Applied Medical Informatics
Original Research Vol. 46, No. 3 /2024, pp: 83-90

Exploring the Significance of Cluster Analysis on Time-Series Measurement of Plasma Cancer Antigen 15-3 in a Patient with Metastatic Breast Cancer

Alexandros CLOUVAS *

Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Egnatia Street, GR-54124 Thessaloniki, Greece.
E-mail : clouvas@ece.auth.gr

*Author to whom correspondence should be addressed.

Received: 25 August 2024/Accepted: 19 September 2024/ Published online: 29 September 2024

Abstract
Cancer Antigen 15-3 (CA 15-3) is a glycoprotein often linked to breast cancer. Elevated levels of CA 15-3, above the normal reference range of 30 U/mL (units per milliliter), are frequently found in the blood of patients with metastatic breast cancer, where the cancer has spread to other parts of the body, such as the bones. This study examines the importance of cluster analysis in evaluating time-series measurements of plasma CA 15-3 in a male patient with metastatic breast cancer that has spread to the trochanteric region of the left leg. Clustering is a statistical method used to organize data based on similarity, though it may not directly reflect underlying physical properties. The trend of CA 15-3 time-series measurement presented here is familiar to oncologists. The novelty of this study lies in evaluating the significance of applying cluster analysis to the specific time-series data. The results indicate that this approach is indeed meaningful. Notably, two distinct clusters were identified within the data, as anticipated. The first cluster corresponds to the period before the recurrence of illness (metastatic breast cancer), while the second cluster reflects the advanced (metastatic) stage of the disease. The boundary between these clusters provides valuable insights into the onset of the metastatic stage. To our knowledge, this is the first study to apply cluster analysis to CA 15-3 time-series data. The results are promising. Its potential use in identifying the onset of the metastatic stage merits further examination.

Keywords: Cluster analysis; K-means; Time-series; Cancer Antigen CA 15-3 (CA 15-3); Male breast cancer

Introduction

Cancer Antigen 15-3 (CA 15-3) is a glycoprotein found in benign and malignant breast disease, as well as in breast cancer with liver or bone metastases. CA 15-3 was initially identified as a potential marker for breast cancer in the 1980s [1]. Since its discovery, extensive research has been conducted, leading to its widespread utilization in clinical practice as a tumor marker for breast cancer (De Cock et al. and references there in) [2]. However, current guidelines [3-5], discourage the serial measurement of CA 15-3 in the follow-up of patients with early breast cancer due to insufficient evidence demonstrating a survival benefit.

Despite the guidelines [3-5], many oncologists routinely conduct serial assessments of blood-based tests for CA 15-3 as part of the standard follow-up for asymptomatic patients with early breast cancer. De Cock et al. [2] revealed that systematic utilization of CA 15-3 in the follow-up of such patients resulted in the diagnosis of metastatic disease in 37% of cases due to an increase in CA 15-3 levels. It should be noted that monitoring CA 15-

3 levels can be particularly valuable for tracking disease progression and assessing the response to treatment in patients with metastatic breast cancer [6].

In the present study, time-series measurements of plasma CA 15-3 in a male patient with metastatic breast cancer were conducted from April 14, 2020 (13 days prior to surgery for the primary breast tumor), until July 1, 2024. Cluster analysis was applied to CA 15-3 time-series data for the first time in this study. Clustering is a statistical method used to group data based on similarity, though it may not directly reflect underlying physical properties. Therefore, it is crucial to carefully interpret the clusters within the specific context of the data and problem domain to determine if they have meaningful implications. It is worth noting that the time-series trend of CA 15-3 measurements presented here is familiar to oncologists. The aim of this study is to assess whether the clustering observed in the CA 15-3 time-series data holds significant meaning.

Materials and Methods

Medical History of the Patient

On April 27, 2020, the author, who is also the patient (a 69-year-old male in 2024), underwent a mastectomy (right breast) due to a cancerous tumor (Grade 2, stage pT1cN0 according to TNM staging; PR and ER 100%, Ki-67: 30%, HER2 negative). Subsequently, he was recommended a treatment regimen comprising Tamoxifen (Novaldex) and radiotherapy, which involved 16 radiations (16/8/2020-1/9/2020), each with a dose of 2.66 Gy, totaling an absorbed dose of 42.56 Gy. Furthermore, based on the results of an Oncotype test, chemotherapy was not deemed necessary in his case.

