for the lowest observed value, rank 2 for the second lowest value and so on up to $N = \sum_{i=1}^{g} n_i$, corresponding to the highest value observed in the complete sample. The Kruskal-Wallis test statistic is given by the expression:

$$H = \frac{12}{N(N+1)} \frac{\sum_{j=1}^{g} n_j \bar{R}_j^2 - 3(N+1)}{1 - \sum_{i=1}^{l} (t_i^3 - t_i)/(N^3 - N)}, \tag{5}$$

where n_j, $j = 1, 2, ..., g$, represents the number of elements from the j-th cluster, $\bar{R}_j$ is the average of the ranks in j-th cluster, l is the number of groups with tied ranks and t_i is the number of ties in the i-th cluster.

For $g > 3$, $n_j > 5$, and assuming true H_0, the H statistic has an approximate chi-square probability distribution, with $k - 1$ degrees of freedom.

In every hypothesis test there is always a probability of making a mistake, called the level of significance of the test (α), associated with the decision to reject H_0.

To decide whether to reject H_0 or not, it is necessary to compare α with p-value $= P(H > h)$, where h is an estimate for H, given in the equation (5). If p-value $> \alpha$, then H_0 is rejected, otherwise the null hypothesis must not be rejected. If the null hypothesis is rejected, then at least one cluster differs from the others, however the Kruskal-Wallis test does not identify which clusters are different.

To identify the different clusters, the difference test is used. The results of the difference test, in general, are presented in tables, whose values are followed by letters. Different letters indicate that, for each variable, there is a significant difference between the averages in each cluster. The test analyzes the difference between clusters two by two, verifying the validity of inequality

$$|\bar{R}_u - \bar{R}_v| \geq z_{\alpha/g(g-1)}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)}\,, \tag{6}$$

where the u and v indices identify the clusters and $z_{\alpha/k(k-1)}$ is the quantile of the standard Normal probability distribution, such that $P(Z \geq z_{\alpha/k(k-1)}) = \alpha/k(k-1)$.

Through the Kruskal-Wallis and differences tests, qualitative results were obtained, in which it assesses the differences between the clusters. To quantify the intensity of statistical dependence between the set of clinicopathological variables, Spearman's correlation is calculated. The Spearman's correlation coefficient is a modification of the Pearson coefficient, in which the observed values of each variable are replaced by ranks (BEST *et al.*, 1975; BOLBOACA *et al.*, 2006; RAMSEY, 1989).

The mathematical expression for the Spearman's coefficient is given by

$$r_s = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{7}$$

where x_i and y_i are ranks related to the i-th element, in relation to the variables X and Y, respectively, and $\bar{x}$ and $\bar{y}$ are the average ranks of X and Y.

The correlation coefficient, r_s, is limited between -1 and 1. Values close to 1 for r_s indicate a strong positive association between the variables X and Y, while values close to -1 represent a strong negative association between these variables. If r_s is identically equal to zero, or assumes values close to zero, it is said that there is no correlation between the variables, or that the correlation is weak.

However, an observed value of r_s may be a mistake due to the randomness of the sample, and therefore, it is necessary to perform a hypothesis test for Spearman's correlation. The test hypotheses are H_0, there is no association between X and Y, against H_1, there is an association between these variables.

The statistics of the Spearman's correlation test are given by

$$T = r_s \sqrt{\frac{N-2}{1-r_s^2}}, \tag{8}$$

where the T random variable has t-Student probability distribution with $N-2$ degrees of freedom. Thus, H_0 will be rejected if p-value $= 2P(T \geq t) < \alpha$ and it is concluded that the variables are correlated.

3 Results

For data statistical analysis, software R was used, R version 4.0.3 (R CORE TEAM, 2020). The clinicopathological variables age, weight, height, LN invasion, TNM staging and BMI of 129 patients were considered. The average age of the patients was close to 57 years. The average weight, when diagnosed, was close to 74 kg, but one of the patients had a weight of 120 kg. The LN invasion variable, which classify patients who have metastasis spread to the lymph nodes (LN positive) and those who do not (LN negative), presented an average of 35.66% of patients with positive lymph nodes. It was found in the patients that the TNM staging varied from IA to IV, but stages II and III were predominant. Table 1 shows some descriptive statistics to these variables.

Table 1 - Descriptive statistics on variables in the sample of 129 patients

	Age	Weight	Height	LN	TNM	BMI
Minimum	32.00	41.00	1.42	0.00	0.00	18.22
Average	56.91	73.87	1.61	0.36	2.09	28.48
Maximum	82.00	120.00	1.79	1.00	7.00	51.26

Considering information about patient variables, the hierarchical formation of the groups was obtained Ward's algorithm, applied by the equation (2). The graph of fusion level, Figure 1, presents jumps in the algorithm steps 125 and 126, suggesting accentuated reduction in the similarity when it is obtained 5 or 4 clusters, respectively, indicating that the algorithm needs to be finished in some of these steps.

To further validate the results shown in Figure 1, it were studied two scenarios including 5 and 4 clusters, but little difference was observed in the created groups, and were kept only four clusters.

