Analysis for Determining Cancer Target Clinical Trial Selection

J. Handzel

3/09/19

Table of Contents

Introduction

Biotechnology companies are developing promising cancer therapeutics to target what is expected to be a nearly $200 billion dollar national cost for cancer care by 2020. The development and approval process is extremely long and the funnel is very narrow. Many therapeutics enter, but very few make it to market. Once a therapeutic is identified, the path for approval is through required FDA clinical trials, long and costly process. For therapeutics with multiple targets, staging trials for efficiency, cost, impact and effectivity becomes a complex problem. Analysis to help determine candidate cancer trial staging would prove advantageous to these companies.

Problem

Many biotech companies are racing to develop cancer therapeutics, both drugs and biologicals, to meet the demand for cancer treatments. Those with multi-target therapeutics that may be effective in multiple types of cancers ask the question, “which target would provide the greatest impact in the U.S. market?” And the follow on question is, “How would you order clinical trials to maximize both ROI and impact?” This ranking or ordering could provide trial staging plans based on cancer types to provide the highest returns early in the product life cycle so that less prevalent or rare cancers could be addressed from profits from the early revenue stream. This is particular importance to smaller (or startup) biotech companies.

The National Institute of Health provides data on cancer and clinical trials including costs, prevalence, deaths and 5 year survival rate projections. This data can be analyzed to provide biotech companies a map to ensure the ROI of costly clinical trials meets their expectations on cost and overall public impact.

The goal in this analysis is to identify these cancer sites using these government datasets and provide combinations based on overall population size, population benefit, and healthcare costs. Companies can utilize the information to formulate ranking plans based on their specific therapeutic pathways to implement the staging of their clinical trials. This type of ranking can then be utilized to identify revenue stream size and trial phase costs to ensure company strength and viability.

Data Sets

The datasets are from the following websites:
https://clinicaltrials.gov/ct2/results?type=Intr&cond=cancer&map=NA%3AUS
https://costprojections.cancer.gov/expenditures.html

The datasets for costs, prevalence, deaths and 5 year survival rate projections are aggregated datasets that provide significant insight for companies seeking to determine which cancer types would provide positive ROI but also a significant pool of patient candidates available for treatment within a clinical trial. Given the data in Figures 1, trials that target breast, prostate, and lung cancer appear to have the highest annual costs in the U.S.. In Figure 2, breast, prostate, lung and lymphoma have the highest prevalence. Target therapeutics in these areas would seem to be good initial targets for clinical trials. Figures 3 and 4 show that the annual deaths and 5 year survival rates for each cancer site.   These provide data for companies that may help them avoid clinical trials on sites where cancer deaths are very high and 5 year survival is very low if the trials length are expected to be very long.

The clinical trial dataset provides historical data on past clinical trials.   Factors from the trials will be included in the analysis to evaluate various attributes in the dataset. An initial review of the data shows the number of trials based on each cancer site (type) in Figure 5 and the end status counts of the trials in the dataset in Figure 6.   From Figure 5 we can see that the largest group of trials is based on leukemia followed by breast cancer and then lymphoma and melanoma.   From Figure 6, we see that most trials are completed but several hundred, about 1/3 of trials are terminated or otherwise stopped.

There is no single variable or factor in the clinical dataset that indicates if the therapeutic had a statistically positive or negative end point on the patient or cancer. The clinicaltrials.gov website defines the status category as follows (ongoing studies have been excluded from the list below and the dataset:

Recruitment status

- **Completed:** The study has ended normally, and participants are no longer being examined or treated (that is, the last participant's last visit has occurred).
- **Terminated:** The study has stopped early and will not start again. Participants are no longer being examined or treated.
- **Withdrawn:** The study stopped early, before enrolling its first participant.
- **Unknown:** A study on ClinicalTrials.gov whose last known status was recruiting; not yet recruiting; or active, not recruiting but that has passed its completion date, and the status has not been last verified within the past 2 years.
- **Suspended:** The study has stopped early but may start again.

A termination is likely indicative of a problem with either the efficacy or side effects.   The other status categories, affecting very small numbers of trials, include withdrawn, suspended or unknown status also likely point to negative indications within the trial or the therapeutic agent itself.

Variables for machine learning

Status: dependent variable, “Completed” is considered a successful trial(1) and the others noted above indicate an unsuccessful trial (0).