On May 18, 2023, about 37 months after the initial breast cancer diagnosis, an increase in CA 15-3 levels (37.2 U/ml) was observed for the first time. Following a continuous rise in CA 15-3 levels, a PET/CT scan was performed on August 11, 2023. The scan revealed metastatic breast cancer in the trochanteric region of the left leg. The trochanteric region refers to the area of the upper thigh near the hip joint, specifically around the greater trochanter of the femur. The greater trochanter is a bony prominence on the femur where several muscles of the hip and thigh attach. This area is commonly involved in conditions affecting the hip and is significant in cases of bone metastasis, particularly in cancers like metastatic breast cancer. This metastasis was classified as grade 3 according to the Elston and Ellis grading system [7] showing ER and PR positivity, HER2 at 20%, and a Ki-67 (MIB1) percentage of 30. Following this diagnosis, the following treatment regimen was given to the patient by the medical team of "Theagenio" Cancer Hospital of Thessaloniki, Greece:

- Letrozole (Letrozole Teva 2.5 mg) in place of Tamoxifen (Novaldex).
- Palbociclib (Ibrance 125 mg).
- Two monthly injections: (XGEVA 120mg/1.7 ml 1 VIAL) (ARVEKAP 3.75 mg INJ).
- Radiotherapy consisting of 10 sessions (10/11/2023-23/11/2023), with each session delivering 3 Gy of radiation, totaling an absorbed dose of 30 Gy.

Additionally, prior to commencing radiotherapy, prophylactic threading with a long gamma nail was performed on September 13, 2023.

Measurement of CA 15-3

The method used to measure patient CA 15-3 levels entails a blood test. Over the last four years, thirty-four blood tests for CA 15-3 were carried out across two microbiological laboratories: twenty-five in the first and nine in the second. The technique employed by the two laboratories to measure CA 15-3 is enhanced chemiluminescence immunoassay and microparticle chemiluminescence immunoassay, respectively. The principle of enhanced chemiluminescence immunoassay involves the binding of CA 15-3 molecules in the sample to specific antibodies that are conjugated with a chemiluminescent compound. Then, in the presence of the enzyme HRP and a substrate, the chemiluminescent compound produces light, which is detected and quantified to determine the concentration of CA 15-3. In the microparticle chemiluminescence method, microparticles coated with specific antibodies are used to capture the target antigen (in this case, CA 15-3) from the sample. Chemiluminescent reagents are then added, and the emitted light is measured, which is directly proportional to the concentration of the target antigen in the sample.

The measurements were cross-checked three times between the two laboratories by performing blood tests at both labs, ensuring that the time difference between the measurements at the two labs was no more than 2 days each time. Discrepancies in the results between the two laboratories were observed at rates of 11.3%, 1.9%, and 3.6%. As previously mentioned, the laboratories use different techniques for measuring CA 15-3 and apply different reference values: the first laboratory uses a reference value of <30 U/mL, while the second uses <31.3 U/mL.

Cluster Analysis

The cluster analysis was conducted on the Wolfram Cloud platform provided by Wolfram Research [8] utilizing the Wolfram Language for computations. In this work was used the built-in function "*FindClusters*" with the method used being the K-means clustering algorithm [9]. *FindClusters* function is designed to cluster data based on the values and their relationships, regardless of whether the data comes from a time series with uniform or non-uniform intervals. In Wolfram Language in case of K clusters the function is:

clusters = FindClusters[data, K, Method -> "KMeans", DistanceFunction -> EuclideanDistance] (1)

The time series data is provided as input to the "*FindClusters*" function in the form of a list of pairs, where each pair consists of a timestamp and a value (e.g., {timestamp, value}). The function internally treats each data point ({timestamp, value}) as an entity to be clustered. The *DistanceFunction -> EuclideanDistance* parameter instructs "*FindClusters*" to compute the Euclidean distance between data points, considering both the timestamp and the value. The non-uniformity of time intervals does not directly affect the clustering process. The algorithm clusters data based on the overall distances in the 2D space defined by {timestamp, value}, effectively handling different time intervals by including them in the distance calculation.

