Proposed Solution Functionalities

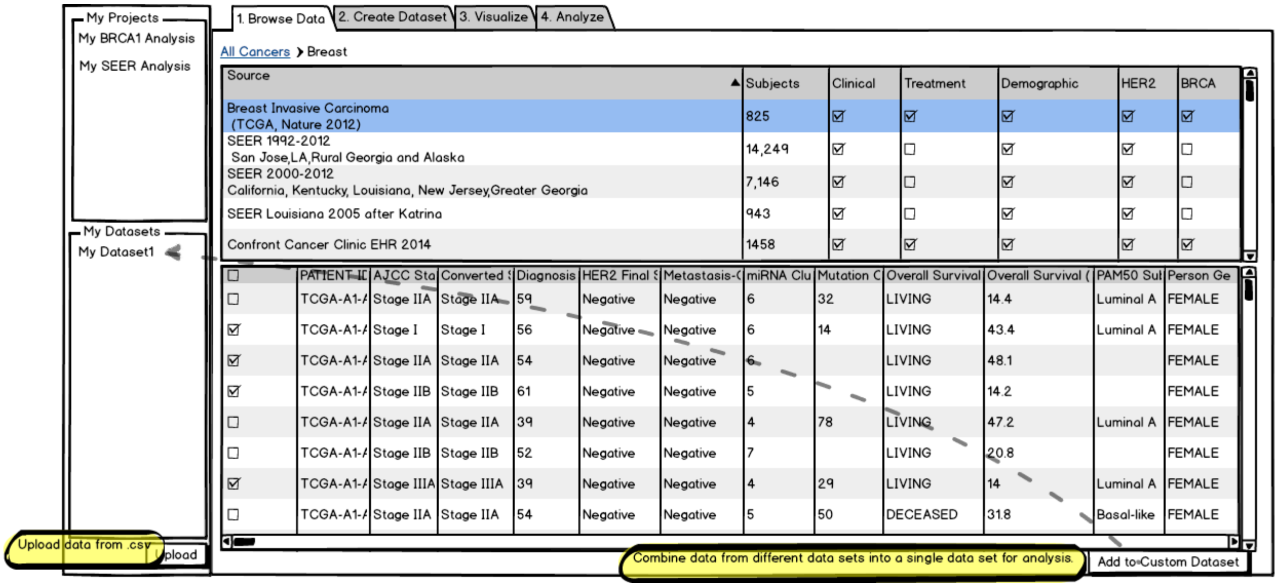
Georgia Tech Analytics Practicum 2015 -- Breast Cancer Research Platform
Sarah Bartlett, Lauren Crabtree, Andrea McCarter, Siddharth Shah

Purpose

This document describes the functionalities required for the proposed breast cancer research platform solution in the areas of Data Exploration, Visualization, and Analysis. It is designed to specify the layout, flow, and user interaction of the proposed system as imagined. A mockup is provided for each specified functionality, where applicable, in order to illustrate the capabilities more clearly than words sometimes can. This document does not detail implementation requirements or technical specifications.

Overview

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

Data Exploration

Connectivity and Data Availability

The system will:

1. Connect, synchronize, and integrate with publicly available data sets.
2. Connect to EHR systems that implement the FHIR API.
3. Accept uploaded data sets in .csv format.
   The user can select the “Upload” button, select a .csv file from their file system, and select “Ok”. When this happens the system will import the file as a custom data set.