Site (type): 18 cancer types:

1	Bladder
2	Brain
3	Breast
4	Cervix
5	Colorectal
6	Esophagus
7	Head_Neck
8	Kidney
9	Leukemia
10	Lung
11	Lymphoma
12	Melanoma
13	Ovary
14	Pancreas
15	Prostate
16	Stomach
17	Uterus

Enrollment: Numeric
Phases: Pre1, 1, 1|2, 2, 2|3, 3, 4
Intervention: Drug,Biological, or Both
Cancercost: Numeric
Estimated New Cases (by site per year): Numeric
Estimated Deaths (by site per year): Numeric
5 year Survival (% by site): Numeric

Data Limitations

The variable most needed to determine trial success would be a positive/negative statistical significance indicator for the intervention.   There were other cancer categories, including liver and thyroid, in the trials dataset but they were eliminated due to lack of consistent data for cost, new cases, deaths and 5 year survival numbers that were readily available in the aggregated National Institute of Health’s datasets.

The clinical trials data set is not straightforward nor entirely consistent. Given that there is a significant trend in increasing clinical trials, much more could be done to provide more robust data and input consistency. Other papers have noted that clinical trials website updates appear slowly and are inconsistent with datasets in other countries. The conclusion seemed to place the onerous on the scientists to repeatedly update rather that have a proactive system. Addressing the database and entry itself would likely be advance trial planning and success and if done correctly, make input by the research teams easier and more failproof.

Data Wrangling

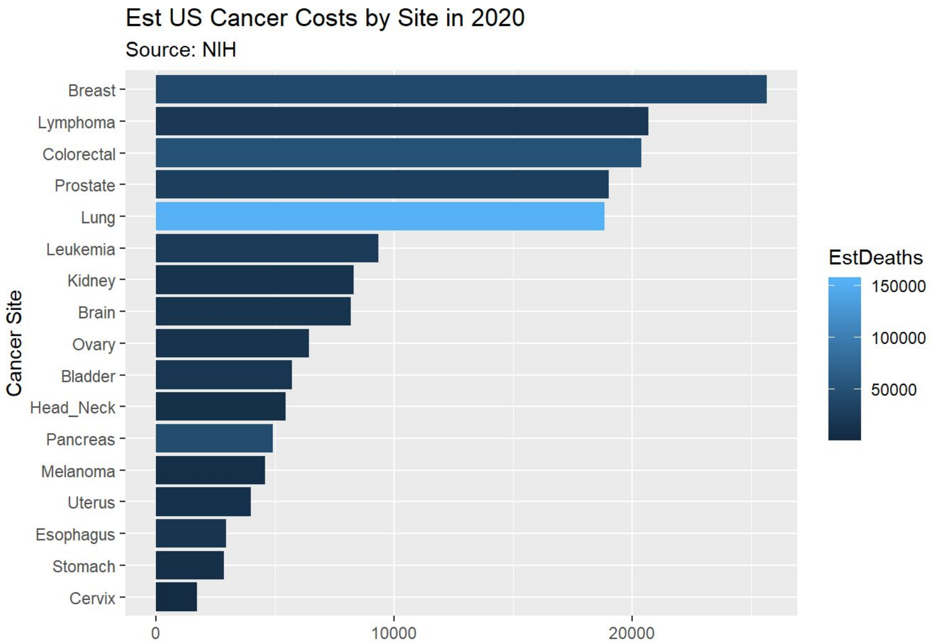
The data contained missing values and inconsistencies as noted. The following steps were completed to provide for a tidy dataset.

1. column names to reflect the data
2. Removal of cancer site variables with incomplete data
3. Clean variables to remove extraneous characters: Site, Drug/Biological, Phase
4. Removal of observations that were incomplete or inconsistent across the dataset
5. Removal of NA’s

This step was extremely time consuming because much of the data was contained in complex character strings. Within each major type of cancer there are many subtypes. For instance, in Brain cancer, there are neuroblastomas and various astrocytomas and gliomas. Each one needed to be categorized within the major subtype.

