Survival Clustering web-based application – A deeper look at risk profiles for METABRIC

Yuan Gu*[1] Mingyue Wang[2] Yishu Gong[3] Song Jiang[4] Zhen Li[5]

1 Department of Statistics, The George Washington University, Washington, DC, USA

2 Department of Mathematics, Syracuse University, Syracuse, NY, USA

3 Harvard T.H. Chan School of Public Health, Harvard University, Boston, NY, USA

4 Department of Biochemistry, Huzhou Institute of Biological Products Co., Ltd. China

5 Booth School of Business, University of Chicago, Chicago, IL, USA

*Corresponding email: uwin@gwu.edu

Abstract:

Medical doctors frequently rely on assistance tools during the decision-making process or when determining suitable chemotherapy options. These tools can take the form of recommendation systems, online test calculators, or web-based applications. They provide support not only in making recommendations but also in conducting thorough profile investigations of patients. Previous researchers have developed web-based survival analysis tools in the cancer survival field. However, many of these tools provide only basic functionality and rely on simplistic models, offering only a superficial understanding of the data. In this study, we undertake a comprehensive analysis of risk profiles using survival clustering techniques applied to a real-world dataset and developed an accessible online Shiny application to facilitate easier utilization of our findings. By leveraging survival clustering, we aim to uncover distinct subgroups based on survival patterns and identify unique risk profiles associated with breast cancer patients. Our online app provides a user-friendly interface for researchers and clinicians to explore the results, enabling them to gain valuable insights into the complex landscape of breast cancer risk profiles.

This interactive tool offers a more accessible means of understanding and utilizing the implications of our research in personalized medicine and clinical decision-making.

Key words: Shiny, Web-based application, Survival, Breast Cancer, Machine Learning

1. Introduction:

Understanding the risk profile and survival outcomes of breast cancer remains a complex and intricate area of research. Despite significant advancements in the field, there are still numerous factors that contribute to the variability in breast cancer progression and patient outcomes. The interplay between genetic predisposition, lifestyle factors, tumor characteristics, and treatment response adds to the challenge of unraveling the intricacies of breast cancer risk profiles. Additionally, the heterogeneity within breast cancer itself further complicates the identification of clear-cut survival patterns. Consequently, there is a pressing need to delve deeper into the complexities of breast cancer risk profiles and survival to uncover new insights and potential avenues for improved diagnosis, treatment, and personalized care for patients.

In this study, we undertake a comprehensive analysis of risk profiles using survival clustering techniques applied to the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset [1]. An unsupervised learning approach was employed to cluster patients based on their survival difference and other relevant clinical variables, by combining k-means clustering and Kaplan-Meier curve (k-m) curves, leveraging the significant risk factors identified through a Cox regression model. Through a deeper exploration of survival clustering, we can discover heterogeneous subpopulations with varying survival characteristics including age, tumor stage and molecular subtypes, etc., providing insights into the underlying heterogeneity of breast cancer and reveal potential biomarkers for risk stratification. Additionally, we have

developed an accessible online application using Shiny to facilitate easier utilization of our findings.

Currently, there is a scarcity of web-based survival analysis tools available for medical researchers conducting survival studies and all previous web applications have primarily focused on Cox regression [2] or genetic analysis [3-4]. In contrast, our study fills an existing void by introducing a unique combination of unsupervised learning techniques and risk factor analysis in clinician side, demonstrates the potential of survival clustering as a valuable tool in uncovering hidden structures based on distinct risk profiles. Overall, our online tool provides a user-friendly interface for researchers and clinicians to explore the results and derive valuable insights on breast cancer. The link for this app is: https://baran-shad.shinyapps.io/breastcancer

2. Literature review

Breast cancer is a complex disease with diverse clinical outcomes, making accurate prediction of survival crucial for personalized treatment and care. Over the years, various survival models have been developed to aid in understanding and predicting breast cancer survival rates.

The most widely used classical approach is the Cox regression model, which assumes proportional hazard rate. Numerous studies have employed this model to identify significant prognostic factors such as age, tumor stage, hormone receptor status, and lymph node involvement [5-6]. However, the Cox model relies on the proportional hazard assumption, which may not always hold true, leading to potential bias in the estimation of survival probabilities.

