

Visualizing NCI Seer Cancer Data

Introduction

I chose to use the National Cancer Institute's Surveillance, Epidemiology and End Results (NCI SEER) database¹ as a basis for my application visualization project. This choice was motivated by my ongoing interest in cancer awareness and the wealth of data that the SEER repository represents. Little did I know that that wealth of data would prove to be the biggest obstacle in my work to create useful visualizations. As background, SEER provides individual data for "more than 3 million in situ and invasive cancer cases"² covering a large portion of the United States, a total of 77 data fields are available for each record, including information: patient demographics, cancer type, progress, survival, and mortality rates. In total the basic SEER database contains approximately 223 million fields of raw data. The wealth of data is too much to tackle without specialized tools to process the information; to that end NCI provides the SEER*Stat program as well as binary and ascii formatted data sets.

Applications

SEER*Stat

SEER*Stat in itself is a necessary component to understanding the data contained in the repository. The program provides basic aggregation functions in the form of: single or multiple incidence frequency, survival rates, and case listings. While the functionality is theoretically competent, the implementation suffers from several drawbacks. First and foremost the application provides only the most basic tools for aggregating and analyzing the data; and several of the tools impose impractical limitations on the operation of the product. All data display is done through tabular matrix display (figure 1); which while it provides very good usability for small detailed examination of data, is somewhat problematic for millions or even thousands of records. SEER*Stat provides no other visualization tools beyond tabular display of raw or aggregated data. Furthermore, impractical limitations limit the utility of even the aggregated data computations. For instance, in determining incidence rates for all types of cancers, across the 50 U.S. states, for the years 1973-2001, the program limits the user to 5 attributes retrievable (out of 77 possible). Even within the confine of selecting just three attributes: cancer site (type), race, and sex yields 1.1 million matrix cells – already in excess of the program's 1 million cell limit. The program does provide a facility for filtering the data-set for output, but unfortunately the resulting tables are the same size, but many of the entries are merely not computed; this effec-

¹ Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Public-Use Data (1973-2001), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2004, based on the November 2003 submission.

² National Cancer Institute. "About SEER" Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute March 2, 2005 <<http://seer.cancer.gov/about/>>

tively reduces computational cost, but does not provide a mechanism for circumventing matrix size limits. Even providing a simple case dump of the data to be export as a comma-separated values file limits the user to 8 attributes to be dumped, but thankfully imposes no limit on the number of data records.

