Summer 15

Breast Cancer Research Platform

Siddharth Shah, Andrea McCarter, Lauren Crabtree, Sarah Bartlett

We propose a system to facilitate the needs of breast cancer researchers and oncologists.

Georgia Tech Analytics Practicum

Introduction

The original task for this practicum was to design an oncology research software platform from scratch. Through our research it became evident how valuable such a platform would be to oncologists and cancer researchers. The challenges that have prevented such a platform from coming into being to date also made themselves evident. Due to the limited time frame and resources of this practicum, we approached this task with the following plan:

1. Research. Talk to oncologists and researchers and find out what kinds of questions they are trying to answer, where the data that they need to answer these questions lives, and what methods they use to solve them.
2. Find Data. Look for publicly available data as well as identify where private data might be made accessible for analysis.
3. Narrow Scope. Look at just the key features that would deliver the most benefit and alleviate the worst pain points of researchers. Focus on one type of cancer.
4. Play the Part. In order to design a system that meets the needs of cancer researchers, we needed to put ourselves in their shoes and do some analysis. This enabled us to discover what the most important usage scenarios were so that we could target our proposed solution to those.
5. Propose a Solution. Based on 1-4, identify and describe the features of a hypothetical system that would address the most pressing needs of a cancer researcher.

Background Information

Available Data sources

- Publicly accessible raw data: Many datasets are available through user upload for individual analysis. A few examples of this raw data are imaging studies, clinical trial information, and genomic datasets. (cBio Portal, TCIA Imaging)
- Patient Records: Patient records are available within health systems via EHR systems and other customized IT systems put in place by individual hospitals and research centers. (Epic, Cerner)
- Shareable Research: Multiple depositories of research protocols as well as medical publications. (PubMed)

Limitations of Data Sources

- No central system: Due to the three different categories of data sources, one central system to allow researchers to easily view the data would be helpful but the lack of this ability is a major limitation of current system architectures.

- Interoperability: Within the patient records systems, the ability to transfer records between competing EHR systems as well as other customized systems being used is not always possible.
- Data formats: Each source collects different pieces of data over different time periods making the data extremely inconsistent from one source to another. The lack of a clear data dictionary with fields that researchers need is a challenge. Much of the consistent data across systems is limited to generic demographic information.
- Lack of analytics and visualizations: Currently there limited analytical capabilities for healthcare data systems and no visualizations to allow a fast depiction of the data and aid in developing research questions.

Analysis Summary

We chose our analysis questions based on topics researchers might want to study when using our proposed platform.

Questions of interest

- How accurately can survival be predicted?
- What are the key factors affecting differences in survival rates?

Exploratory Analysis

- Evaluated survival over time.
- Evaluated survival by demographic factors.
- Analyzed outcomes by subtype.

Survival Analysis[1]

- Three-year survival analysis by stage which resulted in similarly high survival percentages for stages 0-2 and significantly lower survival for patients with stage 4 cancer.
- Three-year survival analysis by subtype showed lowest percentages for subtype triple negative and highest survival percentage for HER2+ subtype.
- Three-year survival analysis by marital status indicated married subjects had highest survival rates. Analysis by race showed Asian Americans had the highest survival rates while African Americans had the lowest.

Logistic Regression and Decision Tree[2]

- Performed logistic regression that predicted survival for stage 4 patients diagnosed in 2010.
- Statistically significant predictors of survival are age, race, tumor subtype, tumor stage, and tumor size.

