International Journal of Public Health Science (IJPHS)
Vol. 10, No. 4, December 2021, pp. 888~897
ISSN: 2252-8806, DOI: 10.11591/ijphs.v10i4.20936

Modeling and analyzing predictive monthly survival in females diagnosed with gynecological cancers

Timothy Samec[1], Raed Seetan[2]
[1]Department of Mathematics and Statistics, Slippery Rock University of Pennsylvania, Slippery Rock, PA, USA
[2]Department of Computer Science, Slippery Rock University of Pennsylvania, Slippery Rock, PA, USA

Article Info

Article history:

Received Mar 11, 2021
Revised Aug 12, 2021
Accepted Aug 25, 2021

Keywords:

Cancer
Clinical data
Gynecological cancers
Prognostics
SEER
Survivability

ABSTRACT

Cancer ranks as a leading cause of death worldwide; an estimated 1.7 million new diagnoses were reported in 2021. Ovarian cancer, the most lethal of gynecological malignancies, has no effective screening with over 70% of patients being diagnosed in an advanced stage. The aim of this study was to determine the most statistically significant contributing factors through a multivariate regression into the severity of female gynecological cancers. Data from the surveillance, epidemiology, and end results program (SEER) cancer database were utilized in this study. Several attempted multivariate linear regressions were implemented with further reduced models; however, a linear model could not be properly fit to the data. Because of unmet assumptions, a nonparametric moving, local regression, locally estimated scatterplot smoothing (LOESS), was performed. After smoothing factors were included to reduced-models, residual information was minimized although few conclusions can be drawn from the resulting statistics. These issues were prevalent mainly because of the massive variability in the data and inherent lack of linearity. This can be a significant issue with clinical data that does not dive deeper into cancer-dependent factors including genetic expression and cell surface receptor overexpression. General patient demographic data and diagnostic information alone does not provide enough detail to make a definite conclusion or prediction on patient survivability. Increased attention to the acquisition of tumor tissue for genomic and proteomic analysis in addition to next-generation sequencing methods can lead to significant improvements in prognostic predictions.

Corresponding Author:

Raed Seetan
Department of Computer Science
Slippery Rock University of Pennsylvania
1 Morrow Way, Slippery Rock, PA 16057, USA
Email: raed.seetan@sru.edu

1. INTRODUCTION

As the second leading cause of death in the United States, cancer extends its effects through many facets of our society [1]. Although it is obvious that significant health effects are the major focus of the cancer disease state progression, several other adverse events are understood to be significant issues in treatment. These include the complexity with which basic science research is required to understand this disease, significant costs for developing new treatment strategies, general healthcare costs, and the previous lack of patient data relevant to cancer diagnosis, treatment, and prognosis [2], [3]. A method to help address some of these concerns is to implement a robust data analysis to begin to answer many research questions. In order to perform analyses of this magnitude, a significant amount of data must be available and, with

consistent additions, can be found using The Cancer Genome Atlas (TCGA), the Catalogue Of Somatic Mutations In Cancer (COSMIC), and the Cancer Gene Census (CGC) [4].

In general, cancer had over 1.7 million new diagnoses in 2021 [1]. Cancers of the uterus, cervix, uterine corpus, ovary, vulva, and vagina in general are the third most common cancers in the United States [5]. Ovarian cancer specifically, the most lethal of gynecological malignancy, has no effective screening with over 70% of patients being diagnosed in an advanced stage and is the 8th leading cause of cancer-related death among women [6]. Although there are technological advances in diagnostics, ovarian cancer still remains a very difficult disease to detect and efficiently treat in advanced stages. While incidences remain high, the three year increase of general cancer mortality rose slightly, going from 595,690 deaths in 2016 to 606,880 in 2019 [7]. These data taken altogether paint a grim picture for continuing the same path of oncological research and, without change, will not improve the current standard of care.

With an increase in disease prevalence comes an increase in research efforts as well as patient data available to research teams. Some of this data can be collected prior to any diagnostics or treatment intervention. General patient information, including age, sex, place of residence, and ethnicity can provide some detail into the prognosis after a patient is diagnosed with cancer [8]. Also, understanding a patients ethnic background, with previous research knowledge, can help develop a more personalized treatment approach that may be more beneficial for one patient over another [9]. Understanding racial disparities in cancer prevalence can help improve clinical understanding of disease presentation, prevalence, and treatment deficiencies between different races [6], [10]. Data like these are readily obtained through national database systems and, with permissions, can be utilized to perform several analyses to determine how specific groups of people may receive and respond to specific treatments.

Although there is a large amount of general patient data, this information can only provide a general understanding for a prognostic outlook for a patient. New therapeutics can possibly be better targeted or transported by different mechanisms based on some of this data, but may still fail to lower the mortality rates of many cancer types. The best method to develop new therapeutics and identify better therapeutics and delivery targets is by using a genetic profile analysis and even expanding to a proteomics based understanding of how certain cancers are activated and respond to treatment by comprehensively characterizing the condition of patients at different cancer stages [11]–[13]. Additionally, novel therapeutic targets can be identified in order to develop cell-specific therapeutics and better develop patient specific medicine [9], [14]–[16]. However, a genomic analysis will not be the implemented in this study as identifying novel therapeutic targets is not within the scope of this work.