Site no with Rajosol and mesothelioma	Year of diagnosis	Place of birth	Race	Sex	Age recode	Marital status at diagnosis	Radiation	Reason no cancer-directed surgery	Survival time recode (Total # of months)
394221	Other MyeloidMx 2000	Unknown	White	Female	45-49 years	Divorced	None	Surgery not recommended	137
394222	Other MyeloidMx 2000	Unknown	White	Male	60-64 years	Married (including common law)	None	Surgery not recommended	040
394223	Other MyeloidMx 2000	Massachusetts	White	Female	60-64 years	Divorced	None	Surgery not recommended	001
394224	Other MyeloidMx 2000	Massachusetts	White	Male	60-64 years	Married (including common law)	None	Surgery not recommended	001
394225	Other MyeloidMx 2000	Connecticut	White	Male	70-74 years	Married (including common law)	None	Surgery not recommended	000
394226	Other MyeloidMx 2000	Connecticut	Black	Male	60-64 years	Married (including common law)	None	Surgery not recommended	013
394227	Other MyeloidMx 2000	Connecticut	White	Male	75-79 years	Married (including common law)	None	Surgery not recommended	019
394228	Other MyeloidMx 2000	Connecticut	White	Female	60-64 years	Married (including common law)	None	Surgery performed	021
394229	Other MyeloidMx 2000	Pennsylvania	White	Female	75-79 years	Divorced	None	Surgery not recommended	013
394230	Other MyeloidMx 2000	Pennsylvania	White	Male	70-74 years	Married (including common law)	None	Unknown	012
394231	Other MyeloidMx 2000	Pennsylvania	White	Female	65+ years	Widowed	Unknown	Unknown	008
394232	Other MyeloidMx 2000	Virginia	White	Male	25-29 years	Single (never married)	Beam radiation	Surgery not recommended	013
394233	Other MyeloidMx 2000	Tennessee	Black	Female	65-69 years	Married (including common law)	Beam radiation	Surgery not recommended	012
394234	Other MyeloidMx 2000	Alabama	White	Female	60-64 years	Divorced	None	Surgery not recommended	001
394235	Other MyeloidMx 2000	Mississippi	Black	Male	75-79 years	Divorced	None	Unknown	009
394236	Other MyeloidMx 2000	Michigan	White	Female	60-64 years	Married (including common law)	None	Surgery performed	019
394237	Other MyeloidMx 2000	Michigan	Black	Female	60-64 years	Married (including common law)	None	Surgery not recommended	000
394238	Other MyeloidMx 2000	Michigan	White	Female	40-44 years	Married (including common law)	None	Surgery not recommended	005
394239	Other MyeloidMx 2000	Ohio	White	Female	70-74 years	Widowed	Beam radiation	Surgery performed	001
394240	Other MyeloidMx 2000	Indiana	White	Female	75-79 years	Widowed	None	Surgery performed	001
394241	Other MyeloidMx 2000	Kentucky	Black	Female	65-69 years	Married (including common law)	None	Surgery not recommended	001
394242	Other MyeloidMx 2000	Minnesota	White	Male	65+ years	Married (including common law)	None	Unknown	002
394243	Other MyeloidMx 2000	Iowa	White	Male	60-64 years	Married (including common law)	None	Surgery not recommended	015
394244	Other MyeloidMx 2000	North Dakota	White	Male	70-74 years	Widowed	None	Surgery not recommended	012
394245	Other MyeloidMx 2000	Illinois	White	Male	65+ years	Widowed	None	Surgery not recommended	011
394246	Other MyeloidMx 2000	Oklahoma	White	Male	70-74 years	Unknown	None	Surgery not recommended	011
394247	Other MyeloidMx 2000	Texas	Black	Female	75-79 years	Divorced	None	Surgery not recommended	005
394248	Other MyeloidMx 2000	Texas	Black	Female	60-64 years	Widowed	None	Surgery not recommended	007
394249	Other MyeloidMx 2000	Ishio	White	Female	70-74 years	Widowed	None	Surgery not recommended	000
394250	Other MyeloidMx 2000	Utah	White	Male	60-64 years	Single (never married)	None	Surgery not recommended	014
394251	Other MyeloidMx 2000	Washington	White	Male	60-64 years	Married (including common law)	None	Surgery not recommended	014
394252	Other MyeloidMx 2000	Washington	White	Male	70-74 years	Married (including common law)	None	Surgery not recommended	012
394253	Other MyeloidMx 2000	California	White	Female	70-74 years	Married (including common law)	Beam radiation	Surgery not recommended	018
394254	Other MyeloidMx 2000	Hawaii	Asian or P	Female	75-79 years	Widowed	None	Surgery not recommended	001
394255	Other MyeloidMx 2000	Urane and Nole	White	Female	75-79 years	Widowed	None	Surgery not recommended	001
394256	Other MyeloidMx 2000	Nd U.S., but no	Asian or P	Female	50-54 years	Divorced	None	Surgery not recommended	016
394257	Other MyeloidMx 2000	Unknown	White	Male	75-79 years	Married (including common law)	None	Surgery not recommended	014
394258	Other MyeloidMx 2000	Unknown	White	Female	60-64 years	Divorced	None	Unknown	001
394259	Other MyeloidMx 2000	Unknown	White	Female	75-79 years	Separated	None	Unknown	001
394260	Other MyeloidMx 2000	Unknown	White	Male	60-64 years	Unknown	None	Surgery not recommended	012
394261	Other MyeloidMx 2000	Unknown	White	Female	60-64 years	Single (never married)	Beam radiation	Unknown	015
394262	Other MyeloidMx 2000	Unknown	White	Male	75-79 years	Unknown	None	Unknown, death certificate only case	018
394263	Other MyeloidMx 2000	Unknown	Asian or P	Male	50-54 years	Married (including common law)	None	Unknown	000
394264	Other MyeloidMx 2000	Unknown	Unknown	Male	65-69 years	Unknown	None	Unknown	019
394265	Other MyeloidMx 2000	Unknown	White	Male	65-69 years	Married (including common law)	None	Surgery not recommended	017
394266	Other MyeloidMx 2000	Unknown	White	Female	65-69 years	Unknown	None	Unknown	021
394267	Other MyeloidMx 2000	Unknown	White	Male	50-54 years	Married (including common law)	None	Surgery not recommended	017
394268	Other MyeloidMx 2000	Unknown	White	Female	35-39 years	Married (including common law)	None	Surgery not recommended	010