To address the limitations, researchers have explored alternative techniques, including machine learning algorithms, such as random forests, support vector machines, and neural networks, which have demonstrated promising results [7-10]. These models can handle complex

interactions between variables and capture non-linear relationships, thus providing improved accuracy in survival prediction. However, their black-box nature often limits interpretability and understanding of the underlying biological mechanisms.

Another noteworthy approach is the use of gene expression data. Gene expression-based models, such as using multi-omics neural networks to make the survival prediction, have shown the ability to provide deeper insight into which types of data are most relevant to improve prognosis [11]. These models provide valuable insights into the underlying biology of breast cancer and offer potential for personalized treatment strategies. However, their reliance on gene expression data may limit their application in clinical settings where gene expression profiling is not routinely performed. Other challenges and limitations exist in the data and visualization. Issues like missing data, limited sample sizes, and the lack of accessible visualization tools for physicians and medical professionals with limited modeling expertise hinder the widespread utilization of the models.

In our study, we address all above challenges by developing a user-friendly interface to uncover hidden structures within breast cancer data and identify unique risk profiles. This intuitive tool enables researchers and clinicians easily explore and interpret the results of our analysis to gain valuable insights into the complex landscape of breast cancer risk profiles, bridging the gap between sophisticated modeling techniques and practical clinical applications.

3. Methods

3.1 Data

The METABRIC dataset is a valuable and publicly accessible resource for researchers. It encompasses a total of 2,506 subjects and 34 variables, providing a comprehensive foundation for studying breast cancer. To ensure the reliability of our analysis, we diligently handled any

missing data and performed thorough feature checks. Following meticulous data preprocessing procedures, our final analysis dataset consists of 1,269 subjects and 23 variables. All analysis is built by R (R 4.3.0), summary of the statistics is depicted in the following Table 1.

	Living(N=548)	Deceased (N= 721)	p-value
Age.at.Diagnosis**	56 (26.72- 85.21)	66 (21.93- 96.29)	<0.0001*
Tumor.Size	23 (1 - 100)	29 (1 - 180)	<0.0001*
Neoplasm.Histologic.Grade			<0.0001*
1	55(10.04%)	46 (6.38%)	
2	228 (41.61%)	269(37.31)	
3	265 (48.36%)	406(56.31%)	
Lymph.nodes.examined.positive	1.15(0-25)	2.45(0-41)	<0.0001*
Mutation.Count	4.9 (1-26)	5.9(1-46)	<0.0001*
Nottingham.prognostic.index	3.9(2-6.19)	4.3(2-6.36)	<0.0001*
Tumor.Stage			<0.0001*
1	227(41.42%)	193(26.77%)	
2	293 (53.47%)	443(61.44%)	
3	27(4.93%)	78(10.82%)	
4	1(0.18%)	7(0.97%)	
ER_Status_bin			0.54
Negative	121(22.08%)	170(23.58%)	
Positive	427(77.92%)	551(76.54%)	
HER2_Status_bin			0.037*
Negative	494(90.15%)	622(86.27%)	
Positive	54(9.85%)	99(13.73%)	
Hormone_therapy_bin			0.39
No	227(41.42%)	281(38.97%)	
Yes	321(58.58%)	440(61.03%)	
PR_Status_bin			0.023*
Negative	242(44.16%)	365(50.62%)	
Positive	306(55.84%)	356(49.38%)	
Relapse_Status_bin			<0.0001*

Not Recurred	479(87.41%)	255(35.37%)	
Recurred	69(12.59%)	466(64.63%)	
Menopausal_bin			<0.0001*
Pre	171(31.20%)	122(16.92%)	
Post	377(68.80%)	599(83.08%)	
Integreative_number	6.30(1-11)	6.32(1-11)	0.89
Chemotherapy_bin			0.12
No	416(75.91%)	574(79.61%)	
Yes	132(24.09%)	147(20.39%)	
Cellularity_bin			0.54
Low	65 (12%)	76 (11%)	
Moderate	218(40%)	275(38%)	
High	265(48%)	370(51%)	
Pam50Claudin_num	3.80(1-7)	3.86(1-7)	0.35
death_reason			<0.0001*
Living	548 (100%)	0 (0%)	
Died of Other Causes	0 (0%)	283 (39%)	
Died of Disease	0 (0%)	438 (61%)	

