**Story**

**2015**

Dr. Grey is a medical oncologist, working at the Confront Cancer Clinic, and specializing in breast cancer. She is frequently faced with questions about how well she is doing with her patients, and must answer increasingly sophisticated questions from her patients as well. She would like to know whether the subtype of breast cancer of her patients is related to their survival, and how these figures compare to other clinics (SEER).

Fortunately, the hospital where she works converted to an electronic medical record system three years ago, and has a “Clinical Data Warehouse” as well, so she thinks it should not be too hard to get these data she needs to answer the question. She asks one of her residents to get the data together for her, and give her the survival analysis figures for the last three years. She checks back with him about the project in a week, and finds him immersed in an Excel spreadsheet, typing in data from PDF files printed from the computer. She says “What on earth are you doing? You know we have an automated health record system!”

The resident explains that, while the data she needed was in the electronic health system and in the data warehouse, he couldn’t get it out without: (1) Querying the system to create a report with the data in it (most of it, anyway); (2) Printing the PDF formatted report out; (3) Creating an Excel spreadsheet with the data fields in it that he needed, in the correct format; and (4) Manually entering the data into the spreadsheet. Also, I did a web search, and SEER has excellent data on this topic for the last three years, but I’ll need to get a data use agreement approved first in order to obtain access. Also, in order to access the Cancer Registry data, I’ll need to specify the fields needed, justify the need, sign a data use agreement, and go through an IRB to obtain access to the data for our state. For any other states that we need, we’ll need to go through the same process. Sorry this is such a difficult and time consuming process! Honestly, I’m doing the best I can!

Dr. Grey goes off, furious, searching for the person in charge of this so-called “Clinical Data Warehouse”, and wondering why no one had yet fixed this problem.