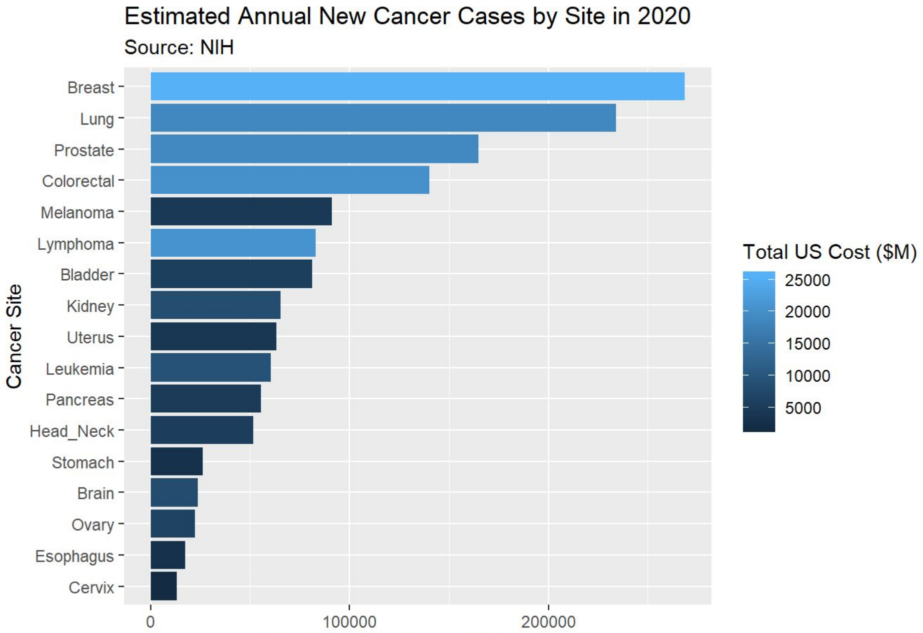
Preliminary Analysis

In the preliminary analysis we first want to explore each of the aggregate independent variables and the relationship to the cancer site.   Figures 1-4 below show these relationships.


