Comparative Study of a Stochastic Pure Death Process and Cox Proportional Hazard Models in analyzing survival of breast cancer in Netherlands Women

by

Oladipo Afolayan, Jeremiah Akpabio, Sandile Ndabezitha, Francis Nkansah, Aaron Niecestro

Department of Biostatistics
School of Public Health, The University of Texas Health Science Center at Houston

PHD 1950: Stochastic Processes in Biostatistics

Dr. Wenyaw Chan

April 2023

Abstract

This study aimed to obtain a pure death process model and Cox proportional hazard model to predict the hazard rates for 272 women with breast cancer in the Netherlands. The death rates were calculated using the breast cancer patient data as crude and time-dependent rates as an initial value for mu (μ) to create the simulations in the pure death process. The pure death process ran ten thousand simulations to calculate the summary statistics (mean, standard deviation, standard error, and coverage probability) of the population size and death rates after approximately 18.4 years. These values were then compared to the Netherlands' actual estimates. We calculated the average yearly survival population for 19 years from the simulations to estimate the Kaplan-Meier and Nelson-Aalen estimators. Furthermore, this paper compared the efficacy of the pure death process and Cox proportional hazard models.

Introduction

The burden of cancer has continued to be on the rise in the last decade. Cancer is the second leading cause of non-infectious disease death globally, second only to cardiovascular disease causes. In 2020, the burden of cancer rose to 19.3 million new cases and over 10 million cancer-related deaths. In the same year, breast cancer became the second most diagnosed cancer globally with about 2.3 million new cases of the disease and about 685,000 mortalities reported in the same year. Globally breast cancer is the prime cause of death in females and the fifth cause of cancer-related death overall.[1] The onset of breast cancer has been highly linked to the rise in human development, with a significant correlation found in nations where there is a visible economic transition. The odds of having a disease and surviving are reduced significantly amongst those who reside in poorer and less developed countries compared to their counterparts in developed countries. Several factors that influence the disparity in prevalence and global survival rate include delay in disease diagnosis, distance and access to a healthcare facility, and paucity of effective treatment, adequate facilities, and personnel. The human development index (HDI) is a combined measure of life expectancy, education, and wealth which is a very useful comparative measure between countries besides just evaluating the country's income by itself. Countries with the highest levels of HDI are observed to have a higher incidence of breast cancer.

The current global age-standardized incidence rate in females is estimated to be 48 per 100,000 of the population, which ranges from under 30 per 100,000 in sub-Saharan Africa to over 70 per 100,000 in Western Europe and North America. Although the relative incidence of breast cancer is highest in the most developed regions of the world, much larger populations in less developed regions mean that over half of all breast cancer cases are diagnosed in low and middle-income countries, creating a significant burden to deal with the disease. The World Health Organization's new Global Breast Cancer Initiative was launched in 2022 to address the urgency of this global health challenge. The main goal of this initiative is to improve survival across the globe through three major strategies: health promotion, timely diagnosis, comprehensive treatment, and supportive care.

The Netherlands ranks third in countries affected by the burden of cancer in Europe, behind Ireland and Denmark compared to what is seen in other countries. Cancers of the lung, esophagus,

[1] "Breast Cancer - Statistics." *Cancer.Net,* ASCO.org, 23 Feb. 2023, https://www.cancer.net/cancer-types/breast-cancer/statistics.

and bladder are noted to be most prevalent amongst the Dutch compared to other Europeans. An estimated 14,000 Dutch women are diagnosed with invasive breast cancer every year, and 2,400 of those cases are breast cancer in situ. The mortality rate of breast cancer in the Netherlands stands at 3,000 per year, with 1 out of every 8 Dutch women being likely to develop breast cancer at one point time in their lifetime, making breast cancer the most prevalent cancer in the country. The average age that a woman is diagnosed with Breast cancer in the Netherlands is approximately 61 years of age. About 100 men are diagnosed with breast cancer every year in the Netherlands, however, treatment involving a multidimensional approach and the participation of healthcare personnel has proven effective in early detection, management, and adequate treatment.

The treatment of breast cancer depends on the stage and nature of the tumor and the prospects of survival. On average, over 87% of women diagnosed with breast cancer survive at least 5 years following this diagnosis and over 77% survive at least 10 years. The overall incidence of breast cancer, in the Netherlands, increased from 103.4 to 153.2 per 100,000 women between 1990 and 2014. The increase was driven by ductal carcinoma in situ (DCIS) and early breast cancer as the incidence of locally advanced and metastatic breast cancer remained stable. Between 1990 and 2014, the ten-year overall survival rate for patients with breast increased from 87% to 93% for those diagnosed with early breast cancer, 41% to 62% for locally advanced breast cancer, and 6% to 9% for those with metastatic disease breast cancer. Annually, breast cancer in the Netherlands is responsible for approximately 3,100 deaths, 26,000 life years of total life the patients could have lived lost, 65,000 Disability Adjusted Life Years (DALYs), and an economic burden of 1.27 billion euros.

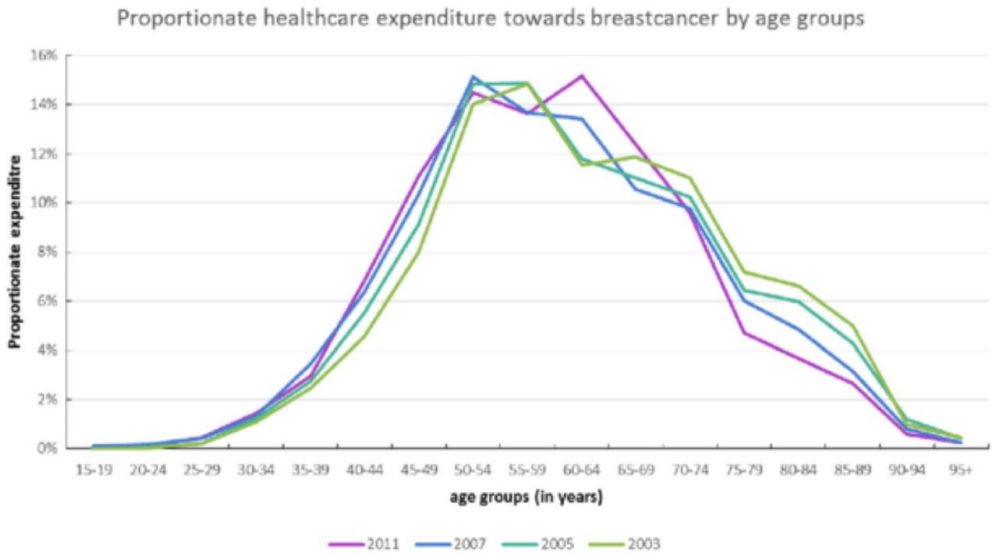

Figure 1: Proportional Healthcare expenditure towards Breast Cancer by Age Group[2]