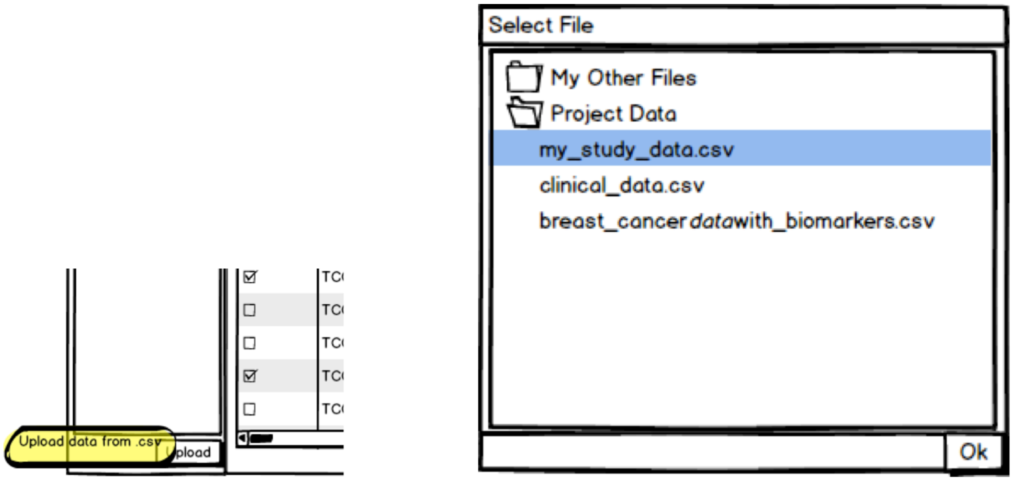
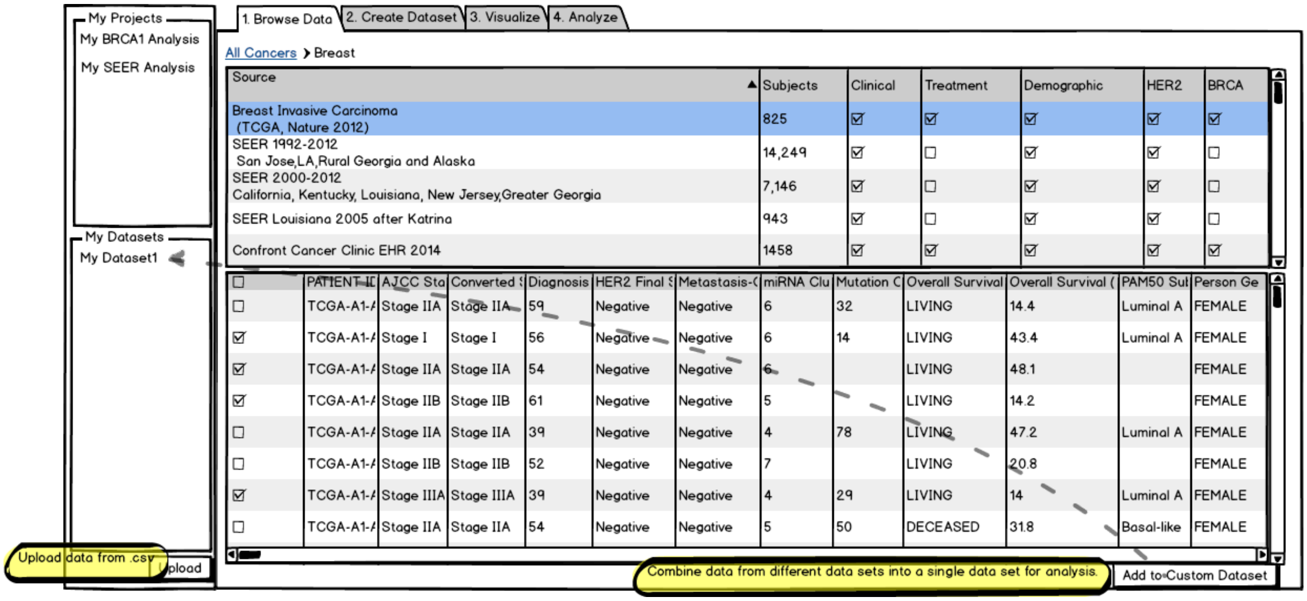


Figure 2 Data Connectivity 1.

Data Browsing and Collecting

In the "Browse Data" tab, the system will support the ability to browse through the available data sets and see relevant information about each data set so that they can select subjects from 1 or more data source for analysis.