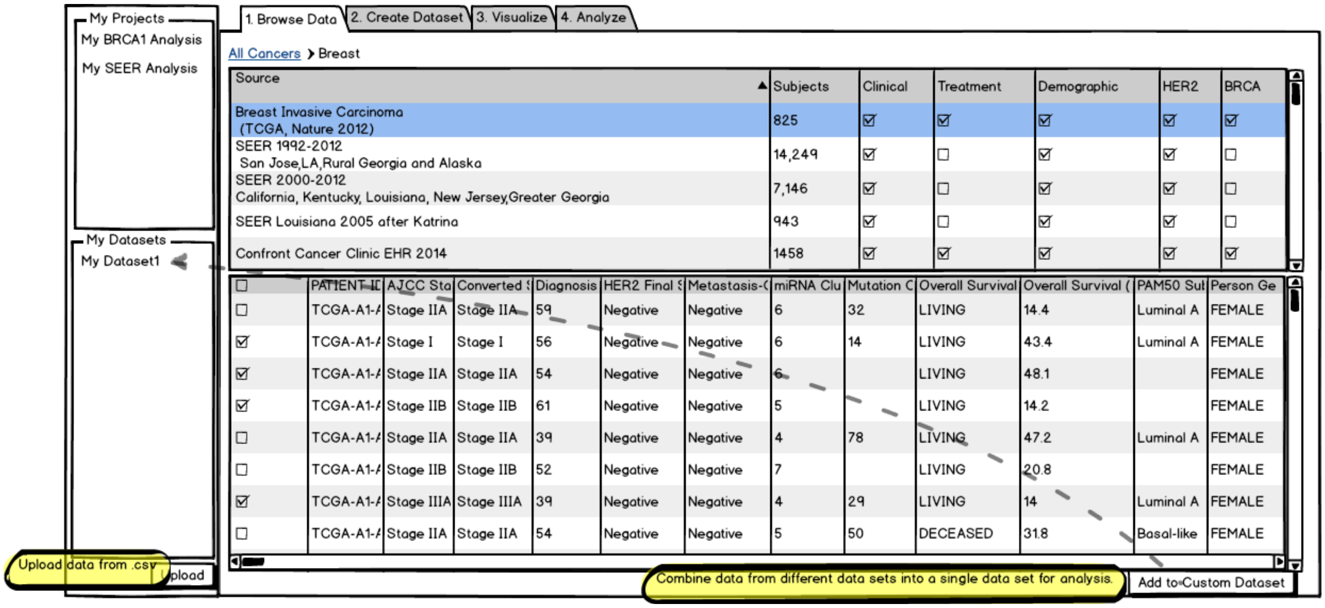
[1] See Survival Models Write-Up.pdf

[2] See Predictive Models Write-Up.pdf

- The decision tree for stage 4 patients resulted in a single split that showed clear differences in survival probability depending on subtype with triple negative having the lowest probability of survival

Proposed Solution Summary[3]

A breast cancer researcher would like to perform a deep-dive analysis to discover survival patterns in the population of women with breast cancer and to determine the key factors affecting differences in survival rates. It is around this high-level usage scenario that we propose a solution consisting of 4 ordered areas: Browse Data, Create Dataset, Visualize, and Analyze. A user may navigate between them at any time and save results in projects that can be accessed at a later time.



Source	Subjects	Clinical	Treatment	Demographic	HER2	BRCA
Breast Invasive Carcinoma (TCGA, Nature 2012)	825	☑	☑	☑	☑	☑
SEER 1992-2012 San Jose,LA,Rural Georgia and Alaska	14,249	☑	□	☑	☑	□
SEER 2000-2012 California, Kentucky, Louisiana, New Jersey,Greater Georgia	7,146	☑	□	☑	☑	□
SEER Louisiana 2005 after Katrina	943	☑	□	☑	☑	□
Confront Cancer Clinic EHR 2014	1458	☑	☑	☑	☑	☑

□	PATIENT ID	AJCC Sta	Converted S	Diagnosis	HER2 Final S	Metastasis-C	miRNA Clu	Mutation C	Overall Survival	Overall Survival (	PAM50 Sul	Person Ge
□	TCGA-A1-A	Stage IIA	Stage IIA	59	Negative	Negative	6	32	LIVING	14.4	Luminal A	FEMALE
☑	TCGA-A1-A	Stage I	Stage I	56	Negative	Negative	6	14	LIVING	43.4	Luminal A	FEMALE
☑	TCGA-A1-A	Stage IIA	Stage IIA	54	Negative	Negative	6		LIVING	48.1		FEMALE
☑	TCGA-A1-A	Stage IIB	Stage IIB	61	Negative	Negative	5		LIVING	14.2		FEMALE
□	TCGA-A1-A	Stage IIA	Stage IIA	39	Negative	Negative	4	78	LIVING	47.2	Luminal A	FEMALE
□	TCGA-A1-A	Stage IIB	Stage IIB	52	Negative	Negative	7		LIVING	20.8		FEMALE
☑	TCGA-A1-A	Stage IIIA	Stage IIIA	39	Negative	Negative	4	29	LIVING	14	Luminal A	FEMALE
□	TCGA-A1-A	Stage IIA	Stage IIA	54	Negative	Negative	5	50	DECEASED	31.8	Basal-like	FEMALE

Upload data from .csv

Combine data from different data sets into a single data set for analysis.

Figure 1 Dashboard

1. Browse Data

   A researcher can view a list of available datasets along with descriptions of the subjects contained in each, as well as a few indicators of whether it contains some research relevant data. Selecting a dataset makes the records in it available for browsing. The researcher can then select all or a subset of subjects to save to a custom data set.
2. Create Dataset

[3] See Clickable Mockups.pdf and Functional Design Proposal.pdf

The researcher can refine their data set, remove and merge variables, and view distributive information about each variable in the data set.

3. Visualize

   The researcher can further explore relationships between variables in their data by creating 2 variable box plots and scatter plots.

4. Analyze

   The researcher can use this customized data set to perform survival analysis, decision tree analysis, and logistic and ordinary regression.