Several studies have identified various risk factors for breast cancer among Netherlands women. These include age, family history of breast cancer, early onset of menstruation, late onset of menopause, nulliparity, and use of hormone replacement therapy. Additionally, lifestyle factors such as alcohol consumption, smoking, and physical inactivity have also been associated with an increased risk of breast cancer. One study found that women with a family history of breast cancer had a higher risk of developing the disease, with a relative risk of 1.8 (95% CI 1.4-2.3) compared to women without a family history. Breast cancer screening is an important tool for early detection and treatment. In the Netherlands, a national screening program is in place for women aged 50-75 years. Several studies have evaluated the effectiveness of this program and found that it has led to a reduction in breast cancer mortality rates. However, there are also concerns about overdiagnosis and overtreatment, particularly in women with low-risk tumors.

Despite decades of research, clinical trials, new drugs, chemotherapy, and advancement in medical treatment, cancer remains one of the major mortalities in the United States and the world at large. For instance, in the year 2023 about 605,213 people died of cancer and this number of deaths will only increase since the year 2023 is not complete yet. Some researchers over the years have used Cox Hazards models to study and predict survival and hazard rates of cancer patients. Some statisticians have studied and modeled cancer survival using stochastic pure death processes.

[2] The source of this graph comes from Vondeling, G.T., Menezes, G.L., Dvortsin, E.P. *et al.* Burden of early, advanced, and metastatic breast cancer in The Netherlands. *BMC Cancer* **18**, 262 (2018). https://doi.org/10.1186/s12885-018-4158-3, Figure 7

Netherland Dataset

The data set was obtained from the Netherlands Research Institute data which included 272 breast cancer patients, with the primary focus being a network built only using gene expressions, and survival time in years was calculated and observed. The observed data is presented below. Our study attempts to carry out a comparative analysis of three separate models which are as follows: 1) a Pure Death Process Model, 2) a Cox Proportional Hazards Model, and 3) A logistic regression Model. The purpose behind these three models is to analyze from different angles the hazard rate as the outcome of interest in breast cancer women in the Netherlands. Besides the pure death process which only uses time as the primary covariate the other two models, the Cox Proportional Hazards Model and the Logistic regression model will use other covariates to predict the probability of death.

Table 1: Censored Observation and Death Statistics

Data Type	Observation	Percentage (%)
Number of Censored Observation	195	71.59
Number of Deaths	77	28.31
Total	272	100

The range of the time that it took for a tumor to recur in a patient after treatment was observed to be from 0.271 to 18.3408 years. The value 18.3408 years is the length of our study's observational time then we had to conclude that when 18.3408 years was observed in the data, one of two outcomes happened. Either every patient had a tumor recurrence observed in the study or when patients were censored out of the study that censored time was the tumor reoccurrence. However, it is unable to determine which of these outcomes was correct since it was not labeled in the data description, and we received no attempts back from the original dataset producers.

The values for the Tumor grades were 1, 2, and 3 which were categorized based on the size of the tumor a patient had. Several treatments including but not limited to Chemotherapy, Hormonal intervention, and Amputation were also described. Age as a primary covariate was recorded, ranging from 25 to 53 years old, and yielded an average of 44.05 years. The data set contained 1555 columns of data containing gene expressions, however, for this paper, only the columns not including gene expressions were accessed and analyzed.

Model description

The three types of models we used in our analysis were the following: 1. A Pure Death Process Model, 2. A Cox Proportional Hazards Model, and 3. A Logistic Regression Model. The first model was the Pure Death Process model, $P(\ X(t+\triangle t)\ ) = k-1 \mid X(t) = k) = k\mu(t) \triangle t + o(t)$, with an exponential fore of mortality, μ(t) = μ. The reason this model was used was to determine and analyze the probability of death over time using stochastic techniques. The second type of model used in our analysis was the Cox Proportional Hazards Model, $h(t) = h_0(t)exp(\beta_1 x_1 + \ ... \ + \beta_p x_p)$. We used this model because we wished to not just analyze the hazard probability but also compare this model with the Pure Death Process model. This analysis was done purposefully because both the Pure Death Process model and the Cox Proportional Hazard Model measure the probability of death overtime with different statistical techniques in two different fields of biostatistics. The last type of model was the Logistic Regression model, $Logit(x) = \beta_0 + \beta_1 x_1 + \ ... \ + \beta_p x_p$. This model was used in our analysis for two reasons. The first is one can measure the odds ratio of death and convert this into the probability of death. The second reason being the pure death process is like 2 state Markov chains, so a logistic regression can be used to parametrize.

