Data Analysis

We used three analytical tools – survival analysis, logistic regression, and a classification tree. The dataset used was the SEER (Surveillance, Epidemiology, and End Results) dataset [1], in which cancer incidence data has been compiled for ten regions across the country for multiple years. The data dictionary and the data use agreement for the SEER data is included in the "survival.zip" file. A cBioPortal dataset was also used for one analysis representing a clinic in the story [2].

Limitations of the SEER dataset:

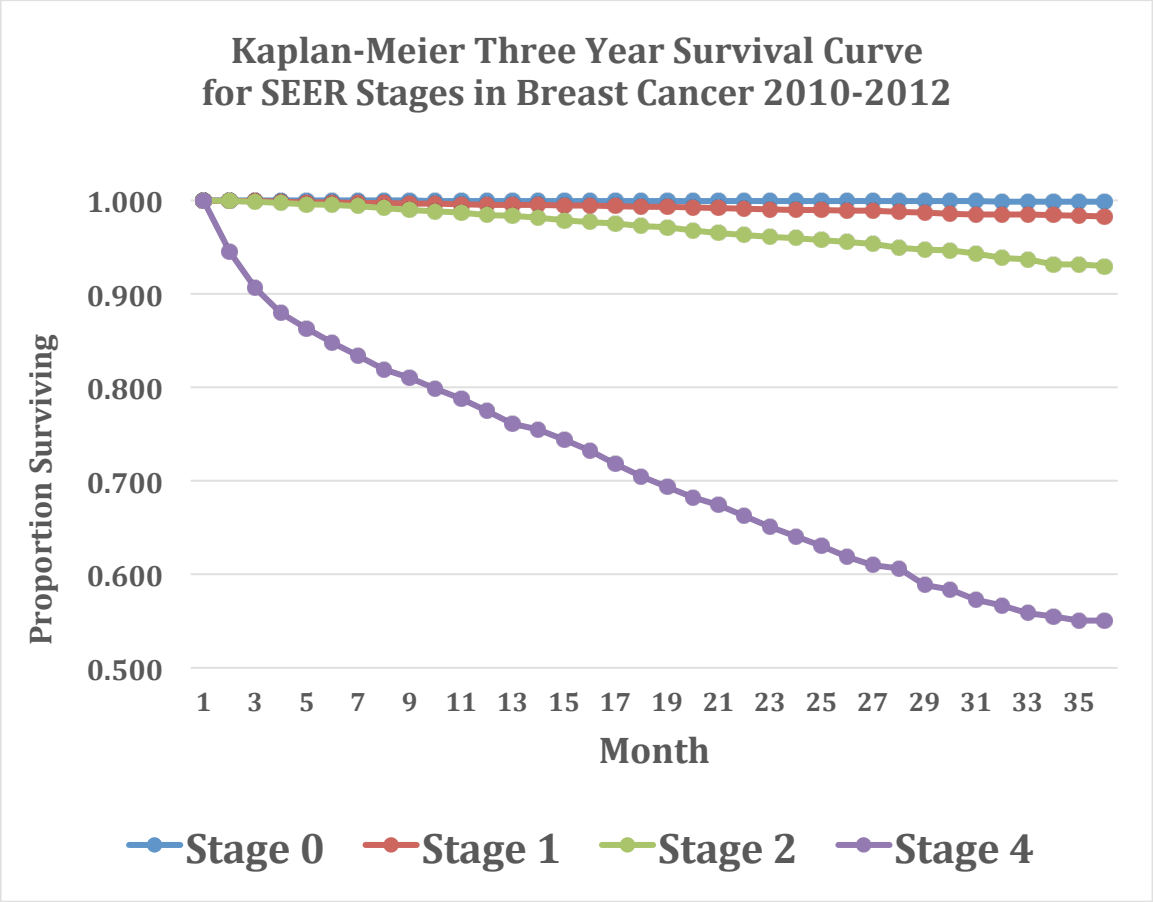
The data we used has certain limitations that may affect our results. This data is only for the years 2010-2012. While the SEER dataset has data starting from 1973, we only used data from 2010 to 2012, since some important variables, such as subtype of tumor, were recorded only from 2010 onwards. Another variable, the size of the tumor, has only been recorded since 2007. Additionally, much research and new treatments have been developed in recent years in breast cancer that make the newer data more relevant. Another limitation is that the SEER dataset is only collected for 10 regions across the US. Therefore, it is possible that, if this data were collected for more regions, other variables may become significant or show a relationship with survival.