Figure 1

The solution is to move to custom processing of the raw data-set in ASCII format. While the binary data-set might be more efficient to manipulate, no data dictionary id provided for that format; the data dictionary for the ASCII format is provided with the raw data. A sample of the raw data is included in figure 2, obviously the data needs to be preprocessed before it is usable in most data processing or statistical packages; since there was only a need for aggregated and mean processing of 12 fields, a simple C program handled the pre-processing of the data for visualization. The resulting data-set yielded approximately 150 thousand records of 12 fields each.

0200002551017 00719130610101400 1975C34998010399 09999048010322030101139 99991629 991
 C3491161162203011022030901 121322030090031

Figure 2.

GeoVISTA

The initial proposal called for processing the data with GeoVISTA³; unfortunately GeoVISTA would proved unwieldy for the number of records used. While the primary reason for visualizing the data using GeoVISTA was to better understand the relationship of cancer mortality to regional area – we hypothesize that economic factors affect cancer survival – the data-set for all 50 states across 28 years proved to be too large for application; additionally the development burden for implementing the necessary visualizations across 12 attributes was considerable.

TimeSearcher

TimeSearcher⁴ seemed like a natural fit to the data since the data set included 28 years of data; here the application proved to be unsuitable on two fronts. First, the visualization does not deal well with high-dimensional data-sets (here we have 12 dimensions), and the application does not handle large data sets well at all; even partial versions of the data set caused frequent crashing. A pared down data set (164 records) visualizing of 28 years of aggregated data for lymphoma, leukemia, and myeloma is provided in figure 3.