The K-means algorithm aims to minimize the within-cluster variance, which is the sum of the squared Euclidean distances between each point in a cluster and its centroid. This minimization process is achieved through iterative refinement until convergence, where the assignments of data points to clusters and the positions of cluster centroids stabilize. In the context of clustering, similarity refers to the degree of resemblance or proximity between data points within the same cluster. Data points that are similar to each other are grouped together into the same cluster, while data points that are dissimilar are placed into different clusters.

In K-means clustering, the number of clusters K must be predefined. In this study, the optimal number of clusters was determined using the elbow method [10]. To determine the optimal number of clusters, the K-means algorithm was run for various values of K (the number of clusters). For each K, the Within-Cluster Sum of Squares (WCSS) — which measures the average squared distance of all points in a cluster from the cluster centroid — was calculated. A plot of WCSS versus K was then generated. The optimal number of clusters corresponds to the point where the rate of decrease in WCSS significantly slows, forming an "elbow" shape on the plot.

Once the optimal number of clusters is identified, the final computation using the K-means clustering algorithm is performed. The quality of the clustering was evaluated using the Silhouette Score [11], which is a valuable metric for assessing clustering performance. This score quantifies how well each data point fits into its assigned cluster and how separate it is from other clusters.

The Silhouette Score for each data point *i* is calculated as follows:

$$s[i] = \frac{\mathrm{b[i]} - \mathrm{a[i]}}{\max(\mathrm{a[i]}, \mathrm{b[i]})} \quad (2)$$

where a[i] represents the average distance from the point i to all other points in the same cluster. It measures how close point i is to other points within its own cluster, giving an idea of how well the point is matched to its cluster. A lower value of a[i] indicates that the point is well clustered with points in its own cluster. b[i] is the smallest average distance from the point i to points in any other cluster, i.e., the distance to the nearest cluster that i is not a part of. It measures how far the point is from the nearest other cluster. A higher value of b[i] indicates that the point is well separated from points in other clusters.

The mean Silhouette Score (s) is given by:

$$s = \sum_{i=1}^{n} \frac{s[i]}{n} \quad (3)$$

The silhouette score (s) ranges from -1 to 1, with higher values indicating better-defined and more appropriate clusters. A score close to +1 indicates that the data points are well clustered. Specifically:

1. $s \geq 0.7$: Strong clustering. The clusters are well-defined, with clear separation between them. Data points are closely grouped within their own cluster and far from neighboring clusters.
2. $0.5 \leq s < 0.7$: Reasonably good clustering. The clusters are fairly well-separated, though there might be slight overlap or proximity between neighboring clusters.
3. $0.25 \leq s < 0.5$: Moderate clustering. The clustering is acceptable but not very distinct. Some points might be close to other clusters, suggesting some overlap or less clear cluster boundaries.
4. $s \approx 0$: The data points are near or on the decision boundary between clusters, being equally close to two or more clusters, which implies overlapping or ambiguous boundaries.
5. s close to -1: Misclassified data points. These points are closer to another cluster's center than their assigned cluster, typically indicating poor clustering where data points do not naturally fit into the assigned cluster.

Results

Time-Series CA 15-3 Results

Table 1 displays the results of 34 plasma CA 15-3 measurements obtained from the two microbiological laboratories: twenty-five from the first laboratory and nine from the second.

Table 1. CA 15-3 (U/mL) measurements

DATE	LAB 1	LAB 2	DATE	LAB 1	LAB 2	DATE	LAB 1	LAB 2
14/4/2020	26.2		23/8/2023	57.5		30/1/2024	29	
20/11/2020	30.6		25/8/2023		56.4	22/2/2024	25.5	
8/3/2021	26.6		1/9/2023	68.2		21/3/2024		27.1
15/7/2021	28.9		27/9/2023		47.4	17/4/2024	33.4	
4/11/2021	28.7		11/10/2023	46.4		18/4/2024		32.2
10/5/2022	24.8		25/10/2023	46.8		1/5/2024		32.9
9/11/2022	28.3		8/11/2023	41.5		9/5/2024	38.1	
18/5/2023	37.2		15/11/2023	40.4		20/5/2024	39.1	
6/7/2023	42.5		29/11/2023	41.6		3/6/2024		38.3
7/7/2023		47.3	13/12/2023	34.4		1/7/2024		35.4
3/8/2023	45.8		28/12/2023	32.2				
16/8/2023		52.7	11/1/2024	26.1				