Kaplan-Meier Survival Analysis

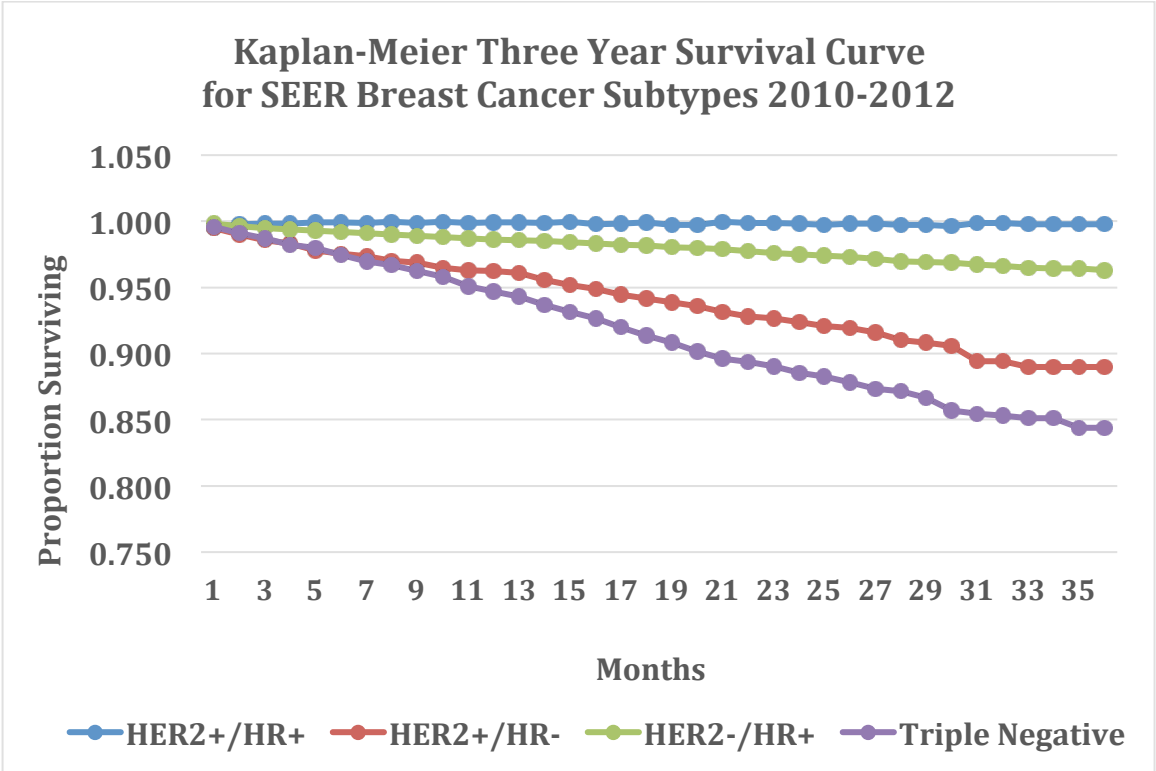
Using the SEER dataset for 2010-2012 and the R "survival" package, Kaplan-Meier exploratory survival analyses were performed for 19 variables (see survival.zip), but the following discussion highlights the results for selected variables of particular interest. This type of analysis uses two primary variables: (1) Event – either death, or still alive (0, 1 variable); and (2) Time – number of months of survival. These variables were then combined with the grouping variables listed into a csv file. A cBioPortal dataset, grouped by tumor subtype into a csv file, was also used to represent the cancer clinic in the story. The survival analysis method is capable of handling the entry and departure of individuals into the analysis as different times, as new patients are diagnosed, deaths occur, and individuals are lost from the study, as occurs in the SEER database. The three year survival curve displays the proportion of patients surviving at each point along with curve, and compares the survival rates of the various groups using the Log Rank statistical test. All of the group comparisons highlighted here were significant at the .01 level or better.