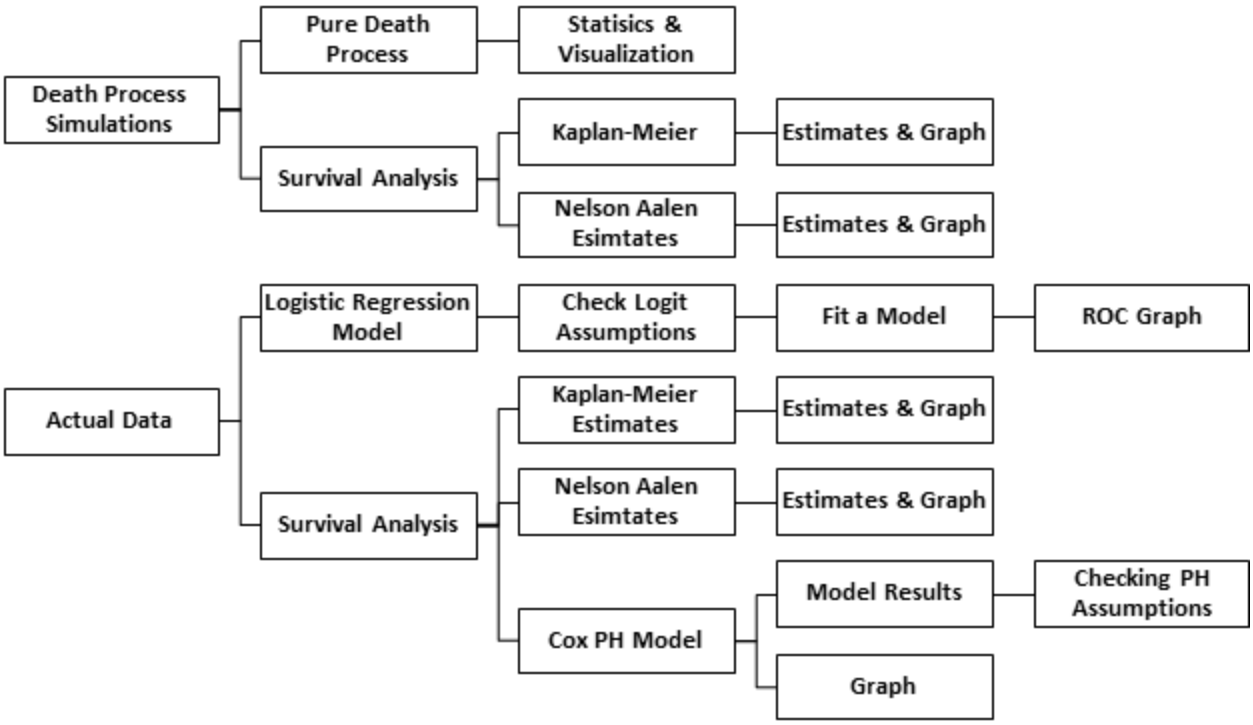

Figure 2: Flow Chart of Research

Methods

In this paper, we carried out several approaches to find a model estimation of our data in predicting the death process. We conducted 10,000 simulations of the pure death process, with several sample sizes derived, and obtained the summary statistics of these simulations. We also performed a survival analysis of the hazard rate of women with breast cancer in the Netherlands, using the Kaplan-Meier and Nelson-Aalen estimators to analyze the data. The real data was compared with the simulation obtained, and we finally used the Cox proportional hazard and Logistic regression models to optimize the model.

Simulation and Results

The tables and figures below give a summary of the conducted Pure death process and Survival Analysis simulation processes as well as an outline to the results of fitting these different models to the Netherlands dataset.

Table 2: Analysis Markers for Results

Categories	Numbers
Number of Simulations	10,000
Number of Simulation Years	18.34
Initial Population	272
True Death Rate	0.018
Significance Level	0.05

Displayed in Table 2 are the initial values used for the Pure death process simulation. We did some exploratory analysis and decided to use 272 as our population size. This may be a very small number for simulation, but we wanted our simulation process to have the same sample size so the results could be compared to the actual data without any issues. For the pure death process simulations, the true death rate used was 0.018 and the number of simulation years was 18.34. These simulations were repeated 10,000 times before obtaining the statistics.

Table 3: Pure Death Process Simulation Results for 0 to 18.4 Years

Time (Years)	True Death Rate	Estimated Death Rate	Standard Deviation	Standard Error	Coverage Probability
0-18.4	0.018	0.01747	0.001905	0.00002	0.988

Table 4: Poisson Death Process Simulation Results for Five-Year Intervals

Time (Years)	True Death Rate	Estimated Death Rate	Standard Deviation	Standard Error	Coverage Probability
0 to 5	0.0397	0.0331	0.00499	0.00005	0.971
6 to 10	0.01249	0.0114	0.00204	0.00002	0.947
11 to 15	0.005	0.0047	0.00107	0.00001	0.962

Displayed in Tables 3 and 4 are the simulation results for the Pure death process. Table 3 shows the simulation results for the 18.34 years duration. Using 0.018 as the true death rate, the estimated death rate was 0.0174, with a standard deviation and coverage probability of 0.001905 and 98% respectively. We were also interested in finding out what was happening between smaller chunks of time, and we conducted the simulation in 5-year intervals (results presented in Table 4). Because these intervals had different true death rates, it is not surprising that the estimated death rates were also different. All the coverage probabilities for the different intervals were above 90%.

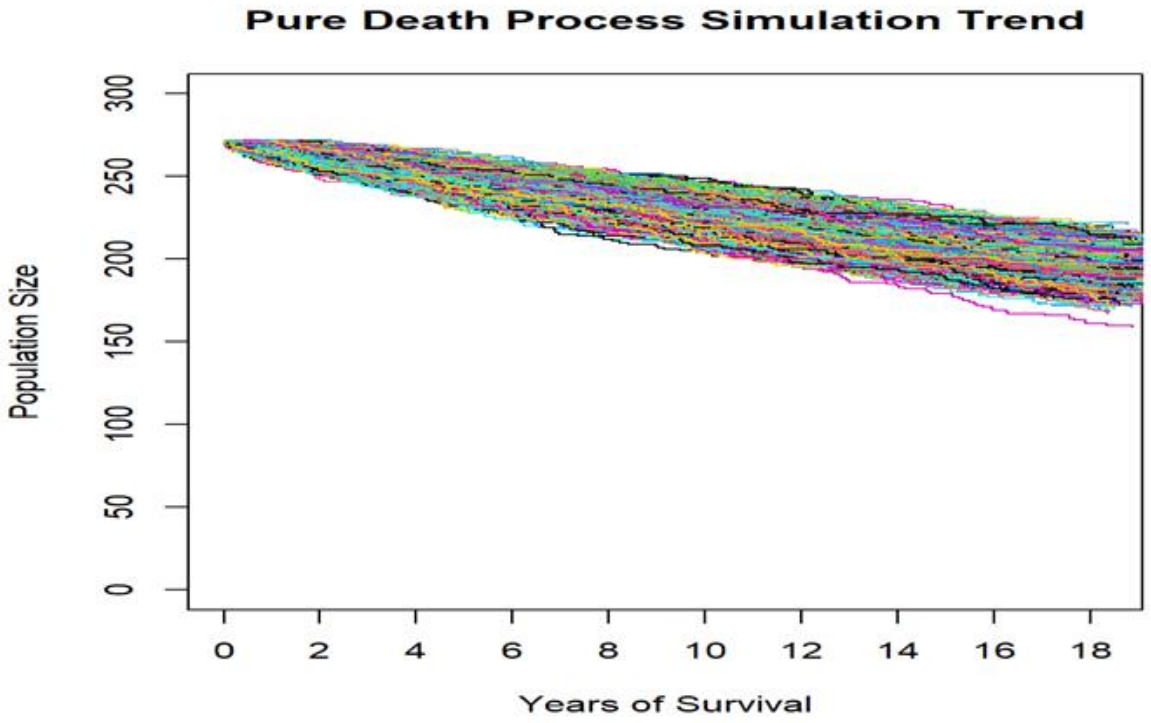

Figure 3: Pure Death Process - Remaining Population

Figure 3 visually displays the simulation process for 18.34 years, which shows that the remaining population decreased over the years. This is understandable because our model used was a pure death process, and we would have been worried if the population ever increased over the years.