Est US Cancer Costs by Site in 2020

Source: NIH

Figure 1: Cost in US Millions


Estimated Annual New Cancer Cases by Site in 2020

Source: NIH

Figure 2: Estimated Annual New Cases

Estimated Annual Cancer Deaths by Site in 2020

Source: NIH

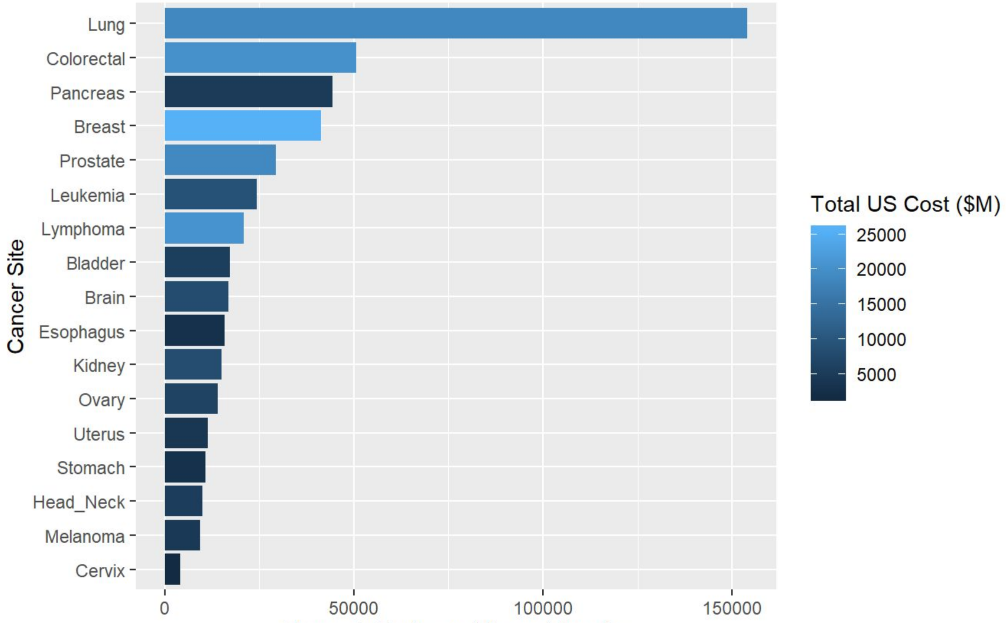


Figure 3: Estimated Annual Deaths

5 year Survival rate by Cancer Site

Source: NIH

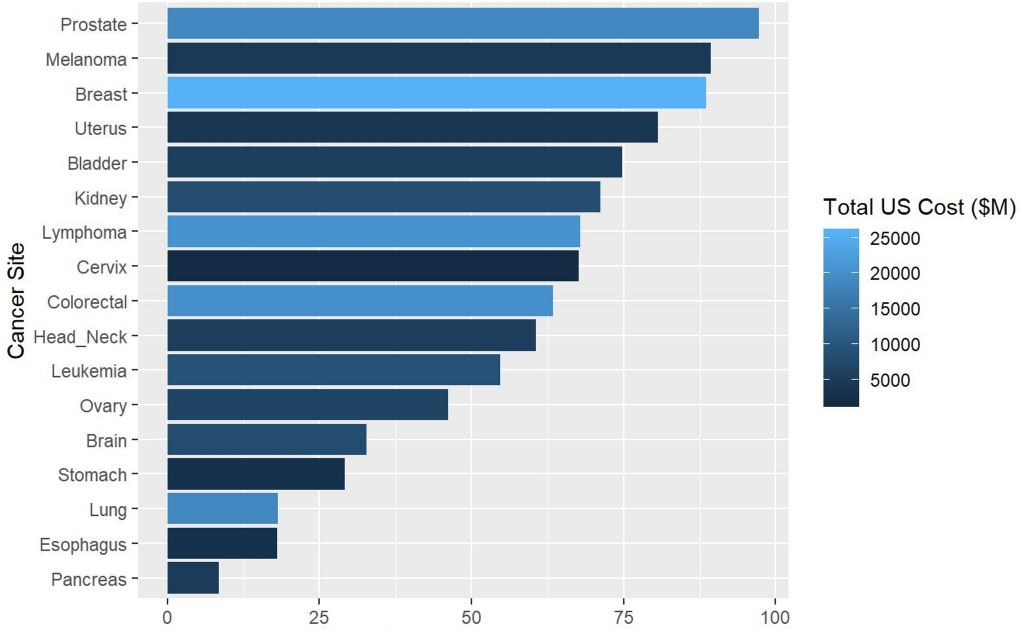


Figure 4: 5 year survival rate - %

Figures 5 and 6 below show the relationship of the large clinical trial dataset and the trial.