Figure 1 depicts the time series measurement of CA 15-3. Closed circles denote the measurements measurement performed in Lab 1, while open circles represent those performed in Lab 2. The dashed line represents the 2-period moving average of the CA 15-3 measurements from Lab 1. The initial measurement was conducted 13 days before the patient's mastectomy. The CA 15-3 marker lacks sensitivity in detecting primary tumors; it is primarily utilized for monitoring metastatic disease [6]. This clarifies why, merely 13 days before the surgery for the primary cancer, the CA 15-3 levels were within the normal range (<30 U/mL). However, the significant increase noted in July-August 2023 can be attributed to the presence of metastatic cancer in the trochanteric region of the patient's left leg. Conversely, the notable decrease is attributed to the effectiveness of the treatment given to the patient by his oncologists, as previously mentioned. Based on the two-period moving average of the CA 15-3 measurements from Lab 1, shown as a dashed line in Figure 1, it can be deduced that the continuous increase in CA 15-3 levels from a normal value (<30 U/mL) to the maximum measured value of 68.2 U/mL began around November 9, 2022. This is approximately nine months before the PET/CT scan revealed metastatic breast cancer in the

trochanteric region of the left leg.

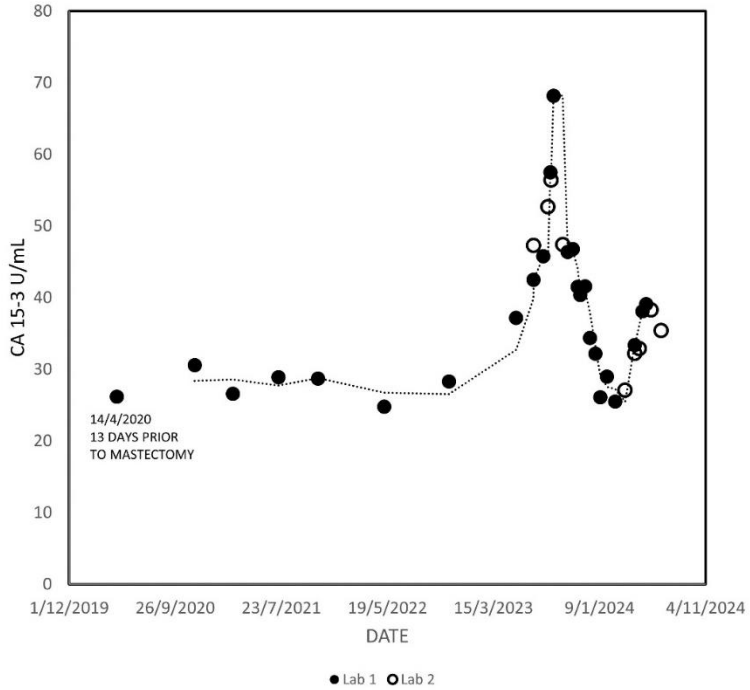

Figure 1. Time series measurement of CA 15-3. Closed circles denote the measurements measurement performed in Lab 1, while open circles represent those performed in Lab 2. The dashed line represents the 2-period moving average of the CA 15-3 measurements from Lab 1.

Cluster Analysis Results

To perform K-means cluster analysis on the time measurements of plasma CA 15-3, the data shown in Table 1 were structured as pairs (x,y), where x represents the number of days since the initial measurement taken on April 14, 2020, and y represents the corresponding value of CA 15-3 in U/mL. As mentioned earlier, even though the data does not have uniform time intervals, the K-means clustering algorithm can still be effectively applied.

The number of clusters (K) must be specified in advance. In this study, the optimal number of clusters was determined using the elbow method, which identifies the most suitable number of clusters based on the data's characteristics. Figure 2 displays the Within-Cluster Sum of Squares (WCSS) values plotted against the number of clusters. The "elbow" point in this plot suggests the optimal number of clusters. The elbow point is where the rate of decrease in WCSS sharply slows down, forming an "elbow" shape on the plot. It is clear from Figure 2 that the optimal number of clusters is K=2.