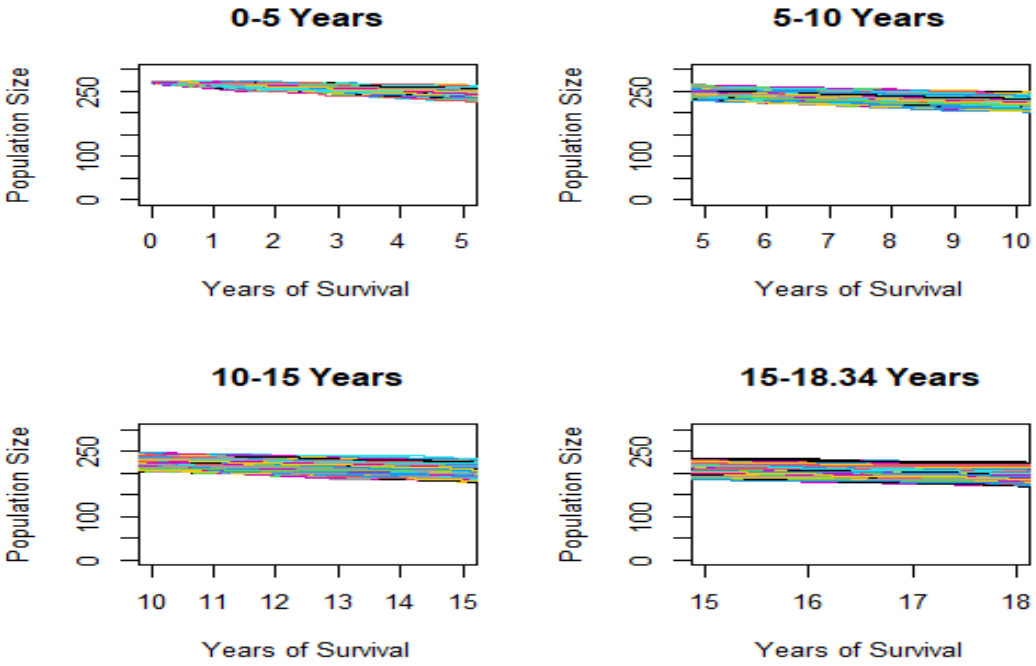

Figure 4: Pure Death Process - 5-Year Intervals

Figure 4 shows the simulation process, broken down into 5-year intervals. Again, the population decreased over time, the first 5 years decreasing at a higher rate, followed by a lower decreasing rate, and reaching almost a flat rate after 15 years.

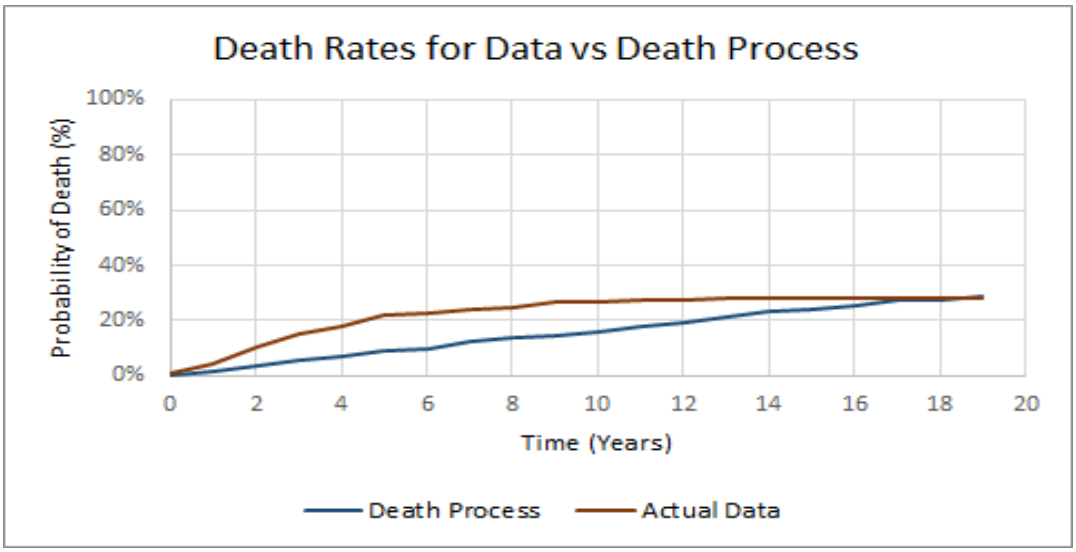

Figure 5: Pure Death Process Simulations

The blue line in Figure 5 depicts the cumulative average death rate from the remaining population at each time point across the simulation, whereas the brown line represents the cumulative change in the actual data. The rate of death seems to flatten out after the 14th year.

In Figure 6, the blue line portrays how the population size changes at 5-year intervals. There was a big drop in the population in the 5th year, but the number of deaths reduced drastically after the 15th year. The dotted line is a depiction of the negative linear trend.

