Optimal Treatment Strategies for Cancer patients in terms of Survival Months and Socio-Economic Factors

by

OMER MOGULTAY

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
May 2020

To my parents Gulsum and Balabey (Kamil),

And my sister Elif and my brother Emre

ACKNOWLEDGEMENTS

It is a genuine pleasure to express my deep gratitude to my advisor Dr. Leili Shahriyari for her expertise, assistance, guidance and patience throughout my Ph.D. study. Her guidance will keep leading me in my further professional life. Without her support, this journey would not have been possible.

I would like to give special thanks to Dr. Tuncay Aktosun, for all his support and help. Besides my advisors, I would like to thank my committee members Dr. Hristo Kojouharov and Dr. Andrzej Korzeniowski for sharing their insight on this subject and for taking time to serve in my comprehensive committee and dissertation committee.

Finally, I would like to thank all my friends at UTA for all the fun and laughter we have shared.

April 27, 2020

ABSTRACT

Optimal Treatment Strategies for Cancer patients in terms of Survival Months and Socio-Economic Factors

Omer Mogultay, Ph.D.

The University of Texas at Arlington, 2020

Supervising Professor: Leili Shahriyari

One of the main challenges of cancer patients and their healthcare providers is making decisions regarding choosing the best treatment option. In the first part of thesis, we analyze breast cancer patients' data to discover characteristics of patients who would benefit from each breast cancer surgical procedure in terms of increasing survival months.

Since the outcome of breast cancer treatments strongly depends on the tumor subtypes, several studies investigated the outcome of surgical procedures for each of these subtypes. On the other hand, it has been shown that the outcome of breast cancer treatments is significantly different between black and white patients. These treatment comparison analyses were mostly done using traditional statistical methods. Here, we integrate statistical methods and machine learning techniques to perform a comprehensive analysis and consider not only patients' clinical data but also demographic information as well as gene expression profile of tumors.

To determine the optimal surgical procedure for each racial group of breast cancer patients with a given tumor subtype, we analyzed clinical and gene expression data sets of 1082 patients with breast invasive carcinoma. We used K-mean clustering with both clinical information and gene expressions to find the best treatment option in the intersections of data sets. We further investigated characteristics of patients' tumors in each group by performing gene set enrichment analysis (GSEA). Our results indicate that the outcome of surgical procedures is a function of race, subtype of the tumor, and gene expression data of primary tumors. Importantly, we also found that radiation therapy have increased survival of white and black patients.

Although survival months is the main factor in making decision regarding cancer treatments, it is not certainly the only important factor. Cancer patients also think about the impact of cancer treatments on their quality of life and careers because of many factors, including side-effects and the cost of treatments. For example, the most common side-effect of cancer treatments is dizziness, which reduces the ability of patients in driving. This minor side effect might completely change cancer patients' life if the only way to get to work is driving. The main goal of the second part of this thesis is to investigate the role of transportation in decision making of cancer patients and their quality of life. To reach this goal, we created a survey and utilize the recent advances in data science to analyze the collected data. We employed machine learning algorithms to identify the characteristics of patients who might benefit from free/discounted rides.

TABLE OF CONTENTS

CHAPTER 1

Introduction

In recent years, rapid advancements in information technology (IT) and computer science led to the formation of various online communication platforms, which altered the way that society functions and transformed a wide variety of industries including healthcare. These advances drove us to a new era, often referred to as the “information age” or “Third Industrial Revolution.” As a result, the healthcare industry has been rapidly evolving, and cloud-based medical records and telemedicine are just two examples of many emerging game-changers. A majority of providers made the switch from paper to electronic medical records (EMR). Moving the records into the cloud and letting patients check and monitor their data allowed the teams of healthcare providers and patients to work together to reach a common goal. Saliently, patients now are able to take a more active informed role in their own care. The main goal of this dissertation is to investigate optimal treatment strategies for cancer patients and to help them in decision making regarding their cancer treatments but at the same time we also want to explore what are the factors important in their quality of life. Is accessibility to the healthcare providers an important factor in the diagnosis or prognosis? Does transportation have any role in making decisions regarding choosing, changing, or quitting a particular treatment?

Although there have been many advances in communication strategies between patients and healthcare providers, recent studies indicate an urgent need for organized, integrated, and patient-centered information and support for cancer patients and their families [1]. Patient-centered care aims to improve clinical practice by building caring relationships between clinicians and patients. Patient-centered care can be achieved by creating environments in which clinicians and patients, and by extension patients' family members, engage in two-way sharing of information to explore patients' values and preferences, help patients and their families in making clinical decisions, and facilitate access to appropriate cares [2]. Furthermore, each patient might have different sets of concerns, some of which might be related to financial problems and the ability of patients to have the treatments. It is important to provide an environment in which patients can freely share their concerns and beliefs so that caregivers would be able to learn about them. The awareness of these concerns and beliefs will help physicians to knowledgeably approach discussions with patients about treatment options.

Scientific decision making is becoming more popular because of the availability of big data and advances in data science. Recent advances in machine learning led to the development of innovative models for analyzing various scenarios and forecasting the possible implications of decisions. The decision to choose a treatment option and choosing between quantity and quality of life is extremely hard for cancer patients and their family members. In most cases, both the patients and their partners are involved in making decisions about treatments' strategies. One of the main goals of this proposed project is to make the process of decision making easier for patients and healthcare providers and help them to make better "scientific" decisions.

Making decisions about cancer treatments is very difficult for some patients, mainly because their definition of effective treatments differs from healthcare providers' definition. While 90% of physicians define the effectiveness of cancer treatment as extending expected survival months, for 45% of patients it means the preservation of quality of life [3]. In one study, fewer than 20% of patients ranked either "effect of treatment on length of life" or "chances of dying of cancer" as one of the four most important factors in making a decision about pursuing a treatment [4]. On the other hand, two-fifths of male patients were unconditionally willing to risk side effects for any potential gain in life expectancy [5]. These studies suggest that treatment efficacy means more than survival months for many patients. In this thesis, we first investigate the optimal surgical procedure for breast cancer patients based on their demographic and clinical information. We then investigate if the ability of patients to get to a particular location, including a healthcare provider or their job, would have an effect on their decision about following a treatment.

CHAPTER 2

Survival Months of Breast Cancer Patients

2.1 Background

Breast cancer is the most commonly diagnosed cancer among women. Although there has been a continuous decline in the death rate of patients with breast cancer in the past decade, the mortality rate is still high due to heterogeneity of the disease [6]. The Surveillance, Epidemiology, and End Results (SEER) program estimates 268,600 new cases of women with breast cancer and approximately 41,760 deaths due to this disease in 2019 in the United States. Importantly, the death rate of black patients with breast cancer is higher relative to white and Asian patients. Between the years 2012 and 2016 in the United States, the death rate of black breast cancer patients was 28.1, while the death rate of white and Asian breast cancer patients were respectively 20.1 and 11.2 in the same period [7].

Several studies have reported a noticeable variation in the rate of occurrences of breast cancer subtypes among racial groups [8, 9, 10, 11]. For example, the distribution of subtypes considerably varies between black and white women. Moreover, significant biological differences have been observed between the tumors of black and white women in clinically defined subgroups [9].

A categorization of breast cancer tumors has been standardized by the study of Sorlie *et al.*[12]. Molecular level analysis of gene expression patterns has revealed five subtypes, which are namely luminal A (LumA), Her2 over-expression (Her2), luminal B (LumB), normal-like and basal tumors, and the most frequent subtypes are LumA and LumB. Although luminal tumors have a poor response to the conventional chemotherapy [13], unlike Her2 and basal subtypes, luminal subtypes carry good prognosis [14]. LumA subtype, which is the most frequent subtype of breast cancer, has the lowest mortality rate among all subtypes [15, 16].

Disparities of breast cancer subtypes were observed by age [17, 18, 19]; young patients have higher proportions of basal-like tumor than older patients. Moreover, subtype proportions considerably varied by race among younger women [10]. In general, young women receive more aggressive treatments than older women [20, 21].

There is an argument that the higher death rate of black patients compared to white and Asian breast cancer patients might be because of the highest rate of the occurrence of basal-like subtype in black patients [22, 23, 24]. However, a study of women after 2 years of diagnosis with ER-positive tumors (e.g. luminal and normal tumors) reveals that white women have higher survival than black women [25].

There are several treatments options for breast cancer that might ultimately effect the outcome. For instance, there are two common surgery options: lumpectomy or breast-conserving surgery (BCS) and mastectomy [26]. The choice of surgical treatment depends on many factors, including socio-demographic status, geographical and personal beliefs [27, 28]. However, older women tend to choose mastectomy while young women are likely to choose lumpectomy [29]. Various randomized controlled trials reported that there is no significant survival difference between lumpectomy and mastectomy treatments [30, 31, 32]. Based on these studies, National Institutes of Health recommended lumpectomy over mastectomy for breast cancer patients with stage I and II in 1990 [33]. The idea was to preserve the breast since survival rates were equivalent. Although there is not enough evidence to say lumpectomy provides better overall survival than mastectomy, Whelan *et al.* [34] suggested radiation therapy after lumpectomy treatment for the long term outcome. Another study for women with breast cancer, in particular localized ductal carcinoma in situ, reported that radiation therapy might be more useful after lumpectomy instead of lumpectomy alone [35].

Since surgical treatment comparisons were mostly analyzed with traditional statistical methods in a similar concept, there is a need for contemporary comparison of surgical procedures utilizing popular data analysis methods including machine learning algorithms. Recently investigators have applied machine learning algorithms on cancer data [36, 37, 38]. Clustering algorithms have shown to be a good approach for breast cancer data sets. In this study, to find optimal treatment strategy for cancer patients, we analyzed the survival months of 1082 patients with breast invasive carcinoma as a function of various factors, including race, surgical procedure, radiation, and gene expression data of primary tumors. We collected the data of The Cancer Genome Atlas (TCGA) project from cBioPortal; Tables in Figures show an overview of demographic and clinical features of the patients. These tables provide the number of patients in each subcategory, including race, type of tumor, surgical procedure, etc. Differences in the numbers are due to missing information for some patients.

2.2 Statistical Analysis

2.2.1 Mann-Whitney Test.

Statistical analysis is a key aspects of many areas including computational biology and clinical research. Statistical analysis helps researchers to make conclusions from their experiments. Often people use parametric tests to evaluate their data but parametric tests are based on certain assumptions such as the assumption of normality and equality of variances. When assumptions are violated one can use non-parametric tests. Non-parametric tests are usually referred as distribution-free and they are appropriate when sample size is small. A popular non-parametric test for two group comparisons is Mann-Whitney's U-test [39]. The Mann-Whitney U-test does not evaluate actual data points rather it considers ordering of observations. The null hypothesis H_0 of the test states that the distribution of two independent groups are identical. The alternative hypothesis H_1 in a two-sided test states that first group of data distribution differs from the second group of data distribution. The Mann-Whitney U-test compares of each data point from two groups and observations must be ordered from lowest to highest. Suppose all a_i belongs to the group A and all b_j belongs to the group B. The total number of possible pairwise comparisons that can be made is $n_a n_b$, where n_a is the number of observations in group A and n_b is the number of observations in group B. For a two-tailed test,

$$H_0 : P(a_i > b_j) = 0.5$$

$$H_1 : P(a_i > b_j) \neq 0.5$$

If the null hypothesis is true, each observation of the group A will have an equal chance of being larger or smaller than each observation from the group B. The test statistics U can be defined as following for each group:

$$\begin{aligned} U_a &= n_a n_b + ((n_a(n_a + 1))/2) - R_a \\ U_b &= n_a n_b + ((n_b(n_b + 1))/2) - R_b. \end{aligned} \tag{2.1}$$

where R_a and R_b are sum of ranks assigned to each group A and group B, respectively. Calculating $U = min(U_a, U_b)$ and using statistical tables for Mann-Whitney U-test, we get the p-value. If p-value is less than a statistical threshold α, we reject H_0. Since the Mann-Whitney U-test [39] and the Wilcoxon's rank sum test [40] use the same statistics, these tests are statistically equivalent.