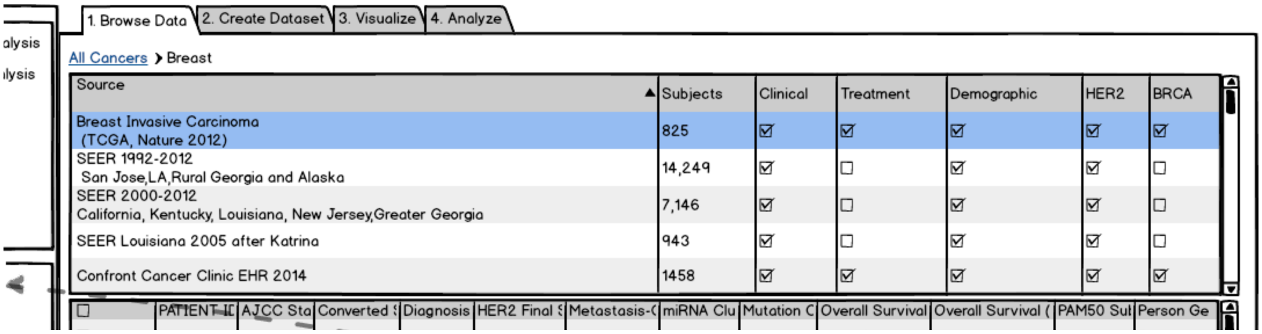
Source	Subjects	Clinical	Treatment	Demographic	HER2	BRCA
Breast Invasive Carcinoma (TCGA, Nature 2012)	825	☑	☑	☑	☑	☑
SEER 1992-2012 San Jose,LA,Rural Georgia and Alaska	14,249	☑	□	☑	☑	□
SEER 2000-2012 California, Kentucky, Louisiana, New Jersey,Greater Georgia	7,146	☑	□	☑	☑	□
SEER Louisiana 2005 after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

Combine data from different data sets into a single data set for analysis.
Add to Custom Dataset

Figure 3 Data Exploration Mockup

1. Browse available data sets and identify:
    - Number of patients
    - Source of data
    - Date of collection
    - Whether the data contains demographic, clinical, treatment, HER2, or BRCA relevant variables.

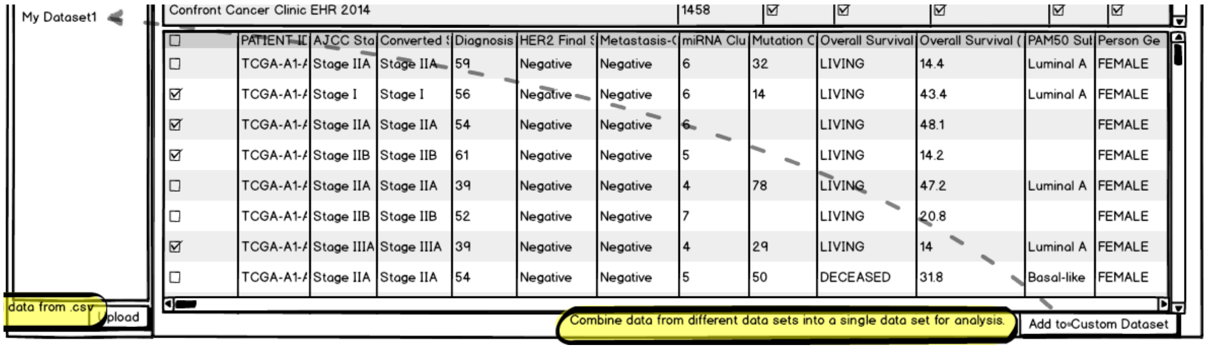


Source	Subjects	Clinical	Treatment	Demographic	HER2	BRCA
Breast Invasive Carcinoma (TCGA, Nature 2012)	825	☑	☑	☑	☑	☑
SEER 1992-2012 San Jose,LA,Rural Georgia and Alaska	14,249	☑	□	☑	☑	□
SEER 2000-2012 California, Kentucky, Louisiana, New Jersey,Greater Georgia	7,146	☑	□	☑	☑	□
SEER Louisiana 2005 after 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 Sub	Person Ge

Figure 4 Data Exploration 1.

2. Select records from one or more datasets and save to custom data set. The user may select all records or only certain cohorts by using check boxes on the left or at the top of the column. Selecting "Add to Custom Dataset" will prompt them to choose an existing data set under "My Datasets" or to create a new one.