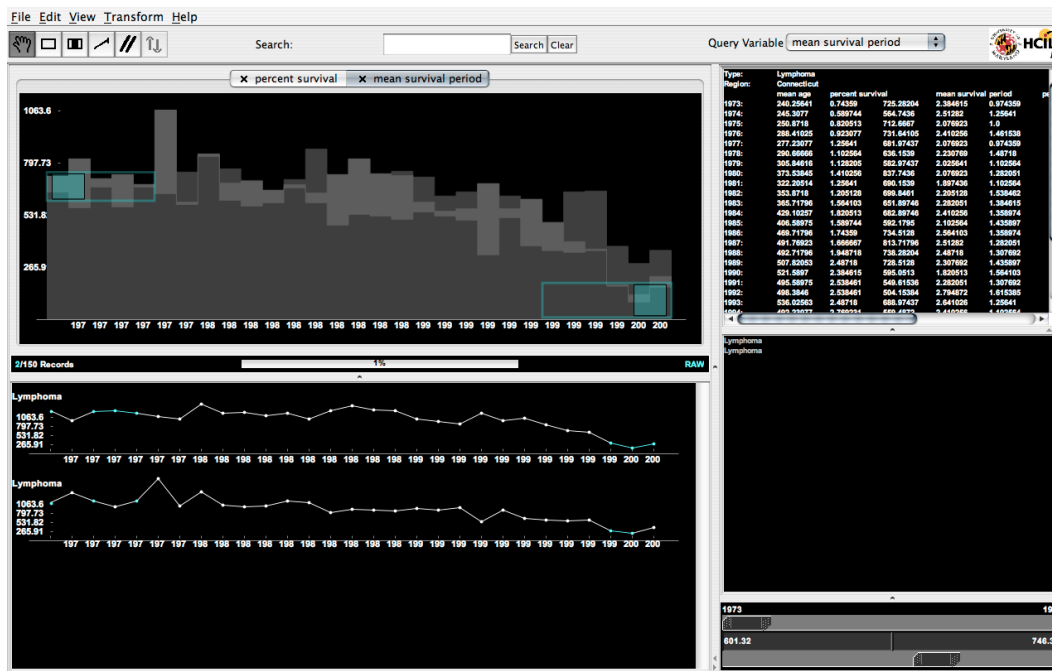


Figure 3

³ GeoVISTA and The Pennsylvania State University. "GeoVISTA Center" GeoVISTA Center March 2, 2005 <<http://www.geovista.psu.edu/index.jsp>>

⁴ Human-Computer Interaction Lab "Visual Exploration of Time-Series Data" University of Maryland Human-Computer Interaction Lab 2005, March 2, 2005 <<http://www.cs.umd.edu/hcil/timesearcher/>>

Here we begin to get some insight into the data. Looking only at survival periods in terms of months we see that there is a pretty consistent mean survival rate with the exception of 1977 (caused by incomplete records), and a recent trend towards shorter survival periods in the last decade. The latter effect is actually an artifact of the data collection methods, all patients still alive in December 2001 were given credit for just the months until then, but not for surviving past that time – in effect all cancer survivors are considered to *only* have lived until december 2001. Unfortunately, much of the interesting data available through SEER cannot effectively be viewed in TimeSearcher since there are so many attributes to analyze.

Hierarchical Clustering Explorer

Since the SEER data is comprised of high-order information, Hierarchical Clustering Explorer⁵ (HCE) was the most logical final choice to visualize the data. In using HCE it once more became necessary to pare down the data for the practical reason that the display becomes difficult to use with too many data records. To this end, we selected just six states: NC, ND, NE, NH, NJ, and NY (NM and NV were not included in the available data-set); and just three years: 1973, 1987, and 2000) to analyze for all forms of cancer. The resulting data-set totaled 18,727 records across 10 attributes: cancer site (type), state of birth, year of diagnosis, sex, age, martial status, cancer stage, surgical intervention, radiation therapy, and survival time in months. Since the 18,000 records would have been difficult to manipulate in HCE the data-set was aggregated by state and year into just 284 records, which HCE handles easily. The resulting data (normalized by mean) and clustered by average group linkage is shown in figure 4.

⁵ [Human-Computer Interaction Lab “Hierarchical Clustering Explorer for Interactive Exploration of Multidimensional Data” University of Maryland Human-Computer Interaction Lab 2005, March 2, 2005 <http://www.cs.umd.edu/hcil/hce/>](http://www.cs.umd.edu/hcil/hce/)



Figure 4

Observations

Within HCE some clear clusters have formed, and some data begins to pop out. Figure 5, provides an enlargement of the vertical clustering of the data.