This study utilized data retrieved from the SEER cancer database [17], [18]. This database is maintained by the National Cancer Institute and required clearance approval prior to receiving patient data. Patient data is acquired through cancer registries and includes 142 variable types that describe patient tumor types and locations. The set selected for this work is the 2005 report of female gynecological cancers. Because of the tumor specificity, files not associated with gynecological disease were omitted from the analysis but have been briefly described in the methods.

The goal of this work is to determine the most statistically significant contributing factors through a multivariate regression into the severity of female gynecological cancers and how they may affect the choice of care and patient survival length as well as forecasting these values according to other variable attributes. Based on prior research, it is hypothesized that a higher number of malignancies, high tumor grade, and age of diagnosis will be the most significant contributors to early patient death, primary and distant lymph surgeries, and low survival time.

2. RESEARCH METHOD

This study utilized a data file accessed from the SEER cancer database from the National Cancer Institute containing data describing females diagnosed with gynecological cancers between the years of 1975 through 2016. The primary goal was to analyze a dataset and determine survival predictability based on currently available demographic and diagnostic information that does not include information related to cellular and molecular level observations. All patients were surveilled until exclusionary criteria were met including dismissal from follow-up care or death due to the primary cancer of death resulting from another cause. Primary data reported in this file include demographic information, diagnostic data including the primary tumor grade and size, and prognostic information including lymph involvement, further surgery, and types of follow-up care. Complete file retrieval included data describing breast, colorectal, leukemia, male genital, respiratory, and urinary malignancies, but the only file implemented in this study was specific to gynecological-related malignancies.

Data was requested through the Surveillance, Epidemiology, and End Results Program, SEER, at the governmental website seer.cancer.gov. SEER is a data repository through the National Institute of Health

– National Cancer Institute, NCI. Because of patient sensitive information and risk of improper use of clinically relevant data, full registration and description of data use was required prior to retrieval of a data file. Upon approval, data file registry access was granted and selection of a file containing patient information from 1976-2016 was pursued. Because of previous research and expertise, data files relevant to gynecological cancer were filtered.

The data folder containing female gynecologic patient data was downloaded in text format and imported into Microsoft Excel. Additionally, the file was imported in Statistical Analysis Software 9.4 (SAS) 9.4, to ensure proper formatting and segmentation. Import into Excel and confirmation of successful SAS reading was confirmed. To better process the full file, numerous variable locations were omitted due to no data entry and repetition. A significant number of variable locations were not applicable for this patient population and did not contain any data and thus should not be included in the file to be processed and analyzed. The final data file to be read and analyzed was 498 rows by 14 columns totaling 6,972 data points.

All calculations and data analysis were performed in SAS 9.4. Descriptive statistics for the quantitative variable locations age of diagnosis, ageDx, and number of malignancies, numMaligTumor, tumor size, tumSize, months of survival after diagnosis, monthSurvival, and total malignancies, totalMalig. Age of diagnosis is a simple numeric value for the age in which the patient was diagnosed with cancer. The number of malignancies is another quantitative variable representing the number of malignant tumors identified at initial diagnosis. Tumor size is the continuous variable that represented the size of the primary malignancy, measured in cm. Months of survival, the responding variable in this study, is a measure used for the amount of time, in months, that a patient has survived or is currently surviving after the initial diagnosis. Total malignancies, is the total number of malignant tumors that were present in initial diagnosis as well as additional recurrent malignancies that may have occurred. Frequency distributions of categorical data were compiled for the variable locations marital status, maritalChar, tumor staging, grade, primary surgery type, surgType, type of lymphovascular/lymph surgery, follow up protocols, followUp, and cause of death, CoD. All quantitative data was tested for normality assumptions using proc univariate normal within the SAS environment. Additionally, descriptive statistics were calculated for data when excluding values of tumor size being equal to 0. Scatter plots, using proc sgscatter, were constructed for all quantitative variables against the dependent variable in further analysis, monthSurvival. Because of the lack of linearity between data plots, logarithmic transformations were retroactively completed for variables that displayed high levels of interactivity with the dependent variable.

Data associated with survival predictability were selected for multivariate regression analysis including ageDx, numMaligTumor, tumSize, and totalMalig. Models were built with single associations prior to reducing the regression to incorporate variable interactions. All independent variables were analyzed using both regression and general linear models, PROC REG and PROC GLM, options within the SAS environment. With significant model influence being considered at $p \leq 0.05$, variables with significant model influence were extracted for further modeling with a reduced multivariate regression (MVR). Here, variable interactions were introduced. Additionally, logarithmic transformed data were included for reduced model analysis.

Categorical variables were extremely limited in the regression analysis. Inability to include these variables without further data shift to modify the categorical values into continuous values reduces the ability to draw conclusions on the weight of these interactions. Data modification for these values may cause loss of value and meaning and should be kept at the categorical entries as provided.

Finally, with complete parametric assumptions not being met and extreme variability present in continuous data, an additional regression model system was introduced. The locally estimated scatterplot smoothing (LOESS) regression system was incorporated for all continuous data as described previously. LOESS provides stability for systems with significant outliers and, when lack of linearity is present, a robust fitting system is necessary. Here, weighted least squares are used to fit a linear function of the independent variables centered at clusters of data termed neighborhoods. Additionally, a smoothing parameter is included to weight each neighborhood in order to control the general model surface smoothness. Input functions for LOESS were identical to regression (REG) and general linear model (GLM) inputs. Output data was exported via the output delivery system command (ODS OUTPUT), and saved for printing and future analysis if necessary.