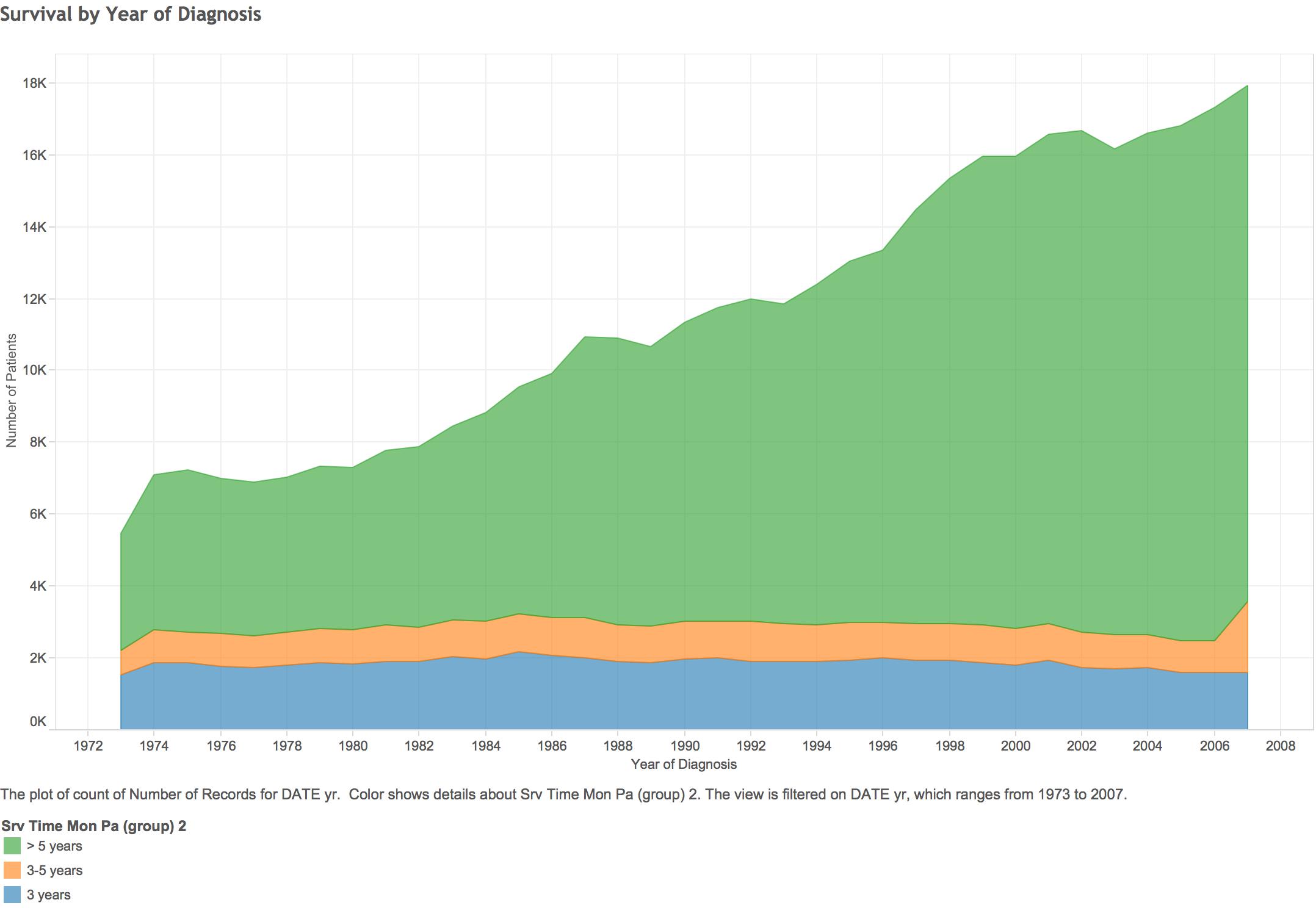
**2016**

Dr. Grey is informed by the director of the “Clinical Data Warehouse” that a new system has been created called the “Breast Cancer Research Platform”, and asked if she’d like to take a look at it for herself. She agrees, follows the instructions