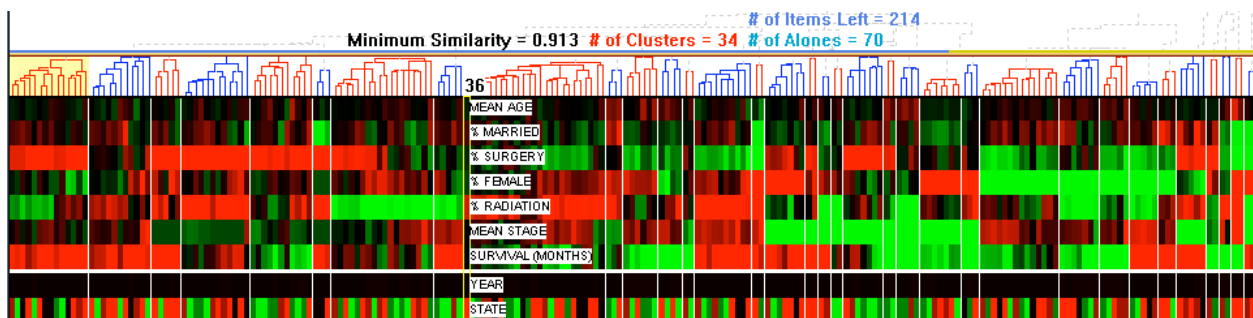


figure 5

Clearly, lack surgical intervention is a marker for a decreased survival times (the leftmost 4 clusters as well as periodically throughout the graphic). Conversely radiation treatment does not seem well correlated to survival. Oddly, catching cancer early doesn't seem to improve survival, but diagnosis late in the diseases progression is clearly detrimental to survival times (

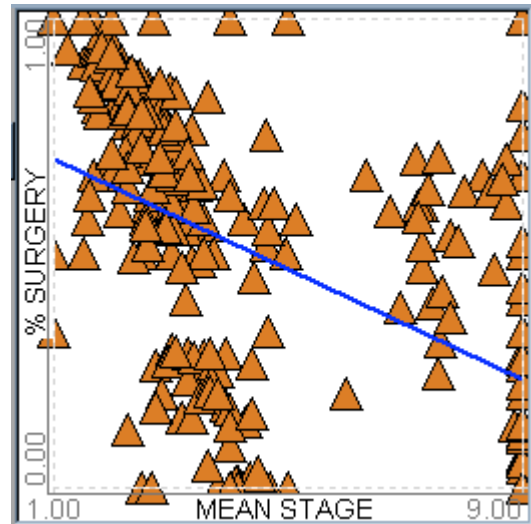
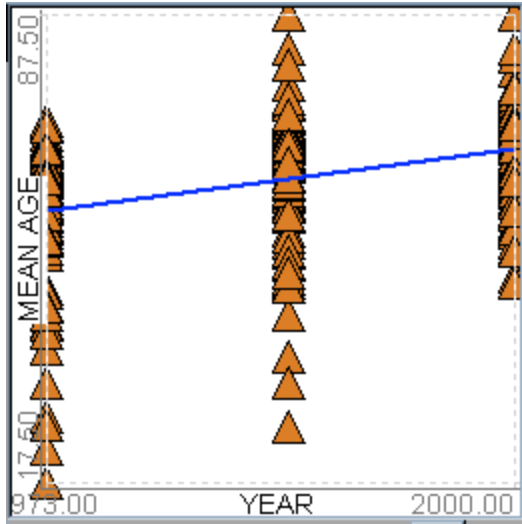
clusters starting under the “Clusters=34” indicator extending to the right). Age, marital status, and gender seem to play little role in determining outcome of cancer treatment, but the limited data set and lack of data resolution might be affecting the results here.

HCE provides scatterplot ordering and parameter correlation visualizations (figures 6-10) that further provide insight into the relationship between the variables. The highest correlation is between mean age and the year of diagnosis; we attribute this to the increased age of the population over the last 30 years. Further strong correlations are between the percentage of women and the incidence of surgical intervention; presumably due to the frequency of mastectomies. The third strongest correlation is between surgical intervention and survival time which confirms the impression from the dendrite visualization. Looking at inverse correlations, we discover a strong inverse relationship between surgical intervention and the stage of cancer progress; this is probably due to the lack of visibility in certain aggressive forms of cancer (lymphoma, pancreatic, and nervous system). There is a trivial relationship between the year of diagnosis and survival time that is once again attributable to the data collection methodology. Unsurprisingly, there is a correlation between the mean age of a group and its survival time; younger people seem to survive cancer more often. The decline of marriage as an institution in the US shows up the inverse relationship between year and marriage percentages. And probably the most telling result is that women are diagnosed earlier than men, probably due to the successful cancer self-exam programs within the last 25 years.