3. RESULTS AND DISCUSSION

3.1. Quantitative and categorical summaries

A small excerpt of the SEER femGen data file is provided in Table 1. The data were imported into the SAS environment and analyzed with UNIVARIATE and MEANS procedures for normality as well as being plotted for general observations. Continuous variables used for UNIVARIATE analysis were ageDx,

numMaligTumor, tumSize, monthSurvival, and totalMalig. Summary statistics are presented in Table 2. Normality assumptions were met for continuous data, however significant degrees of variability existed for a large portion of the data. At this time, the response variable, monthSurvival, was plotted against the other continuous variables ageDx, numMaligTumor, tumSize, and totalMalig. Logarithmic transformation on ageDx was also plotted against monthSurvival. Additionally, frequency plots were developed for categorical data including maritalChar, grade, surgType, lymphSurg, followUp, and CoD. All scatter plots are shown in Figure 1 and histograms in Figure 2. Simple observation of Figure 1 panels show that predictive relationships will be difficult to create between these continuous, predictive variables. In Figure 2, a majority of patients were married, had unknown tumor grade classifications, underwent tumor resection surgery, had no lymph, either regional or distant, surgeries, were undergoing active care management, and were alive or dead of another cause unrelated to the cancer incidence.

Taken from scatter plot images, there were no observable relationships between the data provided; however, this was yet to be confirmed prior to regression analysis taking place. It is important to note that any lack of observable trends within the data, although not providing statistical evidence, does not provide a positive outlook for observation of mathematically based trends, either alone or with data transformation interventions.

Table 1. Sample of data presented in SEER female gynecological cancer file

PatientID	MaritalChar	AgeDx	NumMaligTumor	YearDx	Grade	Tumsize	Surgtype
1	Unknown	48	2	2005	Unknown	0	Tumor resection
2	Married	57	2	2005	Unknown	0	Tumor resection
3	Married	56	3	2005	Grade II	20	Tumor resection
4	Married	56	4	2005	Grade II	35	Tumor resection
5	Divorced	53	2	2005	Grade III	30	None

Table 2. Summary statistics for continuous variables with Shapiro-Wilk Normality scores

Variable	N	Mean	Std Dev	Variance	Shapiro-Wilk
Agedx	252	59.3293651	16.0737423	275.421	0.991554
Nummaligtumor	252	0.3492063	0.7555105	0.58203	0.545962
Tumsize	252	77.5436508	149.9550334	12862	0.311537
Monthsurvival	252	81.0555556	52.1547354	2916.45181	0.818428
Totalmalig	252	1.2341270	0.5250176	0.30787132	0.526392

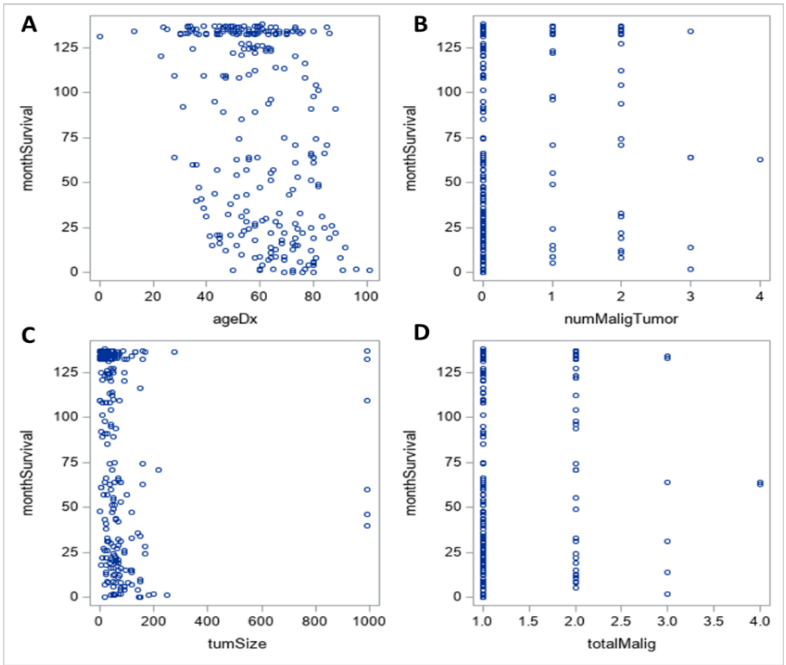

Figure 1. Comprehensive panel of scatter plots characterizing the existence of any relationship between the responding variable, monthSurvival, and (a) the independent variables ageDx, (b) numMaligTumor, (c) tumSize, and (d) totalMalig