Stage of Breast Cancer: For the stage of cancer, there are only four levels: (0) in situ, or non-invasive tumor; (1) localized tumor, which has not spread to the lymph nodes; (2) regional tumor, which has spread to the lymph nodes, but has not yet spread to distant organs; and (4) tumor has spread to distance organs, or metastasized. The SEER dataset does not include any data at all for Stage 3 survival, as presumably this stage is not well defined, being somewhere in between stages 2 and 4. The survival curve for the SEER dataset is steepest for the Stage 4 patients,

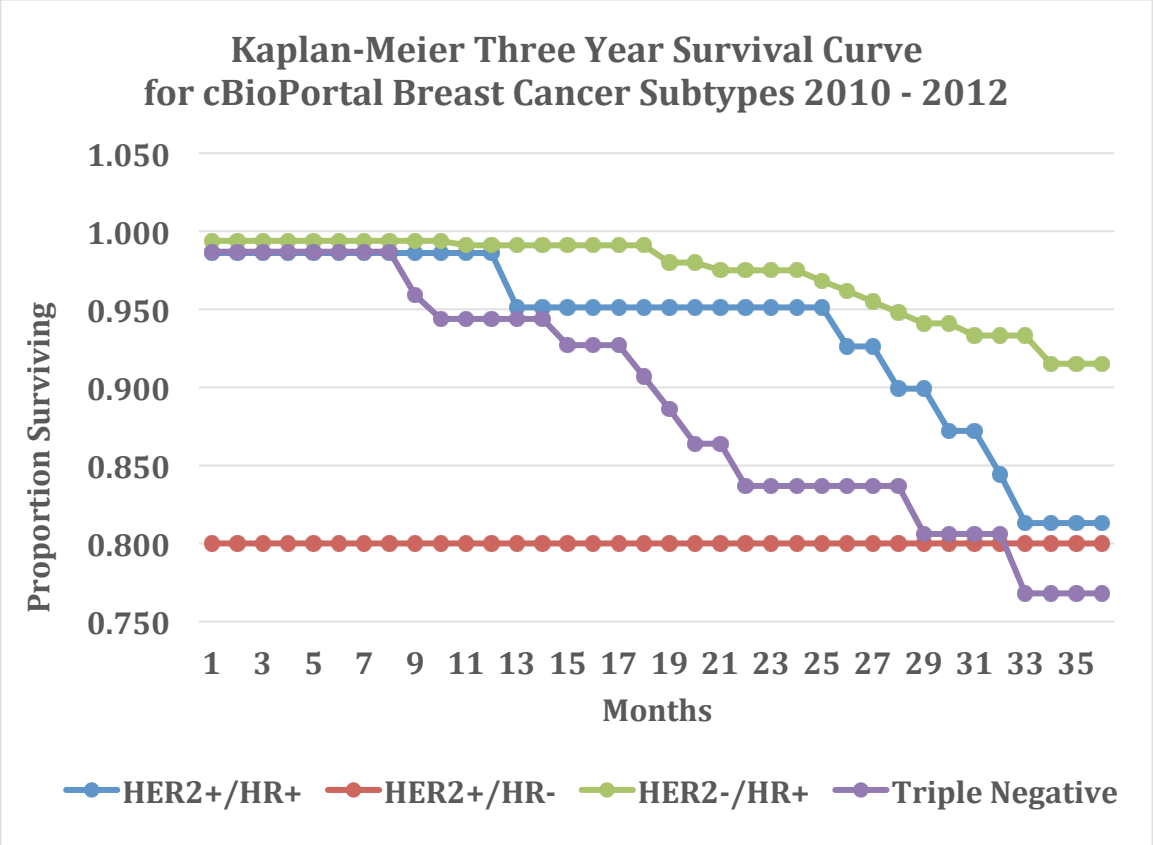
who show the lowest proportion surviving after three years (55%). The other three groups show much better three year survival rates (above 90%), as shown below:



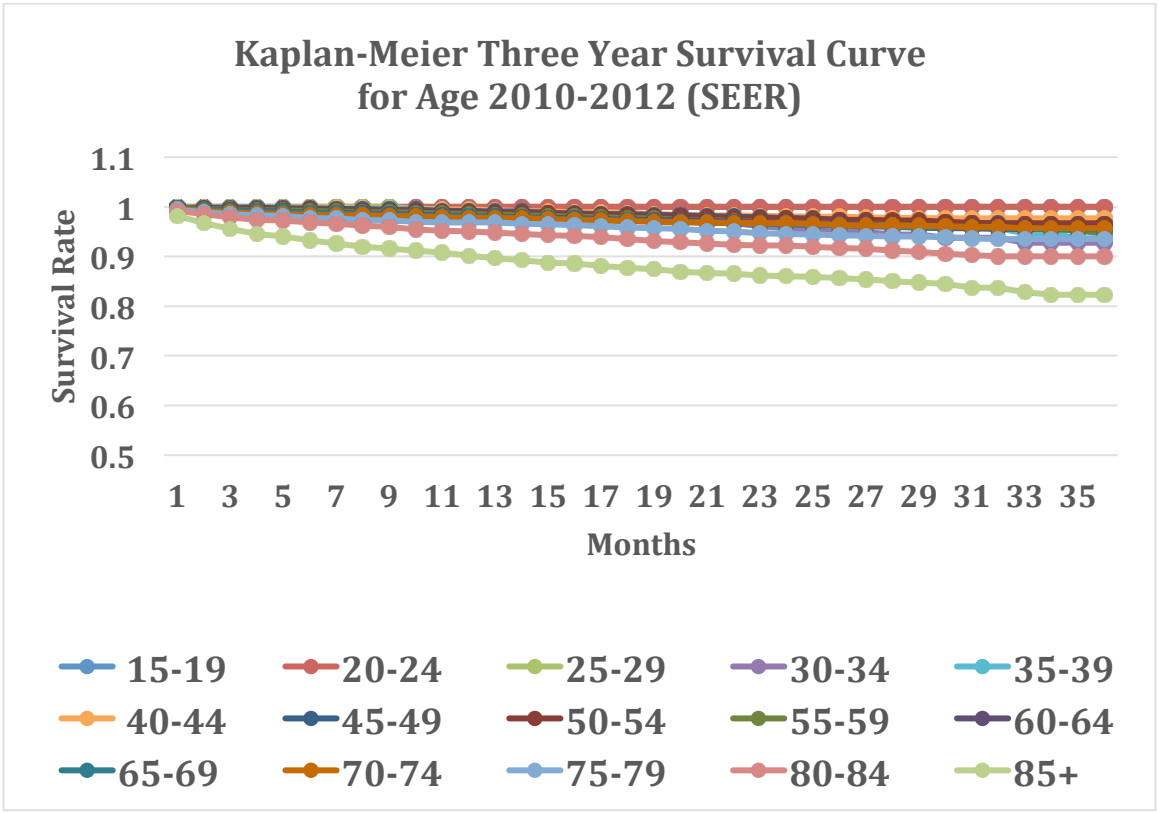
Breast Cancer Tumor Subtype: One factor which has emerged as important in breast cancer is the subtype of the tumor. Tumor subtypes have the following values: (1) HER2+HR+; (2) HER2+HR-; (3) HER2-HR+; (4) Triple Negative. Tumor mutations with the HER2 protein, called the "human epidermal receptor 2", promote the growth of tumor cells. Tumors which receive signals from hormones (estrogen or progesterone), which also promote growth of tumor cells, are called HR ("Hormone Receptor") positive. Triple negative patients have tumor cells which are neither HER2 nor HR positive. Targeted therapies have been developed for both HER2 and HR positive tumor subtypes, but not for Tripe negative tumors. Therefore, the triple negative tumors are much more difficult to treat. Triple negative patients, as would be expected, show a poorer three year survival rate (84%) than the other three groups. The group which is both HER2 and HR positive shows the highest survival rate (99%).