to create a biometric application login, and opens the main dashboard.

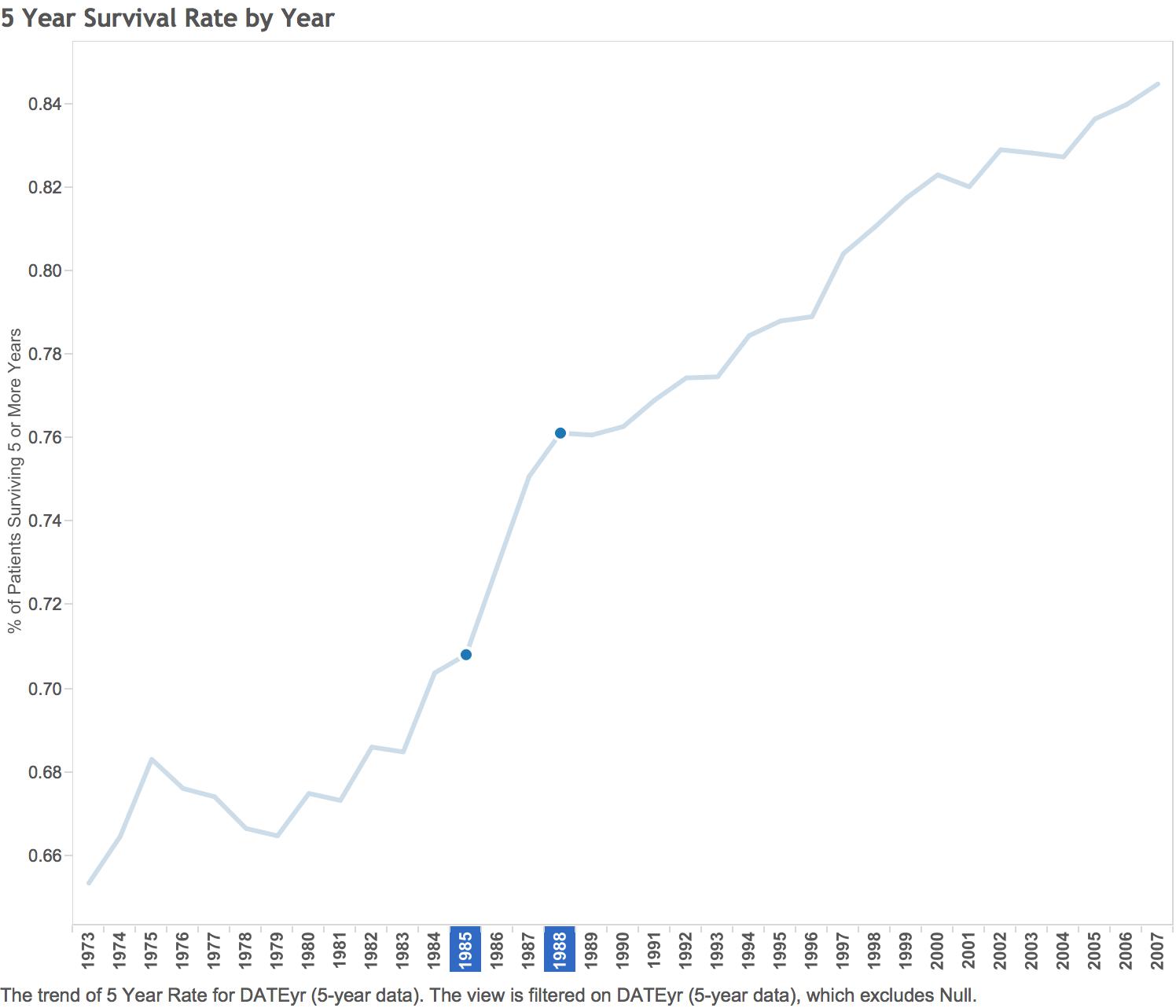
This looks great! First, I’d like to get an idea about the general breast cancer patient population, the number of patients, survival, and some demographics across the nation.