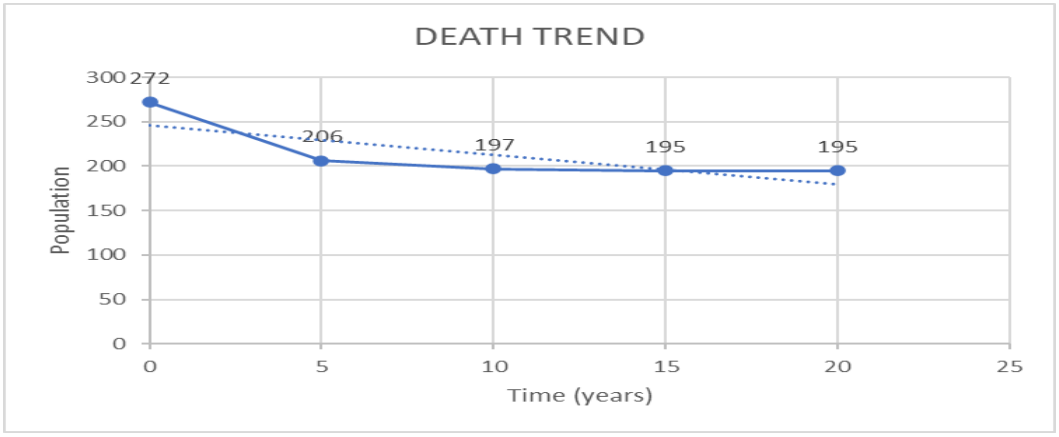

Figure 6: Linear Pure Death Trend

Upon computing the exponential force of mortality for the overall data, the linear death model is given as $P[Y(t) = k] = 0.018k \triangle t$. The data was then subdivided using a 5-year interval, and the death rates for each group were computed. The linear death models for these groups are given as follows:

- The first 5 years: $P[Y(t) = k] = 0.0397k \triangle t$
- Years 6 to 10: $P[Y(t) = k] = 0.01249k \triangle t$
- Years 11 to 15: $P[Y(t + \triangle t) = k - 1 | Y(t) = k] = 0.005k \triangle t$

The death rates in each of the models decrease as time increases. The lower the death rate the lower the number of deaths.

Logistic Regression Model

The logistic regression model was adopted for the pure death model because studies have shown that logistic regression is a good estimator of mortality[3]. In 2019 Minwoo Chan et al[4], not only did they use the logistic regression model to determine transition probabilities, but they also performed error analysis and concluded that logistic regression performed marginally better than traditional methods when estimating the transition probabilities matrix when limited data are accessible.

Table 5: Logistic Regression Model Covariates and their Statistics

Parameter	**Coefficient**	**Exp(Coefficient)**	**Standard Deviation**	**P-value**
Intercept	1.06099	2.89	0.70076	0.13
Time of Tumor Recurrence (Time Rec)	-0.62604	0.53	0.08441	< 0.0001
Grade 2 (Gr2)	1.48784	4.43	0.64553	0.0212
Grade 3 (Gr3)	2.26292	9.61	0.6438	0.0004
The Probability of Death				
$P(Y=1) = \frac{Exp(1.06099 - 0.62604 * Time\ Rec + 1.4878 * Gr2 + 2.2629 * Gr3)}{1 + Exp(1.06099 - 0.62604 * Time\ Rec + 1.4878 * Gr2 + 2.2629 * Gr3)}$				

The logistics regression model is given as:

$$Logit\left(\frac{\pi}{1-\pi}\right) = 1.06099 - 0.62604 * TimeRec + 1.48784 * \mathrm{Gr2} + 2.26292 * \mathrm{Gr3}$$

The dependent variable in this logistic regression model is a function of the probability that a particular observation for a patient will either be alive or dead. Holding the tumor grade constant,

[3] Ho LST, Xu J, Crawford FW, Minin VN, Suchard MA. Birth/birth-death processes and their computable transition probabilities with biological applications. J Math Biol. 2018 Mar;76(4):911-944. doi: 10.1007/s00285-017-1160-3. Epub 2017 Jul 24. PMID: 28741177; PMCID: PMC5783825. Stochastic modeling of bridge deterioration using classification tree and logistic regression Minwoo Chang, Marc Maguire, Yan Sun Journal of Infrastructure Systems 25 (1), 04018041, 2019

[4] Minwoo Chang, Marc Maguire

we will see 47% decrease in dying for a one-year increase in time of tumor. Furthermore, holding the time of tumor recurrence and tumor grade 2 constant, the odds of dying among participants with grade 3 tumor is 860% higher than the odds of dying among participants with grade 0 tumor.

From Figure 7 the area under the curve AUC is equal to 92.3% which means that the model fits the data well and will correctly classify the observation as the right categories (alive or dead).

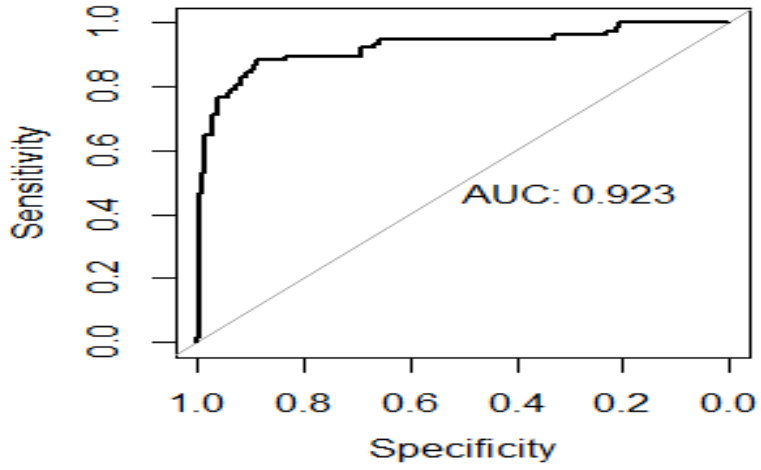

Figure 7: Logistic Regression Model Area Under the Curve

Simulations versus Real Data for Death Rates using Survival Analysis

Once the pure death process simulations were completed for the end-of-year population and pure death process model. We used this simulated data to conduct some simple survival analysis to analyze how mixing two different statistical techniques from two different fields with the same outcome of interest would impact our analysis. Namely we were looking for a comparison between the simulations with the actual data as well as obtain more information about the survival probability and hazard rates before building a Cox Proportional Hazard Model. The survival analysis tools used were the Kaplan-Meier and Nelson-Aalen estimates. The reason both the Kaplan-Meier and Nelson-Aalen estimates were used was because there was no conclusion on the debate within our group about whether 272 observations should be considered as a small sample size. So, a compromise was made where both estimates would be used in our analysis since it would provide more information without negatively impacting our results. After obtaining these estimates we found that the difference between them is negligible and statistically insignificant. (See Table 7 and Figures 8-11)

Table 6: Kaplan-Meier and Nelson-Aalen Death Rates

	Simulation[5]		Actual Data		Difference[6]	
Year	**Kaplan-Meier**	**Nelson-Aalen**	**Kaplan-Meier**	**Nelson-Aalen**	**Kaplan-Meier**	**Kaplan-Meier**
0	0.0000	0.0000	0.0000	0.0000	0	0
3	0.0551	0.0546	0.1066	0.1039	0.0515	0.0493
6	0.0993	0.0984	0.2169	0.2121	0.1176	0.1137
9	0.1471	0.1458	0.2500	0.2452	0.1029	0.0994
12	0.1949	0.1932	0.2721	0.2672	0.0772	0.074
15	0.2390	0.2370	0.2831	0.2783	0.0441	0.0413
18	0.2721	0.2698	0.2831	0.2783	0.011	0.0085