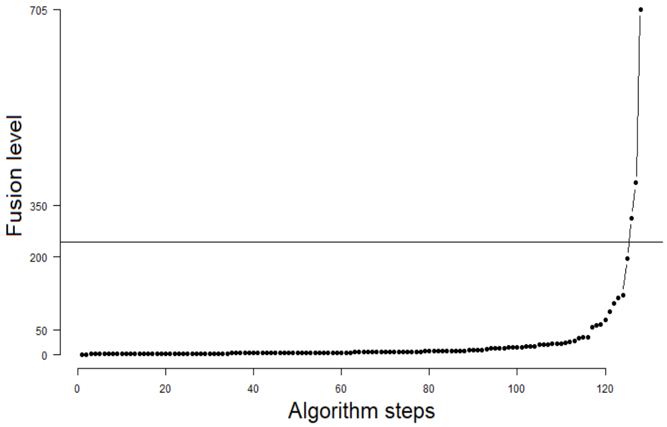


Figure 1 - Fusion level in each step of Ward's algorithm.

The dendrogram, Figure 2, identifies the four main clusters, represented by C_i, $i = 1, ..., 4$, considered from the fusion level calculation, equation (4).

The variable chosen for modeling, age and BMI at diagnosis, were selected based on their contribution to characterize the influence of each variable in breast cancer prognosis. Thus, the identified clusters are different concerning age and BMI variables, with patients with an average age below 50 years and above 70 years and groups with an average BMI below 24.9 kg/m^2 and greater than 30 kg/m^2. Information concerning each variable, in the clusters, are presented in Figure 3.

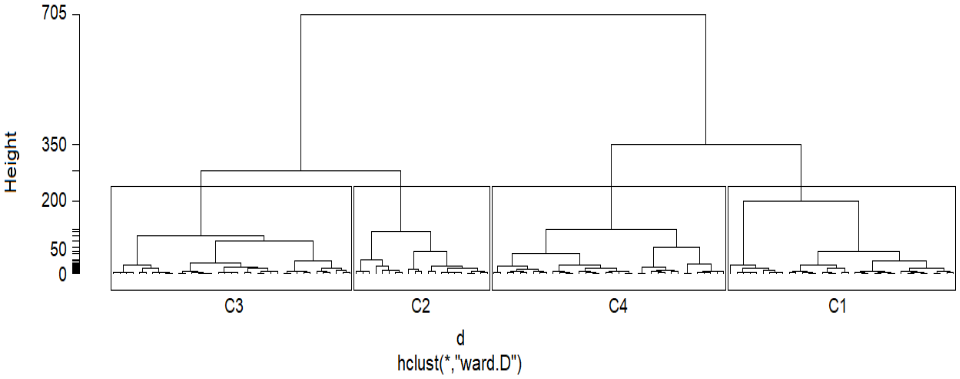
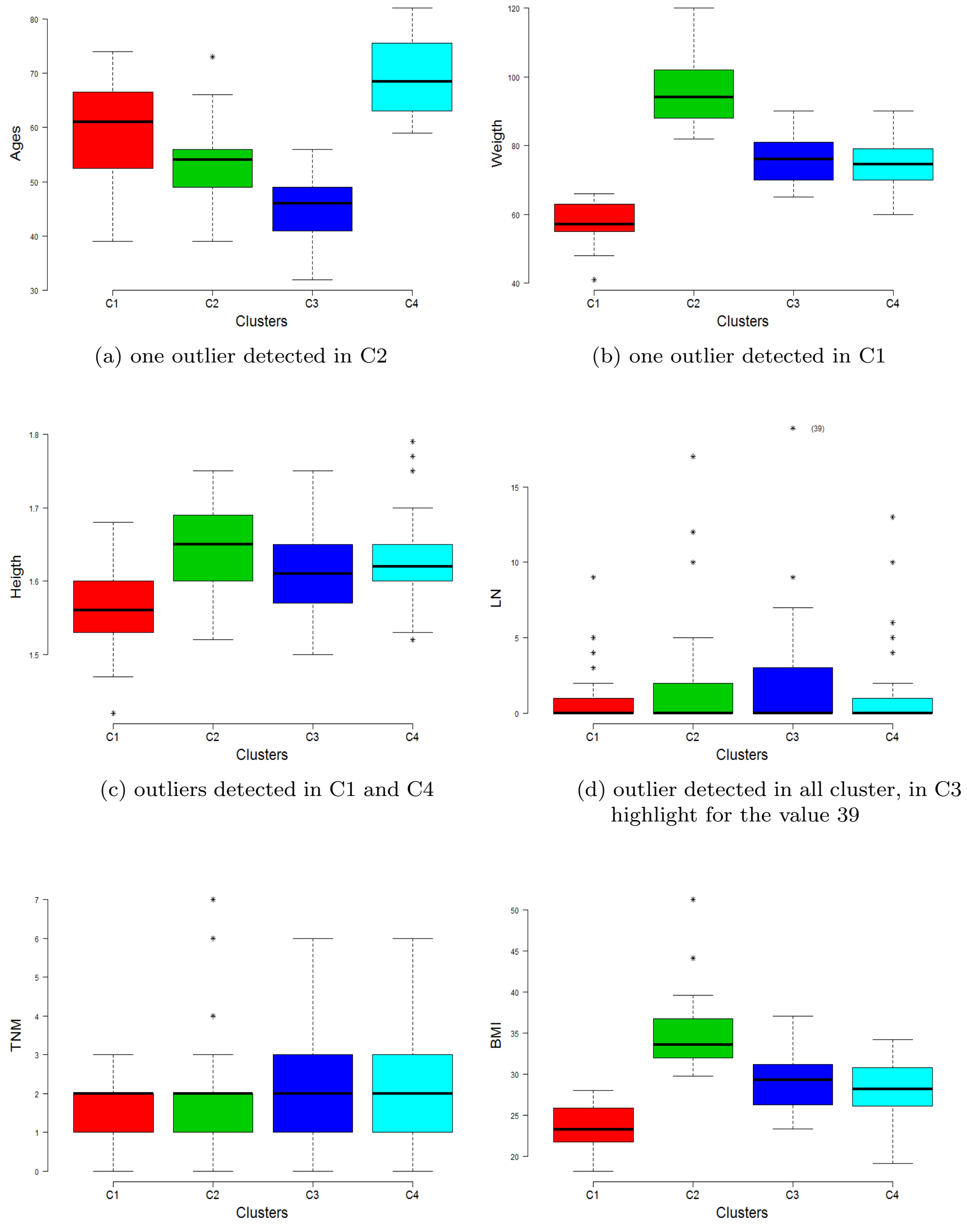