1. Select SEER Dataset from **Browser** tab
2. Select following Variables for all years of data: Cause-Specific Death Class, Months of Survival, Stage, Year of Diagnosis, Race, Age, Marital Status, and Breast Subtype.
3. Click on **Visualization** tab for output

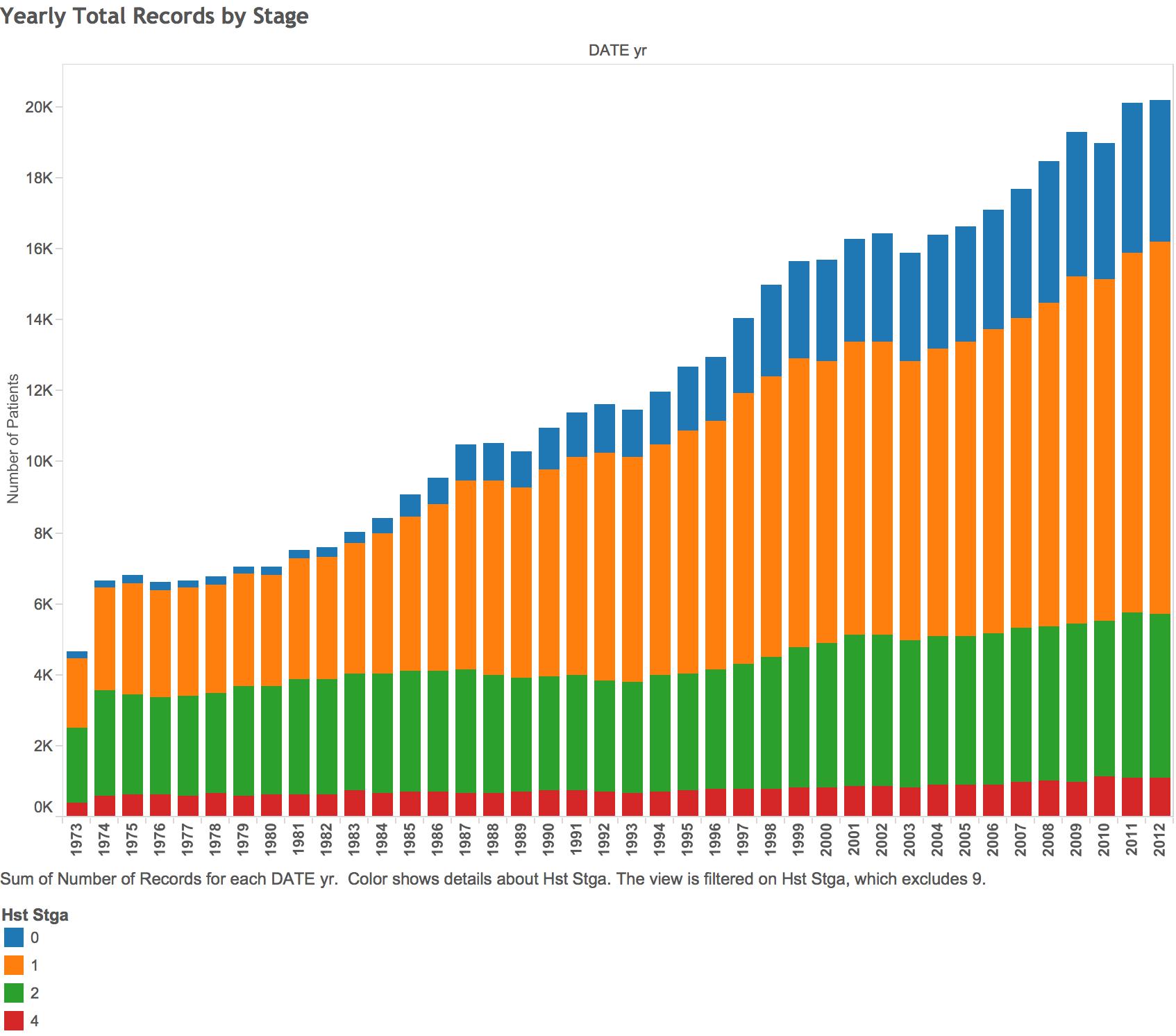
When looking at whole patient counts, the number of patients increases with time. This is most likely due to increases in reporting, population, and screening since 1973. The majority of patients survive more than 5 years.