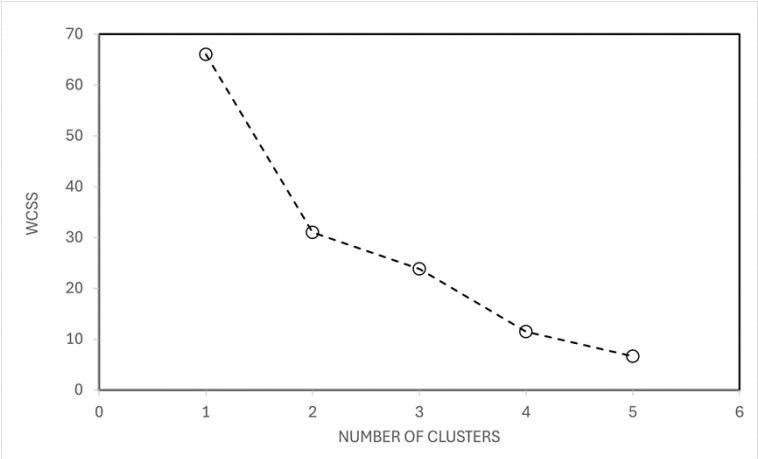

Figure 2. The WCSS values against the number of clusters.

Once the optimal number of clusters K=2 is identified, the final computation using the K-means clustering algorithm is performed. The identification of two clusters within the time series data is depicted in Figure 3. Closed circles indicate measurements associated with cluster 1, while open circles represent those belonging to cluster 2.

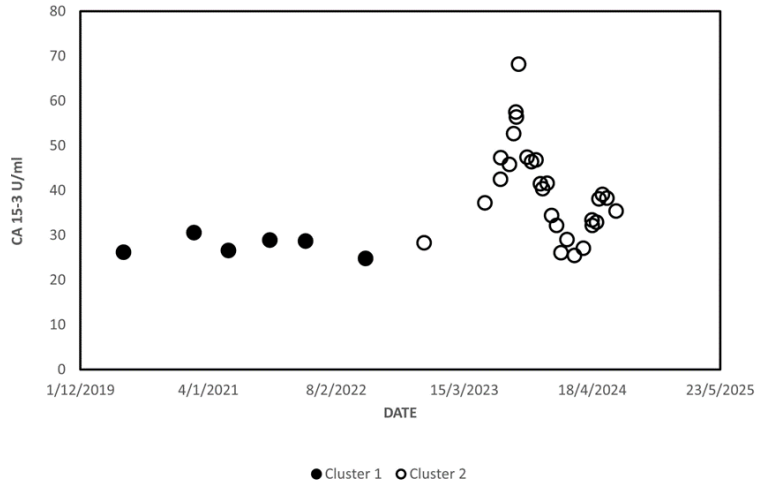

Figure 3. The code's identification of the two clusters within the time series data. Closed circles indicate measurements associated with cluster 1, while open circles represent those belonging to cluster 2.

The Silhouette Score in this study is 0.801. This is a strong indication of a meaningful clustering result, from statistical point of view. The last CA 15-3 measurement belonging to cluster 1 was performed on May 10, 2022, and the first CA 15-3 measurement belonging to cluster 2 was performed on November 9, 2022. There is a relatively long gap of 6 months between these two CA 15-3 measurements. This is expected because systematic serial measurement of CA 15-3 is not recommended for the follow-up of patients with early breast cancer. In contrast, systematic monitoring of CA 15-3 levels is particularly valuable for tracking disease progression and assessing treatment response in patients with metastatic breast cancer.

Since the time period between these two measurements cannot be clearly assigned to either cluster, we will present the silhouette scores of individual data points from both cluster 1 and cluster 2 in Figure 4 to gain a better understanding.