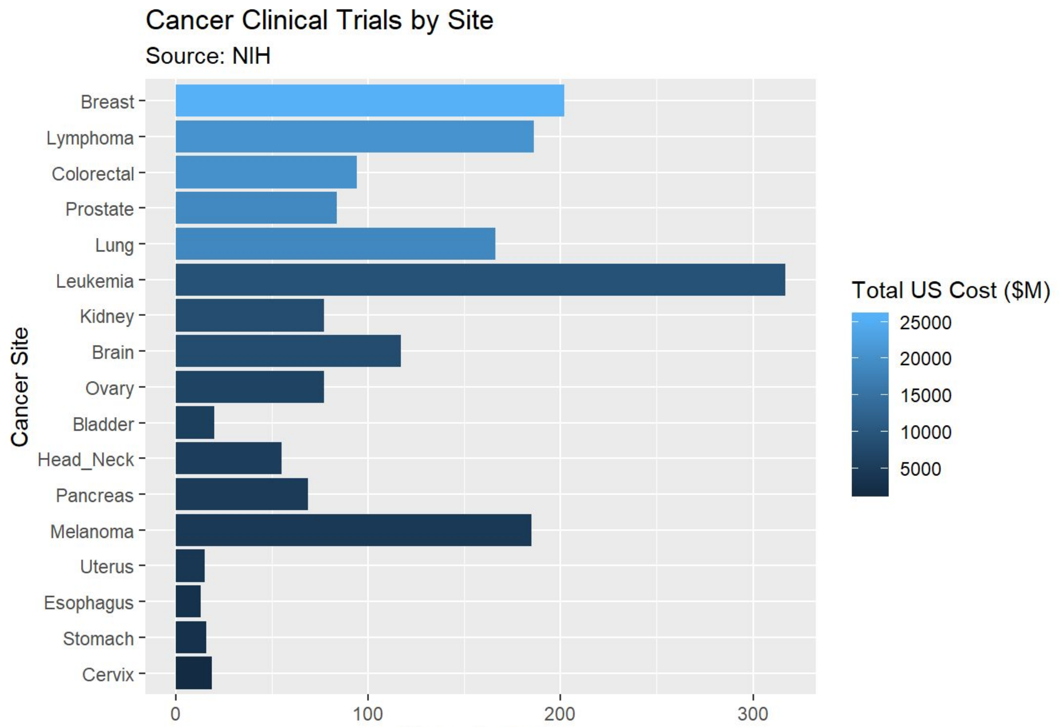


Figure 5: Count

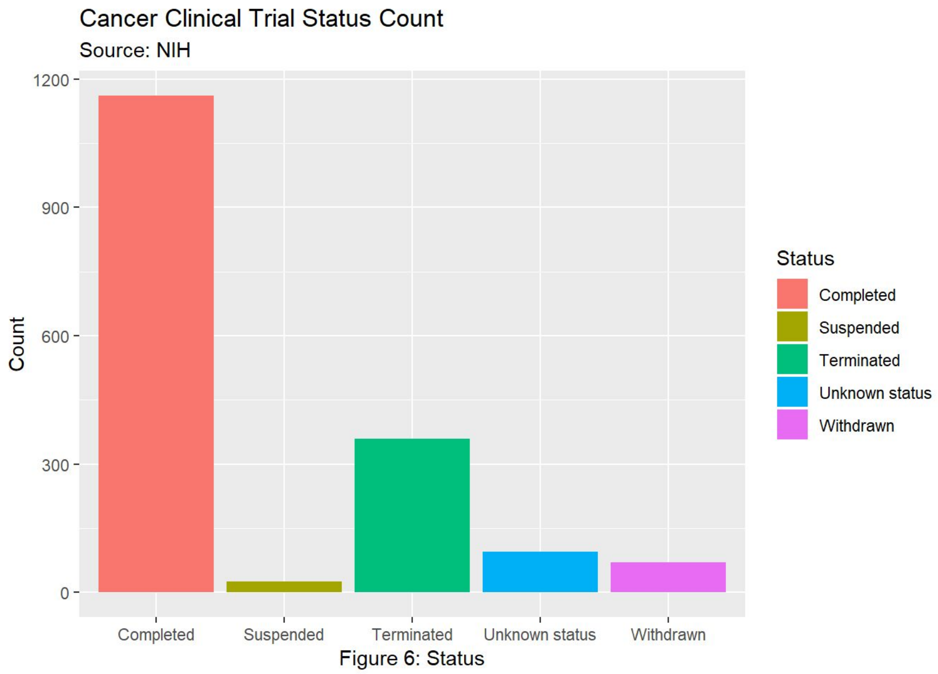


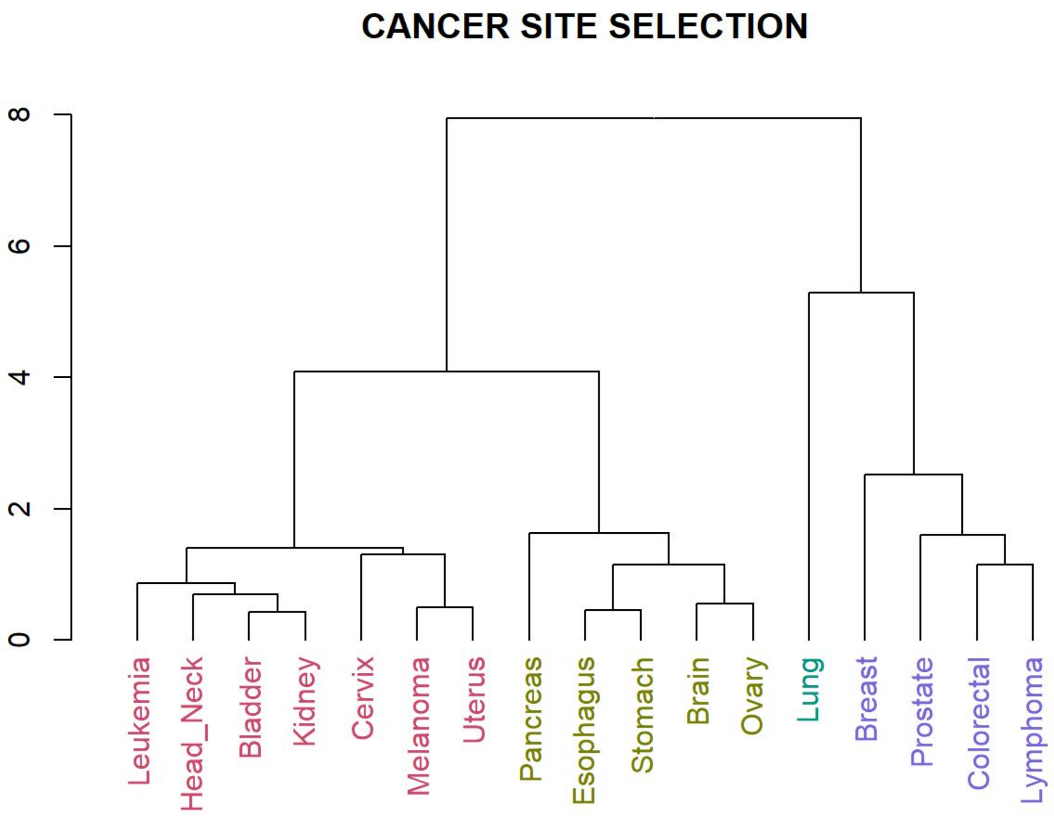
Figure 6: Status

Machine Learning

Now that we have a general view of the variables in clinical trials and an aggregate view of cancer costs, cases, deaths and survival, we will use different machine learning methods to explore the data to assist in the clustering of types of trials and attempt identify key cancer types and variables that indicate significance in the data We will use unsupervised learning to explore hierarchical structure and K-means clusters of the data for cancer selection.