2.2.2 Dimension Reduction and Feature Selection

Principal Component Analysis

High dimensional data sets have been generated from DNA microarrays and RNA-Seq experiments. There is a great need to develop techniques to analyze large amount of data. Principal component analysis (PCA) is a popular approach to reduce dimensionality of such large data sets while preserving most of the information [41]. Reducing dimension can be useful for visualization, exploration of the data and it also decreases computational time when using machine learning algorithms. It transforms data set into principal components (PCs). These new variables (PCs) are uncorrelated with each other and they are linear functions of original data set such a way that first PCs have the largest variance. Let us briefly explain optimization idea behind PCA. Let X be our $n \times p$ data matrix. We seek a linear combination of features $X_1, X_2, ..., X_p$ for the PCs that maximizes variance. For example, for the first PC;

$$C_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p, \tag{2.2}$$

where $a_1 = (a_{11}, a_{21}, ..., a_{p1})'$ referred as PC loading vector and $\sum_{k=1}^{p} a_{k1}^2 = 1$, which is a restriction to avoid large variance. In order to focus on variance, we can center data to make column means of the data matrix X to be zero. We then form the optimization problem using sample feature values as

$$\max_{a_{11},\ldots,a_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\sum_{k=1}^{p} a_{k1} x_{ik})^2 \right\}$$

subject to $\sum_{k=1}^{p} a_{k1}^2 = 1$. It is important to note that columns have zero mean, the mean of $c_{11}, ..., c_{n1}$ will also have zero mean. That means our objective maximizes sample variance of the n values of c_{i1}. A common approach to solve this optimization is to use eigenvalue decomposition that can be found in more details here [41]. Similarly, we can find other PCs.

A typical PCA algorithm follows these steps:

1. Create $n \times p$ dimensional data matrix X, where n and p are the number of samples and features, respectively.
2. Subtract mean from each row vector x_n in X.
3. Calculate covariance matrix of X.
4. Find eigenvalues and eigenvectors of the covariance matrix.
5. Sort eigenvalues in ascending order to get largest eigenvalues and corresponding eigenvectors.
6. Output these eigenvectors as PCs.

Variance Threshold.

Variance threshold is a basic approach to select variables. It is very common to do feature selection to focus on relatively more important features in data analysis. The underlying idea in variance threshold is that high variant variables contain more information. Selecting a threshold depends on data and researcher.

2.2.3 K-mean Clustering

One of the most common unsupervised clustering algorithm is K-mean. K-mean has a rich and long history in different scientific disciplines such as biology, statistics, computer science and psychology [42, 43, 44]. The algorithm is simple and efficient, therefore, it has been used to find underlying structure and classification purposes [45]. Given a data set of n data points $H = x_1, x_2, ..., x_n$ such that each data points is in R^D. Let $P_1, P_2, ..., P_K$ denote mutually exclusive clusters such that $\bigcup_{i=1}^{K} P_i = H$, $Pi \cap Pj = \emptyset$ for all $i \neq j$. Let μ_i be the empirical mean of cluster P_i, also known as cluster centroids. Main problem is to find the minimum within-cluster variation for cluster P_i.

$$\sum_{k=1}^{K} \sum_{x_j \in Pi} \|x_j - \mu_i\|^2 \tag{2.3}$$

The main steps of K-means algorithm are as follows [46]:

1. Cluster assignments P will be done by randomly assigning a number from 1 to K to each data points. Repeat steps 2 and 3 until the cluster assignments do not change.
2. Compute the centroids μ_i for each cluster K.
3. Assign each observation to the closest cluster centroid.

This algorithm will reduce the total within-cluster variance at each step. Since algorithm converges to a local optima, one must run the algorithm many times with different random partitions. It is important to find optimal number of clusters. A method called *elbow* can help us to find optimal K which basically shows decrease in sum of square distances while K increases.

2.2.4 Survival Analysis.

Microarrays and RNA sequencing technologies have created various opportunities to explore many different complicated tumor types including breast cancer. Studies have shown that survival of cancer patients can be predicted not only using clinical features of patients but also genes sets [47, 48]. Some survival analysis techniques, such as the Kaplan-Meier and the log-rank test have been used to explore predictive bio-markers.

The Kaplan-Meier (KM) method is a non-parametric technique to analyze time-to-event data [49]. Most often survival times are visualized by KM curves. KM does not require the knowledge of the underlying distribution of patients' survival times, but it requires censoring information. Then the log-rank test is used to compare survival times of two or more groups.

2.2.5 Gene Set Enrichment Analysis (GSEA)

High-throughput measuring technologies such as DNA microarrays and RNA-Seq have become powerful tools in recent years. These technologies can be used to study changes in gene expression profiles of thousands of genes simultaneously. Initially, a substantial number of papers has been proposed to explore underlying biological mechanism by evaluating differential expression of single genes. Among these studies a similar statistical approach can be seen in the most of the gene expression microarray experiments by starting with a null hypothesis to test individual gene at a time and determining a p-value for differential expressions. Then, a penalty method can be applied to p-values for multiple hypothesis testing. These gene-gene methods have been useful to explore many significant biological changes in gene expression data sets, however, some concerns were raised due to the high variability in microarray data set and the difficulty of getting useful biological insight from the long list of differentially expressed genes [50]. Multiple approaches have been utilized to address these problems. A common method is gene set enrichment analysis(GSEA) [50, 51] that based on the Kolmogorov Smirnov (K-S) test and it has already been extended and generalized in a various ways [52, 53].

Mootha *et al.* [50] first introduced GSEA by using a Kolmogorov-Smirnov (K-S) like statistic. Given a pre-defined genes set S such as genes in a known metabolic pathway, or sharing the same Gene Ontology (GO) category, GSEA aims to find the elements of S tend to be found at the top (or bottom) of a ranked list L or randomly distributed throughout L. Furthermore, a signal-to-noise ratio (SNR) is implemented for ranking list L. Since the null distribution of the K-S statistic was based on various size of gene sets, this method was not sensitive enough for each gene statistics. A remedy seems to be provided for this limitation in [51]. We give more detail in the next paragraph.

Subramanian *et al.* [51] proposed a GSEA method that uses very similar technique with slightly modified version of initial GSEA method. The idea is to see the difference between distribution of the genes and uniform distribution. This modified version of original method has three steps. Firstly, an enrichment score (ES, the maximum deviation from zero) needs to be calculated by walking down the ranked list L. If a gene is in pre-defined gene set S then a running-sum statistic goes up and otherwise it goes down. Each increment is evaluated by correlation of the gene with phenotype. Then a weighted version of K-S statistic is performed to get the ES which was a modification of the initial method. In the next step, in order to get a null distribution of the ES, a permutation of the phenotype labels is applied and the ES is re-assessed for each set. After that, by using the null distribution, they were able to get nominal p-values of the ES. Since many hypothesis testing have been performed, a correction method is also applied after normalization of the test statistics. The false discovery rate (FDR) is used to control the proportion of false positives as a correction method for each gene set. In the original method, Mootha *et al.* [50] used equal weights for each gene, however, Subramanian *et al.* [51] used a weighted genes according to their correlation with phenotype. Let us briefly show the difference between two methods.

Let the genes be ordered by SNR (basically take the average value of control group and disease group and calculate the difference between average value of each groups then divide by sum of standard deviations of each group) and in [50] K-S statistic can be defined by the following equations:

$$X_i = \sqrt{\frac{N-G}{G}}$$

if the gene is a member of the gene set S, and

$$X_i = -\sqrt{\frac{G}{N-G}}$$

if the gene is not member of the gene set S, where G is the number of genes in the S and N is the total number of genes in the dataset. Let gene list $L = \{g_1, ...g_N\}$ and the correlation $r(g_j) = r_j$ with phenotype. The enrichment score ES is calculated as

$$\max_{i \leq j \leq N} \sum_{i=1}^{j} X_i$$

In [51], a weighted K-S is defined by the following equations:

$$P_{\text{hit}}(S,i) = \sum_{g_i \in S, j \leq i} \frac{\mid r_j \mid^p}{N_R}$$

where

$$N_R = \sum_{g_i \in S} \mid r_j \mid^p$$

is the number of genes in S that are present before position i in L.

$P_{\text{miss}}(S,i) = \sum_{g_i \notin S, j \leq i} \frac{1}{N-N_V}$, where $N - N_V$ is the number of genes in N but not in S.

Here, ES is the maximum of $P_{\text{hit}}(S,i) - P_{\text{miss}}(S,i)$. And when weighting factor $p = 0$, it reduces to same K-S statistic in [50]. In [50] a procedure similar to below is followed:

- Normalize the expression data set D.
- Calculate SNR for each of N gene with w samples in D.

- Sort genes by their SNR difference metric (or any other metric) in order to produce L.
- Evaluate the ES for every gene set.
- Permute randomly the class labels of w samples 1000 times to estimate significance
- Re-evaluate ES for gene set S and record them. So that we can get a distribution of ES values.

In [50], the null hypothesis is given as *"no gene set is associated with the class distinction"*. However, in [51] slightly different procedure can be seen as following:

- Create the gene list L by sorting genes using their correlation r_j with phenotype.
- Determine the difference between P_{hit} (the fraction of genes in S that are present before position i) and P_{miss} (the fraction of the rest of the genes which is $N - N_V$ that are present before position i up to position i in the list L).
- a weighting factor p helps for controlling weight of each step
- Permute randomly the class labels of w samples 1000 times to estimate significance
- Re-evaluate ES for gene set S and plot a histogram of these ES values.
- Compute the p-values and estimate FDR for multiple hypothesis testing.

2.3 Methods

We downloaded the clinical and gene expression data sets of 1082 patients with breast invasive carcinoma (BRCA) from the cBioPortal. We combined BRCA TCGA PanCancer Atlas data set and BRCA TCGA Provisional data set. The RNA-Seq data, subtype and radiation therapy information were taken from BRCA TCGA PanCancer Atlas data. Race, tumor status, tumor stages, surgical procedure, histological diagnosis, tumor pathologic, primary site patient, overall survival months and status were taken from provisional data. All data sets were combined based on unique patient identifier (patient id).

We excluded missing values from this study. Due to missing information of tumor subtypes and very small sample size, we did not consider 12 male patients and one case of “American Indian or Alaska Native” in the analysis. Hence, in the present study patient refers to female patients. We also combined tumor stages such as “Stage IA” and “Stage IB” into “Stage IAB”. We ended up having 844 patients to continue further analysis. Number of patients depended on features used in analysis. Small tables were given under graphs to see exact number of patients.

In order to find statistically significant differences between two samples in each box plots, we performed the two-sided Mann-Whitney U-test for testing if values of one sample are less than or greater than values of another sample. The significance levels were shown by stars. For example, “*” shows that the difference is significant with p-value less than 0.05. “**” shows that p-value is less than 0.01 and so on.