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 Sub	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

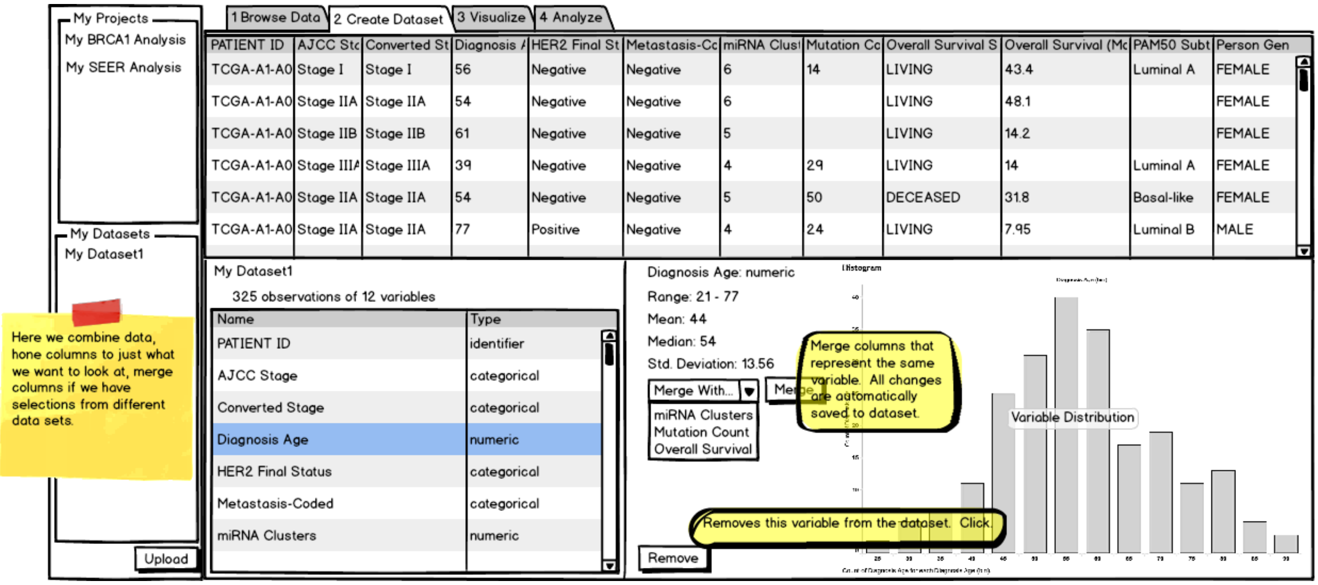
Combine data from different data sets into a single data set for analysis.

Figure 5 Data Exploration 2.

Prepare Data for Analysis

From the "Create Dataset" tab the user can:

1. Reduce and combine subsets into a single data set suitable for analysis.
2. Remove variables that are not of interest.
3. Merge variables of the same type for subjects originating in a different data set.



PATIENT ID	AJCC Stc	Converted St	Diagnosis A	HER2 Final St	Metastasis-Cc	miRNA Clust	Mutation Cc	Overall Survival S	Overall Survival (Mc	PAM50 Subt	Person Gen
TCGA-A1-A0	Stage I	Stage I	56	Negative	Negative	6	14	LIVING	43.4	Luminal A	FEMALE
TCGA-A1-A0	Stage IIA	Stage IIA	54	Negative	Negative	6		LIVING	48.1		FEMALE
TCGA-A1-A0	Stage IIB	Stage IIB	61	Negative	Negative	5		LIVING	14.2		FEMALE
TCGA-A1-A0	Stage IIIA	Stage IIIA	39	Negative	Negative	4	29	LIVING	14	Luminal A	FEMALE
TCGA-A1-A0	Stage IIA	Stage IIA	54	Negative	Negative	5	50	DECEASED	31.8	Basal-like	FEMALE
TCGA-A1-A0	Stage IIA	Stage IIA	77	Positive	Negative	4	24	LIVING	7.95	Luminal B	MALE

My Dataset1

325 observations of 12 variables

Name	Type
PATIENT ID	identifier
AJCC Stage	categorical
Converted Stage	categorical
Diagnosis Age	numeric
HER2 Final Status	categorical
Metastasis-Coded	categorical
miRNA Clusters	numeric

Diagnosis Age: numeric
Range: 21 - 77
Mean: 44
Median: 54
Std. Deviation: 13.56

miRNA Clusters
Mutation Count
Overall Survival

Here we combine data, hone columns to just what we want to look at, merge columns if we have selections from different data sets.

Merge columns that represent the same variable. All changes are automatically saved to dataset.