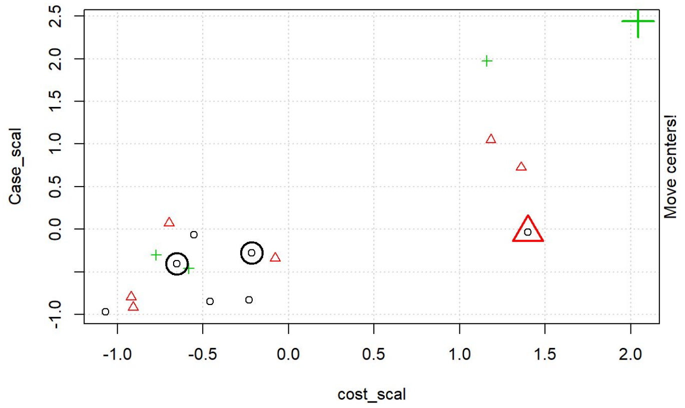
Hierarchical Clustering

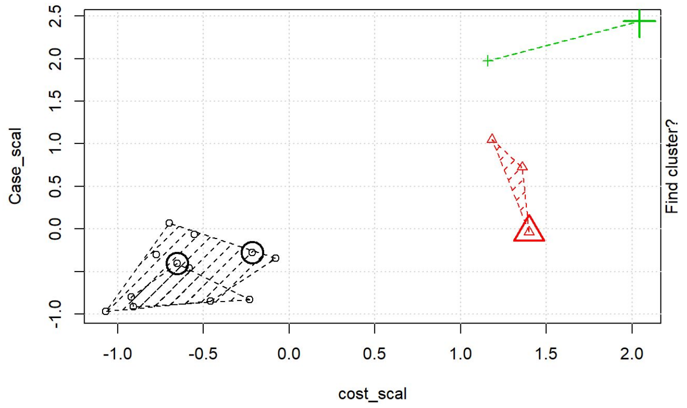
This analysis produces a nested set of clusters by analyzing similarities between pairs of sites and grouping objects into a hierarchical tree.   The resulting dendrogram showing the hierarchical relationship between clusters of cancer sites. Lung cancer is a single cluster due to high costs, cases, and attributable deaths.   Another cluster contains breast, prostate, colorectal and lymphoma likely due primarily to cost and number of cases.