To apply appropriate methods for analyzing gene expression data, we first visualized the data and looked at its statistical information. We calculated the average and standard deviation of each gene across patients as well as average and standard deviation of expressions in each primary tumors. In other words, if we denote the expression value of gene j in patient i by g_{ij}, we obtained the average and standard deviation of expression values in each patient i by calculating $\bar{P}_i = \text{avg}(\{g_{ij} : 1 \leq j \leq m\}) = \sum_j g_{ij}/m$ and $\sigma_{P_i} = \text{std}(\{g_{ij} : 1 \leq j \leq m\})$, respectively, where m is the total number of genes. We also obtained the average and standard deviation of expression levels of each gene j across patients by calculating $\bar{G}_j = \text{avg}(\{g_{ij} : 1 \leq i \leq n\}) = \sum_i g_{ij}/n$ and $\sigma_{G_j} = \text{std}(\{g_{ij} : 1 \leq i \leq n\})$, respectively, where n is the total number of patients.

Analogous to a recent paper [54] by Shahriyari, we observed that the distributions of whole genome expressions of patients' primary tumor are very similar to each other, but the distribution of expression levels of each gene across patients are quite different from one gene to another. Figure 2.1 shows that how the average expression values of one gene (e.g. $\bar{G}_j$) is very different from another gene ($\bar{G}_{j'}$), while the average values of expression level of genes in primary tumor of patients ($\bar{P}_i$) are quite similar to each other. Also the level of the changes (standard deviation) of each gene across patients (σ_{G_j}) are more extreme than variance of genes in primary tumors (σ_{P_i}).

For this reason to avoid losing statistical information of the data, we scaled the gene expression levels of primary tumors for each patient separately instead of scaling each gene. In other words, we found maximum expression level of genes for each patient, and we then divided the values of each gene by the maximum value of gene expression of the same patient.

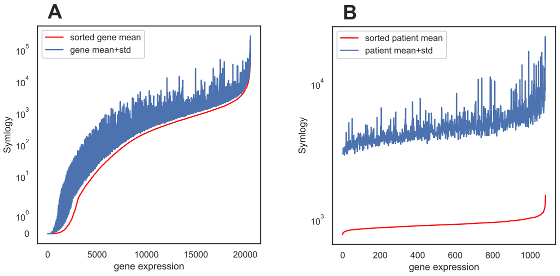


Figure 2.1. (**A**) means and standard deviations of each gene. (**B**) means and standard deviations of gene expression levels of each primary tumor.

Furthermore, after normalizing the data, we utilized dimensionality reduction techniques, because the number of genes (20,531) was much more than the number of patients (1,082). Herein, we implemented a variance threshold method from scikit learn library [55] to get highly variant genes. Firstly, we found that ACTB, ADAM6, COL1A1, COL1A2, COL3A1, CPB1, EEF1A1, FN1, IGFBP5 and MGP are ten most variant genes across patients. Secondly, we set another threshold to get top 200 variant genes.

Another popular dimensionality reduction method is PCA. Main idea is to reduce the number of variables by maintaining most of the information in the data set. So, we performed PCA twice, initially to find first ten principal components, then to find PCs that are responsible for 95% of variability in RNA-Seq. Interestingly, among top 10 variant genes, we found five of them (COL1A2, COL1A1, COL3A1, ACTB, and EEF1A1) highly correlated with PCs. For example, ACTB is positively correlated with PC1 (Figure 2.2A) and COL1A2 is negatively correlated with PC2 (Figure 2.2D).

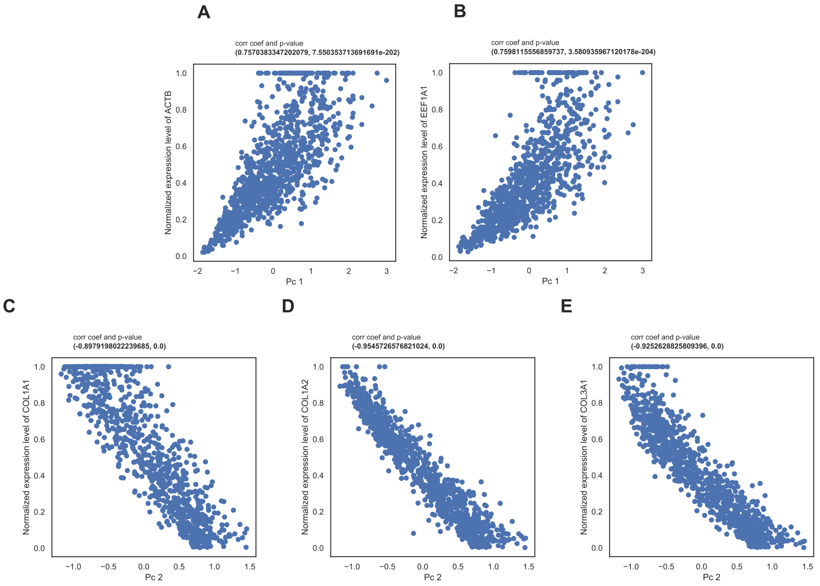


Figure 2.2. Top variant genes that are highly correlated with principal components.

We used K-mean algorithm to cluster patients. The optimal number of clusters were chosen by *elbow* method. We first clustered patients based on “age”, “ajcc_pathologic_tumor_stage”, “race” and ”subtype” from clinical data referred as limited clinical information, then we added “ajcc_tumor_pathologic_pt”, “histological_diagnosis” and “primary_site_patient” where we called it as extra clinical information. We also clustered patients using PCA and high variant genes. All statistical analyses and computations were performed in Python v. 3.7.

We censored patients using clinical features OS_MONTHS and OS_STATUS of patients. After censoring each patient, we created survival curves by KM technique and analyzed differences by log-rank test. All KM related analysis and figures were produced using R v. 3.6.1.

We used the java-GSEA v 4.0.3 desktop application with graphical user interface from the GSEA web site (www.broadinstitute.org/gsea/index.jsp). We prepared our data for GSEA using intersections between K-mean clusters with PCs and K-mean cluster with extra clinical information. We labeled these intersections as SMAS when simple mastectomy is better (i.e. higher survival rate) and as LUMP when Lumpectomy is better to indicate superior treatment. By doing this, we generated two phenotypes.

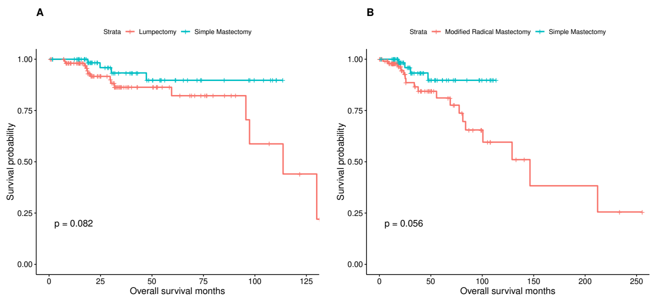


Figure 2.3. Comparison of KM curves for surgical procedures (**A**) lumpectomy versus simple mastectomy, (**B**) lumpectomy versus modified radical mastectomy in *SMAS*.

Then we combined with RNA-Seq data to get the genes. KM curves of phenotypes showed significant differences over modified radical mastectomy but difference was marginal for lumpectomy versus simple mastectomy.

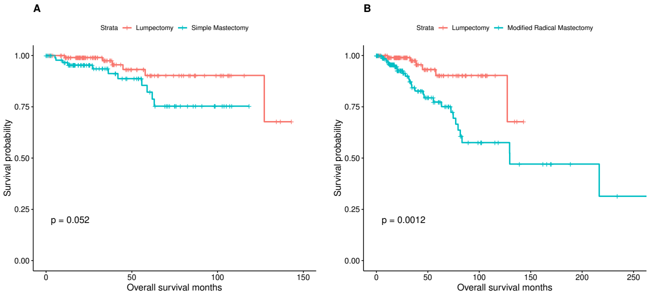


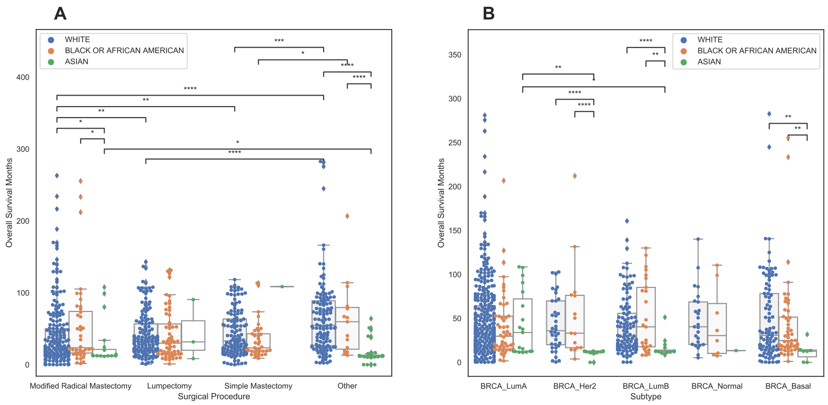
Figure 2.4. Comparison of KM curves for surgical procedures (**A**) lumpectomy versus simple mastectomy, (**B**) lumpectomy versus modified radical mastectomy in *LUMP*.

2.4 Results

Asian patients have the lowest survival months compared to black and white patients for all tumor subtypes except LumA.

The median age at diagnosis of women who had undergone modified radical mastectomy treatment was 58, it was 60 for lumpectomy and 54 for simple mastectomy. Figure 2.5A indicates that black women have higher overall survival months compared to white and Asian women who had undergone modified radical mastectomy. The difference is significant between white and Asian women (P=0.046) as well as black and Asian women (P=0.014). However, black women have the lowest survival months among patients who have had simple mastectomy.

The comparison of survival months for different surgical procedures for all races shows that lumpectomy (P=0.002) and simple mastectomy (P=0.003) are probably better surgical procedures than modified radical mastectomy in terms of overall survival months for white patients (Figure 2.5A). Furthermore, Figure 2.5B shows that for all tumor subtypes, except LumA, Asian patients have the lowest survival month compared to the other races.