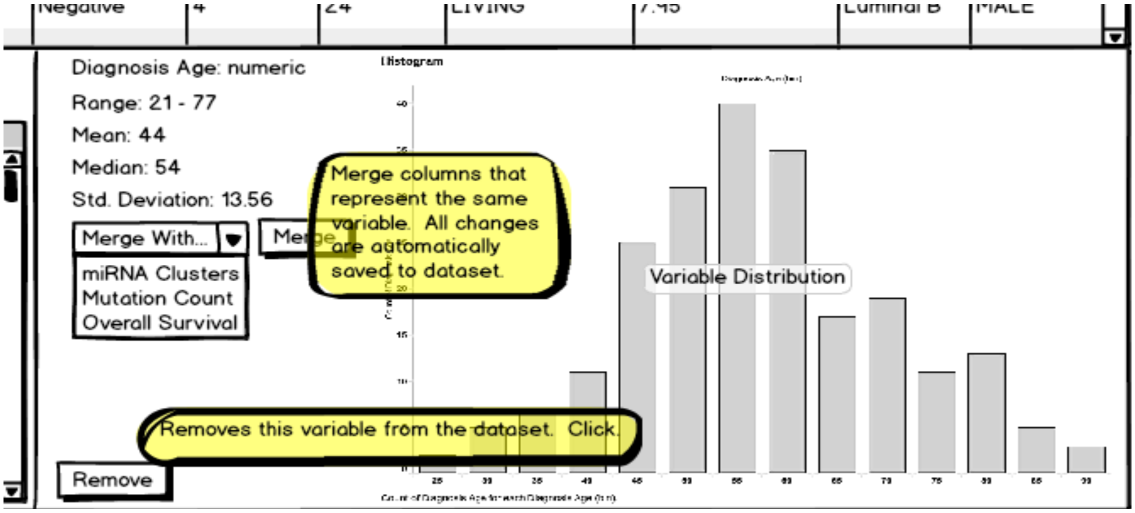
Removes this variable from the dataset. Click.

Figure 6 Prepare Data for Analysis.

Visualization

Descriptive Statistics

While still in the “Create Dataset” view, each variable in the data set may be selected in order to view descriptive statistics about it. For numeric variables this would be the mean, median, standard deviation, and data range. For categorical variables this would be a list of each category present in the data. For both numeric and categorical variables a visual display of the distribution of each will be shown in the form of a histogram or a bar chart.



Diagnosis Age: numeric

Range: 21 - 77

Mean: 44

Median: 54

Std. Deviation: 13.56

Merge columns that represent the same variable. All changes are automatically saved to dataset.

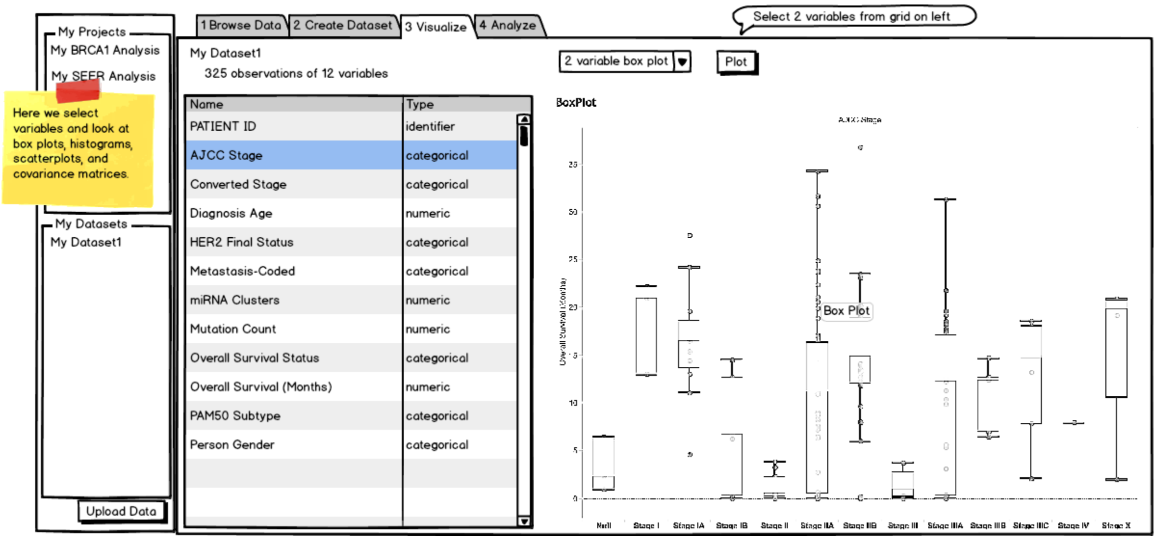
Removes this variable from the dataset. Click.