Figure 2 - Hierarchical formation of each group by similarity, obtained through Ward's algorithm.



(a) one outlier detected in C2

(b) one outlier detected in C1

(c) outliers detected in C1 and C4

(d) outlier detected in all cluster, in C3 highlight for the value 39

(e) outlier detected in C2

(f) outlier detected in C2

Figure 3 - Boxplot distribution of variables in the analyzed clusters.

It was observed that Figures 3(a)-3(f) present data regarding age at diagnosis and BMI, that were chosen for cluster modeling. It was also verified that some clusters presented discrepant values in relation to its variable, that is, values that were very small or very big in relation to the other values. These atypical observations alter significantly the averages and the variability of the groups that it belongs, and can even distort the conclusions obtained from statistical analysis. The variable that shown the biggest distort was LN invasion, Figure 3(d), and for this reason, it will be considered as 0 if the patient did not have LN invasion and 1 for LN positivity.

3.1 Kruskal-Wallis test

The information on the averages of the clusters' variables, obtained in the hierarchical formation of patient groups, are described in Table 2. Still, in order to assess the differences between the clusters, the Kruskal-Wallis test was performed, equations (5)-(6), where a, b, c and d letters indicate significant difference between the means of the variables in each cluster.

Table 2 - Information on the means of the variables in the clusters

	Total patients	**Age**	**Weight**	**Height**	**LN**	**TNM**	**BMI**
C1	35	58.66^b	57.78^c	1.56^b	0.34^a	1.60^a	23.70^c
C2	21	52.95^c	95.81^a	$1{,}65^a$	0.33^a	2.38^a	35.41^a
C3	37	45.16^d	75.91^b	1.61^a	0.38^a	2.05^a	29.34^b
C4	36	69.58^a	74.62^b	1.63^a	0.36^a	2.42^a	28.18^b

Different letters indicate a significance level of 5%, difference between the averages in each cluster.

In Table 2, by the Kruskal-Wallis test, the differences in the means of the variables age, weight and BMI, between the clusters, were considered statistically significant, while the differences of the other variables were not considered significant. Moreover, considering the complete sample and the means of the variables, it is observed that patients with lower and higher BMI are allocated in clusters C1 and C2, and younger and older, in clusters C3 and C4, respectively, as described below.

Cluster C1: It was the cluster with the lowest BMI, below 29 kg/m^2. With a mean age at diagnosis similar to that described to the Brazilian population (SANTOS, 2018), fro breast cancer incidence, approximately 57.78 years. These patients have normal average weight (eutrophic, BMI $<$ 25 kg/m^2). On average, patients in this cluster have small tumors ($<$ 2 cm), without spreading to lymph node chains and distant organs, which categorize them as stage IA in TNM staging.

Cluster C2: In this cluster, the mean age at diagnosis of disease incidence is around the fifth decade of life, but patients are obese, with an average BMI of 35.41 kg/m^2. It is known that obesity is an important risk factor, not only for the development, but also for the worsening of breast cancer prognosis due to activation of additional biological mechanisms that enhance tumor aggressiveness (SUN *et al.*, 2017). On average, these patients were classified as stage IIA in TNM staging, and may present locally advanced disease, with invasion of lymph node chains near the breast.