Race	Modified Radical Mastectomy	Lumpectomy	Simple Mastectomy	Other
White	226	159	125	125
Black	40	59	37	15
Asian	20	3	1	34

Race	LumA	Her2	LumB	Normal	Basal
White	359	36	117	25	98
Black	52	15	25	8	51
Asian	19	15	16	1	7

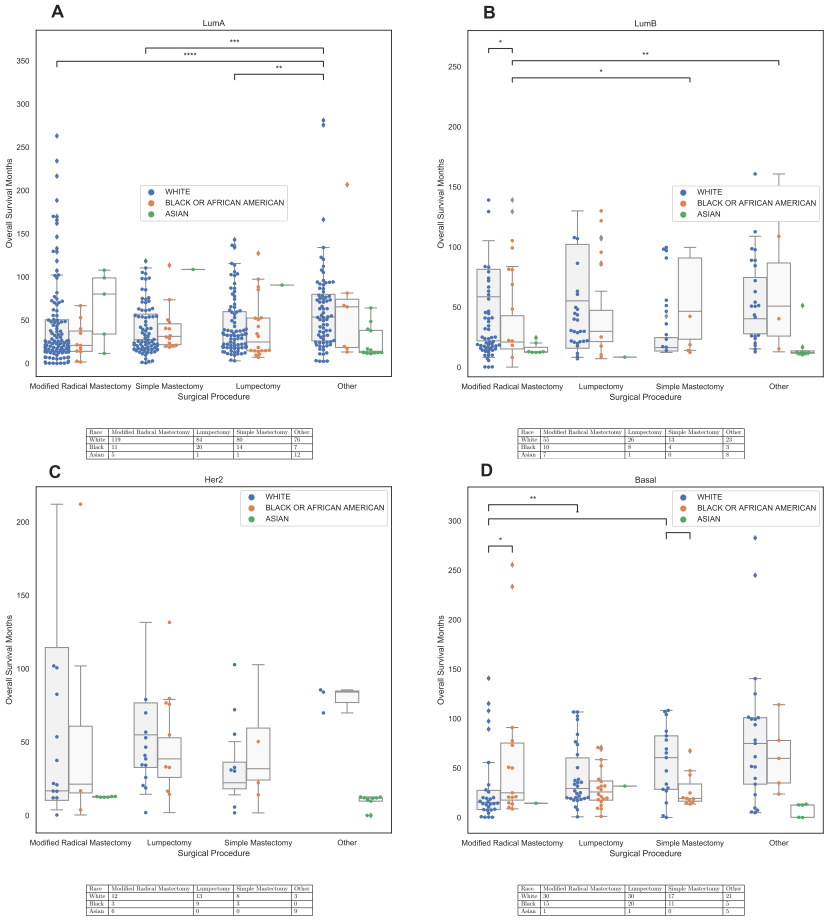
Figure 2.5. Overall survival months as a function of race for (**A**) different surgical procedures and (**B**) for tumor subtypes.

Figure 2.5B shows overall survival months of patients as a function of tumor subtypes and race. Overall survival month has been observed to be substantially different among subtypes for different races. Among white patients, females with normal-like subtype have the highest overall survival duration, but the differences are not significant. Asian women with LumA subtype have considerably higher overall survival months than other subtypes. Figure 2.5B also indicates differences within the same subtype. For patients with Her2, LumB and basal subtypes, important differences are observed between white and Asian women as well as between black and Asian women.

Optimal surgical procedure depends on tumor subtype and race.

Among patients with LumA tumors, those who had modified radical mastectomy have the lowest overall survival months compared to the other surgical procedures for both white and black but not Asian patients (Figure 2.6A). However, among white patients with LumB subtype, females who had modified radical mastectomy have the highest overall survival months compared to the other surgical procedures (Figure 2.6B). The difference between white and black patients within the modified radical mastectomy treatment is significant (P=0.046, number of black women < 10). Although among patients with Her2 tumors, females undergone lumpectomy have higher overall survival compared to the other surgical procedures, the differences are not statistically significant possibly due to the small number of patients (Figure 2.6C).

Finally, among patients with basal subtype, modified radical mastectomy leads to the lowest overall survival months compared to the other surgical procedures for white patients (Figure 2.6D). For black patients with basal tumors, females undergone simple mastectomy have the lowest overall survival months compared to the other surgical procedures. Among patients with basal tumors who had modified radical mastectomy, the overall survival months of white females is significantly lower than black patients (P=0.01) whereas if we look at the simple mastectomy, we see the reverse pattern (P=0.03).



Race	Modified Radical Mastectomy	Lumpectomy	Simple Mastectomy	Other
White	119	84	80	76
Black	11	20	14	7
Asian	5	1	1	12

Race	Modified Radical Mastectomy	Lumpectomy	Simple Mastectomy	Other
White	55	26	13	23
Black	10	8	4	3
Asian	7	1	0	8

Race	Modified Radical Mastectomy	Lumpectomy	Simple Mastectomy	Other
White	12	13	8	3
Black	3	9	3	0
Asian	6	0	0	9

Race	Modified Radical Mastectomy	Lumpectomy	Simple Mastectomy	Other
White	30	30	17	21
Black	15	20	11	5
Asian	1	1	0	5

Figure 2.6. Survival months as a function of surgical procedure for each subtype.

Using the KM method [49] and log-rank test, we examined whether overall survival probabilities differed by each of subtype. We made same comparisons where we saw significant differences in Figure 2.6 excluding “other” surgical procedure. Although patients who underwent modified radical mastectomy had the lowest survival for each subtype, there were no significant survival probability differences between surgical procedures for LumA, LumB and Her2 subtypes. However, we detected significant differences in survivals within basal subtype. White patients had significantly higher survival than black patients who underwent simple mastectomy 2.7A.

Moreover, patients who underwent modified radical mastectomy had significantly lower survival than those who underwent simple mastectomy 2.7B.

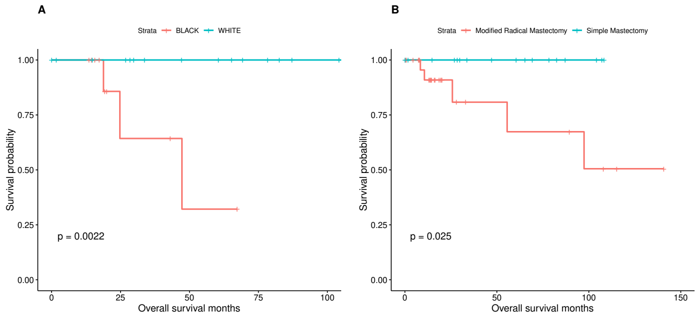
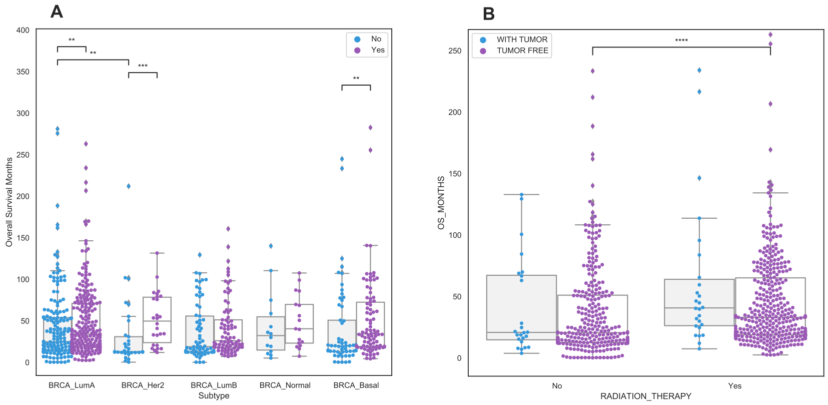


Figure 2.7. KM curves for significant comparisons within Basal subtype. (**A**) White and black patients who underwent simple mastectomy (**B**) Patients who underwent modified radical mastectomy versus simple mastectomy.

Radiation therapy increases survival months for all tumor subtypes.

We evaluated radiation therapy effect on survival months with two different techniques. In the first analysis, we did not censor our data and we assessed group differences by Mann-Whitney test [39]. It is clear that radiation therapy makes a significant difference in overall survival months of breast cancer patients (P<0.001, Figure 2.8B), especially for LumA, Her2 and basal tumor subtypes (P=0.003, P=0.0001, P=0.003 respectively, Figure 2.8A).



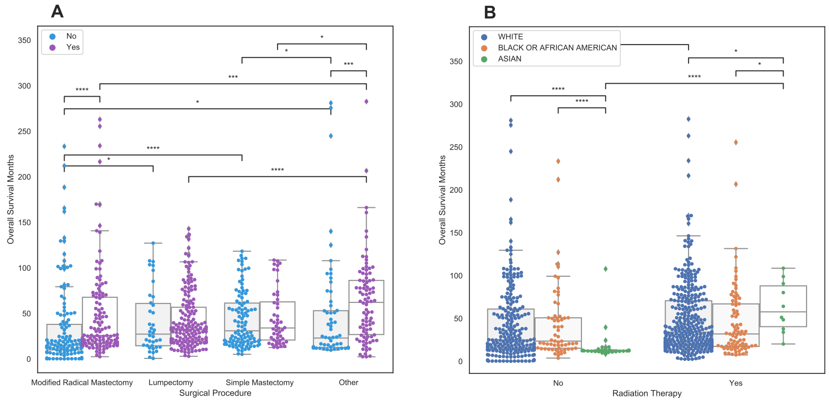
Radiation Therapy	LumA	Her2	LumB	Normal	Basal
Yes	224	26	81	17	86
No	176	33	65	14	57

Radiation Therapy	Tumor Free	With Tumor
Yes	353	26
No	293	24

Figure 2.8. Effect of radiation therapy (**A**) in each subcategory and (**B**) tumor cases.

Radiation therapy significantly increases the survival months of white and black patients.

Radiation therapy significantly increases the survival months of white women with breast cancer (P=0.001). Although the radiation therapy increases the survival month for all surgical procedures, it seems it has no significant effect on the survival months of black patients in this first analysis (Figure 2.9B). However, it has significant effect on patients who underwent modified radical mastectomy (P<0.001). Among patients who did not go through radiation therapy, Asian women's overall survival is significantly less than the survival months of white and black patients (P<0.001, Figure 2.9B).



Radiation Therapy	Modified Radical Mastectomy	Lumpectomy	Simple Mastectomy	Other
Yes	125	171	48	90
No	137	38	107	63

Radiation Therapy	White	Black	Asian
Yes	342	82	10
No	240	62	43

Figure 2.9. Effect of radiation therapy on overall survival months for all (**a**) surgical procedures and (**b**) races.

In the second analysis with censored patients, radiation therapy made a substantial effect on white and black patients' survival. One might conclude that radiation therapy increases overall survival months of most BRCA patients.

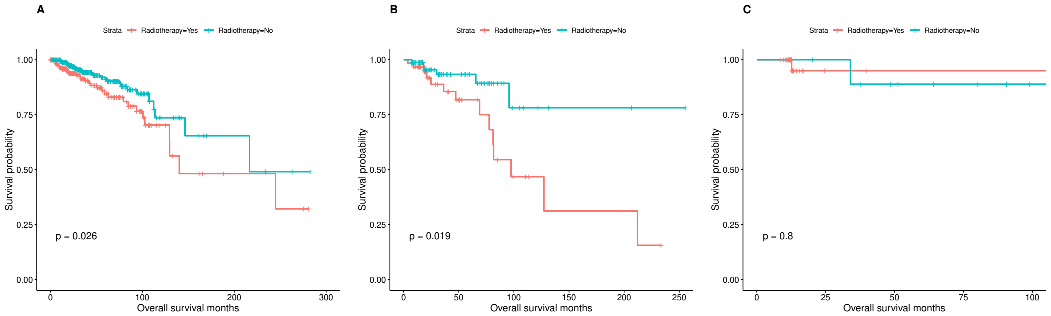


Figure 2.10. KM curves for radiation therapy. (**A**) Comparison of overall survival for presence and absence of radiation for white patients. (**B**) Comparison of overall survival for presence and absence of radiation for black patients. (**C**) Comparison of overall survival for presence and absence of radiation for Asian patients.

Survival Analysis of Breast Cancer patients Using K-mean Clustering and Kaplan-Meier

We applied K-mean clustering algorithm on clinical and RNA-Seq data. In the first analysis, we applied K-mean on limited clinical information data to cluster patients. We chose the optimal number of cluster as k=3 utilizing *elbow* method in an attempt to find groups of cancer patients with distinct survival characteristics. Then using the Kaplan–Meier method [49] and log-rank test, we examined whether each cluster has a significantly different overall survival probability compared to another cluster. There were no significant survival differences between clusters and surgical procedures, however, we observed statistical significant overall survival months between patients undergone modified radical mastectomy and lumpectomy within cluster 0 (P=0.024, Figure 2.11B).