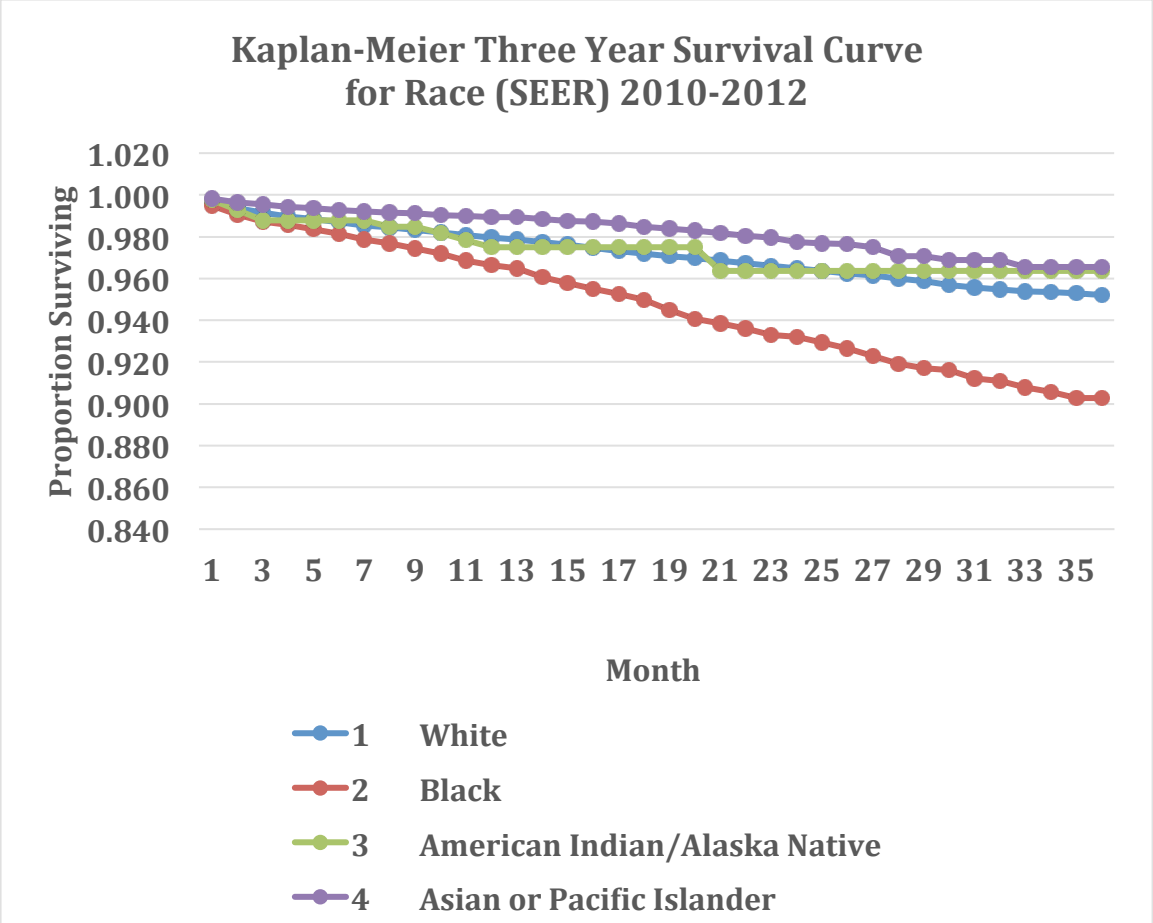
If these results are compared with a clinic (cBioPortal dataset), the results are quite different, as shown below. The clinic shows an even lower level of survival for triple negative patients (77% vs 84%), and for HER2+ cases (< 85%). This type of comparison would be useful for clinics to give them indications of challenges to work on. It should be noted that there was only one patient in this dataset with HER2+HR- tumor, so that subtype result is not reliable.



Breast Cancer by Age: As might be expected, the three survival curve for breast cancer is lower for higher ages. However, what is surprising is that the age groups under 75 years all show survival curves above 95%, and the decline does not appear to begin until patients are 75 or older.