Cluster C3: In this cluster, patients have early age at diagnosis, with an average age of 45 years, presenting overweight with an average BMI of 29.35 kg/m^2. These patients are categorized in stage IIA-IIIC of TNM staging, and may have lymph node metastases. Note that the positive LN for this group is the highest. Verify that the measured data of this group are in accordance with the information described in the literature that confers disease poor prognosis, since the combination of young age at diagnosis and overweight results in important risk factors for breast cancer, leading to a reduction of disease-free survival and overall survival of these women (GAJDOS *et al.*, 2000; SANTOS, 2018; LAUBY-SECRETAN *et al.* 2016).

Cluster C4: This cluster has patients with older age at diagnosis (mean age of 69.58 years) and overweight (mean BMI of 28.18 kg/m^2). The TNM staging for these patients is the highest among the clusters, classified between IIA and IV.

The analysis of the results presented in Table 2 shown that overweight is a factor present in all analyzed clusters, with the exception of cluster C1, where patients have normal average weight, setting that obesity is an important risk factor for breast cancer patients, regardless of age. About average of lymph node invasion (LN positive), it was shown similar for all groups. On the other hand, despite the average lymph node invasion pattern being close between the groups, we observed that there are variations in relation to BMI and TNM staging. An alternative to quantify the intensity of statistical dependence between the set of variables age, weight, height, positive LN, TNM staging and BMI is the calculation of Spearman's correlation.

3.2 Spearman's correlation

The purpose of the Spearman's correlation analysis was to compare, for every two variables, the dependence between them. Thus, it was calculated the Spearman's rank coefficient for the complete sample and for each of the cluster, equations (7) and (8).

3.2.1 Complete sample

Considering the information from the complete sample, the results of Spearman's correlations are presented in Table 3.

Table 3 - Spearman's correlation coefficients of complete sample

	Age	**Weight**	**Height**	**LN**	**TNM**	**BMI**
Age	1.00	-0.16	-0.04	0.04	0.01	-0.16
Weight		1.00	**0.37†**	0.00	0.16	**0.90†**
Height			1.00	-0.10	**0.20‡**	-0.02
LN				1.00	**0.68†**	0.02
TNM					1.00	0.06
BMI						1.00

Significant correlation at the level of 1% (†) and 5% (‡).

According to Table 3, considering a significance level at 5%, it is concluded that the variables weight, height, positive LN, TNM and BMI present very different Spearmen's correlation coefficients.

The strongest correlation occurs between BMI and weight, $r_s = 0.9$ ($p < 0.0001$), thus, when considering the complete sample data, the BMI and weight values are strongly positive-correlated. Weight and height also result in a significant posi-tive correlation, $r_s = 0.37$ ($p < 0.0001$). It is known that both weight and height influence the BMI value, which was expected due to their dependence on the BMI calculation, equation (1).

The correlation between positive LN and TNM staging is $r = 0.68$ ($p < 0.0001$). Thus, it could mean that patients with positive LN tend to be in a more advanced stage of the disease. This finding is in accordance to TNM categorization. All correlations obtained have a significance level of 1%, Table 3. It is observed that the correlation between height and TNM, despite having a significance level below 5%, $r_s = 0.2$ ($p = 0.0265$), indicating that TNM is weakly influenced by the height of the patients.

Further, it can be seen in the results of the complete sample, Table 2, that patients with different ages, weight, height, and BMI have the same TMN staging on average, in spite of no correlations found among LN invasion and TNM, and the variables age at diagnosis, weight, height and BMI. For a better understanding of these results, the analysis of the Spearman's correlation coefficients, for each cluster, is presented below.

3.2.2 Cluster C1

Cluster C1 contains 35 patients and stands out for having the lowest averages of weight, height, TNM and BMI as can be seen in Table 2. Therefore, here we have the patients that are eutrophic and have been diagnosed in the fifth decade of life, which are parameters related to good disease prognosis.

Considering this information, the results of Spearman's correlations of cluster C1 are presented in Table 4.

Table 4 - Spearman's correlations of cluster C1

	Age	**Weight**	**Height**	**LN**	**TNM**	**BMI**
Age	1.00	0.19	-0.11	0.04	0.14	**0.29ł**
Weight		1.00	0.08	0.13	**0.36‡**	**0.74†**
Height			1.00	-0.05	0.13	**-0.54†**
LN				1.00	**0.66†**	0.05
TNM					1.00	0.14
BMI						1.00

Significant correlation at the level of 1% (†), 5% (‡) and 10% (ł).

Table 4 shows that Spearman's correlation for the variables weight, height, BMI, positive LN and TNM staging, resulted in some significant correlations. The BMI variable, ranging from 18.22 to 28.07 kg/m^2, has a significant correlation with the variables weight and height, which vary between 41 kg to 66 kg and 1.42 m to 1.68 m, respectively. This correlation is expected, since weight and height are inside the BMI calculation formula. The positive LN and TMN staging variables also present significant correlation, with TNM between IA e IIA stages, but on average, the stage is IA. This is another expected correlation, because TNM categorization consider LN invasion as a parameter (N category) in association with tumor size and metastasis.