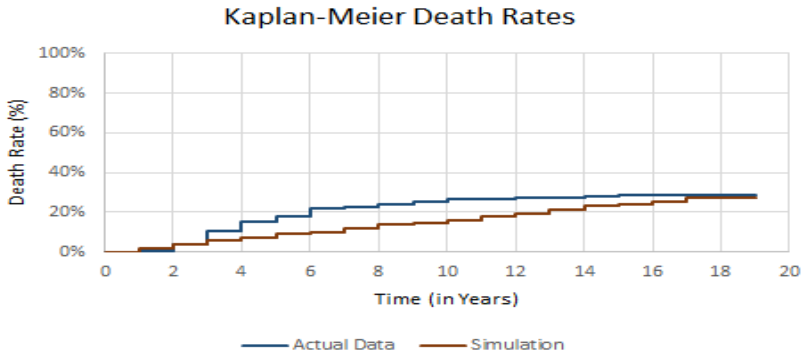

Figure 8: Kaplan-Meier Death Rates for Simulation versus Actual Data

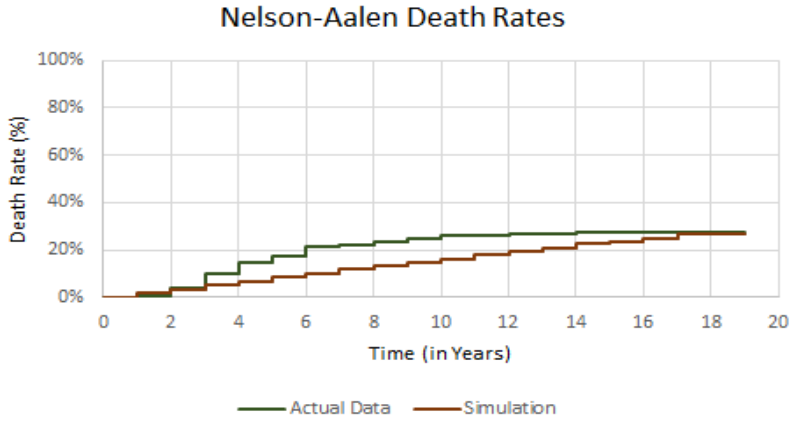

Figure 9: Nelson-Aalen Death Rates for Simulation versus Actual Data

[5] To get the numbers in the table first the pure death process simulation for the population at the end of each year through 10,000 simulations was recorded. These numbers were then averaged early and used to calculate the Kaplan-Meier and Nelson-Aalen estimates you see in the table.
[6] This is the absolute value difference between the simulation and actual data Kaplan-Meier Nelson-Aalen death rate estimates.

In Table 7 the results between the pure death process simulated data and the actual data Kaplan-Meier and Nelson-Aalen estimates are very close to one another. The difference between the simulation and actual data estimates starts out as a very large gap where the simulation is approximately a little more than half of the actual data hazard rates. However, as time goes by the difference starts to get smaller until the end where it is approximately 0.011[7], and 0.0085, respectively. In conclusion although the simulation could not accurately predict the death rate for almost half the observational time in the beginning[8], by the end of our observation time it did do a good job of predicting the estimates since the difference in the two estimates was very tiny and statistically insignificant.

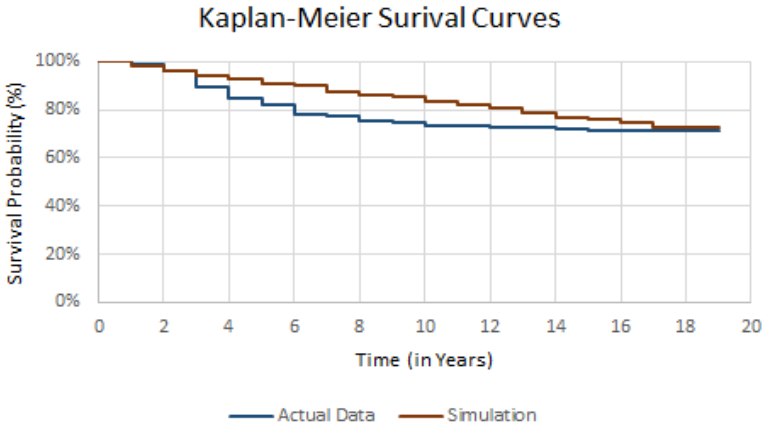

Figure 10: Kaplan-Meier Survival Probability for Simulation versus Actual Data

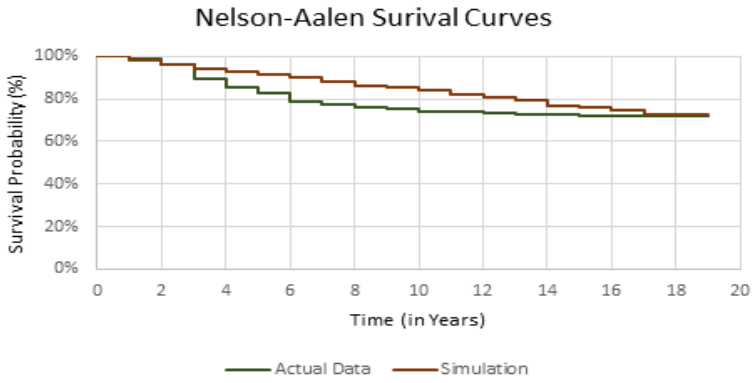

Figure 11: Nelson-Aalen Survival Probability for Simulation versus Actual Data

[7] These numbers are from Table 6
[8] See Figures 8 & 9

Through Figures 9 and 10, one can see that the survival probability decreases over time. However, the simulation survival probability decreases at a much slower rate than the observed data survival probability does. It should also be noted that after 15 years the observed survival probability does not decrease any further which was not captured in the data from the pure death process simulation. The cause for this difference in the simulation and actual data Kaplan-Meier and Nelson-Aalen estimates has to do with the pure death process predicting the death rate to be linearly decreasing until the last observation time while Kaplan-Meier and Nelson-Aalen estimate the survival probability directly from the data. This was of no cause of concern for us in our analysis and was expected to happen since at this current time we have no way of letting the pure death process be linearly decreasing from 0 to 15 years and stop from 15-18 years without jeopardizing the analysis.