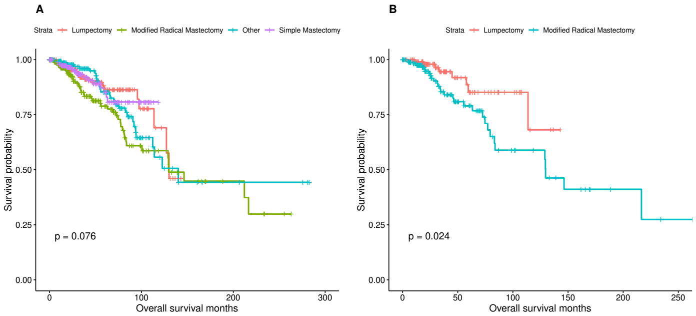


Figure 2.11. Comparison of KM curves in K-means with limited clinical data. (**A**) Comparison of overall survival between surgical procedures. (**B**) Comparison of overall survival between the modified radical mastectomy and lumpectomy treatment within cluster 0.

In the second analysis, we added extra features from clinical data. Here again, the optimal number of cluster was set to k=3 using *elbow* method. Similar to first analysis, there were no significant survival differences between neither clusters nor surgical procedures. However, interestingly, there were significant associations in the survivals of different surgical procedures within cluster 0 and as well as cluster 1. Survival probabilities of patients who underwent simple mastectomy were significantly lower than patients who underwent lumpectomy treatment within cluster 0 (P=0.024, Figure 2.12A). For the same cluster, patients who had modified radical mastectomy surgery had significantly lower survival than patients who had lumpectomy surgery (P=0.018, Figure 2.12B). There was also significant differences in survival of patients undergone simple mastectomy and modified radical mastectomy within cluster 1 (P=0.002, Figure 2.12C). Patients who had simple mastectomy surgery had higher survival probability than patients who undergone modified radical mastectomy surgery.

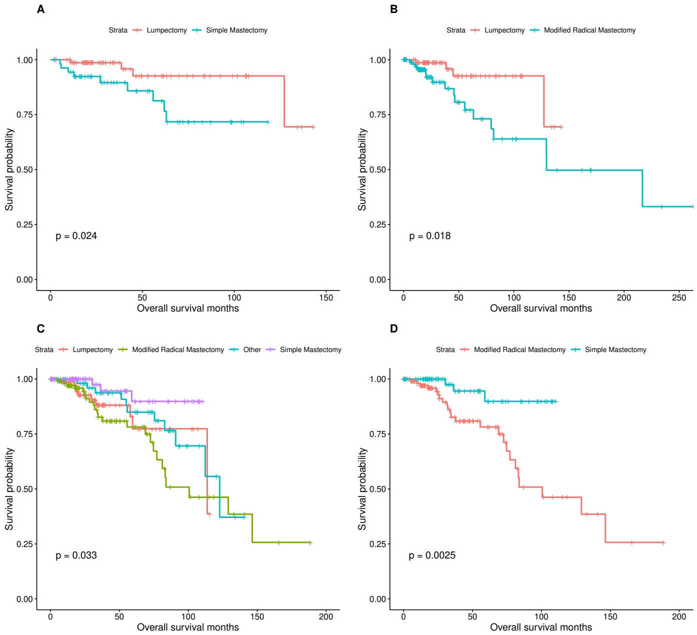


Figure 2.12. Comparison of KM curves in K-means with extra clinical data. (**A**) Comparison of overall survival between the simple mastectomy and lumpectomy treatment within cluster 0. (**B**) Comparison of overall survival between the modified radical mastectomy and lumpectomy treatment within cluster 0. (**C**) Comparison of overall survival in cluster 1. (**D**) Comparison of overall survival between the modified radical mastectomy and simple mastectomy treatment within cluster 1.

In the next analysis, we applied PCA on RNA-Seq data. We selected first 202 principal components that accounted for 95% of variability in the RNA-Seq data. This time optimal number of cluster was 2. We ended up having more patients because we did not include any clinical features in clustering. After adding survival related clinical features for KM curve comparisons, the number of patients was 1023. Contrary to first two analysis, we observed significant survival differences between surgical procedures regardless of clusters (P=0.042). Detailed examination for each cluster showed that significant differences in survival of surgical procedures occurred in cluster 1 (P=0.027). In particular, survival probability of patients who had modified radical mastectomy surgery was considerably lower than patients who had lumpectomy in cluster 1 (P=0.001).

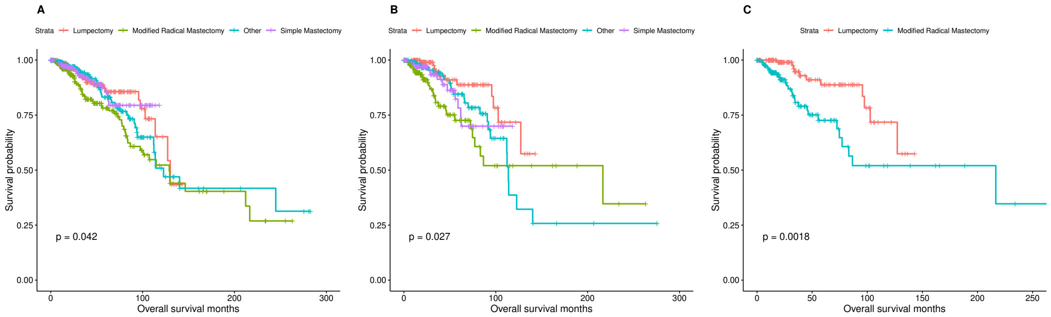


Figure 2.13. Comparison of KM curves in K-means with PCs. (**A**) Comparison of overall survival between surgical procedures. (**B**) Comparison of overall survival between surgical procedures within cluster 1. (**C**) Comparison of overall survival between the modified radical mastectomy and lumpectomy treatment within cluster 1.

In the analysis that follows, we surveyed highly variant genes and we clustered breast cancer patients with respect to the top 200 most variant genes but 8 genes were excluded due to high correlation (more than 90%). There were apparent statistically significant differences in surgical procedures of patients in cluster 1

(P=0.038). Patients who underwent lumpectomy had significantly higher survival probability comparing to patients who underwent modified radical mastectomy. Eventually, we first combined clinical features with highly variant genes, then clustered patients but result did not change.

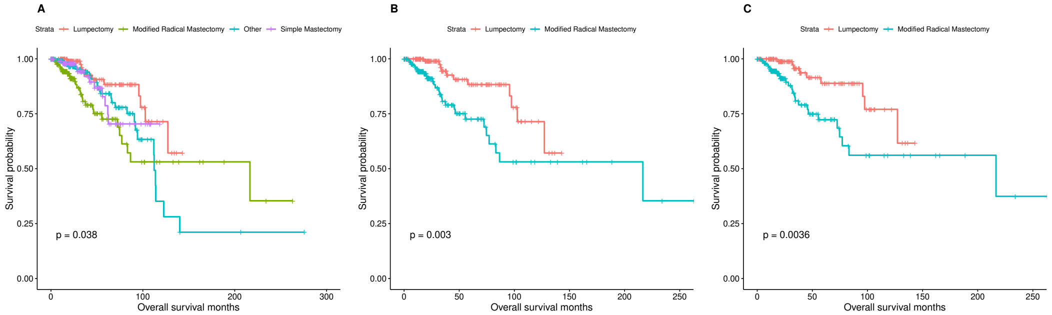


Figure 2.14. Comparison of KM curves in K-means with variant genes. (**A**) Comparison of overall survival between surgical procedures for cluster 1 in clustered data based on top genes. (**B**) Comparison of overall survival between modified radical mastectomy and lumpectomy treatment within cluster 1 in clustered data based on top genes. (**C**) Comparison of overall survival between modified radical mastectomy and lumpectomy treatment within cluster 1 in clustered data based on combined data.

Final clustering analysis was devoted to 3 genes that we observed high correlation with PCs among top 10 variant genes (Figure 2.2). Patients who underwent lumpectomy and simple mastectomy had higher survival probabilities than modified radical mastectomy. However, there was no significant survival difference between lumpectomy and simple mastectomy in any of three clusters.

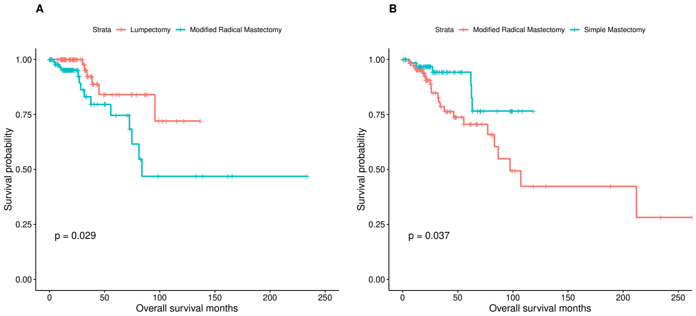


Figure 2.15. of KM curves in K-means with 3 genes. (**A**) Comparison of overall survival between modified radical mastectomy and lumpectomy treatment within cluster 0 in clustered data based on 3 genes. (**B**) Comparison of overall survival between modified radical mastectomy and simple mastectomy treatment within cluster 1 in clustered data based on 3 genes.

n, best surgical procedure, p-value		**K-mean cluster with limited clinical features**			**Best Surgical Procedure**
		C0	C1	C2	
	C0	238, Lumpectomy,p=0.01	5	58, Lumpectomy, p=0.06	301, Lumpectomy, p=0.018
K-mean cluster with extra clinical information	C1	260, S Mastectomy, p=0.003	6	70, Lumpectomy, p=0.67	336, S Mastectomy, p=0.002
	C2	0	146, S Mastectomy, p=0.12	51, Lumpectomy, p=0.16	197, S Mastectomy, p=0.72
K-mean cluster with PCs	C0	196, S Mastectomy, p=0.35	71, S Mastectomy, p=0.4	98, S Mastectomy, p=0.66	443, S Mastectomy, p=0.069
	C1	302, Lumpectomy, p=0.003	86, S Mastectomy, p=0.32	82, Lumpectomy, p=0.11	580, Lumpectomy, p=0.0018
K-mean cluster with top genes	C0	206, S Mastectomy, p=0.38	70, S Mastectomy, p=0.23	99, S Mastectomy, p=0.74	452, S Mastectomy, p=0.1
	C1	292, Lumpectomy, p=0.005	87, M R Mastectomy, p=0.34	81, Lumpectomy, p=0.11	571, Lumpectomy, p=0.003
K-mean cluster withcombination of extra clinical information	C0	190, S Mastectomy, p=0.06	69, S Mactectomy, p=0.23	98, S Mastectomy, p=0.74	357, S Mastectomy, p=0.061
	C1	308, Lumpectomy, p=0.005	88, M R Mastectomy, p=0.36	81, Lumpectomy, p=0.11	477, Lumpectomy, p=0.003
	C0	138, Lumpectomy, p=0.078	54, S Mastectomy, p=0.15	57, S Mastectomy, p=0.35	309, Lumpectomy, p=0.029
K-mean cluster with only 3 genes	C1	199, Lumpectomy, p=0.004	57, S Mastectomy, p=0.19	45, Lumpectomy, p=0.02	378, Lumpectomy, p=0.003
	C2	160, S Mastectomy, p=0.16	46, Lumpectomy, p= 0.005	78, Lumpectomy, p=0.87	335, S Mastectomy, p=0.6
Best Surgical Procedure		498, Lumpectomy, p=0.024	157, S Mastectomy, p=0.2	179, Lumpectomy, p=0.19	

