Analysis for Determining Cancer Target Clinical Trial Selection

J. Handzel
3/9/2019

Introduction - Cancer Therapeutics

Biotechnology companies are developing promising cancer therapeutics at record paces but approvals are long and costly:

- Global pharmaceutical market share for cancer therapeutics: 18%.
- Cancer care national cost by 2020: ~$200 Billion.
- Approval is through required FDA clinical trials: ~$80 Million.
- New challenges in finding enough patients to recruit for clinical trials.

Problem: Staging trials for multi target therapeutics based on cost, impact and effectivity becomes a complex problem. This analysis helps determine candidate cancer trial staging based on cost and impact.

Data Sets

The datasets are from the following National Institute of Health 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 combined with the clinical trial dataset to provides a data set for machine learning analysis.

Variables for machine learning:

- Sites: (Cancer types)s, Numerics
- 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

Preliminary Analysis

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.

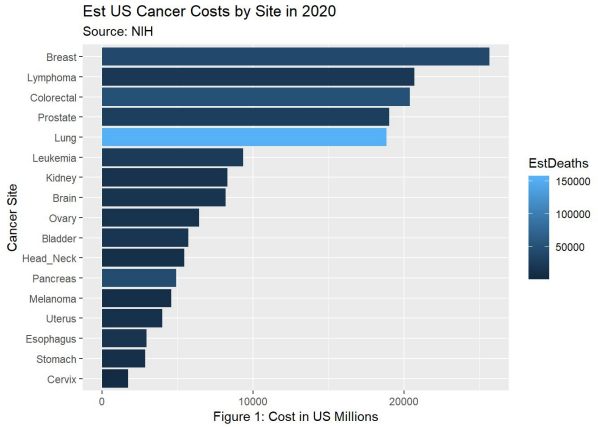


Figure 1: Cost in US Millions

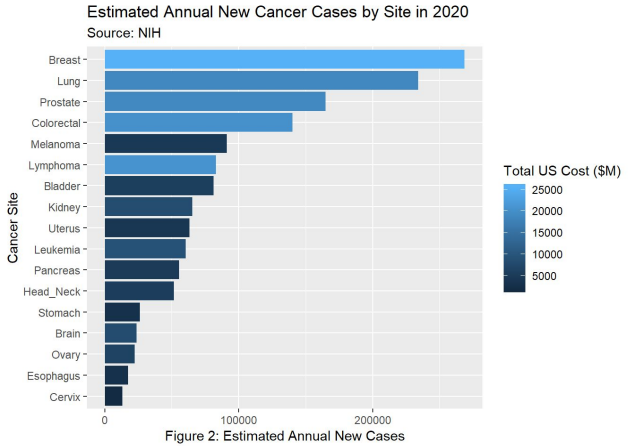


Figure 2: Estimated Annual New Cases

Preliminary Analysis (con't)

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.

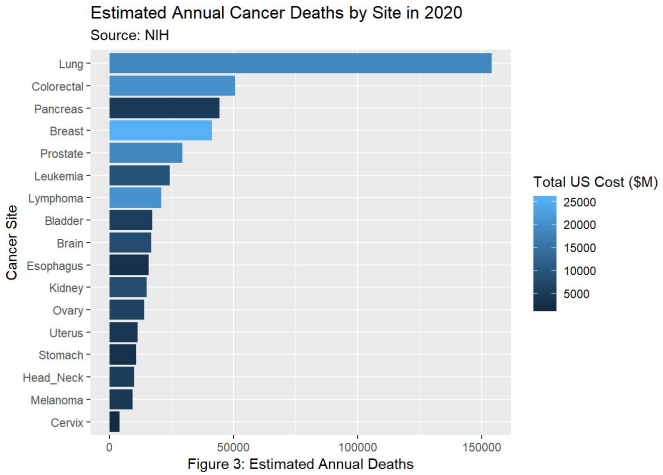


Figure 3: Estimated Annual Deaths

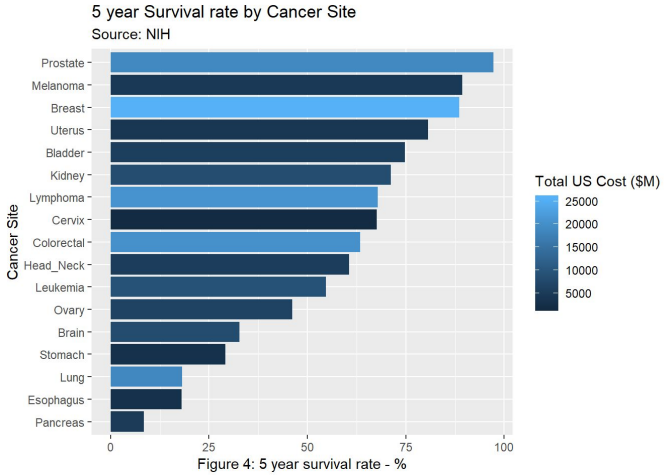


Figure 4: 5 year survival rate - %

Preliminary Analysis (con't)

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.