Table 1 Basic descriptive statistics. * p value denotes significant level when p<0.05
** continuous variables are summarized by mean (min-max)

3.2 Model

The analysis of the breast cancer data encompassed three distinct phases. In the first phase, Cox regression with stepwise AIC selection was employed to identify statistically significant risk factors. This approach allowed us to determine the variables that most significantly influenced breast cancer outcomes among the 23 variables. Following this, K-means clustering was conducted based on the selected risk factors from the Cox model in the first phase. By grouping similar individuals together, this clustering analysis provided insights into distinct subgroups within the dataset. In the final phase, a Kaplan-Meier model was constructed using the predicted clusters, enabling a deeper exploration of the risk profiles associated with each cluster. This

comprehensive approach allowed for a thorough examination of the relationship between risk factors, clustering patterns, and breast cancer outcomes, ultimately enhancing our understanding of the disease.

The first step is the Cox regression with stepwise AIC selection. Cox regression, assumes proportional hazards, is specifically designed for survival analysis, allows us to assess the impact of various variables on the time until death occurs. By employing stepwise AIC selection, the model identifies the subset of variables that provide the best fit for the data, while controlling for the risk of overfitting. This approach considers the trade-off between model complexity and goodness of fit, selecting a parsimonious model that optimizes the AIC criterion.

Once significant risk factors were identified, the next step is the k-means clustering to group individuals based on those factors. To determine the optimal number of clusters, various methods such as the elbow method, Silhouette coefficient, and gap statistics were considered. However, given that our ultimate objective was to examine the survival risk profile, we employed the Kaplan-Meier (KM) model, which incorporates the log-rank test to identify differences in survival curves among clusters. By leveraging the insights provided by the log-rank test, we were able to select a different number of clusters and visualize the results. The visualized clusters were also presented in the web-app, allowing for interactive exploration.

For each number of clusters, we conducted an analysis of the basic characteristics associated with the predicted clusters. By assessing these characteristics, we gained insights into the distribution and patterns within each cluster. For continuous risk variables, the mean values provided an indication of the average risk level within that cluster. Meanwhile, for categorical risk factors, the frequency analysis allowed us to identify the prevalence of specific risk factors

within each cluster. This comprehensive examination of basic characteristics facilitated a deeper understanding of the distinct profiles.

4. Results

The Cox regression analysis identified several significant risk factors associated with breast cancer outcomes; the result is presented in Table 2.

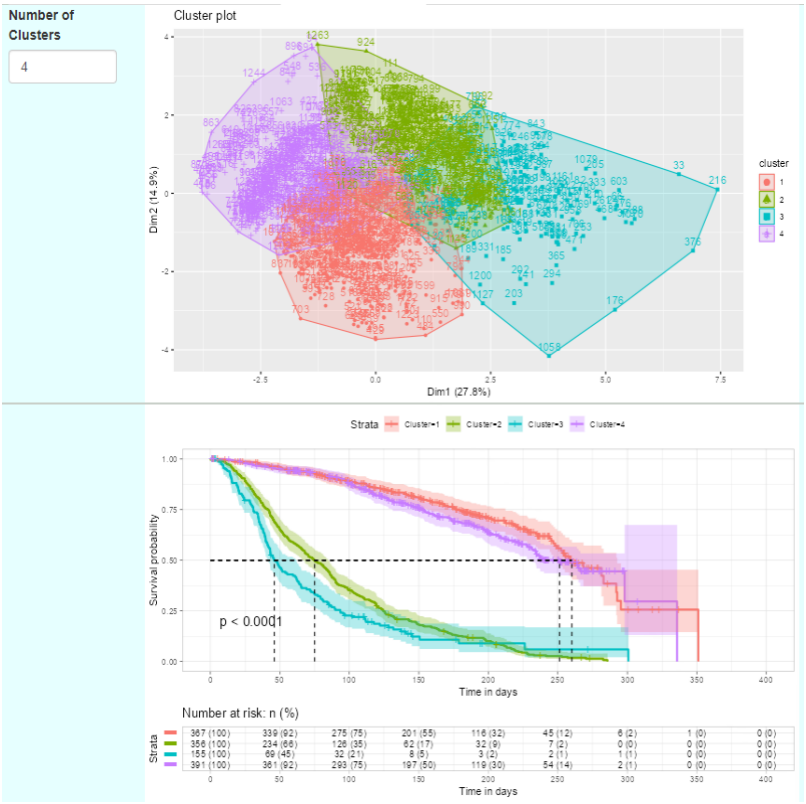
Age at diagnosis ($p = 0.002147$) showed a significant positive relationship with death, suggesting that older age is associated with increased risk, cohort membership (p-value < 0.000001) was found to have a protective effect on breast cancer outcomes. Neoplasm Histologic Grade ($p = 0.015497$), the presence of positive lymph nodes ($p = 0.018251$), and a higher Nottingham Prognostic ($p = 0.031113$) Index were associated with increased hazard rates. Longer relapse-free periods (p-value < 0.000001), smaller tumor size (p-value = 0.000248), and lower tumor stage (p-value < 0.000001) were associated with decreased hazard rates. ER-negative status ($p < 0.000001$) and experiencing a relapse event ($p < 0.000001$) were strongly associated with lower hazard rates. And the cause of death ($p < 0.000001$) also had a substantial positive effect on hazard rates. The model has a concordance of 0.937, demonstrating a very strong discriminatory power to distinguish between individuals with different survival outcomes with a high degree of accuracy.