Figure 7 Descriptive Statistics.

Comparative Visualization

A researcher can further explore their data in the “Visualize” view. Initially the system will allow the visual comparison of 2 variables but it may be expanded in time to include more extensive visualizations. The proposed “out of the box” charts are box and whisker plots (box plots) and scatter plots. In either case, the user will select the two variables they wish to compare, select the chart type, and click “Plot”.

1. Box plots show the distribution of a numeric variable against a categorical variable's values. There are horizontal ticks for the minimum, median, and maximum values, and the box contains the values that lie within the 1st quartile to the 3rd.



Here we select variables and look at box plots, histograms, scatterplots, and covariance matrices.

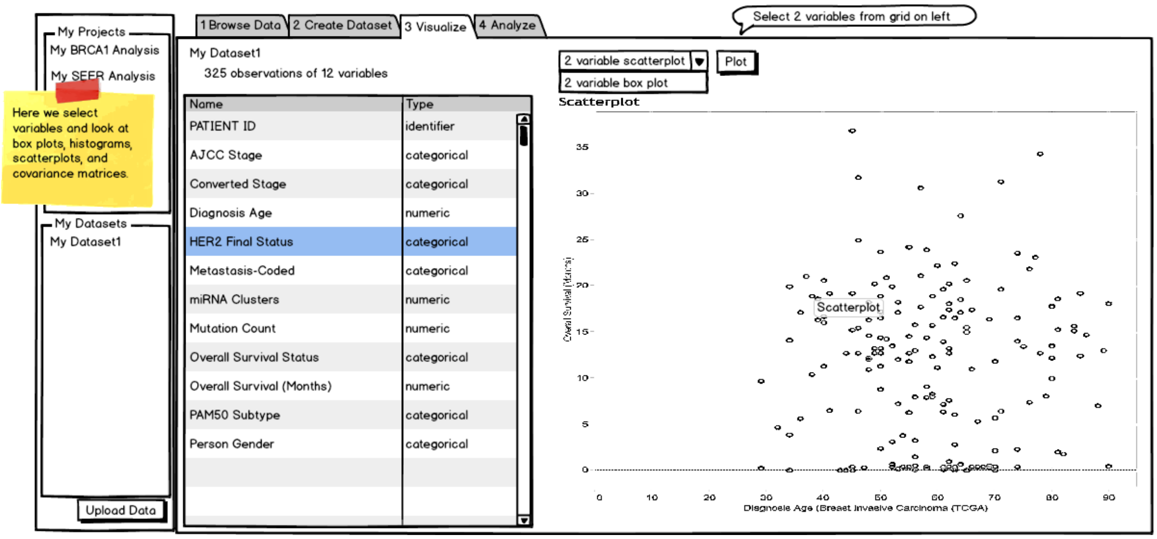
My Dataset1
325 observations of 12 variables

2 variable box plot ▾

Name	Type
PATIENT ID	identifier
AJCC Stage	categorical
Converted Stage	categorical
Diagnosis Age	numeric
HER2 Final Status	categorical
Metastasis-Coded	categorical
miRNA Clusters	numeric
Mutation Count	numeric
Overall Survival Status	categorical
Overall Survival (Months)	numeric
PAM50 Subtype	categorical
Person Gender	categorical

Figure 8 Box Plots.

2. Scatter plots compare two numeric variables. Each data point is plotted individually on the graph based on its x and y variables, for x and y representing the two numeric variables.



Here we select variables and look at box plots, histograms, scatterplots, and covariance matrices.