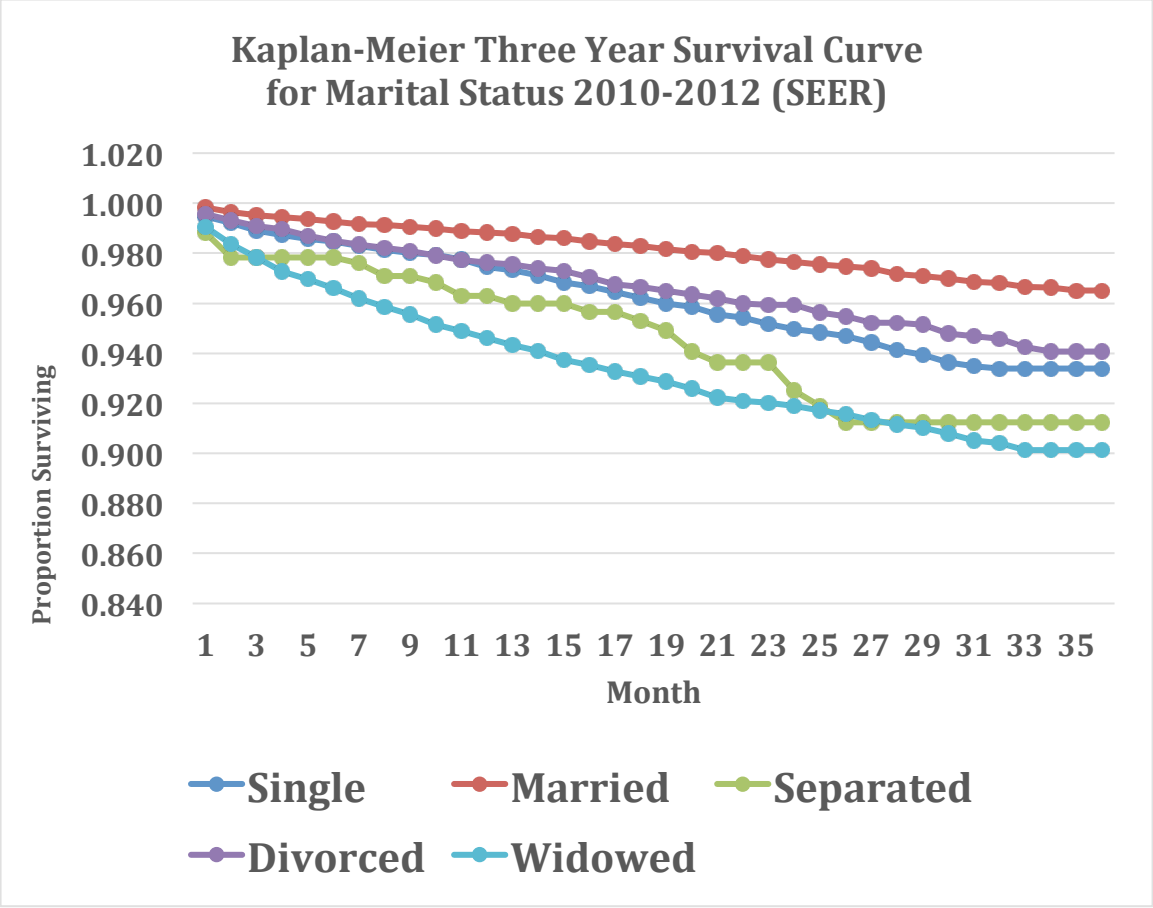


Breast Cancer by Race: Asians and American Indians/Alaska natives show the best three survival curve (97%). White are just below, with a (95%) survival rate, and Blacks are lowest with a 90% survival rate. Unfortunately, socioeconomic data for this group of patients was unavailable, but this may be a confounding factor. Health insurance for this data is also not yet reliable for all patients, but this may be an important factor as well.



Breast Cancer by Marital Status:

Married women consistently show higher survival rates than all of the other statuses in this category, regardless of age. When first reviewing these overall findings, the question arose of whether the age variable was confounding the results, and was an intervening variable accounting for the findings. This suspicion led to further examinations in this area.