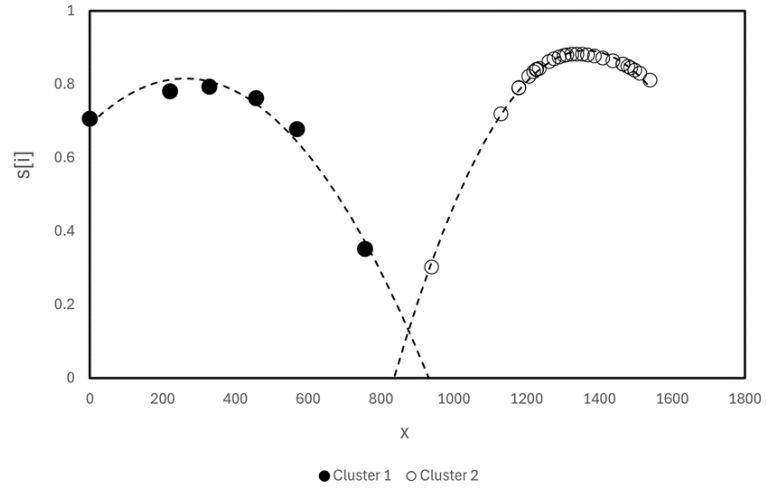

Figure 4. Silhouette scores s[i] for i=34 individual data points from both cluster 1 and cluster 2, where x represents the number of days since the initial CA 15-3 measurement taken April 14, 2020.

The dashed lines in Figure 4 represent polynomial trend lines with R^2 values of 0.977 for Cluster 1 and 0.994

for Cluster 2. The boundary between the two clusters is estimated at the intersection of these trend lines, which occurs 880 days after the initial CA 15-3 measurement, corresponding to September 11, 2022. The following section will discuss the physical interpretation of the two clusters.

Discussion

The cluster analysis identified (Figure 3) two distinct clusters on the time series CA 15-3 data: the first spanning from April 14, 2020, to May 10, 2022, and the second extending from November 9, 2022, to the most recent CA 15-3 measurement on July 1, 2024. The boundary between these clusters is estimated to be in September 2022. In Figure 3 cluster 2 encompasses the period of illness recurrence (metastatic breast cancer). One of the data points of this cluster corresponds to the initial elevated CA 15-3 measurement (37.2 U/mL on May 18, 2023). Approximately two weeks later, the patient reported slight discomfort in the trochanteric region of the left leg while walking. Additionally, the significant increase in CA 15-3 levels is attributed to metastatic cancer in the trochanteric region of the left leg, identified by a PET/CT scan in August 2023. This increase was followed by a notable decline after effective therapy was administered. It is evident that the data points in cluster 2 correspond to the advanced (metastatic) stage of the disease.

The data points in cluster 1 (Figure 3) show a relatively narrow range, from 24.8 to 30.6 U/mL, with an average of 27.6 ± 1.6 U/mL. These values fall within the normal range for CA 15-3 (<30 U/mL). Since the patient exhibited no symptoms or clinical signs of disease recurrence during the period corresponding to cluster 1, it is reasonable to conclude that the data points in this cluster likely correspond to the period before illness recurrence (metastatic breast cancer.)

It is important to note that the initial data point in cluster 2 corresponds to a CA 15-3 measurement of a normal value of 28.3 U/mL (<30 U/mL) taken on November 9, 2022, about seven months before the first symptom appeared. Additionally, within cluster 2, during the period from November 1, 2024, to March 21, 2024, there is a subgroup of four CA 15-3 measurements ranging between 26 and 29 U/mL, with a mean value of 27 U/mL. Despite these values having similar variation and mean as those in cluster 1 (27.6 ± 1.6 U/mL), the code accurately categorized them into the appropriate cluster.

In summary, the main outcomes of the present work are:

- The code successfully identified two clusters within the time series data, as expected. Data points in cluster 1 correspond to the period before the recurrence of illness (metastatic breast cancer), while those in cluster 2 correspond to the advanced (metastatic) stage of the disease.
- The Silhouette Score is 0.801. This is a strong indication of a meaningful clustering result, from statistical point of view.
- The boundary between the two clusters is estimated to occur around 11 September 2022 (Figure 4). The continuous increase in CA 15-3 levels from a normal value (<30 U/mL) to the maximum measured value of 68.2 U/mL began two months later in November 2022 (dashed line in Figure 1). This rise occurred 9 months before the recurrence was diagnosed by PET/CT scan, which aligns well with the recent findings of De Cock et al. [2], who observed a gradual increase in CA 15-3 levels 6–12 months prior to the detection of the first metastases.

Conclusions

Our study provides evidence supporting the meaningful application of clustering analysis on time-series CA 15-3 data in a patient with metastatic breast cancer. However, the study is limited to a single patient, and as a result, the quantitative findings such as the time between CA 15-3 elevation and cancer recurrence cannot be generalized. To our knowledge, this is the first application of cluster analysis to CA 15-3 time-series data. Its potential to detect the onset of the metastatic stage warrants further investigation.