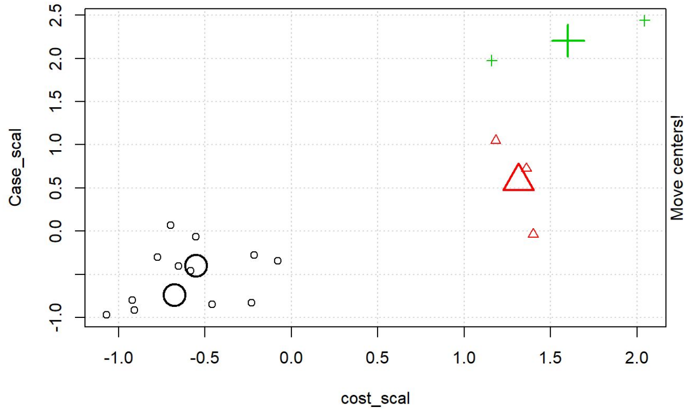


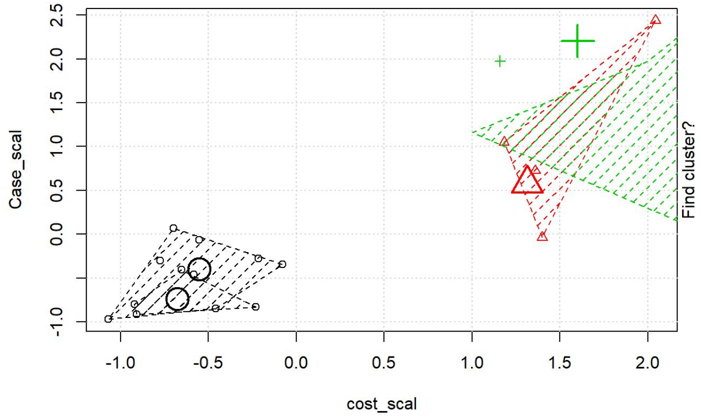
K-Means Clustering

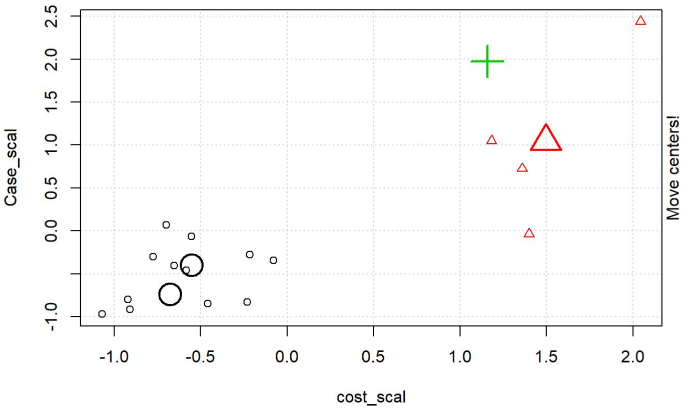
In this second method, we partition the full data into k number of mutually exclusive clusters. How well a point fits into a cluster is determined by the distance from that point to the cluster's center. Given the structure found above in the hierarchical clustering, we set k=4. Using the animations package, we observe the convergence of the clusters in the following figures. Note that the simulation can only display a two dimensional (2 variables) output but it is clear that more than just the variables cost and cases are significant. Convergence happens quickly and just like above, lung cancer observations, appears in a single cluster (green).

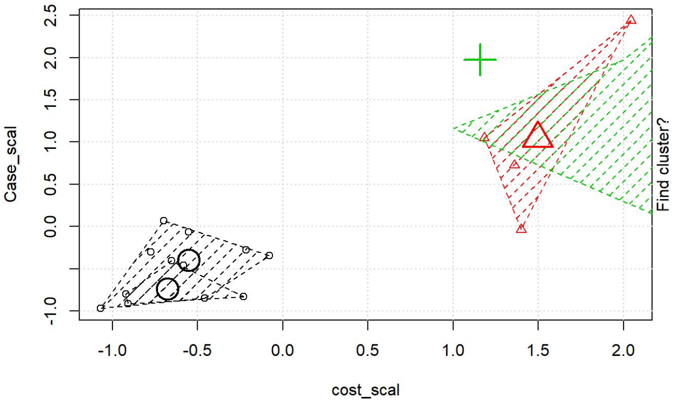





Case_scal
cost_scal
Move centers!
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5


Case_scal
cost_scal
Find cluster?
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5