Table 2.1. Table of intersections between clusters

Furthermore, these clusters have many patients in common. We first found intersecting patients in clusters for each data set and we assessed survivals of patients that were in the intersection of clusters. Then using these clusters we generated two phenotypes. From Table 2.2, we combined patients in the intersections of K-mean clusters with extra clinical data and K-mean cluster with PCs where simple mastectomy is indicated, and named as phenotype *SMAS*. Similarly, using same data sets we combined patients where is lumpectomy indicated and named as phenotype *LUMP*.

n, best surgical procedure, p-value		**K-mean cluster with extra clinical information**			**Best Surgical Procedure**
		C0	C1	C2	
	C0	114, Lumpectomy, p=0.44	154, S Mastectomy, p=0.023	96, S Mastectomy, p=0.45	443, S Mastectomy, p=0.069
K-mean cluster with PCs	C1	187,Lumpectomy, p=0.016	182, Lumpectomy, p=0.034	101, S Mastectomy, p=0.28	580, Lumpectomy, p=0.0018
K-mean cluster with top genes	C0	123, Lumpectomy, p=0.28	156, S Mastectomy, p=0.02	95, S Mastectomy, p=0.45	452, S Mastectomy, p=0.1
	C1	178, Lumpectomy, p=0.02	180, S Mastectomy, p=0.05	102, S Mastectomy, p=0.3	571, Lumpectomy, p=0.003
K-mean cluster with combination of extra clinical information	C0	115, Lumpectomy, p=0.44	148, S Mastectomy, p=0.019	94, S Mastectomy, p=0.45	357, S Mastectomy, p=0.061
	C1	186, Lumpectomy, p=0.005	188, S Mastectomy, p=0.06	103, S Mastectomy, p=0.31	477, Lumpectomy, p=0.003
	C0	88, S Mastectomy, p=0.86	96, Lumpectomy, p=0.01	65, S Mastectomy, p=0.61	309, Lumpectomy, p=0.029
K-mean cluster with only 3 genes	C1	106, Lumpectomy, p=0.009	126, S Mastectomy, p=0.024	68, S Mastectomy, p=0.13	378, Lumpectomy, p=0.003
	C2	107, Lumpectomy, p=0.13	113, S Mastectomy, p=0.003	64, M R Mastectomy, p=0.051	335, S Mastectomy, p=0.6
Best Surgical Procedure		301, Lumpectomy, p=0.018	336, S Mastectomy, p=0.002	197, S Mastectomy, p=0.72	

Table 2.2. Table of intersections between clusters

Gene Set Enrichment Analysis (GSEA)

GSEA detected that 4,665 genes were up-regulated in phenotype *SMAS* and 265 gene sets were significant at FDR $<$ 25%. Three gene sets with the smallest FDR were “NADH Dehydrogenase Activity”, “Positive Regulation of Cellular Amide Metabolic Process” and “Rna Phosphodiester Bond Hydrolysis”. While the number of up-regulated genes were 470 in phenotype *LUMP* and only 3 gene sets were significant at FDR $<$ 25% namely “Odorant Binding”, “Ligand Gated Anion Channel Activity” and “Inhibitory Extracellular Ligand Gated Ion Channel Activity”.

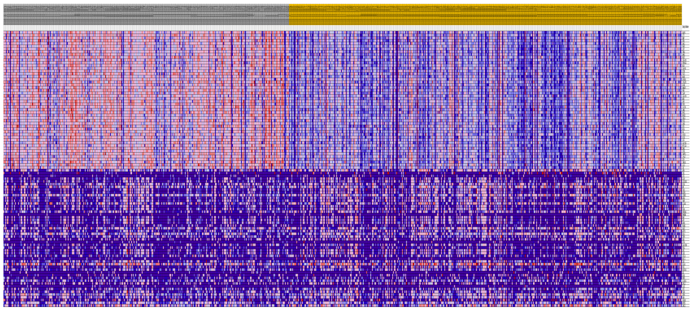


Figure 2.16. Heat map of the top 50 genes for each phenotype (i.e. *SMAS* vs *LUMP*).

Heat map indicates 50 genes were over-expressed in *SMAS*. All these 50 genes have differentially expressed between two phenotypes (all p-values were less than 0.001). We plotted pairwise density plots. Among these 50 genes, we took a closer look at 6 genes with race and subtype information. CLIC1 (chloride intracellular channel 1) is in a set of proteins that responsible for main cellular processes [57]. STK25 (Serine/Threonine Kinase 25) encodes enzyme that works in pathway of serine-threonine liver kinase B1 (LKB1) [58]. PREB (Prolactin regulatory element-binding) is plays a role in gene expression of prolactin [59]. PCGF1 (Polycomb Group Ring Finger 1) is a protein coding gene and related to nervous system [60]. METTL11A (Methyltransferase-Like Protein 11A) is a protein coding gene that catalyze chemical reactions by methylation in cell [61]. IDH3G (Isocitrate Dehydrogenase 3 (NAD(+)) Gamma) is one of the candidate gene for periventricular heterotopia [62]. These 6 genes showed high distribution differences between phenotypes. Black patients with basal subtype have high expression level of these genes. Additionally, normalized expression level of 6 genes were higher in *SMAS* for all subtypes.

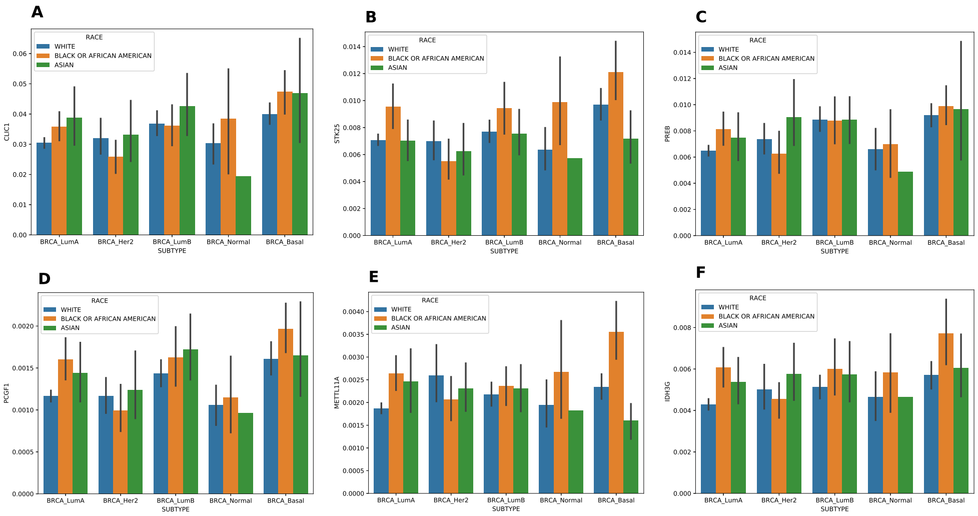


Figure 2.17. Comparison of six highly differentially expressed genes in *SMAS* (**A,B,C**) versus *LUMP* (**D,E,F**) phenotypes.

The number of patients in *LUMP* group was 483, while it was 351 in *SMAS* group. Majority of patients with basal and Her2 subtypes were in *SMAS* that indicates for patients with more agressive tumor subtypes, simple mastectomy could give better results.

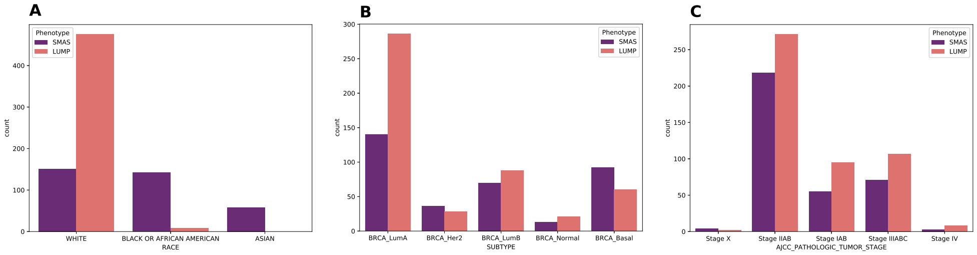


Figure 2.18. Number of patients for some clinical features. (**A**) Comparison of number of patients between race by phenotype. (**B**) Comparison of number of patients between subtypes by phenotype. (**C**) Comparison of number of patients between tumor stages by phenotype.

COL1A2 is a predictor of survival months of black patients undergoing simple mastectomy, and ACTB is a predictor of survival months of white patients undergoing modified radical mastectomy.

We analyzed the gene expression data of primary tumors to find possible signatures of treatments' outcomes in the expression levels of genes. After normalizing gene expression data of primary tumors, we found the most variant genes (see the Methods section for more details).

Among ten most variant genes across breast cancer patients, ACTB, EEF1A1, COL1A1, COL1A2 and COL3A1 are highly correlated with the principal components (Figure 2.2). Importantly, there is no correlation between any of these genes and age at diagnosis, overall survival months of patients (Figure 2.19C).

We further investigated if any of these genes would predict the outcome of surgical procedures for any of races. We found that COL1A2, which is highly correlated with COL1A1 and COL3A1, might be a predictor of survival months of black patients who had undergone simple mastectomy (Figure 2.19A). Additionally, survival months of white patients who had undergone modified radical mastectomy is positively correlated with the normalized expression level of ACTB (Figure 2.19B).