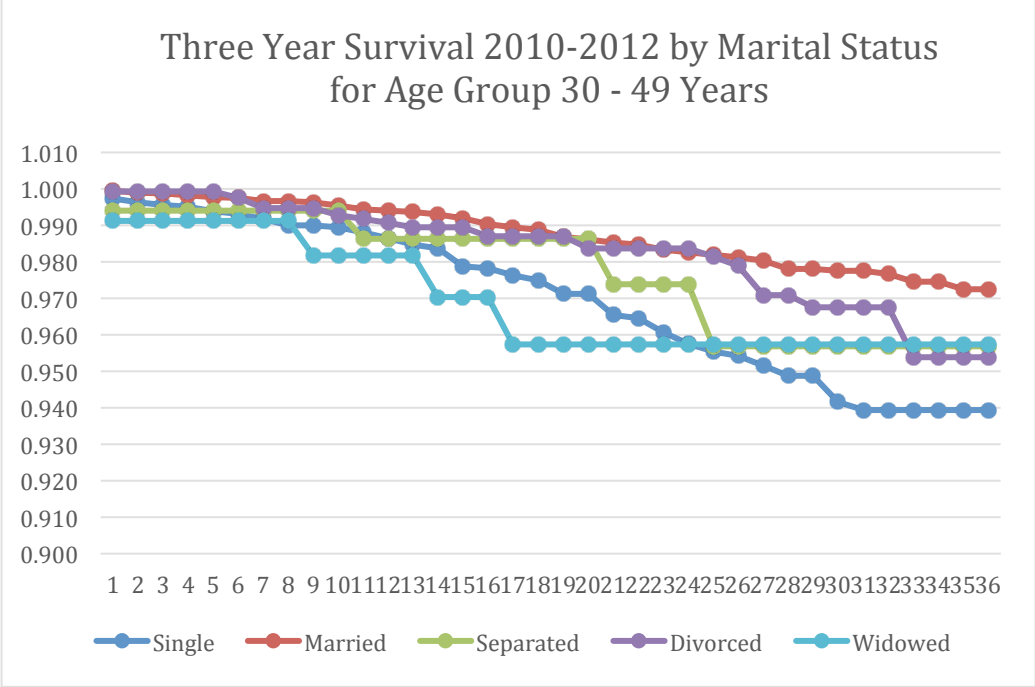


First, a logistic regression was done with marital status and age at diagnosis as predictor variables for survival response variable (0, 1) for the 2010 year data. The results were that both age and marital status were significant predictors. Married status clearly provides an advantage over the other situations, although the other statuses are not significant.

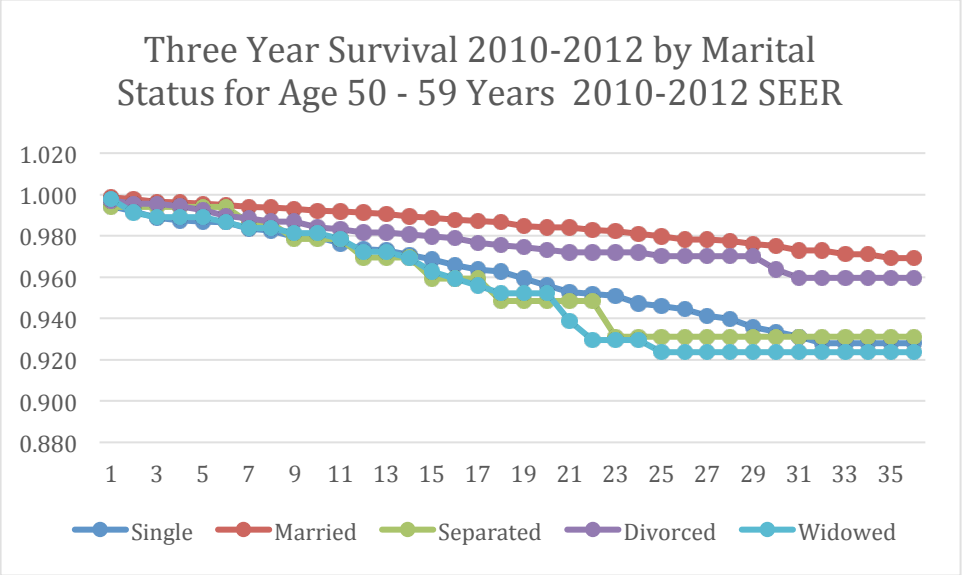
Model parameters (Variable EVENT):