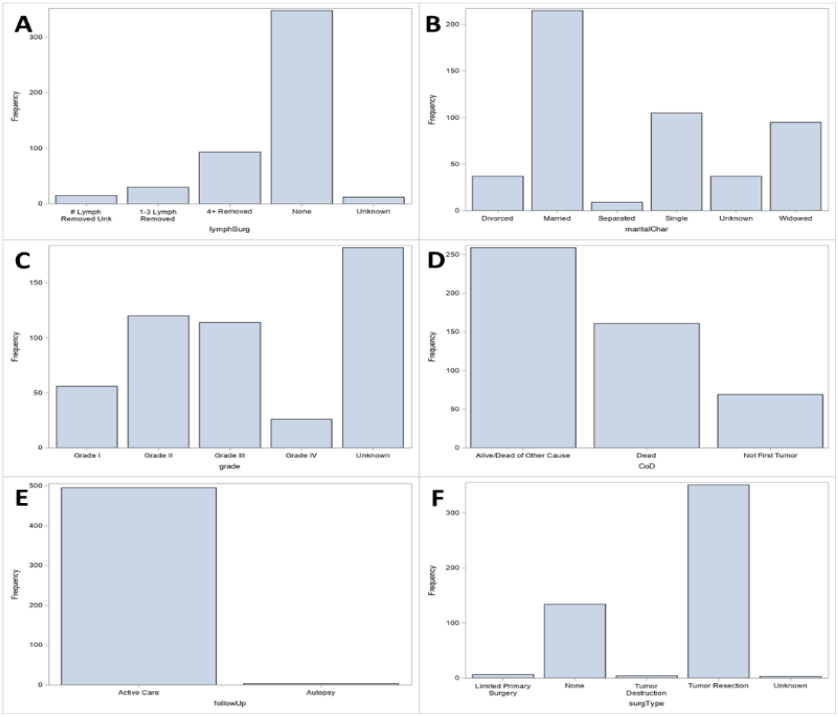

Figure 2. SGPLOT depictions of categorical femgen data in frequency distributions highlighting the most common occurrences for the cluster of patients included within this data set; (a) Data displayed high prevalence of no lymph surgeries, (b) married status, (c) primary tumor resection surgery, (d) tumor of unknown grade, (e) active care follow up plans, and (f) patients remaining alive or dying of other causes not associated with the primary tumor

3.2. Multivariate regressions with transformations

Regression results were formulated from both PROC REG and PROG GLM statements in order to optimize the conclusions from two tests that explain the relationship between multivariate influence on a dependent variable as well as including a robust analysis that also includes a variance analysis across data that is unbalanced. Initial analysis was completed with independent variables ageDx, numMaligTumor, tumSize, and totalMalig. With this model, an R^2 value of 0.1414 was obtained. The incorporation of each of these independent variables does not explain a large majority of the monthSurvival response data and was further reduced to the single statistically significant variable, ageDx. The residual plots are shown in Figure 3. It is important to note the cylindrical appearance of the ageDx residuals that depict a large, irregular residual values from the current predictive model. Furthermore, the residual values for other variables numMaligTumor, tumSize, and totalMalig are distributed across the residual values for each sample variable data point. Within the GLM procedure, ageDx again was the only variable that provided statistical significance to explain the variability in the response variable. Table 3 is the resulting variance analysis results for each continuous variable contained within the regression model.

Further regression models were built around the two variables that had high F statistics, reflected with significant and nearly significant p-values, in the larger regression analysis. AgeDx and totalMalig were further studied for thein influence on monthSurvival. Figure 4 presents the regressors prior to the reduction model but have included the logarithmic transformation of ageDx. Similar observations can be made compared to the untransformed data, but it can be seen that the residuals appear more concentrated within the two linear areas. This is due to the data transformation causing a lower spread along the x-axis for the ageDx data. The reduced model testing the untransformed and transformed ageDx data as well as totalMalig had R^2 values of 0.14 and 0.12 respectively. Still, a large amount of variability was not explaining using these variable measures. Interestingly, logAge, shown in Table 4 displays a higher F statistic and more significant p-value than ageDx when placed in the regression model, suggesting that the transformation provided more evidence for a definite influence in the expected number of months for survival in a patient.

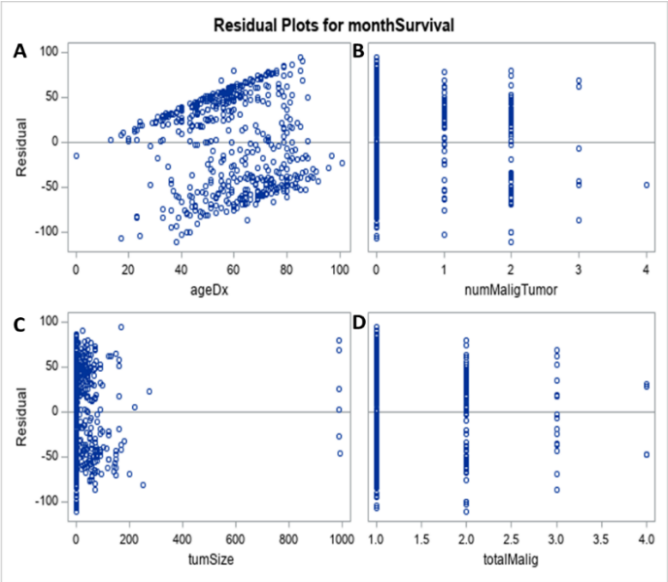

Figure 3. Comprehensive panel of residuals plots exhibiting the predictive ability of the current model provided by PROC REG. Residuals were centered around 0 with a cylindrical and two linear increasing trends with increasing age of diagnosis (A). No conclusions can be drawn from panels B-D with a large portion of residuals being distributed across the range for each predictive data point