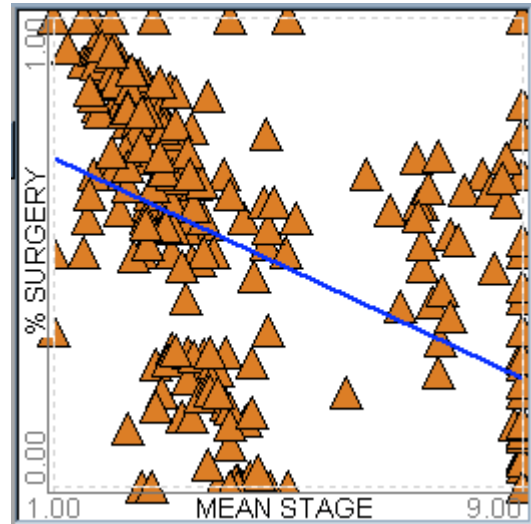
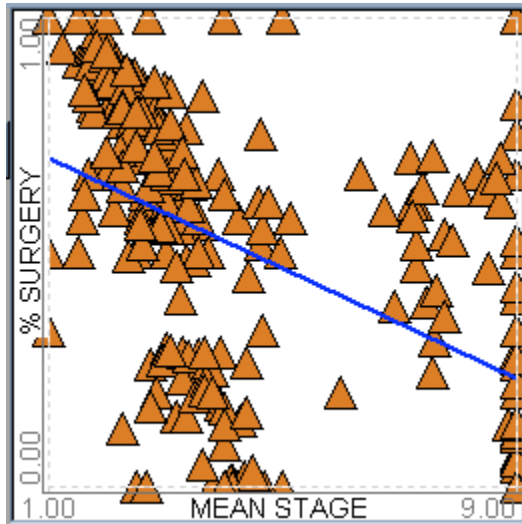
It is worth noting that we assumed the profile search visualization in HCE would prove useful, but the granularity of control made this tool cumbersome to use for this data-set, and provided no new insights into the data; although there was a brief time when we were convinced otherwise by some interface quirks.

Conclusion

There are several insights to be gained from this particular endeavor; primarily working with large data sets is particularly difficult, and that the choice of visualization tool can drastically impact the feasibility and insight gained from the data. In particular working with large data-sets seems to require preprocessing the data into a tractable quantity before visualization can effectively be employed. Additionally, using a variety of visualization tools that emphasizing the strength of each tool might be a useful – if time consuming – tactic.



Figures 6 & 7



Figures 8 & 9

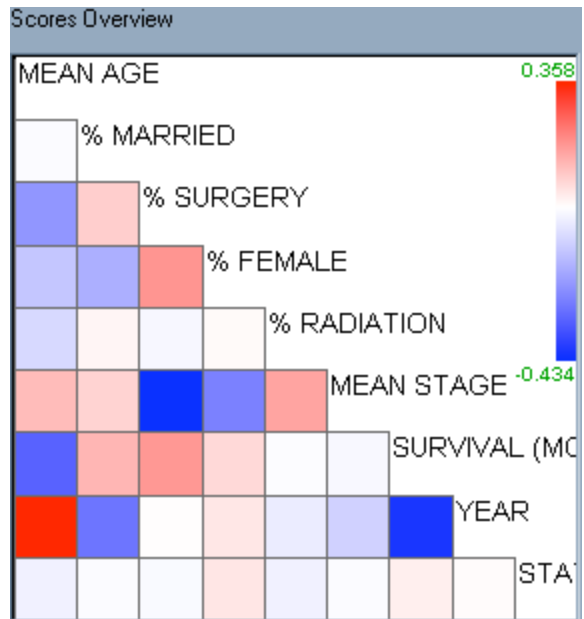


Figure 10