The weight and BMI variables show $r_s = 0.74$ ($p < 0.0001$), a positive correlation, which also occurs in the complete sample. On the other hand, height and BMI variables have $r_s = -\,0.54$ ($p = 0.0009$), a negative correlation, note that this correlation was not observed in the complete sample. This correlation describes that taller patients tend to have a lower BMI. Note that these correlations have a significance level of 1%, Table 4.

The significant correlation, level of 5%, in this cluster occurs between the variables TNM staging and weight, $r_s = 0.36$ ($p = 0.00336$). This correlation, despite being a weak correlation, shows that, for increasing values of the patient's weight, there are higher values for the stage of cancer. This finding is extremely important, since weight is not considered when performing TNM staging of breast cancer patients. Note that this correlation was observed in the complete sample, Table 3, but with a correlation considered insignificant, whereas in this cluster, although weak, it is considered acceptable.

Another significant correlation occurred between TNM stating and positive LN with $r_s = 0.66$ ($p < 0.0001$), corroborating that patients with positive LN tend to be in a more advanced stages of the disease.

In conclusion, the correlations obtained in this cluster, cluster whose patients have an average age of 58 years and are categorized as normal weight, identify that the increase in body weight is associated with the presence of more advanced disease. It appears that the increase in TNM staging (positively correlated with the LN variable) suggests spread of the cancer, from the primary site to other sites such as lymph nodes and distant organs. The analysis of this cluster strengthens the idea that excessive body weight, could be a factor related to the progression of breast cancer for higher stages (TRETLI, 1989).

3.2.3 Cluster C2

Cluster C2 has the lowest number of patients ($n = 21$). This cluster stands out because it has the highest BMI average, and the weight of the patients differed significantly from the other clusters, ranging from 82 kg to 120 kg, with an average close to 96 kg (Table 2). This BMI characterizes these patients as obese, but they are in the mean age at diagnosis for breast cancer. In addition, this cluster has patients who are in advanced stages of the disease, between IA e IV. Considering this information, the results of Spearman's correlations of cluster C2 are presented in Table 5.

Table 5 - Spearman's correlations of cluster C2

	Age	**Weight**	**Height**	**LN**	**TNM**	**BMI**
Age	1.00	-0.28	-0.33	0.16	0.00	-0.15
Weight		1.00	0.14	0.08	0.07	**0.83†**
Height			1.00	-0.03	0.28	-0.36
LN				1.00	**0.74†**	0.12
TNM					1.00	-0.10
BMI						1.00

Significant correlation at the level of 1% (†).

It can be seen, Table 5, that this cluster results in two significant correlations, with a significance level of 1%. A correlation between BMI and weight, $r_s = 0.0083$ ($p < 0.0001$), and a correlation between positive LN and TNM staging, $r_s = 0.74$ ($p = 0.0001$).

Here, the cluster analysis did not identify any significant correlation concerning LN invasion and TNM with the other parameters (age at diagnosis, weight, height and BMI). A possible explanation to the lack of correlation could be the small number of patients in this cluster, and the wide variability of TNM data (Ia to IV), as shown in Figure 3(e).

3.2.4 Cluster C3

Cluster C3 contains the largest number of patients ($n = 37$). This cluster is characterized by patients with the lowest average age at diagnosis, ranging between 32 years and 56 years, high mean BMI, with a value of 29.34 kg/m^2, and the highest average of positive LN, with a value of 0.38. Therefore, this cluster is formed by patients with early age at diagnosis, and also with overweight.

Both parameters are determinant of poor prognosis in breast cancer. Considering this information, Table 6 shows the respective Spearman's correlations of the cluster.

Table 6 - Spearman's correlations of cluster C3

	Age	**Weight**	**Height**	**LN**	**TNM**	**BMI**
Age	1.00	-0.08	-0.09	0.09	0.00	0.04
Weight		1.00	-0.10	-0.09	0.15	**0.77†**
Height			1.00	-0.08	0.17	**-0.69†**
LN				1.00	**0.72†**	0.03
TNM					1.00	0.03
BMI						1.00

Significant correlation at the level of 1% (†).

It is observed, Table 6, that three correlations stand out for presenting statistical significance, with a significance level of 1%. Again, there is a significant correlation between weight and BMI, $r_s = 0.77$ ($p < 0.0001$), between height and BMI, $r_s = -\ 0.69$ ($p < 0.0001$), and between TNM and LN, $r_s = 0.72$, as already observed in cluster C1.

No significant correlations were identified regarding LN invasion and TNM with other parameters (age at diagnosis, weight, height and BMI), Table 1.

3.2.5 Cluster C4

Cluster C4, contains 36 patients, characterized by older age at diagnosis, with an average of 69.58 years. In this cluster, the average BMI does not differ significantly from the average BMI of cluster C3, both presenting overweight patients, although C3 has younger patients, with an average age of 45.16 years. Considering this information, Table 7 describes the Spearman's correlations of cluster C4.