Table 3. PROC GLM statistics for influence of variance on dependent monthSurvival variable

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ageDx	1	194343	194343.6	76.99	<.0001
numMalig	1	8.8261	8.8261	0.00	0.952
tumSize	1	1879.615	1879.615	0.74	0.388
totalMalig	1	3475.415	3475.41	1.38	0.241

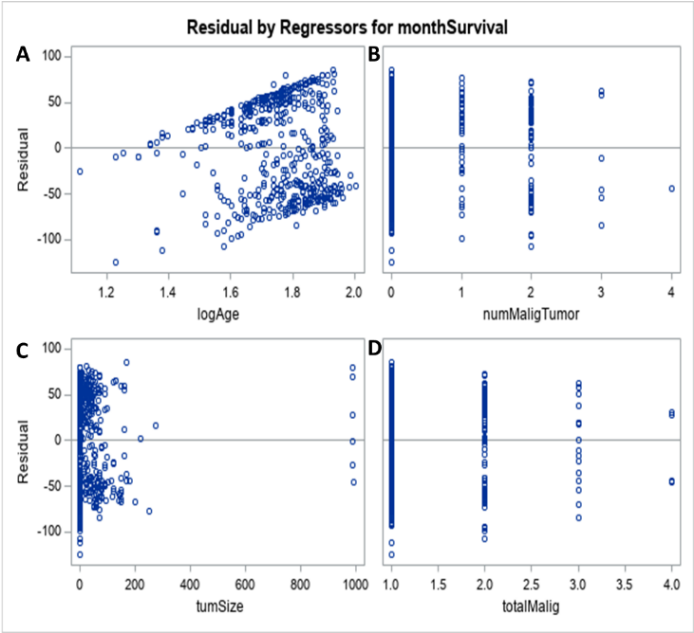

Figure 4. Comprehensive panel of residuals plots exhibiting the predictive ability of the reduced model provided by PROC REG. Residuals were centered around 0 with a cylindrical and two linear increasing trends with increasing logarithmic transformed age of diagnosis but appear to be more clustered than non-transformed data (A). No conclusions can be drawn from panels B-D with a large portion of residuals being distributed across the range for each predictive data point

Table 4. PROC GLM statistics for influence of variance on dependent monthSurvival variable

Source	DF	Type III SS	Mean Square	F Value	Pr > F
logAge	1	14839.23	14839.23	5.89	0.0156
totalMalig	1	5263.38	5263.38	2.09	0.1490
Cross	1	1863.37	1863.37	0.74	0.3902

3.3. Nonparametric methods

Due to the large degree of variability through all variables selected for a parametric regression analysis, a nonparametric model was also constructed using the LOESS procedure and algorithm. With LOESS, an optimal smoothing parameter is selected for the total number of data points within a certain quantity of data 'neighborhoods' or clusters. The smoothing parameter is selected to minimize the AICC value, so to strike a balance between the residual sum of squares and model fit complexity. Output from LOESS is displayed in graphical representation of the smoothing parameter selection, residual values, residual normality curves, and predictive values. These representations are shown in Figures 5 and 6 for the optimum model system of ageDx and totalMalig, based on F and R^2 findings from previous models.