List of Abbreviations: CA 15-3: Cancer Antigen 15-3; TNM: T (tumor) N(node) M (metastasis); pT1cN0: describes a tumor that is between 1 and 2 cm in size, has been confirmed through pathological examination, and has not spread to nearby lymph nodes; PR: Progesterone Receptor; ER: Estrogen Receptor; HER2: Human Epidermal Growth Factor Receptor 2; WCSS: Within-Cluster Sum of Squares; HRP: Horseradish Peroxidase; PET-CT: Positron Emission Tomography/Computed Tomography.

Author Contributions: Not applicable, as there is only one author.

Funding: This research received no funding.

Ethics Statement: Not applicable.

Data Availability Statement: The author declares that the data supporting the findings of this study are available within the paper. In particular Table 1 displays the values of the CA 15-3 time-series. The computations concerning the elbow method (Figure 2), K-means clustering (Figure 3) and Silhouette Scores (Figure 4) were performed on the Wolfram Cloud platform utilizing the Wolfram Language.

Acknowledgments: The author would like to thank the medical staff of 1) the C' Department of Clinical Oncology and Chemotherapy at Theagenio Cancer Hospital of Thessaloniki and 2) the 3rd Department of Orthopedics at Papageorgiou Hospital of Thessaloniki for the high-quality care provided during his treatment.

Conflict of Interest: The author declares no conflict of interest.

References

1. Gang Y, Adachi I, Ohkura H, Yamamoto H, Mizuguchi Y, Abe K. [CA 15-3 is present as a novel tumor marker in the sera of patients with breast cancer and other malignancies]. Gan To Kagaku Ryoho. 1985;12(12):2379-86. Japanese.
2. De Cock L, Heylen J, Wildiers A, Punie K, Smeets A, Weltens C, Neven P, Billen J, Laenen A, Wildiers H. Detection of secondary metastatic breast cancer by measurement of plasma CA 15.3. ESMO Open. 2021;6(4):100203. doi: 10.1016/j.esmoop.2021.100203.
3. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, Zackrisson S, Senkus E; ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019 Aug 1;30(8):1194-1220. doi: 10.1093/annonc/mdz173.
4. Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, Somerfield MR, Hayes DF, Bast RC Jr; American Society of Clinical Oncology. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol. 2007 Nov 20;25(33):5287-312. doi: 10.1200/JCO.2007.14.2364. Epub 2007 Oct 22. PMID: 17954709.
5. Khatcheressian JL, Hurley P, Bantug E, Esserman LJ, Grunfeld E, Halberg F, Hantel A, Henry NL, Muss HB, Smith TJ, Vogel VG, Wolff AC, Somerfield MR, Davidson NE; American Society of Clinical Oncology. Breast cancer follow-up and management after primary treatment: American Society of Clinical Oncology clinical practice guideline update. J Clin Oncol. 2013 Mar 1;31(7):961-5. doi: 10.1200/JCO.2012.45.9859. Epub 2012 Nov 5. PMID: 23129741.
6. Duffy MJ, Evoy D, McDermott EW. CA 15-3: uses and limitation as a biomarker for breast cancer. Clin Chim Acta. 2010 Dec 14;411(23-24):1869-74. doi: 10.1016/j.cca.2010.08.039. Epub 2010 Sep 8. PMID: 20816948.
7. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991 Nov;19(5):403-10. doi: 10.1111/j.1365-2559. 1991.tb00229. x. PMID: 1757079.
8. Wolfram Research, Inc., Wolfram|Alpha Notebook Edition, Champaign, IL. 2024.
9. MacQueen J.B. Some methods for classification and analysis of multivariate observations Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability 1967 University of California Press 281-297 http://projecteuclid.org/euclid.bsmsp/1200512992
10. Kodinariya TM, Makwana PR. Review on determining number of Cluster in K-Means clustering. Int J Adv Res Comput Sci Manag Stud.2013; 1:2321–7782
11. Rousseeuw PJ. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics. 1987; 20:53–65. doi:10.1016/0377-0427(87)90125-7.