My Dataset1
325 observations of 12 variables

2 variable scatterplot ▾

Name	Type
PATIENT ID	identifier
AJCC Stage	categorical
Converted Stage	categorical
Diagnosis Age	numeric
HER2 Final Status	categorical
Metastasis-Coded	categorical
miRNA Clusters	numeric
Mutation Count	numeric
Overall Survival Status	categorical
Overall Survival (Months)	numeric
PAM50 Subtype	categorical
Person Gender	categorical

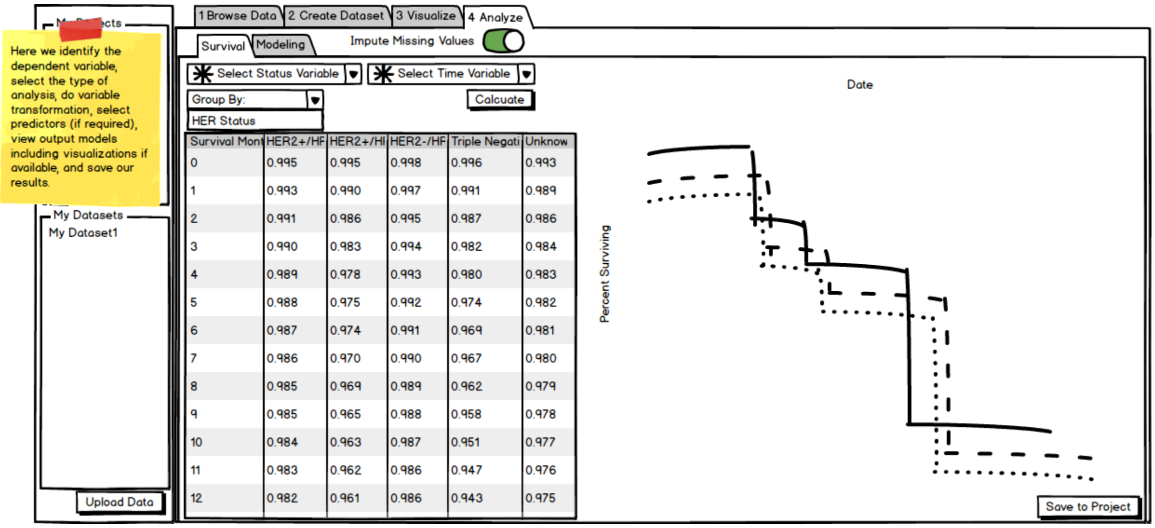
Figure 9 Scatter Plot.

Analysis

A researcher can conduct in-depth analyses on their custom data set from the "Analyze" view. Initially, the platform will provide two main types of analysis: survival analysis and predictive modeling. Predictive modeling will consist of logistic regression, ordinary regression, and decision trees. For all analyses, an option will be provided for imputing missing values using a probabilistic algorithm. Results of any analyses may be saved to the project and viewed later.

Survival Analysis

The system will support basic Kaplan-Meier survival analysis with up to one grouping variable.



Here we identify the dependent variable, select the type of analysis, do variable transformation, select predictors (if required), view output models including visualizations if available, and save our results.