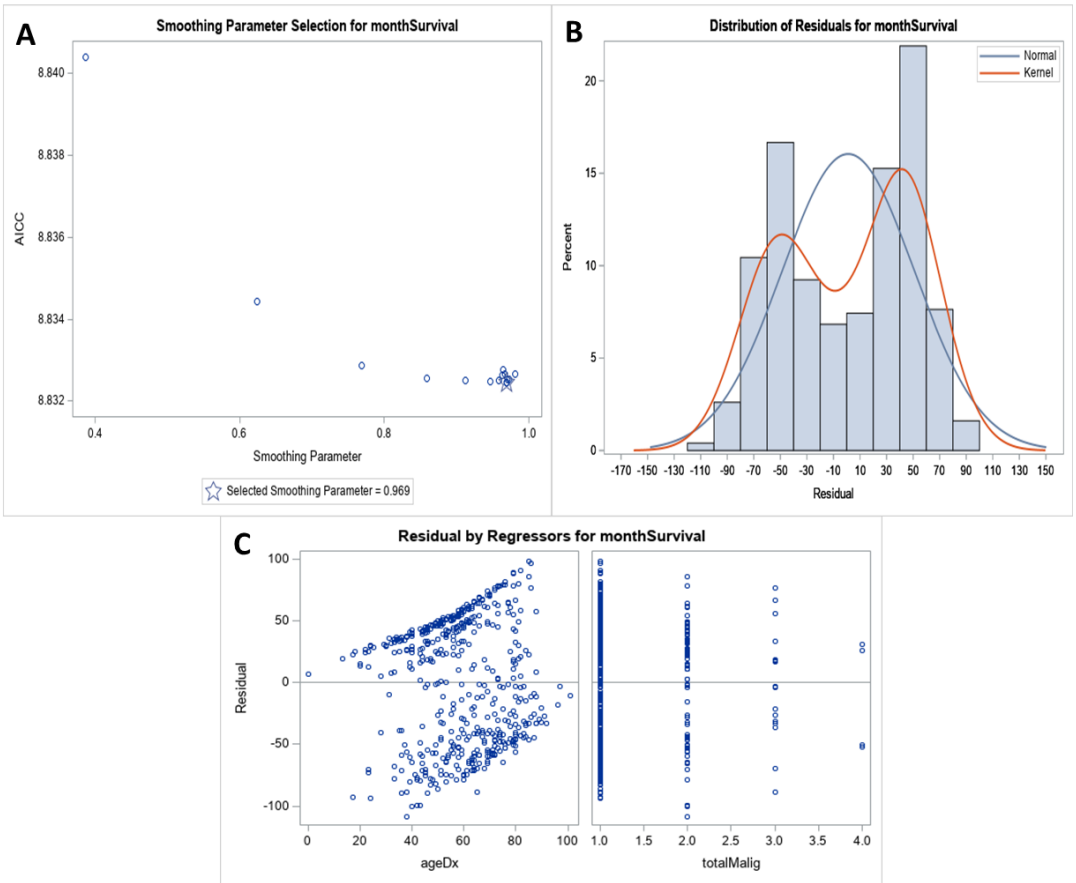

Figure 5. (a) Smoothing parameter optimization to minimize AICC in LOESS, Residual distribution for the dependent variable monthSurvival when analyzed with LEOSS, (b) A bimodal distribution is present suggesting data irregularities and possible incompatibility with predictive model systems. (c) Residuals by regressors, ageDx and totalMalig, for LOESS model. Residual trends show similarities to parametric methods with less linearity shown in ageDx

For the chosen model variables, the resulting smoothing parameter was optimized at a value of 0.969 and consequential residual-regressor distributions are given in Figure 5, (a) and (c). With LOESS, the residual values appear to have a similar distribution to the parametric testing but with less linearity. This can be explained by the resulting predicted values interaction with the smoothing parameter. The two focal areas of residual values are still present and are also seen in the bimodal residual distribution shown in Figure 5, (b). Finally, a scatter plot depicting the relationship between the observed and predicted values of month

Survival is shown in Figure 6. Two focal areas are seen yet again, with few data points falling along the diagonal, representing low residual values. These data together do not show support for any of the parametric and nonparametric analyses to accurately predict the number of months survived post-diagnosis with the currently provided data.

Developing model systems to assist in determining prognostic outlooks is important in updating survival, 5-year survival, and long-term outlook for patients with both local and metastatic disease. This study was able to describe the most prevalent patient characteristics in data acquired during the 1975-2016 time period with most patients being married, having unknown tumor grade classifications, undergoing tumor resection surgery, having no lymph, either regional or distant, surgeries, undergoing active care management, and are either alive or dead of another cause unrelated to the cancer incidence. Continuous variables thought to be involved in the predictability of the survival time after diagnosis were age of diagnosis, number of malignancies, tumor size, and total number of malignancies. Previous works have shown limitations in using demographic, diagnostic, and death related data as survivability predictors [8], [19]. Additionally, several studies have noted discrepancies in cause of death provided on a patient-to-patient basis, as well as increased racial bias due to dramatic variations in post-surgery or treatment follow-up in several other cancer types [20], [21]. Difficulties and complexities in developing predictive models for cancer survivability are continuing to be addressed and, through this study, further confirmation of the degree of prediction variability is shown when only using general cancer patient data. Additionally, statistical regressions have shown limitations in cancer survival predictions further compounding the poor predictability with generalized patient data, so it is recommended to instead explore machine learning applications to improve prediction reliability [22]–[24].

The data provided and analyzed are extremely general and do not report values that would be expected to have a significant effect on survivability. Factors not included in this study that would be expected to display significant effects are lymphovascular invasion rates, blood serum tumor marker levels, and hormone marker levels [23]–[25]. Studies including blood serum markers, such as erythropoietin, vascular endothelial growth factor, apolipoprotein, and high-density lipoprotein, have shown high levels of statistical significance on cancer survival impact in lung and biliary malignancies [26], [27]. Additionally, genetic information about patient tumors is invaluable when determining the possibility for recurrence, ability to be successfully treated, and determining the future treatment protocol and survival outlook [28], [29]. Genetic profiling can also make a large impact on developing and delivering precision medicine to help increase survival rates and drop rates of cancer recurrence in several tumor types [30]. Without these factors, models described in this work can be extremely limited in their predictive abilities.

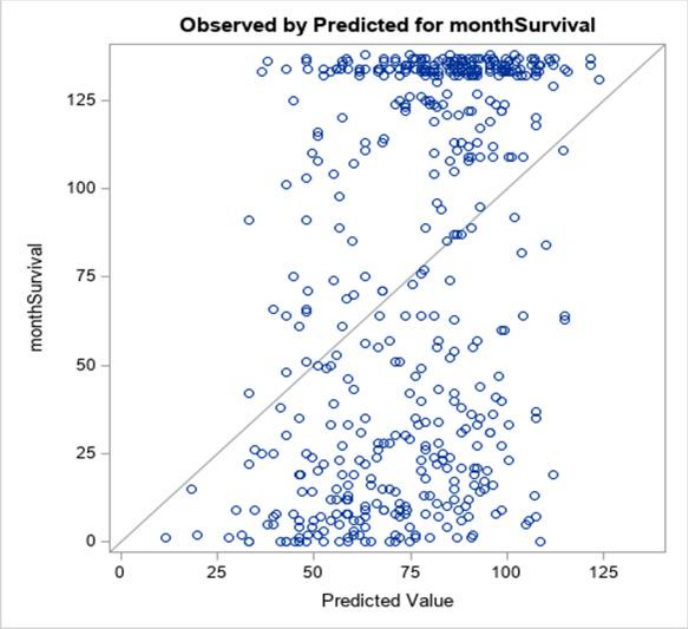

Figure 6. Observed and predicted monthSurvival values. Significant differences are thought to be brought about by observed 0 values and high monthSurvival values. As observed in residual plots and in the bimodal distribution, two focal areas are seen in the predictive plot