Table 7 - Spearman's correlations of cluster C4

	Age	**Weight**	**Height**	**LN**	**TNM**	**BMI**
Age	1.00	-0.15	0.18	0.12	-0.07	-0.27
Weight		1.00	0.06	-0.04	-0.08	**0.81†**
Height			1.00	-0.21	0.10	**-0.47†**
LN				1.00	**0.68†**	0.00
TNM					1	-0.16
BMI						1.00

Significant correlation at the level of 1% (†).

It was observed that three correlations stand out for presenting statistical significance Table 7, with a significance level of 1%.

Again, there is a significant correlation between weight and BMI, $r_s = 0.81$ ($p < 0.0001$), between height and BMI, $r_s = -0.47$ ($p = 0.4$), and between TNM and LN, $r_s = 0.68$ ($p < 0.0001$). Thus, considering that the difference between C3 and C4 is the age at diagnosis, it can be concluded that overweight is an important risk factor for breast cancer patients, independent on their age.

4 Conclusion

In this article, information on six clinicopathological variables of 129 women diagnosed with breast cancer was analyzed: age, height, weight, lymphnodal invasion, TNM staging and BMI.

For data analysis, multivariate statistical methods were used, such as construction of dendrogram and cluster analysis. These clusters were studied individually, and compared among themselves and with the data of the complete sample, using the Kruskal-Wallis test and the calculation of the Spearman's correlation coefficient.

It was found that the Kruskal-Wallis test showed significant differences between the clusters in the variables age, weight and BMI. The calculations of Spearman's correlations showed significant positive correlations between weight and BMI, and between positive LN and TNM staging. There were also significant negative correlations between BMI and height.

The results obtained reinforces the importance of bring together mathematical tools and medical information with the aim of identify new ways to categorize patients' prognosis regarding classical standards used for a long time in clinical practice.

The main finding of this study is that cluster C1 presented a correlation that is still little discussed in the literature, showing that patient's weight may be directly related to TNM staging, and, consequently, to the spread of the disease.

Further, the comparison between C3 and C4 shown that overweight is an independent factor observed in breast cancer patients, in spite of their age at diagnosis. Excessive body fat is well established as a fuel for metastasis in cancer, and the result obtained here strengthens high BMI as a factor that may affect breast cancer staging.

ROMEIRO, N. M. L.; SANTOS, M. C. T.; PANIS, C.; SANTANA, T. V. F.; NATTI, P. L.; RECH, D.; CIRILO, E. R. Análise de cluster identifica variáveis relacionadas ao prognóstico da doença de câncer de mama. *Rev. Bras. Biom.,* Lavras, v.39, n.4, p.536-555, 2021.

▪ *RESUMO: Este trabalho apresenta uma abordagem de análise de cluster com o objetivo de determinar grupos distintos com base em dados clínico-patológicos de pacientes com câncer de mama (CM). Para tanto, foram consideradas as variáveis clínicas: idade ao diagnóstico, peso, altura, invasão linfonodal (LN), estadiamento tumor-nódulo-metástase (TNM) e índice de massa corporal (IMC). O algoritmo de agrupamento hierárquico de Ward foi usado para formar grupos específicos. Com base nos resultados, as pacientes com CM foram separadas em quatro grupos. O teste de Kruskal-Wallis foi realizado para avaliar as diferenças entre os clusters. A intensidade da influência das variáveis no prognóstico do CM também foi avaliada pelo cálculo da correlação de Spearman. Correlações positivas foram obtidas entre peso e IMC, invasão TNM e LN em todas as análises. Correlações negativas entre IMC e altura foram obtidas em algumas das análises. Por fim, uma nova correlação foi obtida, entre o peso e o TNM, demonstrando que o estado trófico-adiposo das pacientes com CM pode estar diretamente relacionado ao estadiamento da doença.*

▪ *PALAVRAS-CHAVE: Cluster; hierárquico; teste de Kruskal-Wallis; correlação de Sperman; câncer de mama.*

References

AL-AMMAR, Y. AL-MANSOUR, B.; AL-RASHOOD, O.; TUNIO, M. A.; ISLAM, T. AL-ASIRI, M.; AL-QAHTANI, K. H. Impact of body mass index on survival outcome in patients with differentiated thyroid cancer. *Brazilian journal of otorhinolaryngology*, v.84, n. 2, p.220-226, 2018.

AZRAD, M.; BLAIR, C. K.; ROCK, C. L.; SEDJO, R. L.; WOLIN, K. Y.; DEMARK-WAHNEFRIED, W. Adult weight gain accelerates the onset of breast cancer. *Breast cancer research and treatment*, v.176, n.3, p.649-656, 2019.

BEST, D. J. ROBERTS, D. E. Algorithm AS 89: the upper tail probabilities of Spearman's rho. *Journal of the Royal Statistical Society*. Series C (Applied Statistics), v.24, n.3, p.377-379, 1975.

BEWICK, V.; CHEEK, L.; BALL, J. Statistics review 10: further nonparametric methods. *Critical care*, v.8, n.3, p.196, 2004.