Group By: HER Status

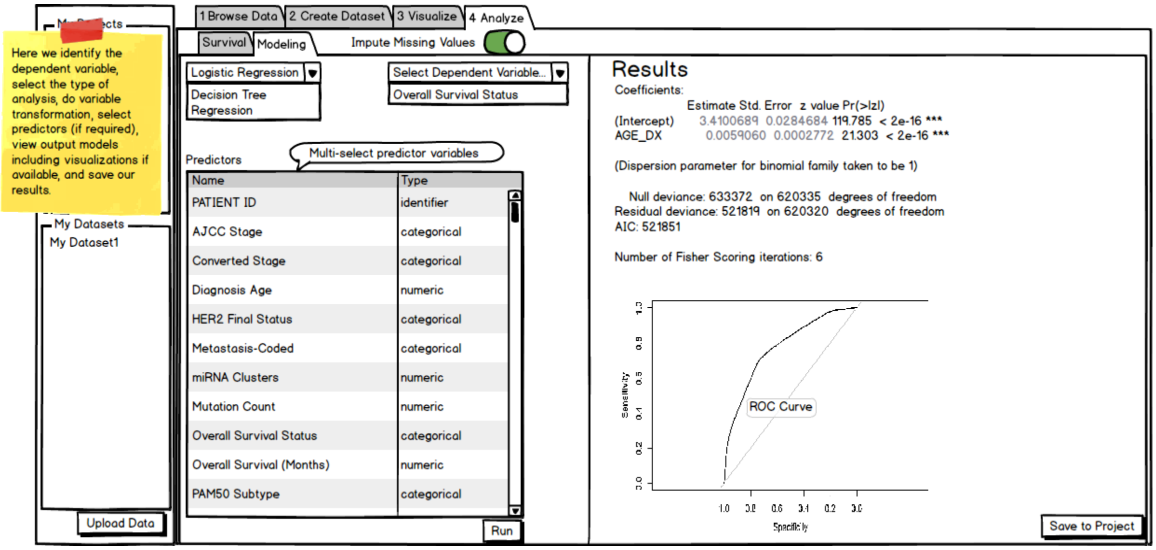
Survival Mont	HER2+/HF	HER2+/HI	HER2-/HF	Triple Negati	Unknow
0	0.995	0.995	0.998	0.996	0.993
1	0.993	0.990	0.997	0.991	0.989
2	0.991	0.986	0.995	0.987	0.986
3	0.990	0.983	0.994	0.982	0.984
4	0.989	0.978	0.993	0.980	0.983
5	0.988	0.975	0.992	0.974	0.982
6	0.987	0.974	0.991	0.969	0.981
7	0.986	0.970	0.990	0.967	0.980
8	0.985	0.969	0.989	0.962	0.979
9	0.985	0.965	0.988	0.958	0.978
10	0.984	0.963	0.987	0.951	0.977
11	0.983	0.962	0.986	0.947	0.976
12	0.982	0.961	0.986	0.943	0.975

Figure 10 Survival Analysis.

1. Select survival status variable, time variable, and an optional grouping variable and select "Calculate".
2. Survival Tables are calculated and shown under the variable selection.
3. The survival curve is plotted to the right. If grouping was selected, a line is plotted for each series belonging to a category.

Predictive Modeling

Initially, the system will provide the ability to do 3 types of predictive modeling: logistic regression, decision trees, and ordinary regression (for non-classification variables such as number of months surviving).



Here we identify the dependent variable, select the type of analysis, do variable transformation, select predictors (if required), view output models including visualizations if available, and save our results.

Name	Type
PATIENT ID	identifier
AJCC Stage	categorical
Converted Stage	categorical
Diagnosis Age	numeric
HER2 Final Status	categorical
Metastasis-Coded	categorical
miRNA Clusters	numeric
Mutation Count	numeric
Overall Survival Status	categorical
Overall Survival (Months)	numeric
PAM50 Subtype	categorical

Results

```
Coefficients:
               Estimate Std. Error  z value Pr(>Izl)
(Intercept)    3.4100689 0.0284684 119.785  < 2e-16 ***
AGE_DX          0.0059060  0.0002772  21.303  < 2e-16 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 633372  on 620335  degrees of freedom
Residual deviance: 521819  on 620320  degrees of freedom
AIC: 521851

Number of Fisher Scoring iterations: 6
```

Figure 11 Predictive Modeling Logistic Regression.

1. Select type of model, dependent variable, and predictors. Predictors are selected via multi-select in the grid containing all variables except the variable selected for prediction. Select "Run".
2. Results of the model are displayed.
3. If applicable, the system will produce a visualization of the results.
    a. Logistic Regression (shown): ROC curve
    b. Decision Tree: tree visualization
    c. Standard Regression: a plot of the regression line through a scatter plot of the actual values.

Appendix

A. Table of Figures

B. FHIR

FHIR is an API specification designed to help all health are systems speak a common language in order to make it easier to share data between them. It contains:

Model specifications for common health artifacts
Serialization formats
RESTful API query specification
A framework for extending the specification as needed
Resources for clinical, administrative, and infrastructure
Clinical resources of interest (not comprehensive)
DeviceObservation
Observation
DiagnosticReport
CarePlan
Medication

It is a trial standard, widely embraced but still early stage of adoption. It is not a software system but a set of rules for implementers of a software system to follow

Wherever we have a requirement to specify a common data format or any communication interface between systems we should defer to the FHIR standard and favor data sources that plan to implement it.

How FHIR fits into an EHR http://www.hl7.org/fhir/ehr-fm.html
Technical documentation for SMART on FHIR http://docs.smarthealthit.org