4. CONCLUSION

The models used in this system were only able to identify one factor providing statistical significance being the age of diagnosis. Other variables did not show significance and did not display significant contributions to the longevity of a patient's survival time. However, this must be not be taken as the entire picture of prognostics. The data acquired displayed significant variability and, because of this, data may be skewed to not report variables that may actually have an influence on the patient outlook. Multiple focal areas of residual values were seen across many regression models, including parametric and nonparametric methods. These areas were not able to be unified and linearized with a logarithmic transformation, so it is concluded that the models provided do not sufficiently explain the response variable of survival time after diagnosis of any gynecological cancer.

Optimization and further modifications to the provided model can be completed to better predict and minimize residual values for the survivability with the data provided by SEER. Expanding this work to a larger team could provide insights and additional tools to better process the current data file and alleviate some issues present in the regression system or implement robust machine learning algorithms. Additionally, it would be advantageous to include genetic information, if available, as well as other diagnostic information that is commonly used in predicting recurrence rates and survivability profiles including blood-serum tumor marker levels, lymphovascular invasion, hormone levels, and in-depth cytology and pathology reports for each patient. With more data and more detailed information, better models can be developed and produced for this patient population to assist with analyzing risk factors and determining prognostic values.

ACKNOWLEDGEMENTS

Special thanks to the National Institute of Health – National Cancer Institute for permissions to use the SEER database and acquire clinically relevant patient data including demographic and diagnostic information.

REFERENCES

[1] R. L. Siegel and K. D. Miller, "Cancer Statistics, 2021," *CA. Cancer J. Clin.*, vol. 71, no. 1, pp. 7–33, 2021, doi: 10.3322/caac.21654.
[2] J. Zugazagoitia, C. Guedes, S. Ponce, I. Ferrer, S. Molina-Pinelo, and L. Paz-Ares, "Current Challenges in Cancer Treatment," *Clin. Ther.*, vol. 38, no. 7, pp. 1551–1566, 2016, doi: 10.1016/j.clinthera.2016.03.026.
[3] A. Zarros *et al.*, "Evolution of Cancer Pharmacological Treatments at the Turn of the Third Millennium," *Front. Pharmacol*, vol. 9, pp. 1-26, 2018, doi: 10.3389/fphar.2018.01300.
[4] J. G. Tate *et al.*, "COSMIC: The Catalogue Of Somatic Mutations In Cancer," *Nucleic Acids Res.*, vol. 47, no. 1, pp. 941–947, 2019, doi: 10.1093/nar/gky1015.
[5] S. S. Faubion, K. L. Maclaughlin, M. E. Long, S. Pruthi, and P. M. Casey, "Surveillance and Care of the Gynecologic Cancer Survivor," *J. Women's Heal.*, vol. 24, no. 11, pp. 899–906, 2015, doi: 10.1089/jwh.2014.5127
[6] S. K. Srivastava *et al.*, "Racial health disparities in ovarian cancer: not just black and white," *J. Ovarian Res.*, vol. 10, no. 1, pp. 1-9, 2017, doi: 10.1186/s13048-017-0355-y.
[7] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics," *CA Cancer J Clin*, vol. 66, no. 1, pp. 7–30, 2016.
[8] N. Howlader, L. A. G. Ries, A. B. Mariotto, M. E. Reichman, J. Ruhl, and K. A. Cronin, "Improved estimates of cancer-specific survival rates from population-based data," *J. Natl. Cancer Inst.*, vol. 102, no. 20, pp. 1584–1598, 2010, doi: 10.1093/jnci/djq366.
[9] F. F. Costa, "Big data in biomedicine," *Drug Discov. Today*, vol. 19, no. 4, pp. 433–440, 2014, doi: 10.1016/j.drudis.2013.10.012.
[10] Y. Collins, K. Holcomb, E. Chapman-Davis, D. Khabele, and J. H. Farley, "Gynecologic cancer disparities: A report from the Health Disparities Taskforce of the Society of Gynecologic Oncology," *Gynecol. Oncol.*, vol. 133, no. 2, pp. 353–361, 2014, doi: 10.1016/j.ygyno.2013.12.039
[11] G. Taglang and D. B. Jackson, "Use of big data in drug discovery and clinical trials," *Gynecol. Oncol.*, vol. 141, no. 1, pp. 17–23, 2016, doi: 10.1016/j.ygyno.2016.02.022.
[12] E. Visser, I. A. Franken, L. A. A. Brosens, J. P. Ruurda, and R. van Hillegersberg, "Prognostic gene expression profiling in esophageal cancer: A systematic review," *Oncotarget*, vol. 8, no. 3, pp. 5566–5577, 2017.
[13] A. K. Das, S. Mishra, D. K. Mishra, and S. S. Gopalan, "Survival prediction for bladder cancer using machine learning: Development of BlaCaSurv online survival prediction application," *medRxiv*, pp. 1-19, 2020, doi: /10.1101/2020.11.13.20231191.
[14] K. Xiao, T. Lin, K. Lam, and Y. Li, "A facile strategy for fine-tuning the stability and drug release of stimuli-responsive cross-linked micellar nanoparticles toward precision drug delivery," *Nanoscale*, vol. 9, no. 23, pp. 7765–7770, 2017, doi: 10.1039/c7nr02530k
[15] F. M. Drawnel *et al.*, "Molecular Phenotyping Combines Molecular Information, Biological Relevance, and Patient Data to Improve Productivity of Early Drug Discovery," *Cell Chem. Biol.*, vol. 24, no. 5, pp. 624-634, 2017, doi: 10.1016/j.chembiol.2017.03.016.