BOLBOACA, S. D.; JANTSCHI, L. Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, v.5, n.9, p.179-200, 2006.

BURFORD, B.; GAZINSKA, P.; MERA, A.; NOOR, A.M.; MARRA, P.; CHERYLGILLETT, A.; GRIGORIADIS, A.; PINDER, S.; TUTT A.; DE RINALDIS, E. Splicing imbalances in basal-like breast cancer underpin perturbation of cell surface and oncogenic pathways and are associated with patients' survival. *Scientific reports*, v.7, n.1, p.1-14, 2017.

CARELS, N.; SPINASSE L. B.; TILLI, T. M.; TUSZYNSKI, J. A. Toward precision medicine of breast cancer. *Theoretical Biology and Medical Modelling*, v.13, n.1, p.1-46, 2016.

CASTILLO-OTINIANO, C. C. YAN-QUIROZ, E. F. Blood hypertension and diabetes mellitus as risk factors for breast cancer, 2019, *Revista Del Cuerpo Médico Del HNAAA*, v.12, n.1, p.35-39.

CHATFIELD, C.; COLLINS, A. Introduction to multivariate analysis, Springer, US, 1980.

CHEN, Y.; LIU, L.; ZHOU, Q.; IMAM, M.U.; CAI, J.; WANG, Y.; QI, M.; SUN, P.; PING, Z.; FU, X. 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*, v.17, n.1, p.1-11, 2017.

CHRISTIAN, N.J.; HA, I.D.; JEONG, J.-H. Hierarchical likelihood inference on clustered competing risks data, *Statistics in Medicine*, v.35, n.2, p.251-267, 2015.

COX, C. E. *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 journal*, v.8, n.2, p.88-91, 2002.

FRANCESCHINI, J.; JARDIM, J. R.; FERNANDES, A. L. G.; JAMNIK, S. SANTORO, I.L. Relationship between the magnitude of symptoms and the quality of life: a cluster analysis of lung cancer patients in Brazil. *Jornal brasileiro de pneumologia*, v.39, n.1, p.23-31, 2013.

GAJDOS, C.; TARTTER, P. I.; BLEIWEISS, I. J.; BODIAN, C.; BROWE, S. T. Stage 0 to stage III breast cancer in young women. *Journal of the American College of Surgeons*, v.190, n.5, p.523-529, 2000.

HECKE, T. V. Power study of anova versus Kruskal-Wallis test. *Journal of Statistics and Management Systems*, v.15, n. 2-3, p.241-247, 2012.

HORIGOME, A.; OKUBO, R.; HAMAZAKI, K.; KINOSHITA, T.; KATSUMATA, N.; UEZONO, Y.; XIAO, J. Z.; MATSUOKA, Y. J. Association between blood omega-3 polyunsaturated fatty acids and the gut microbiota among breast cancer survivors. *Beneficial Microbes*, v.10, n.7, p.751-758, 2019.

HUANG, R.; XIANG, L.; HA, I. D. Frailty proportional mean residual life regression for clustered survival data: A hierarchical quasi-likelihood method. *Statistics in medicine*, v.38, n.24, p.4854-4870, 2019.

JATOI, I.; MILLER, A. B. Why is breast-cancer mortality declining?. *The lancet oncology*, v.4, n.4, p.251-254, 2003.

JOHNSON, R. A; WICHERN, D. W. Applied Multivariate Statistical Analysis, PEARSON, Prentice hall Upper Saddle River, NJ, 2007.

KRUGER, D. T.; BEELEN, K.J.; OPDAM, M.; ANDERS, J. S.; van der NOORT, V.; BOVEN, E.; LINN, S. C. Hierarchical clustering of activated proteins in the PI3K and MAPK pathways in ER-positive, HER2-negative breast cancer with potential therapeutic consequences. *British journal of cancer*, v.119, n.7, p.832-839, 2018.

KULLDORFF, M.; FEUER, E. J.; MILLER, B. A.; FREEDMAN, L. S. Breast cancer clusters in the northeast United States: a geographic analysis. *American journal of epidemiology*, v.146, n.2, p.161-170, 1997.

KUROZUMI, S.; INOUE, K.; MATSUMOTO, H.; FUJII, T.; HORIGUCHI, J.; OYAMA, T.; KUROSUMI, M.; SHIRABE, K. Clinicopathological values of PD-L1 expression in HER2-positive breast cancer. *Scientific reports*, v.9, n.1, p.1-8, 2019.

KYUWAN LEE, K.; KRUPER, L.; DIELI-CONWRIGHT, C. M.; MORTIMER, J. E. The impact of obesity on breast cancer diagnosis and treatment. *Current oncology reports*, v.21, n.5, p.41, 2019.

LAMINO, D. A.; MOTA, D. D. C. F.; PIMENTA, C. A. M.; Prevalence and comorbidity of pain and fatigue in women with breast cancer. *Revista da Escola de Enfermagem da USP*, v.45, n.2, p.508-514, 2011.