Cox Proportional Hazards Model

The Cox Proportional Hazards model is a semiparametric model that makes no assumptions about the form of h(t) (nonparametric part of model) and assumes the parametric form for the effect of the predictors on the hazard. Many studies have conducted breast cancer and cancer research using a Cox Proportional Hazards Model for predicting the hazard rate and survival probability of patients with different parameters of interest. Since hazard rates were the focus of our research, using a Cox Proportional Hazards model was best suited to our needs. Before building the model though we conducted a very rigorous and extensive analysis to build our model while emphasizing statistical significance or the variables and proportional hazards assumption hold true. All this was conducted using the following steps: 1) Which variables need transformation through martingale residual plots, 2) which variables stay in the model through variable selection using AIC, BIC, and hypothesis testing, 3) checking whether the Proportional Hazards model assumptions holds for the individual variables and overall model, 4) examining the overall fit of the Cox PH model through a Cox-Snell residual plot, 5) checking for outliers using Deviance residual plot, and 6) checking the influence of the outliers and data points using transformed score residual. The Cox Proportional Hazards model we built was as follows:

$$h(t) = -1.5998 * \text{Log}(\text{Time Rec}) + 1.2082 * \text{Gr2} + 1.5915 * \text{Gr3}.$$

Although we did find outliers which had a high influence rate, we believed that this was caused by the small sample size we had in our data. This was noted in our analysis report, but nothing was done to these influential points since we thought it would compromise our research objective.

Table 7: Cox Proportional Hazards Model

Risk Factor	Parameter Estimate	Hazards Ratio	Standard Error	P-value	Confidence Interval
Log(Time of Tumor Recurrence)[9]	-1.5998	0.2019	0.1175	< 2e-16	(0.1604, 0.254)
Grade 2 (Gr2)	1.2082	3.3473	0.5424	0.0259	(1.1561, 9.6916)
Grade 3 (Gr3)	1.5915	4.9113	0.5384	0.0031	(1.7096, 14.1091)
Model Fit					
Concordance (C-Statistic)	0.949 (+/- 0.008)				

One could interpret the Cox Proportional Hazard Model after adjusting for all other variables by the following:

1. The relative risk associated with a log transformation of time to a tumor recurrence with a 1-year increase is 0.2019.
2. The hazard ratio for patients with a tumor grade of 2 relative to a patient with tumor grade of 1 is 3.3473, which indicates that the probability of a patient dying from a grade 2 tumor is approximately 3.473 times more than a patient with a grade 1 tumor.
3. The hazard ratio for patients with a tumor grade of 3 relative to a patient with tumor grade of 1 is 4.9113, which indicates that the probability of a patient dying from a grade 2 tumor is approximately 4.9113 times more than a patient with a grade 1 tumor.

Therefore, the longer the patient does not have the tumor reoccur, the higher probability of them living another each increase but as the tumor grade increases the higher probability, they are likely to die increases too. In conclusion the Cox Proportional Hazards model predicts the hazard rate well at 94.9% with a standard error of 0.8%.

[9] This is abbreviated as Log(Time Rec)

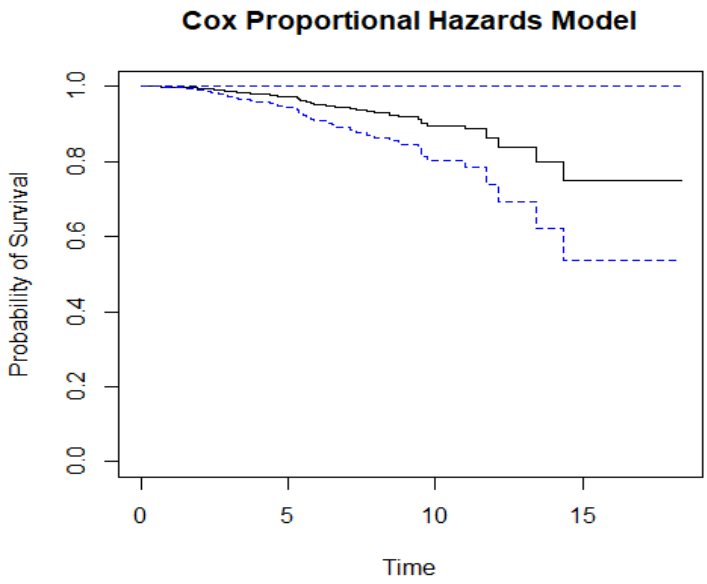

Figure 12: Cox PH Model Graph

The survival probability has a slow decrease at the beginning between 0 and 5 years and starts to have a sharper decrease between 5 to 10 years. After 10 years the survival probability drastically decreases where it remains constant at approximately 78% after 15 years. Overall, the survival probability is very high, which could be caused by the average of the patient's age being approximately 44.05 years old. This age is in the lower quartile for which women are either checking for breast cancer or having been diagnosed with breast cancer.

Discussion

Our Pure Death process simulations had over 90% coverage probability, showing the strength and usefulness of the model. The logistic regression had an Area under the curve (AUC) of 92.3% while the COX proportional hazard model had a C-Statistic of 94.9%. Based on these results, we can infer that our models predict the deaths and survival rates well enough.

Both Pure death process and Survival analysis predicted higher death rates in breast cancer patients for the first years, and later lower rates as the years increase. This may be caused by many factors including time-to-case finding/diagnoses of cancer as well as treatment plans. There is a need to strengthen such areas to increase the survival of cancer patients.

To our knowledge, this is the first study to use a Pure death process to explain survival functions. Also, we utilized the strengths of three different models to explain the death and survival rates of breast cancer. This also shows the novelty of our study.

Due to time constraints, we could not create an R code to cater for a full analysis of our data using the pure death process. This then led us to use the logistic model as an approximation to the pure death process. As much as this process was backed by literature, we would like to fully utilize the Pure death process approach. We plan to include more covariates such as age, treatments, and recurrence index in the future to try and come up with more precise predictions and inference. Our dataset had only 272 patients, which we considered to be a small sample. With several breast cancer research data all over the world, attempts to use other survival data with larger sample size would be more encouraged. As much as the Cox PH models is widely used and accepted as the gold standard when it comes to survival analysis, we could be happy to do a comparison of these stochastic models and the Cox PH model to come up with newer techniques of analyzing survival data, as well as adding to the body of knowledge.