[16] E. D. Karagiannis, C. A. Alabi, and D. G. Anderson, "Rationally designed tumor-penetrating nanocomplexes," *ACS Nano*, vol. 6, no. 10, pp. 8484–8487, 2012, doi: 10.1021/nn304707b.
[17] National Cancer Institute and DCCPS, "Surveillance, Epidemiology, and End Results (SEER) Program," [Online]. Available: www.seer.cancer.gov.
[18] National Cancer Institute, "Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1975-2016)," National Cancer Institute, DCCPS, Surveillance Research Program, released April 2019, based on the November 2018, [Online]. Available: www.seer.cancer.gov.
[19] B. A. Mahal *et al.*, "Incidence and determinants of 1-month mortality after cancer-directed surgery," *Ann. Oncol.*, vol. 26, no. 2, pp. 399–406, 2015, doi: 10.1093/annonc/mdu534.
[20] M. D. Pineda, E. White, A. R. Kristal, and V. Taylor, "Asian breast cancer survival in the US: A comparison between Asian immigrants, US-born Asian Americans and Caucasians," *International Journal of Epidemiology*, vol. 30, no. 5. pp. 976–982, 2001, doi: 10.1093/ije/30.5.976
[21] C. Zeng, W. Wen, A. K. Morgans, W. Pao, X. O. Shu, and W. Zheng, "Disparities by race, age, and sex in the improvement of survival for major cancers: Results from the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) program in the United States, 1990 to 2010," *JAMA Oncol.*, vol. 1, no. 1, pp. 88–96, 2015, doi: 10.1001/jamaoncol.2014.161.
[22] A. Kazarian *et al.*, "Testing breast cancer serum biomarkers for early detection and prognosis in pre-diagnosis samples," *Nat. Publ. Gr.*, vol. 116, pp. 501–508, 2017, doi: 10.1038/bjc.2016.433.
[23] A. Rodriguez, Z. Blanchard, K. Maurer, and J. Gertz, "Estrogen Signaling in Endometrial Cancer: a Key Oncogenic Pathway with Several Open Questions," *Horm Cancer*, vol. 10, no. 2–3, pp. 51–63, 2019.
[24] Y. J. Song, S. H. Shin, J. S. Cho, M. H. Park, J. H. Yoon, and Y. J. Jegal, "The role of lymphovascular invasion as a prognostic factor in patients with lymph node-positive operable invasive breast cancer," *Journal of Breast Cancer*, vol. 14, no. 3. pp. 198–203, 2011.
[25] Y. Ueda *et al.*, "Serum biomarkers for early detection of gynecologic cancers," *Cancers (Basel)*, vol. 2, no. 2, pp. 1312–1327, 2010, doi: 10.4048/jbc.2011.14.3.198.
[26] L. Sun *et al.*, "Integrated analysis of serum lipid profile for predicting clinical outcomes of patients with malignant biliary tumor," *BMC Cancer*, vol. 20, no. 1, pp. 1–14, 2020, doi: 10.1186/s12885-020-07496-8.
[27] R. Suwinski *et al.*, "Blood serum proteins as biomarkers for prediction of survival, locoregional control and distant metastasis rate in radiotherapy and radio-chemotherapy for non-small cell lung cancer," *BMC Cancer*, vol. 19, no. 1, pp. 1–12, 2019, doi: 10.1186/s12885-019-5617-1.
[28] J. E. Dancey, P. L. Bedard, N. Onetto, and T. J. Hudson, "The genetic basis for cancer treatment decisions," *Cell*, vol. 148, no. 3. pp. 409–420, 2012, doi: 10.1016/j.cell.2012.01.014.
[29] C. B. Meador and G. R. Oxnard, "Effective cancer genotyping-many means to one end," *Clin. Cancer Res.*, vol. 25, no. 15, pp. 4583–4585, 2019, doi: 10.1158/1078-0432.CCR-19-1233.
[30] E. R. Malone, M. Oliva, P. J. B. Sabatini, T. L. Stockley, and L. L. Siu, "Molecular profiling for precision cancer therapies," *Genome Med.*, vol. 12, no. 1, pp. 1–19, 2020, doi: 10.1186/s13073-019-0703-1.