When looking at the percentage of patients who survive 5 or more years for each year of diagnosis, we do see an increase over time. This would indicate that the odds of surviving more than 5 years after a breast cancer diagnosis have improved with time and medical treatment advances. There is a steep jump between 1985 and 1988 of about 6 points to 76%, and then a slower ascent through the 90s and early 2000s to 84%.



While it looks like the overall survival rates have increased over the years, caution is needed because there may be significant differences in the patient populations over the years. One possibility is that with increased screening more patients were diagnosed earlier and so the ratio of early stage patients to late stage patients increased, thus creating an increase in the survival rates. Let’s look at the stage now, and how that changes over the years.

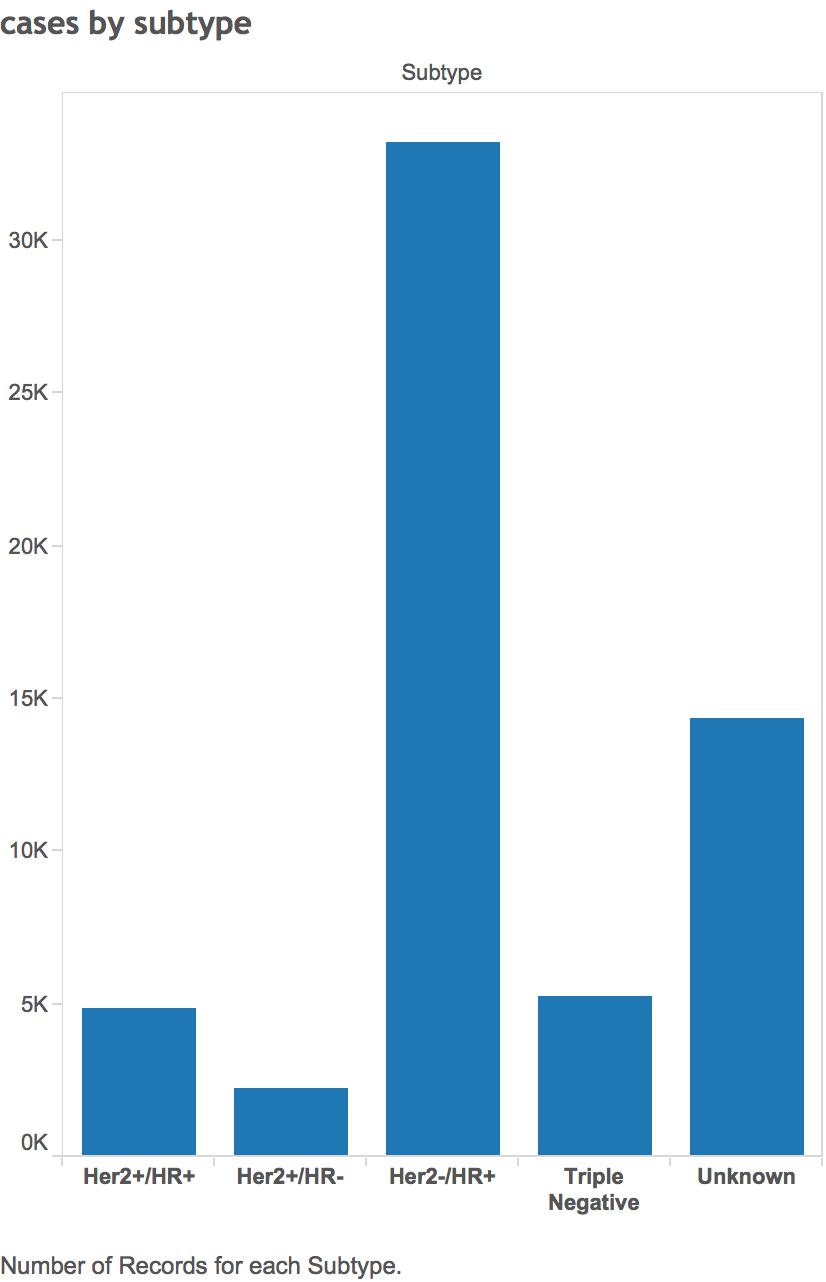


As can be seen from the chart, the more advanced stages, Stage 2 (extended to the lymph nodes) and 4 (extended to distant organs) remains fairly stable, while the least advanced stages, Stage 0 (non-invasive) and Stage 1 (confined to within the breast) increase over the years. This change is probably related to the increase in the number of patients surviving over 5 years. The increase in the number of patients being diagnosed initially at the lower stages is also probably related to the increase preventive options, including mammograms and other imaging methods, along with self-examinations being used more frequently.

Let’s move on to the **Analysis** tab for a moment to select the Kaplan Meier **Survival Analysis** by breast cancer stage at initial diagnosis. We will select the Cause Specific Death Class as the **event**, the **Survival Months** as the **time**, and the Stage as the **group** variable. Improved medical treatments have certainly played a role in the increased length of survival as well. As can be seen from the chart below, Stages 0, 1, and 2 from 2010-2012 have a 90% or more chance of survival after three years. For Stage 4 patients, however, the survival rate is only 55% after three years. According to the Log Rank test, these survival curves are significantly different at the .001 level.

Let’s not forget my original question! We know that tumors that are HER2 positive and HR positive have targeted treatments, whereas Tripe Negative tumors have fewer options right now. How do patients with the various subtypes fair across the nation? Then, we need to take a look at how our clinic is doing with these patients, by pulling the data from the electronic records.

Going back to the **Describe** tab, the HER2-/HR+ positive subtype is the most common of all of the subtypes, and has targeted treatments available for it. The HER2+/HR+ and HER2+/HR- negative subtypes also have targeted treatments available, but the Triple Negative is still the most challenging to treat, although much research is currently going on in this area.



Moving on to the **Analyze** tab, and selecting the **Kaplan-Meier** **Survival Analysis** option, the Cause Specific Death Class will be the **event**, the **Survival Months** the **time**, and the Breast Subtype the **group** variable. The Triple Negative subtype has the lowest survival likelihood at three years, with an 84% chance of survival. The two HR+ subtypes, HER2+/HR+ and HER2-/HR+ have the best chance at survival, with both being above 95%. In the middle is the HER2+/HR- subtype, which has an 89% chance of survival.

Let’s compare these subtype survival rates with those from my patients.\* We’ll need to create another dataset from the electronic record system.

1. Select Clinic Electronic Dataset from **Browser** tab
2. Select following Variables from electronic records system for the last three years: Cause-Specific Death Class, Months of Survival, Stage, Breast Subtype. Click on **Visualization** tab for output

In terms of the distribution of subtype, the most frequent subtype in the clinic is the HER-/HR+, just like in the SEER dataset. The number of HER+/HR- and Triple Negative subtypes are similar, and then the HER+/HR- subtypes are the lowest frequency.

Moving on to the **Analyze** tab, and selecting the **Kaplan-Meier** **Survival Analysis** option, the Cause Specific Death Class will be the **event**, the **Survival Months** the **time**, and the Breast Subtype the **group** variable. As in the SEER dataset, the Triple Negative subtype has the lowest survival likelihood at three years, with a 77 % chance of survival. The HER+ subtypes, HER2+/HR+ and HER2+/HR- are in the middle, with both being close to 80%. The HER-/HR+ subtype has the best chance of survival (91.5%). Overall, the clinic has lower survival rates than shown in the SEER dataset, but this may reflect the smaller sample size, although the clinic curves are still significantly different (Log Rank, p < .01). In particular, the HER+ subtypes are about 10% different from the SEER dataset, so targeted therapies for these subtypes need to be made available to the patients in the clinic.

**DEMOGRAPHIC VARIABLES**

Now, I’d like to look at some demographic variables, such as marital status, race, and age, so I can get an idea of how my patients will do. Staying on the **Analyze** tab, and re-selecting the **Kaplan-Meier** **Survival Analysis** option, the Cause Specific Death Class will once again be the **event**, the **Survival Months** the **time**, and Marital Status the **group** variable.