All the above significant risk factors are standardized by simply converting into a z-score and then used as the input variables into the k means clustering. Standardization helps ensure that all input variables are on a similar scale, preventing variables with larger magnitudes from dominating the clustering process. In the clustering phase, K-means will be used to partition the dataset into K clusters. The algorithm assigns each data point to the cluster with the nearest mean value.

	Survival model		
Predictors	Estimates	Confidence Interval	p-value
Age.at.Diagnosis	1.01	1.00- 1.02	0.002
Cohort	0.81	0.74-0.89	<0.001
Neoplasm.Histologic.Grade	1.31	1.05-1.62	0.015
Lymph.nodes.examined.positive	1.03	1.01-1.06	0.018
Nottingham.prognostic.index	0.85	0.73-0.98	0.031
Relapse.Free.Status..Months	0.97	0.97-0.97	<0.001
Tumor.Size	0.99	0.99-1.00	<0.001
Tumor.Stage	1.81	1.54-2.13	<0.001
ER_Status_bin	0.50	0.41-0.61	<0.001
Relapse_Status_bin	0.01	0.01-0.02	<0.001
death_reason	29.38	20.49-42.13	<0.001
Integreative_number	0.97	0.95-1.00	0.050

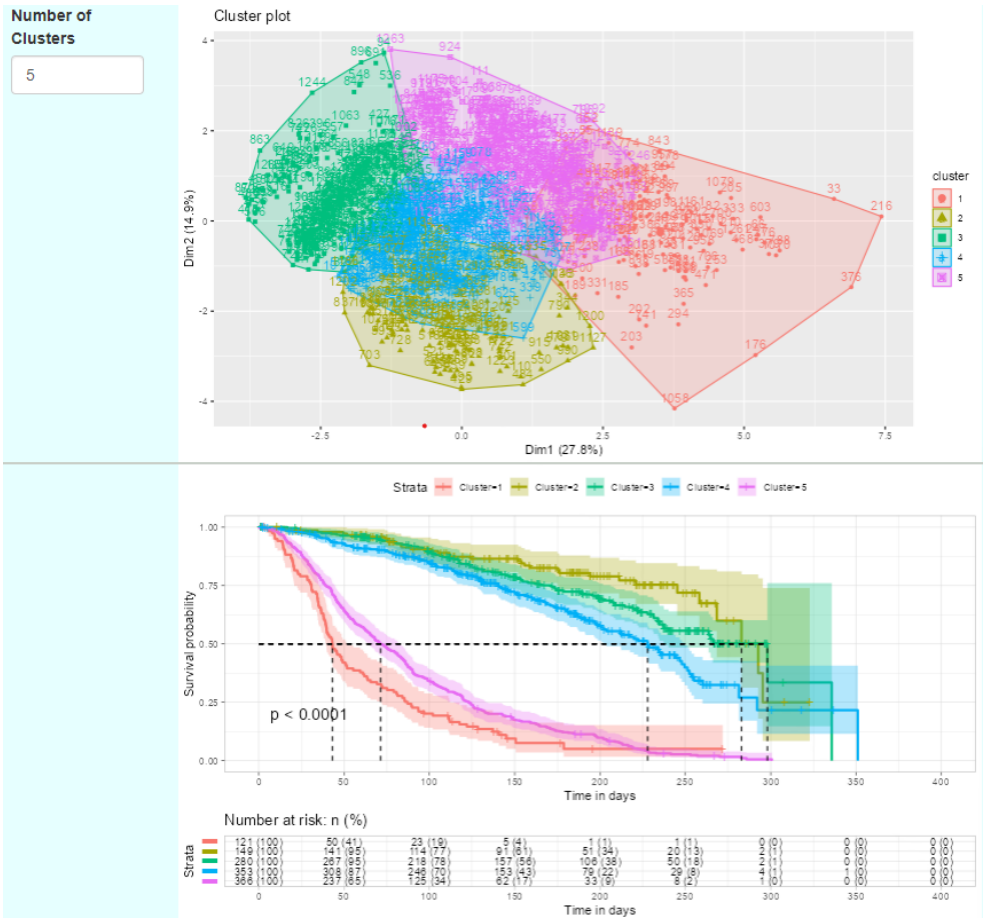
Table 2 Cox regression results

Although traditional k-means is unsupervised learning without predefined labels or target variables, in our case, we improved it by incorporating a pseudo-supervised label, such as the K-means survival difference. We used the survival difference between different clusters to visually guide the selection of the appropriate number of clusters. This approach adds an extra layer of information beyond the traditional K-means, enabling the model to assess the quality of clustering based on survival differences. The following Picture 1 presents the clustering points and the survival curve difference by selecting k = 4:



Picture 1 The clustering points and the survival curve for 4 clusters

In the above depicted picture, four distinct clusters are discernible, each highlighted with a unique color. The accompanying graph below represents the survival plot. The sizes of these clusters vary, with 367 samples in the first cluster, 356 in the second, 155 in the third, and 391 in the fourth. Significant disparities can be observed when comparing clusters 1 and 4 to clusters 2 and 3. Additionally, within these groupings, minor survival differences are also evident between clusters 1 and 4, as well as between clusters 2 and 3. However, when we select k = 5,



Picture 2 The clustering points and the survival curve for 5 clusters

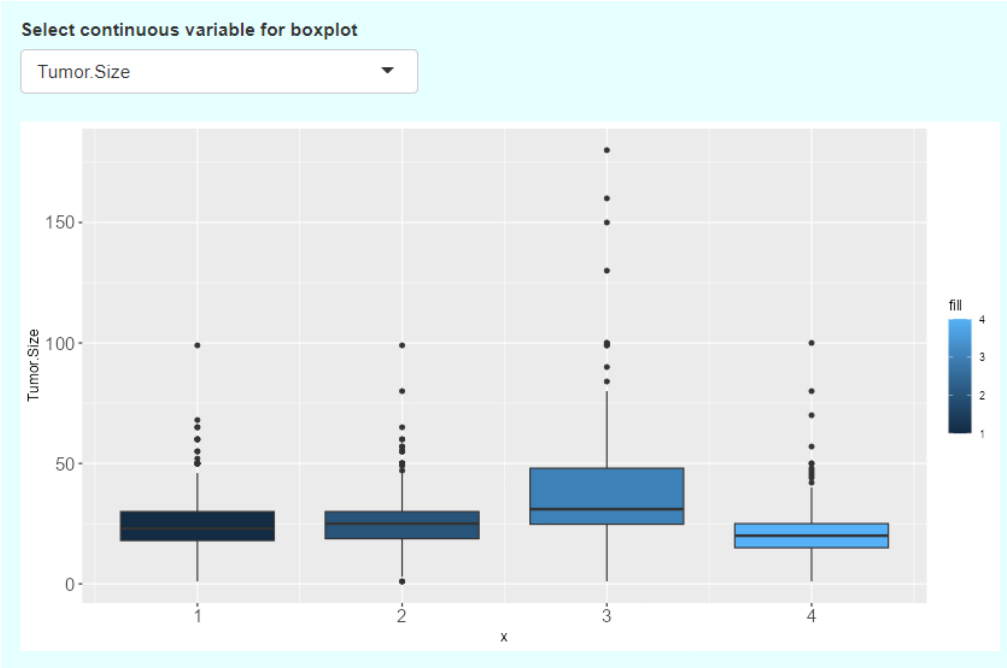
Upon reconfiguration, we now observe clusters 1 and 5, formerly known as clusters 2 and 3 when k was set to 4. Notably, there is a significant survival difference between these two clusters (1 and 5) compared to the remaining three clusters (clusters 2, 3, and 4). In terms of sample size, clusters 1 through 5 consist of 121, 149, 280, 353, and 355 samples, respectively. Furthermore, the p-values obtained from the log-rank test are both less than 0.0001, as indicated in the accompanying plots. The median survival time, represented by the horizontal dashed line in the plots, is also an important metric to consider.