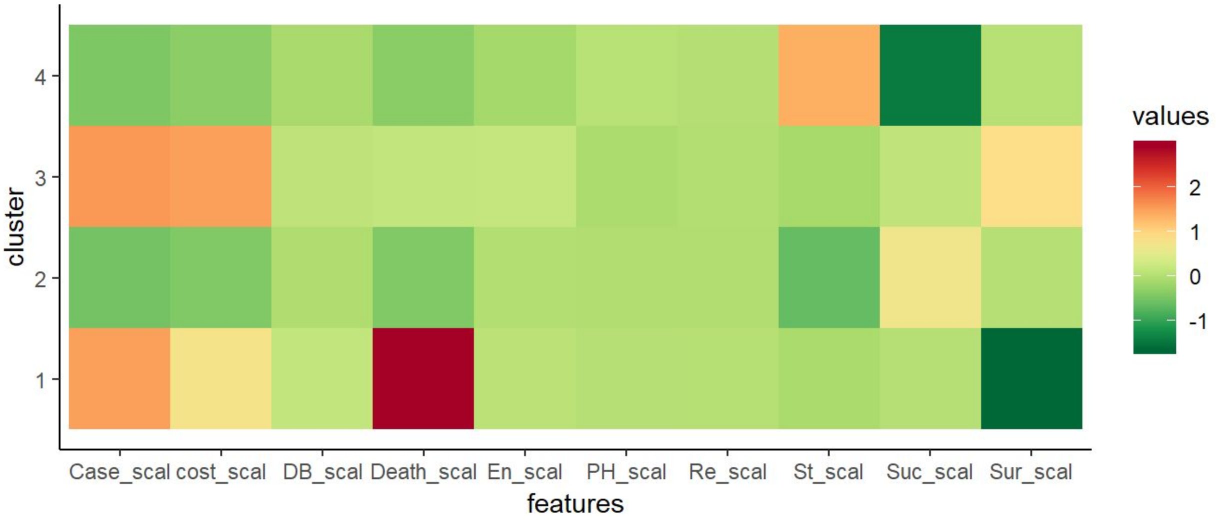




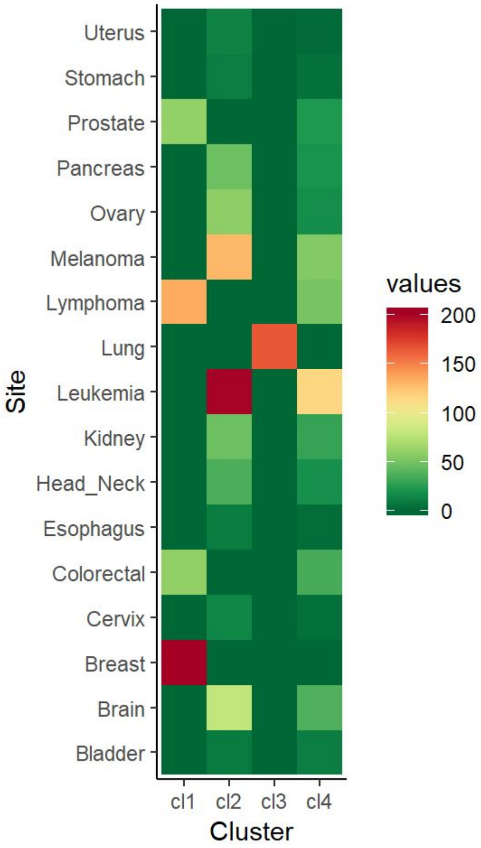
The red cluster, as similar to the dendrogram above, contains observations of breast, prostate, colorectal cancers and lymphomas.   The two black clusters at the bottom, appear to overlap but the heatmaps identify the difference between the two clusters.

Heatmaps from the K-means Clustering

A heatmap of the variables and their relationship with each cluster is shown below. The darker the color the higher the significance of the variable. The map shows that the variables: estimated deaths, 5-yr survival rates and the trial success (completion), had the most significance in the dataset. Lighter green and yellow across all clusters indicate that the particular variable was not significant and could be removed from the dataset. The difference between clusters 2 and 4 are related to trial status and number of trials. The data suggests that there are no other variables in the clinical trials dataset that would provide additional information or yield a more robust model to predict clinical trial success.



A second heatmap shows a relative frequency of the observations within the cluster and cancer site type. The darker the color the higher the relative frequency of observations with that cluster and cancer site type. The map shows that cluster 3, correlated with the observations with the highest estimated deaths, assigned all lung cancers to this cluster. Cluster 1 contained observations with high cost and new cases variables. Clusters 2 and 4,are related to be related to trial frequency (counts).



Conclusion

Both unsupervised learning methods provided the useful information for the selection of clinical trial targets. As expected, both clustering methods indicated that lung cancer with the corresponding mortality rate and prevalence put it at the top of the rankings to target for any multi-cancer therapeutic clinical trials. They both also showed that breast, prostate, colorectal and lymphoma all ranked in the next tier and would be good candidates for trials.

While the two methods, after a simple inspection, appear to have identified the two obvious clusters noted in the previous paragraph, the hierarchical method broke down the remaining sites (lower left of the tree) in interesting and very logical ways. The first level of lower grouping appears to split this group into high and low survivability clusters. The high survivable group is split again into two clusters based on lower mortality and lower overall costs. These divisions provide additional information for companies that may have therapeutics that target cancers in these low cluster levels but need additional differentiation information.

Given the high cost of treatment for cancer, Biotech companies developing cancer therapeutics can use these clusters to identified cancer targets and provide a ranking that would help maximize both early profits and populations affected. Of the two clustering methods, the hierarchical method provided a deeper insight, and a useful dendrogram for ranking target candidates.

The resulting information, in the form of the hierarchical structure dendrogram, is also of great use to public health experts that are working to reduce overall healthcare cases and costs related to cancer trial funding. NIH is a major contributor to cancer trials. Given that funding comes from the public, cluster information such as this, would be of great significance in lowering the overall costs to U.S. healthcare.

Recommendations

1. There are several areas that would benefit from further analysis and additional data.
2. Additional cost, estimated cases and, deaths and survival rates on other types of cancers, such as liver and thyroid, would make this information more complete.
3. Similarly, detailed data within each high level category would be of use to researchers or companies that target a specific area such as brain or breast cancer.
4. The cancer trials dataset needs to include success factors (positive or negative) for each primary, secondary (and beyond) outcome measures.
5. The K-means analysis could be repeated with a k=3 and k=5 and non-significant variables removed.
6. Ultimately a tool could be developed where researchers/companies could specify a number of types of cancers and the output would order inputs based on the hierarchical system above.