Conclusion

Based on our analysis, although the Cox PH Model outperformed the logistic regression model with their statistics (AUC, and C-statistic), we believe that the pure death process model approach using logistic regression was more efficient and better than the survival model. This is because, in the Cox PH model, the tradeoffs of taking away the treatments and the subject's age variables had a much higher and steeper price than what we originally thought. We also should have evaluated the sample size of 272 patients with a higher weight and taken this into account for our analysis much more seriously than we originally planned.

References

1. Bray F, Laversanne M, Weiderpass E, Soerjomataram I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer. 2021;127(16):3029–30.
2. Pocock SJ, Gore SM, Kerr GR. Long term survival analysis: The curability of breast cancer. Stat Med. 1982;1(2):93–104.
3. Neyman J. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, June 30-July 30,l960. University of California Press; 1961. 436 p.
4. Lum, P., Singh, G., Lehman, A. *et al.* Extracting insights from the shape of complex data using topology. *Sci Rep* **3**, 1236 (2013). https://doi.org/10.1038/srep01236
5. Nicolau, Monica, et al. "Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival." *Cloudfront.net*, Proceedings of the National Academy of Sciences of the United States of America, 11 Apr. 2011, https://d1bp1ynq8xms31.cloudfront.net/wp-content/uploads/2015/02/Topology_Based_Data_Analysis_Identifies_a_Subgroup_of_Breast_Cancer_with_a_unique_mutational_profile_and_excellent_survival.pdf.
6. Chang J, Chan HK, Lin J, Chan W. Non-homogeneous continuous-time Markov chain with covariates: Applications to ambulatory hypertension monitoring. Stat Med. 2023 Mar 10. doi: 10.1002/sim.9707. Epub ahead of print. PMID: 36896833.
7. Niese, Birgit. "A Martingale Characterization of Polya-Lundberg Processes." *Journal of Applied Probability*, vol. 43, no. 3, 2006, pp. 741–54. *JSTOR*, http://www.jstor.org/stable/27595769. Accessed 10 Apr. 2023.
8. Pfeifer, Dietmar, and Ursula Heller. "A Martingale Characterization of Mixed Poisson Processes." *Journal of Applied Probability*, vol. 24, no. 1, 1987, pp. 246–51. *JSTOR*, https://doi.org/10.2307/3214076. Accessed 10 Apr. 2023.
9. Barraza NR, Pena G, Moreno V. A non-homogeneous Markov early epidemic growth dynamics model. Application to the SARS-CoV-2 pandemic. Chaos Solitons Fractals. 2020 Oct;139:110297. doi: 10.1016/j.chaos.2020.110297. Epub 2020 Sep 18. PMID: 32982083; PMCID: PMC7500902.
10. Korosteleva, O. (2022). Stochastic Processes with R: An Introduction (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003244288
11. Fontes, Luiz Renato, and Rinaldo B. Schinazi. "Implosion of a Pure Death Process." Physica A 523 (2019): 1171–1174. Web.
12. Barraza NR, Pena G, Moreno V. A non-homogeneous Markov early epidemic growth dynamics model. Application to the SARS-CoV-2 pandemic. Chaos Solitons Fractals. 2020 Oct;139:110297. doi: 10.1016/j.chaos.2020.110297. Epub 2020 Sep 18. PMID: 32982083; PMCID: PMC7500902.
13. Korosteleva, Olga. "Birth-and-Death Process." Stochastic Processes with R. United Kingdom: CRC Press LLC, 2022. Print.
14. Ma J, Chan W, Tsai CL, Xiong M, Tilley BC. Analysis of transtheoretical model of health behavioral changes in a nutrition intervention study--a continuous time Markov chain model with Bayesian approach. Stat Med. 2015 Nov 30;34(27):3577-89. doi: 10.1002/sim.6571. Epub 2015 Jun 29. PMID: 26123093; PMCID: PMC4626363.

15. Ho LST, Xu J, Crawford FW, Minin VN, Suchard MA. Birth/birth-death processes and their computable transition probabilities with biological applications. J Math Biol. 2018 Mar;76(4):911-944. doi: 10.1007/s00285-017-1160-3. Epub 2017 Jul 24. PMID: 28741177; PMCID: PMC5783825.
16. Stochastic modeling of bridge deterioration using classification tree and logistic regression Minwoo Chang, Marc Maguire, Yan Sun Journal of Infrastructure Systems 25 (1), 04018041, 2019
17. Siddiqui, A., Siddiqui, A., Maithani, S., Jha, A. K., Kumar, P., & Srivastav, S. K. (2018). Urban growth dynamics of an Indian metropolitan using CA Markov and Logistic Regression. *The Egyptian Journal of Remote Sensing and Space Science*, *21*(3), 229-236.
18. Verkooijen HM, Peeters PH, Buskens E, et al. Breast cancer screening in the Netherlands: a concise history of its introduction and implementation. Breast Cancer Res Treat. 2018;168(2):247-254. doi:10.1007/s10549-017-4627-5
19. Van den Broek JJ, van Ravesteyn NT, Mandelblatt JS, et al. Comparing CISNET Breast Cancer Incidence and Mortality Predictions to Observed Clinical Trial Results of Mammography Screening from Ages 40 to 49. Med Decis Making. 2020;40(1):3-14. doi:10.1177/0272989X19892608
20. Van den Berg MM, van der Pol CC, van den Broek JJ, et al. Overdiagnosis and overtreatment of breast cancer in the Netherlands: a population-based study of incidence trends and identification of determinants. Eur J Cancer Prev. 2020;29(2):97-104. doi:10.1097/CEJ.0000000000000529
21. Van de Water W, Bastiaannet E, Dekkers OM, et al. Adjuvant chemotherapy and cognitive function in breast cancer patients: a systematic review. Crit Rev Oncol Hematol. 2018;126:181-191. doi:10.1016/j.critrevonc.2018.04.011
22. Van den Berg MM, van der Pol CC, van den Broek JJ, et al. Breast cancer treatment in the elderly: a population-based study in the Netherlands. Breast Cancer Res Treat. 2019;177(3):677-686. doi:10.1007/s10549-019-05323-9
23. "Breast Cancer - Statistics." *Cancer.Net*, ASCO.org, 23 Feb. 2023, https://www.cancer.net/cancer-types/breast-cancer/statistics.