Our final step involves a comprehensive examination of the risk profiles across the different clusters. After assigning clusters, we evaluated the mean values of various input variables. These included age at diagnosis, tumor size, Nottingham prognostic index, mutation counts, and the

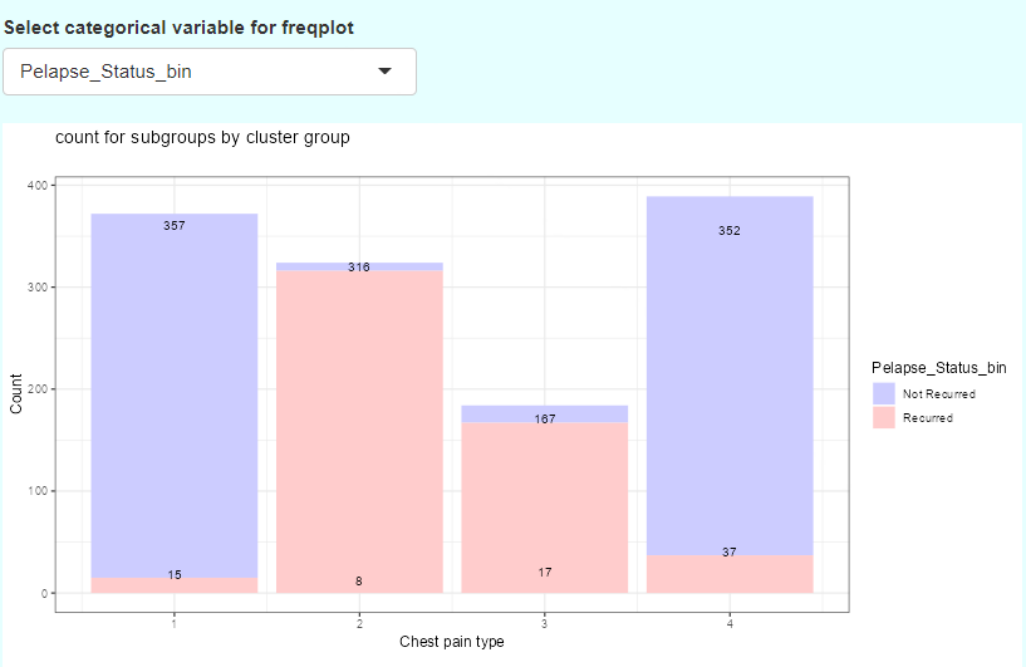
number of positively examined lymph nodes. The table and illustrations below provide summaries corresponding to 4 clusters. In terms of continuous risk factors, such as tumor size, we observe variances in range and mean values across different cluster groups. Table 3 and Picture 3 illustration provide more specific details. Furthermore, we also depicted the frequency distribution of each subgroup within categorical variables (e.g., recurrence vs. no recurrence for the 'Relapse Status') stratified by clusters in the following illustration (Picture 4).