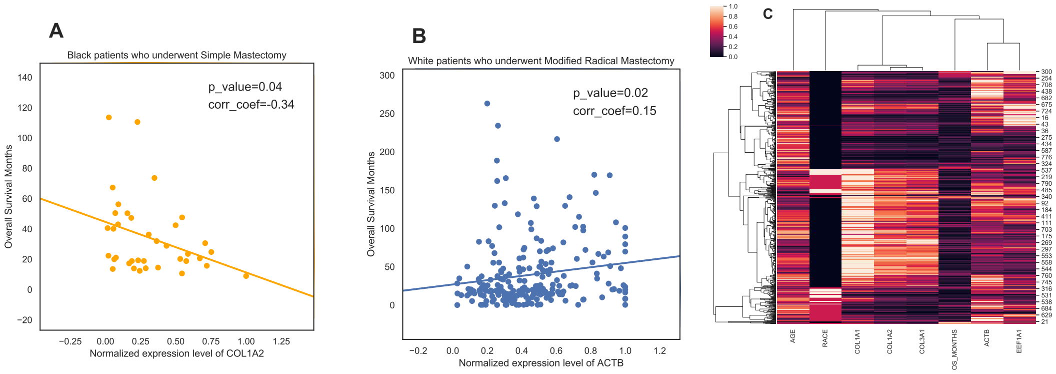
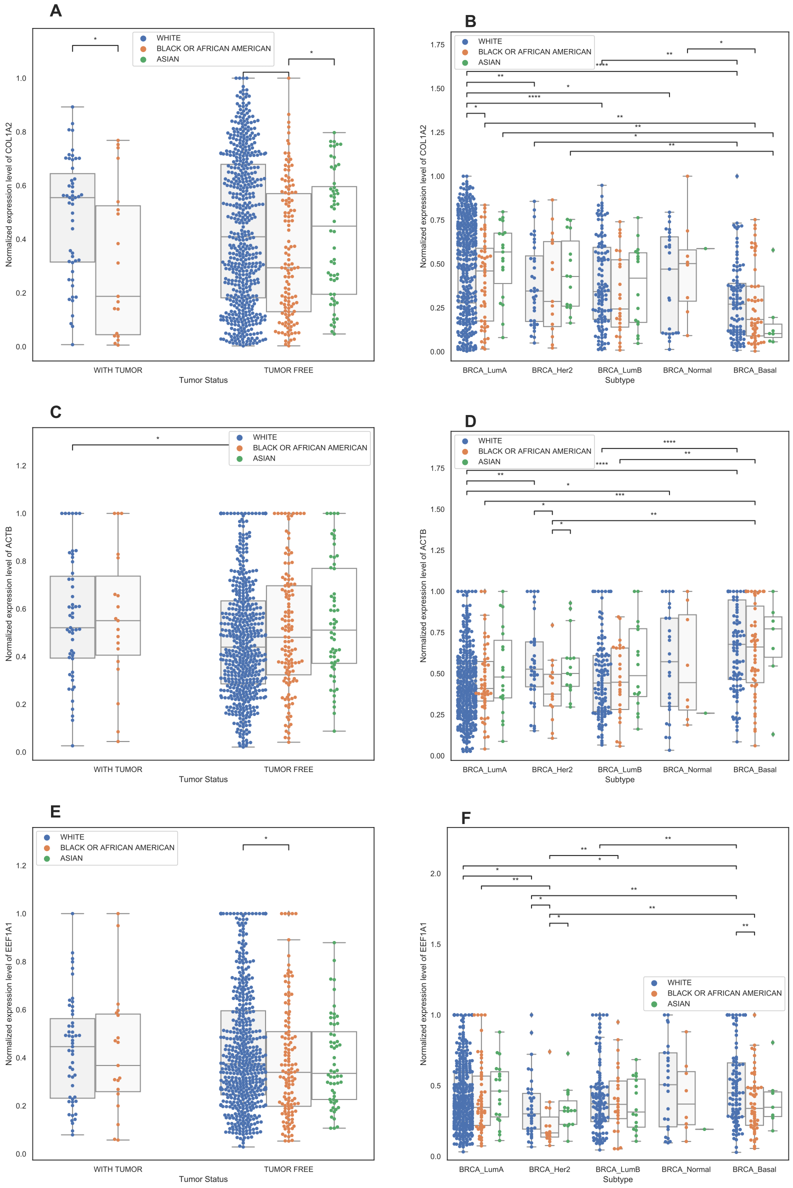


Figure 2.19. Relationship between survival months of two genes by race and correlated based heat-map. (**A**) Survival months of black women who had undergone simple mastectomy as a function of normalized level of COL1A2. (**B**) Survival months of white women who had undergone modified radical mastectomy as a function of normalized ACTB expression. (**C**) Hierarchically-clustered heat-map of top variant genes with some demographic and clinical attributes.

Among patients with LumA, black females have significantly lower COL1A2 expression than white women.

Black females in both with tumor and tumor free groups have a lower level of COL1A2 compared to the other races. Additionally, among tumor free patients, white females have a significantly higher COL1A2 expression than black women (P=0.001, Figure 2.20A). After further investigation, we observed that black patients have lower expression level of COL1A2 than white patients for all subtypes except normal-like.

This difference in COL1A2 levels between white and black patients is only significant within LumA subtype (P=0.03, Figure 2.20B). Furthermore, among black patients, females with LumA tumors have a significantly higher expression level of COL1A2 than women with basal tumors (P=0.004). White patients with LumA breast cancer subtype have significantly higher expression level of COL1A2 than Her2 (P=0.005), LumB (P<0.001), normal-like (P=0.05), and basal subtypes (P<0.001). For Asian patients with LumA breast cancer subtype have a significantly higher expression level of COL1A2 than basal subtype (P=0.003).



Race	Tumor Free	With Tumor
White	534	51
Black	129	19
Asian	56	0

Race	LumA	Her2	LumB	Normal	Basal
White	384	37	126	25	106
Black	56	15	26	9	53
Asian	20	15	16	1	7

Figure 2.20. (**A**,**B**): normalized expression level of COL1A2, (**C**,**D**): ACTB with tumor status and subtypes, and (**E**,**F**): EEF1A1 with tumor status and subtypes.

Regardless of race, basal tumors have a higher level of ACTB than other tumor types

Figure 2.20C shows a slight difference in the normalized level of ACTB between tumor free and with-tumor groups. However, women with basal tumors have a higher normalized level of ACTB than other tumor types regardless of race. The differences between LumA, Her2, LumB and basal subtypes are substantial for black women. Although Figure 2.20D shows similar differences for Asian women, however, number of Asian patients with normal-like and basal subtypes are small. In particular, within Her2 subtype, black women have a significantly lower level of ACTB compared to white (P=0.02) and Asian (P=0.04) women.

White patients have significantly higher level of EEF1A1 than black patients for Her2 and basal subtypes.

For normalized expression level of EEF1A1, there is a considerable difference between white and black women among tumor free patients. White patients have higher level of EEF1A1 than black patients for all subtypes. In particular these differences are significant for basal (P=0.008) and Her2 (P=0.022) subtypes. Moreover, we observe significant difference between black and Asian patients (P=0.018) within Her2 subtype.

2.5 Discussion

Clustering algorithms have been implemented on various cancer data sets including breast cancer [36, 37, 38, 63, 64, 65, 66] to classify patients with a focus on finding distinct survival characteristics and/or high classification accuracy. Yu *et al.* [64] showed that K-mean algorithm can be an efficient approach to cluster breast cancer patients to evaluate survival rates. In a study which set out to determine prognostic factors for breast cancer, Jones *et al.* [67] showed that the use of clustering on luminal and myoepithelial (basal-like) genes is helpful to find independent prognostic information in breast cancer. In present study, one of our primary goal was to explore different partitions of both clinical and RNA-Seq data to find possible underlying structures in data sets that may help to find survival characteristics of breast cancer patients for optimal surgical treatment.

We clustered patients based on clinical features, PCs and highly variant genes. Particularly, clustering clinical features were based on limited clinical information and extra clinical information. We found that the best surgical procedure depends on cluster where different prognostic factors were used. Lumpectomy treatment was better surgical procedure for cluster 0 (n=498) in terms of survival probabilities from KM analysis in clustering with limited clinical information. For the other two clusters we did not observe significant survival differences in surgical treatments. By adding extra clinical information, lumpectomy was the best surgical options for patients in cluster 0 (n=301) and simple mastectomy was better surgical option for cluster 1 (n=336) in terms of survival probabilities from KM analysis, respectively. Cluster analysis using only PCs and clustering with top 200 variant genes both suggested Lumpectomy as better surgical procedure for cluster 1 (n=571 for clustering with PCs and n=477 for clustering with top variant genes). When we cluster with only 3 genes which are ACTB, COL1A2 and EEF1A1, we observed that Lumpectomy was the best surgical procedure in terms of survival probabilities not only for cluster 1 (n=309) but also cluster 0 (n=378). Finally, lumpectomy was still better surgical treatment for cluster 1 (n=477) in the data where we combined top variant genes with extra clinical information. These well separated survival curves can be utilized as a prognostic tool [38]. Our findings are similar to those of Agarwal *et al.* [68], who reported that breast cancer patients who have undergone lumpectomy had higher survival rate than those treated with mastectomy.

Several studies that have examined the impact of different surgical procedures [69, 70] and subtype [71, 72, 70] on mortality but the effect of race with combination of these factors have not been investigated to the best of our knowledge. In this study, we investigated the outcome of various surgical procedures for patients with breast cancer. We evaluated the effect of surgical procedures for four subtypes (LumA, LumB, Her2, and basal). We did not consider normal-like tumor subtype because of limited data; very small number of patients had normal-like tumors.

We found that the outcome of surgical procedures is a function of race, subtype of the tumor, and possibly gene expression data of primary tumors. We observed that simple mastectomy is the poorest surgical choice in terms of overall survival months for black women with basal tumors. However, for white patients with basal tumors, modified radical mastectomy leads to the lowest survival months compared to the other surgical procedures. Modified radical mastectomy is also the poorest surgical procedures in terms of survival months for white women with LumA tumors.

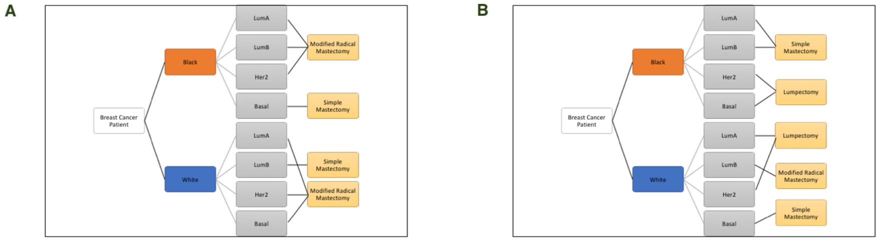


Figure 2.21. (**A**) Poorer surgical treatment choices by subtype and race. (**B**) Better surgical treatment choices by subtype and race.

White women with LumA tumors have a better survival with lumpectomy surgery than simple or modified radical mastectomy. On the other hand, modified radical mastectomy is the best surgical procedure for white patients with LumB tumors, while it is the worst surgical procedure for black patients with LumB and Her2 tumors. There was a substantial racial disparity among breast cancer patients with basal subtype from KM analysis. White patients had higher survival than black patients. Additionally, patients are more likely to survive under simple mastectomy treatment than under modified radical mastectomy treatment.

We also found that radiation therapy has a positive effect on survival months regardless of surgical procedures. Our study suggests that women with LumA, Her2, normal-like and basal subtypes are more likely to have higher overall survival months with the radiation therapy, as compared with no radiation therapy. Moreover, survival curves of white and black patients who treated with radiation therapy showed that radiation therapy considerably increased survival of these patients.

Previous studies have indicated the potential relationship between breast cancer and COL1A2 [73, 74]. COL1A2 has a key role in collagen production and development of tumor in breast cancer [75]. A prior study by Lin *et al* [74] that have noted the importance of type I collagen genes such as COL1A1 and COL1A2, were shown up-regulated in a comparison of invasive breast cancer and normal breast cancer. However, they did not specify race of patients in this comparison. We observed that black patients have a lower expression level of COL1A2 compared to the white patients regardless of tumor status and Asian patients in tumor free cases. Additionally, the expression level of COL1A2 varies by subtypes; basal tumors express the lowest level of COL1A2 compared to the other subtypes (Figure 2.20B). We also found that normalized expression level of COL1A2 could be used to predict the outcome of simple mastectomy surgical procedure for black patients; for these patients the survival months is a decreasing function of the expression level of COL1A2, which is highly correlated with COL1A1 and COL3A1. Hence, higher amount of COL1A2 corresponds to lower survival (Figure 2.19).

Although Beta-actin gene ACTB has been explored in some cancer studies, its key role in breast cancer has remained unclear [76, 77]. Our analysis shows that ACTB level is a good predictor of the overall survival months of white women who had undergone modified radical mastectomy; higher level of ACTB corresponds to higher survival months. For all races, ACTB level is higher in basal subtype compared to the other breast cancer subtypes (Figure 2.20C). Note that, among patients with basal tumors who had undergone modified radical mastectomy, black patients have a significantly higher survival months than white patients. In fact, white patients with basal tumors have a significantly better survival with simple mastectomy than modified radical mastectomy (Figure 2.6D).