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²
Intercept	-3.778	0.178	448.697	< 0.0001
Age_Diagnosis	0.020	0.003	53.074	< 0.0001
Marital_Status-1 (Single)	0.000	0.000		
Marital_Status-2 (Married)	-0.818	0.094	75.074	< 0.0001
Marital_Status-3 (Separated)	0.285	0.282	1.021	0.312
Marital_Status-4 (Divorced)	-0.218	0.122	3.187	0.074
Marital_Status-5 (Widowed)	-0.007	0.116	0.003	0.953

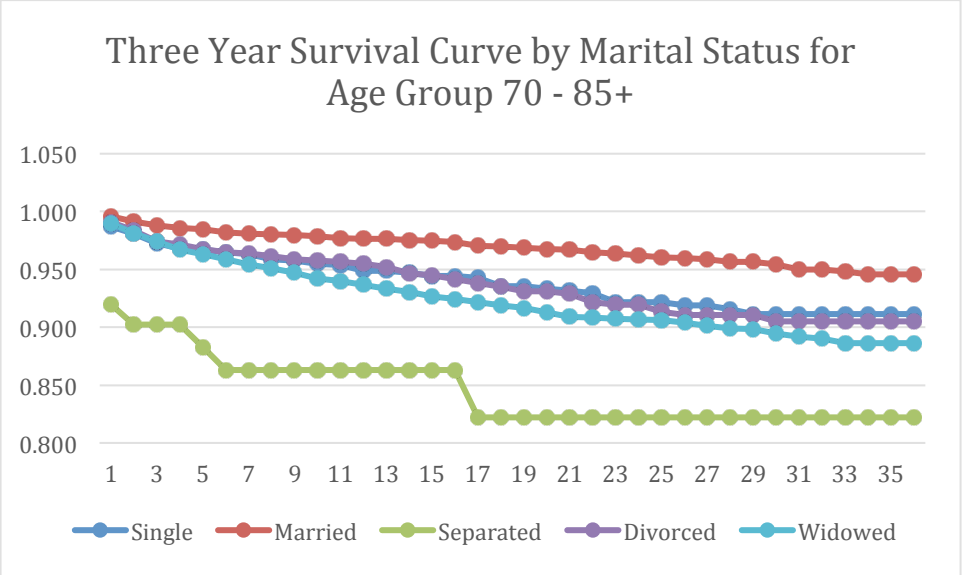
To further explore marital status is relation to survival other three years, multiple age groups were examined. The under 30 year age category could not be examined, as there were only three deaths in this small group of 321 patients. However, for the 30 - 49 year old category, marital status was a significant factor (Log Rank = 41.5, $p < .0001$). However, in this case, the order is somewhat different from the overall findings. As before, married patients have the longest survival, but single patients have the lowest survival rather than widowed patients. This may be because widowhood is relatively rare in this age group, and the sample of deaths is small in the widowed group (4 deaths out of 134 patients).



In contrast, in the 50 to 69 year old group, single, widowed, and separated patients were very similar, and were the lowest surviving groups. Once again the married were the highest surviving group, with divorced was just below the married group, as in the overall findings (Log Rank = 118.5 , $p < .0001$).



Looking at the 70 – 85+ year old group, the separated was the lowest group, followed by the widowed, with divorced and single closely following. Once again, the married group has the best outcome. The marital differences for this age group are significant (Log Rank = 122.8 , p < .0001).



With this amount in variation among the age groups, perhaps the most definitive conclusion is that the married group will always fare better than the other four groups. Oddly, of the other four groups, the divorced group generally does the best. Perhaps they have been able to form new supportive relationships, whereas the other groups (single, separated, and widowed) do not have the supports that they need to do well.

Citations

[1] SEER Research Data 1973-2012 -- ASCII Text Data: Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2012), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2015, based on the November 2014 submission.

[2] Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct 4;490(7418):61-70. doi: 10.1038/nature11412. Epub 2012 Sep 23. 357 Authors collaborated on this project.