As shown in the above chart, married patients have the best survival chances after three years (96%), and widowed the poorest chances (still 90%). The remaining situations are in between. These curves are significantly different (Log Rank, p < .001). These differences make sense in that married women would have the support of a spouse to help them through, and make sure that all their treatment appointments are kept.

The age survival curves (see below) were also significantly different (Log Rank, p<.001). As expected, the 20-24 yr old age group has the best chance of survival, and the oldest group (85+) the worst. The next three groups (80-84, 75-79, and 70-74) follow the same pattern, but the rest of the age groups are bunched together, with the same survival rate. So the decrease in survival rate does not begin until after age 70, which is surprising in comparison with other age changes being more gradual, beginning in early adulthood.

As found in other studies, the Asian group has the best chance of survival (97%), and the black group (90%) the worst. The white group (95%) and the Indian group (96%) have survival rates in between. Socioeconomic information was unavailable, but this may be a mediating variable in this relationship. These race survival curves are significantly different (Log Rank, p <.001).

**Integrated Analysis**

Is there any way to develop a model for the prediction of breast cancer? That would help us in developing preventive programs, so that the initial stage of presentation of breast cancer could be even more restricted to Stages 1 and 0, and the general incidence may be reduced as well.

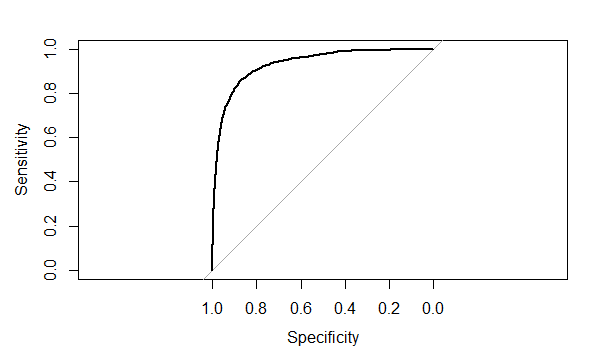
In our research system, we have two general types of predictive models: Regression and Decision Trees. The regression model has some critical assumptions which make its use not as general as decision trees, such as linearity, but we should, nevertheless, take a look at it first.

**Regression Analysis**

First, we’ll need to return to the **Browser** tab to select all of the variables we’ll need for the regression. Based on general knowledge about breast cancer, the following variables should be candidates for prediction:

Marital Status, Race (Including Asian breakdown), Stage, Age at Diagnosis, Tumor Subtype, Number of Nodes Positive, Tumor Size, and Reason for No Surgery.

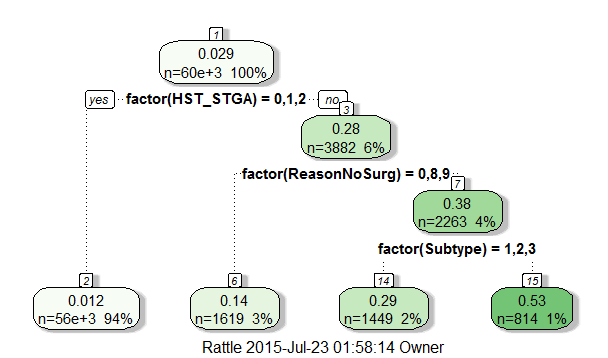
Clicking on the **Analysis** tab, select **Logistic Regression** from the options. The **Response Variable** will be the Cause Specific Death Class, and the **Predictor Variables** will be the ones that we just selected. Then we’ll need to go over to the **Visualization** tab to see the results.

****

The area under the ROC curve is 0.9318, indicating that the likelihood of accurately predicting that the person is deceased is very good, so this is a highly sensitive model, even though there are just eight variables.

**Decision Tree**

For this type of model, we’ll need to select the **Analysis** tab, and then **Decision Tree** from the list of options. Also, we’ll need to specify several variables, although default settings will automatically apply if these are not modified: (1) Minimum split size into groups, with 20 as the default; (2) Minimum bucket size for the output of the split, with 7 as the default; and the (3) Maximum depth of the tree, with 25 as the default. The resulting simple decision tree is shown below.



This research system is definitely superior to the previous one available to me. The options are incredible! When can get it installed for everyone in our clinic?

\*This data is actually from the following source, and not from electronic medical records of an oncologist in practice: 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.