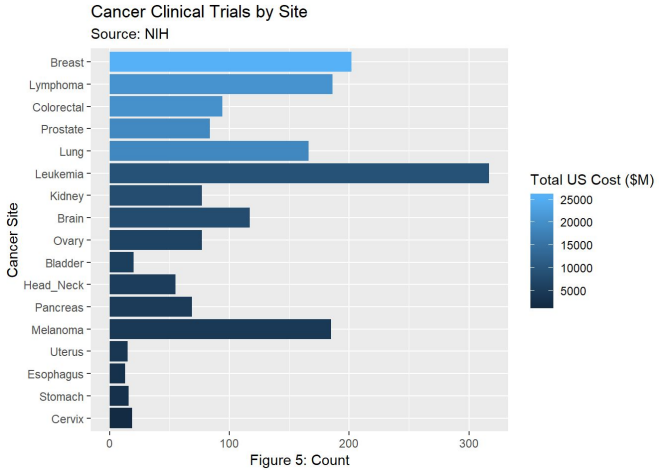


Figure 5: Count

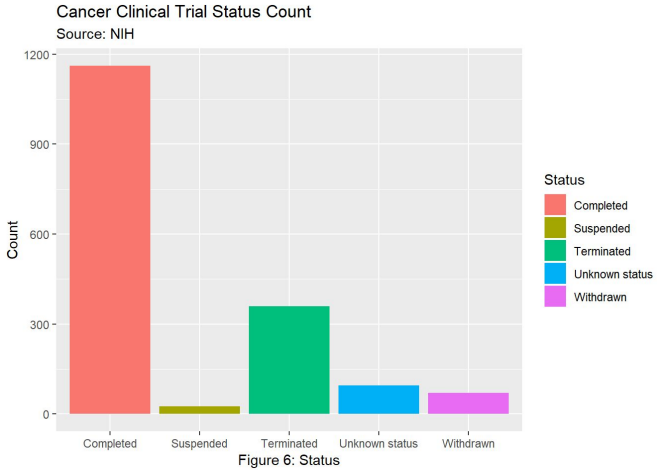
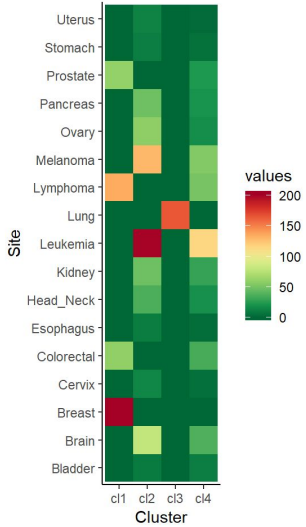


Figure 6: Status

Machine Learning - K-means Clustering

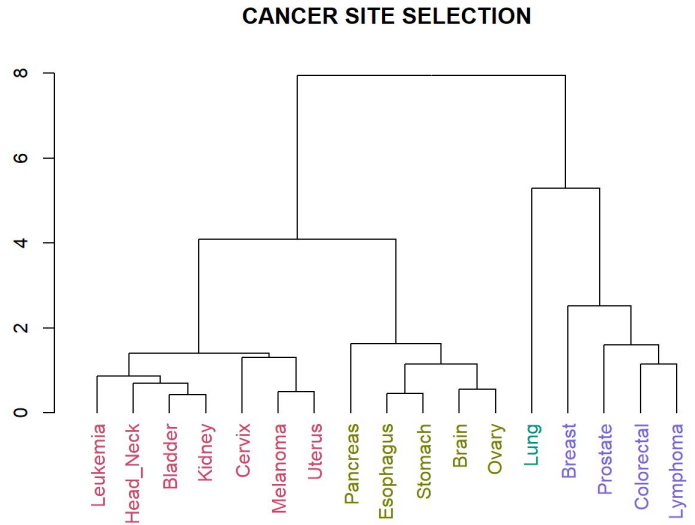
In this Machine Learning method, we partition the full data into k number of mutually exclusive clusters. How well a point fits into a cluster is determined by the distance from that point to the cluster's center. We set k=4. Below is resultant heatmap.



This heatmap shows a relative frequency of the observations within the cluster and cancer site type. The darker the color the higher frequency of observations with that cluster and cancer site type.

- **Cluster 1 contained most of the prostate, lymphoma colorectal and all breast cancers.**
- **Clusters 2 and 4 contained the observations not in cluster 3 and are related to be trial frequency (Leukemia the highest) per site.**
- **Cluster 3: assigned all lung cancers to this cluster.**

Machine Learning - 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 estimated deaths.**
- **Breast, prostate, colorectal and lymphoma due primarily to cost and number of cases.**
- **Remaining cancers are in low incidence clusters. One with low survivability (Pancreas) and one with higher survivability (Leukemia).**

Conclusion

The analysis showed:

1. **Both unsupervised learning methods provided the useful tiered information for the selection of clinical trial targets.**
2. **Both clustering methods indicated that lung cancer with the corresponding mortality rate and prevalence put it at the top.**
3. **Both methods showed that breast, prostate, colorectal and lymphoma all ranked in the next tier would be good candidates for trials based on high prevalence and costs.**
4. **The hierarchical method broke down the remaining low prevalence sites (lower left of the tree)**
    a. **The first level of lower grouping splits this group into a low survivability cluster (Pancreas, Ovary, Stomach, Brain, Esophagus).**
    b. **The high survivable group is split again into two clusters based on medium overall cost (Leukemia, Head and Neck, Bladder and Kidney) versus the group with lower overall costs (Cervix, Uterine, Melanoma).**

Given the high cost of treatment for cancer and clinical trial costs, biotech companies developing cancer therapeutics can use these clusters to identified cancer targets rankings to maximize both early profits and populations impact. Of the two clustering methods, the hierarchical method provided a deeper insight, and a useful dendrogram for ranking target candidates.