LAUBY-SECRETAN, B.; SCOCCIANTI, C.; LOOMIS, D. GROSSE, Y.; BIANCHINI, F.; STRAIF, K. Body fatness and cancer-viewpoint of the IARC Working Group.*New England Journal of Medicine*, v.375, n.8, p.794-798, 2016.

MADHULATHA, T.S. An Overview on Clustering Methods, *IOSR Journal of Engineering*, 2, 4, 719-725, 2012.

MAEHLE, B. O.; TRETLI, S.; THORSEN, T. The associations of obesity, lymph node status and prognosis in breast cancer patients: dependence on estrogen and progesterone receptor status. *Journal of Pathology, Microbiology and Immunology*, Apmis, v. 112, n.6, p.349-357, 2004.

MARTEL, S. *et al.* Impact of body mass index on the clinical outcomes of patients with HER2-positive metastatic breast cancer. The Breast, v.37, p.142-147, 2018.

NATAL, R. A. Exploring collagen parameters in pure special types of invasive breast cancer. *Scientific reports*, v.9, n. 1, p.1-11, 2019.

NEVO, D.; ZUCKER, D. M.; TAMIMI, R. M.; WANG, M. Accounting for measurement error in biomarker data and misclassification of subtypes in the analysis of tumor data. *Statistics in medicine*, v.35, n.30, p.5686-5700, 2016.

NEWMAN, S. C.; MILLER, A. B.; HOWE, G. R. A study of the effect of weight and dietary fat on breast cancer survival time. *American journal of epidemiology*, v.123, n.5, p.767-774, 1986.

ORNSTEIN, P.; LYHAGEN, J. Asymptotic properties of Spearman's rank correlation for variables with finite support. *PloS one*, v.11, n.1, p.1-7, 2016.

PAPA, A. M.; PIRFO, C. B. L.; MURAD, A. M.; RIBEIRO, G. M. Q.; FAGUNDES, T. C. Impact of obesity on prognosis of breast cancer, *Revista Brasileira de Oncologia Clínica*, v.9, n.31, p.25-30, 2013.

PINEDA-HIGUITA, S. E.; ANDRADE-MOSQUERA, S. M.; MONTOYA-JARAMILLO, Y. M. Factors Associated with Quality of Life in Women with Breast Cancer. Medellin 2013. *Revista Gerencia y Políticas de Salud*, v.16, n.32, p.85-95, 2017.

PINTO, F. G.; CURI, P.R. Mortality from neoplasms in Brazil (1980/1983/1985): grouping by State, behaviors and tendencies. *Revista de Saúde Pública*, v.25, n.4, p.276-281, 1991.

R CORE TEAM R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, 2020.

RAMSEY, P.H. Critical values for Spearman's rank order correlation.*Journal of Educational Statistics*, v.14, n.3, p. 245-253, 1989.

RENDA, I.; BIANCHI, S.; VEZZOSI, V.; NORI, J.; VANZI, E.; TAVELLA, K.; SUSINI, T. Expression of FGD3 gene as prognostic factor in young breast cancer patients. *Scientific reports*, v.9, n.1, p.1-8, 2019.

RODRIGUEZ, M. Z. *et al.* Clustering algorithms: A comparative approach. *PloS one*, v.14, n.1, p.1-34, 2019.

SANTOS, M.O. Estimate/2018 - Cancer Incidence in Brazil, Estimate/2018 - Cancer Incidence in Brazil, *Revista Brasileira de Cancerologia*, V.64, 1, p.119-120, 2018.

SCHMID, F.; SCHMIDT, R. Multivariate extensions of Spearman's rho and related statistics. *Statistics & probability letters*, v.77, n.4, p.407-416, 2007.

SMETANA JR.; K.; LACINA, L.; SZABO, P.; DVORANKOVA, B.; BROŽ, P.ŠEDO, A. Ageing as an important risk factor for cancer. *Anticancer research*, v.36, n.10, p.5009-5017, 2016.

SUN, H.; ZOU, J.; CHEN, L.; ZU, X.; WEN, G.; ZHONG, J. Triple-negative breast cancer and its association with obesity (Review). *Molecular and clinical oncology*, v.7, n.6, p.935-942, 2017.

TRETLI, S. Height and weight in relation to breast cancer morbidity and mortality. A prospective study of 570,000 women in Norway. *International Journal of Cancer*, v.44, n.1, p.23-30, 1989.

WHO, *Physical Status*: The Use and Interpretation of Anthropometry, Report of a WHO *Expert Committee*, Tech. Rep.; Series No. 854. Geneva: WHO. 1995.

XIA, J.; TANG, Z.; DENG, Q.; WANG, J.; YU, J. Being slightly overweight is associated with a better quality of life in breast cancer survivors. *Scientific Reports*, v.8, n.1, p.1-8, 2018.

ZAPATA, C. S.; ROMERO, H. G. Calidad de vida y factores asociados en mujeres con cáncer de mamaen Antioquia, Colombia, 2010. *Revista Panamericana de Salud Pública*, v.28, p.9 - 18.

Received on 18.10.2020.

Approved after revised on 03.05.2021.