Cluster	n	Age at diagnosis	Tumor Size	Nottingham prognostic index	mutation counts	number of positively examined lymph nodes
1	367	59.73	24.11	4.51	5.55	1.03
2	356	59.91	25.80	4.14	5.92	1.03
3	155	61.15	43.22	5.72	4.83	9.57
4	391	61.08	21.52	3.12	5.27	0.44

Table: 3 Mean values of continuous risk factor for different clusters



Picture 3 Box plot for Tumor size group by 4 clusters



Picture 4 Relapse status subgroup frequency group by clusters

5. Conclusion

In conclusion, our web app has demonstrated significant advantages, creativity, and contributions in analyzing breast cancer data. Through a comprehensive three-phase approach, we have provided valuable insights into the risk factors, clustering patterns, and outcomes associated with breast cancer. Our advantage lies in employing Cox regression with stepwise AIC selection as the first phase of analysis. The final model identifies statistically significant risk factors for breast cancer with a concordance value of 0.937. The second phase involves k-means clustering, which groups individuals based on the selected risk factors from the Cox model. By identifying similar individuals within the dataset, this clustering analysis reveals distinct subgroups and provides a deeper understanding of the data. We consider the optimal number of clusters by log rank test according to the KM model to explore the risk profiles associated with each cluster. Our web app's contribution lies in the comprehensive examination of basic characteristics associated with each predicted cluster. This in-depth examination enhances our knowledge of the

distinct profiles and risk factors associated with each predicted cluster, ultimately contributing to our understanding of breast cancer.

In summary, our web app excels in its advantage of utilizing Cox regression, k-means clustering, and the KM model to analyze breast cancer data. The interactive visualization and comprehensive examination of risk factors and clustering patterns contribute to our understanding of the disease. By providing valuable insights, our web app empowers healthcare professionals and researchers to make informed decisions and advance their knowledge in the fight against breast cancer.

However, additional research is necessary to achieve a broader application. The current app is designed specifically for this dataset and lacks a comprehensive investigation of generalizability. To enhance usability for medical professionals, it would be beneficial to develop a more convenient pipeline that allows users to import and download their own datasets. Another limitation is the relatively small number of input variables in this dataset. Therefore, the development of additional models or tools capable of handling larger datasets should be considered for future endeavors.

References

[1] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403):346–352, 2012.

[2] Lánczky A, Győrffy B. Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. Journal of medical Internet research. 2021 Jul 26;23(7):e27633.

[3] Dwivedi B, Mumme H, Satpathy S, Bhasin SS, Bhasin M. Survival Genie, a web platform for survival analysis across pediatric and adult cancers. Scientific Reports. 2022 Feb 23;12(1):3069.

[4] Zhao T, Wang Z. GraphBio: a shiny web app to easily perform popular visualization analysis for omics data. Frontiers in Genetics. 2022:2265.

[5] Hajihosseini M, Faradmal J, Sadighi-Pashaki A. Survival analysis of breast cancer patients after surgery with an intermediate event: application of illness-death model. Iranian Journal of Public Health. 2015 Dec;44(12):1677.

[6] Vahdaninia M, Montazeri A. Breast cancer in Iran: a survival analysis. Asian pacific journal of cancer prevention. 2004 Apr 1;5(2):223-5.

[7] Tong L, Mitchel J, Chatlin K, Wang MD. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. BMC medical informatics and decision making. 2020 Dec;20:1-2.

[8] Cure-Cure CA, Cure P, Gu Y, Tian X, Patel T, Wu CO, Sviglin H, Sopko G, Csako G, Cody S, Dandi G. Predictors of all cause mortality and their gender differences in a hispanic population from barranquilla-colombia using machine learning with random survival forests. Circulation. 2018 Nov 6;138(Suppl_1):A16252-.

[9] Song Jiang. Yuan Gu. Ela Kumar Magnetic Remaging (MRl) Brain Tumor lmaae Classification Basedive Machine learnina Alaorithms. Cloud Computina ancData Science [lnternet]. 2023 May21:4(2):122-33.

[10] Chi CL, Street WN, Wolberg WH. Application of artificial neural network-based survival analysis on two breast cancer datasets. InAMIA annual symposium proceedings 2007 (Vol. 2007, p. 130). American Medical Informatics Association.

[11] Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, Zhang J, Salama P, Rizkalla M, Han Z, Huang K. SALMON: survival analysis learning with multi-omics neural networks on breast cancer. Frontiers in genetics. 2019 Mar 8;10:166.