Contrary to our findings, a recent study reported that the difference is not significant between the expression level of EEF1A1 and breast cancer subtypes [78]. However, we observed differences in subtypes by race. For black and white women, significant differences were found between subtypes except for normal-like subtype. On the other hand, there were differences between races in Her2 and basal subtypes. Furthermore, black women have lower expression of EEF1A1 than white women among tumor free cases.

GSEA detects potential biological characteristics of two sub-groups of patients. We also observed differences especially for black patients with basal subtype. Genes CLIC1, STK25, PREB, PCGF1, METTL11A and IDH3G showed high differences in distributions between phenotypes. Expression level of these 6 genes were higher in *SMAS* regardless of tumor subtype. In phenotype *SMAS*, NADH dehydrogenase activity gene set was up-regulated. NADH dehydrogenase is an enzyme that responsible for catalysis of the nicotinamide adenine dinucleotide (NAD) to NADH. It is an important component of mitochondrial electron transport complex I [79]. The relationship between NADH dehydrogenase activity and breast cancer cell proliferation has been investigated in the work of Li *et al.* [80]. They showed that the NDUFB9 gene from the gene set was down-regulated in highly metastatic breast cancer cells. Other gene set, positive regulation of cellular amide metabolic process, includes many genes related to pathways and chemical reactions of amides. Third gene set is related to phosphodiester bond in rna metabolic process.

In phenotype *LUMP*, two enriched gene sets are related to anion-ion channel activity and one gene set is related to odorant binding genes which forms the basis of the sense of smell. A study evaluates ion channel genes as prognostic factor in breast cancer [81]. Olfactory receptor expression profiles have been shown to function as biomarkers in many carcinoma tissues including breast cancer [82].

We would like to emphasize that this study should be interpreted in the light of several limitations. First, intrinsic subtype differentiations were constructed at molecular level analysis but due to complicated tumor structure of breast cancer, categorizations might not be perfectly correct. Second, since some clinical attributes are not available in the PanCancer Atlas data set, we combined this data set with provisional data using patient's identifier. Third, the number of black and Asian patients were relatively smaller than white patients in the data set especially for Her2 and normal-like subtypes, which limits the statistical significance of results related to these subtypes. We did not include Cox regression model result due to violation of model assumptions [83].

In conclusion, the choice of surgical procedure can have a significant effect on overall survival months of patients, and these effects vary by race and tumor subtypes. Moreover, gene expression data might be a good resource to predict the outcome of some of surgical procedures. However, further studies are needed to have a better understanding of the role and significance of clinical and demographic attributes in predicting the outcome of surgical procedures for breast cancer.

CHAPTER 3

Quality of Life of Cancer Patients

3.1 Background

Transportation is essential and important problem for patients [84, 85, 86, 87]. According to a comprehensive review study by Syed *et al.* [88], transportation is an issue for healthcare access for at least half of patients. This problem may reduce the effect of treatments for patient with chronic diseases and also it may trigger an increase of visiting emergency rooms as found in the work of Heckman *et al.* [89]. Saliently, 52% of cancer patients worry about time away from their work and 22% of patients worry about transportation; these percentages may vary depending on different racial or ethnic groups [90]. For example, Asian and Hispanic cancer patients worry more about transportation than other groups. For low-income cancer patients in Harris County, Houston the relationship between transportation issue and non-compliance with treatment has been found to be significant [91]. In a recent study [92], it has been observed that stresses related to the transportation are the only non-medical factors that increase the overall stress of head and neck cancer patients.

In the most literature, majority of papers investigate the role of transportation in accessing to the healthcare providers. However, a much more important issue is the ability of patients to continue to work that would ultimately affects their mental health and their ability to pay for their future treatments. Note, some patients might not be able to drive while following a treatment, therefore it might be difficult for them to get to their office and continue their work. We created a survey for cancer patients to ask if transportation and location of healthcare providers had any effect on their decisions regarding their cancer treatments and the stage of their disease at the time of diagnosis. The results will assist leaders and scientists, including healthcare professionals to hear patients' opinion about the role of transportation in the quality of their life during and after treatments and to find an optimal solution. It will also assist patients and caregivers to knowledgeably decide about treatment strategies.

In the last few decades, literature has grown up around the theme of assessment of cancer treatment outcome not only using survival and tumor response but also patient's well-being. Health-related Quality of life (QoL) in cancer patients has become increasingly significant part of measuring patient's well-being and treatment effectiveness. Although cancer treatments including surgery, chemotherapy and radiotherapy can be effective, they are burdened by many side effects. These side effects with severity of symptoms considerably reduce the QoL of patients. Severity of symptoms can vary by different types and stages of cancer but mostly advanced cancer stages produce more symptoms. However, physician might not recognize all these symptoms, in that case patient QoL surveys can be a good indicator of patient's needs and these surveys can be a useful source of information to show treatment effectiveness. In a study performed by Nayak *et al.* [93], 82.3% of the study population in a single state had low QoL scores because of the treatment symptoms. Heydarnejad [94] found that pain intensity significantly decline QoL of cancer patients undergoing chemotherapy. In another study, reduction in QoL in breast cancer patients and chemotherapy side effects found to be correlated with early treatment discontinuation [95]. A study conducted by Detmar *et al.* [96] by using step-wise linear regression to analyze the group differences in patients' self-reported QoL assessment, patients in the intervention group significantly improved in mental health and role functioning over time (P=0.04, P=0.05 respectively) as compared to control group. In the same study, 79% of patients expressed that with QoL summary profile their physicians became more aware of their health problems. 87% of patients and all physicians considered that QoL can be beneficial for a routine part of clinical experience of outpatients. Although there is no substantial effect on expected survival, 68% of patients preferred a treatment that potentially improve their QoL as reported by Silvestri *et al.* [97].

In this project, we propose an innovative approach to investigate the role of transportation in cancer patients' decision making and find ways to improve their quality of life. We created a survey to explore the role of transportation in making decisions, keeping or changing their job, or moving to another location. The results will help us to identify the ways that we could improve the cancer patients' quality of life such as providing a share ride app or any other means. We performed the project as described in the following steps. First step was to create the survey. We provided several closed questions in the first part of survey. In closed questions, respondents have a fixed number of possible responses to choose from, such as yes/no, multiple choice, or check boxes. A closed question can be answered in a short or single-word answer. Closed questions were used to obtain facts and specific pieces of information, for example questions about gender, race, age at diagnosis, initial treatment, the time of starting and ending initial treatment, the stage of tumor at the time of diagnosis, and the stage of tumor before starting this treatment. Importantly, some of the closed questions will be about patients' quality of life such as"the quality of your life during the treatment" from *strongly disagree* to *strongly agree.* Furthermore, a subset of questions, would be related to the transportation, including the location of their house, work, and health care providers and the role of transportation in their quality of life and making decisions regarding following their treatments. In the final part of survey, we asked participant about usefulness of free/discounted rides between home and work as well as home and health care provider.

We first investigate the role of each factor such as existence of free/discounted ride in patients' quality of life and their decisions about continuing treatments. To find the most important features in predicting an outcome such as the quality of life, we performed a classification analysis using tree-based model. Decision trees, which are non-parametric supervised learning algorithms, are mostly used for classification purposes. A decision tree consists of nodes and branches. Decision tree paths start with the top-most node which is called root to terminal nodes or leafs. The leafs demonstrate a decision or classification. Several decision tree algorithms have been proposed to handle categorical and numerical target variables [98, 99] Decision trees can handle both numerical and categorical variables. However single decision tree models usually suffer from high variance in prediction. Therefore we implemented extremely randomized tree model which was shown to address this problem. Importantly, we assessed importance of features in the classification.

3.2 Methods

We first re-coded categorical/nominal features in the survey. Re-coding variables is done using by IBM SPSS Statistics for Windows, version 23 (IBM Corp., Armonk, N.Y., USA). In order to understand data better, we visualized data using pyplot python library [100]. We used spearman correlation test to evaluate high correlation between binary and ordinal variables [101]. We did not include variables in the model if the correlation coefficient is higher than 0.75. When modeling for *usefulness(1)*, we filter data based on their answer to employment status since question was about travel from home to work. We took only working or temporarily laid off from a job. We used MinMax scaler [55] before applying machine learning algorithm.

Extra Tree Classifier

This algorithm builds multiple decision trees by using ensemble method [102]. Algorithm randomly chooses cut points and grows the trees from whole learning sample. The main steps are as follows [102]:

Generating a tree (D)

Let D be input data. Algorithm returns a leaf labeled by class frequencies in D:

- If $|D| < n_{min}$ or all variables are constant in D or if the output is constant in D.

Otherwise :

- Randomly choose T variables $v_1, \cdots, v_T$ from non constant candidate variables in D;
- Draw T splits $d_1, \cdots, d_T$ where d_i = random_split (D, v_i), ∀i=1, $\cdots, T$;
- Select a split d_* such that Score$(d_*, D) = \max_i(D_i, D)$;
- Based on d_*, generate two subsets D_{left} and D_{right} and do the same steps above for these two subsets;
- Tie D_{left} as left subtree and D_{right} right subtree to the generated node with split d_*.

Make a random_split(D,v)

When variable v is numerical:

1. Let v_{min} and v_{max} denote minimal and maximal value of v in D.
2. Draw a random cut point v_c uniformly in $[v_{min}, v_{max}]$;
3. Return the split $[v < v_c]$.

When variable v is categorical, let P be set of possible values:

1. Calculate P_D the subset of P of values of a that appear in D;
2. Randomly draw a proper non empty subset P_1 of P_D and a subset P_2 of P \P$_D$;
3. Return the split $[v \in P_1 \cup P_2]$.

where T is the number of variables randomly selected at each node, the default value of $T = \sqrt{n}$ for classification problem and n_{min} is the minimum sample size for splitting a node. Algorithm follows these steps every for $i = 1$ to M, where M is the number of tree.

The score measure used in this algorithm is a modified(normalized) information gain. For sample D and split d, measure is calculated by:

$$\text{Score}_C(d, D) = \frac{2I_c^d(D)}{H_d(D) + H_c(D)}$$

(3.1)

where $I_c^d(D)$ is the mutual information, $H_d(D)$ is the split entropy and $H_c(D)$ is the log entropy [102]. All statistical analyses and computations were performed in Python v. 3.7.

3.3 Results

Participant were adult female (51.2%) and male (48.8%) cancer patients and predominantly white (85%), black (6.2%), Hispanic or Latino (5.2%) and others (3.6%). Participant demographic characteristics included age, gender, race, education, marital status, income, employment status, health coverage, number of cars for households, driving licence and home residency. Clinical information were based on self-report including type of cancer, age at diagnosis, year of diagnosis, type of treatment, tumor grade and tumor stage. Then based on treatment choice, participant were asked treatment related questions such as year of treatment, duration of treatment, frequency and side effects as well as travel behaviour related questions such as travel mode to health care and work during treatment, frequency of travel to health care, average cost of treatment and insurance coverage of treatment. Details of data can be found in *Codebook, Appendix A*.

Figure 3.1 shows cancer types of the respondents. We provided a list of 16 types of cancer and a comment box if respondents' cancer type is not in the list. Breast cancer was the most frequent cancer type among survey respondents. Majority of breast cancer patient had radiotherapy and chemotherapy. Additionally, according to the responses, a high percentage of breast cancer patients would like to have free/discounted ride from home to healthcare provider as well as from home to work during their treatments. Interestingly, if breast cancer respondents had higher tumor grade at diagnosis, they were more willing to have free/discounted ride. Their response to overall quality of life questions showed that they were likely to be happy with their overall physical conditions after treatments.

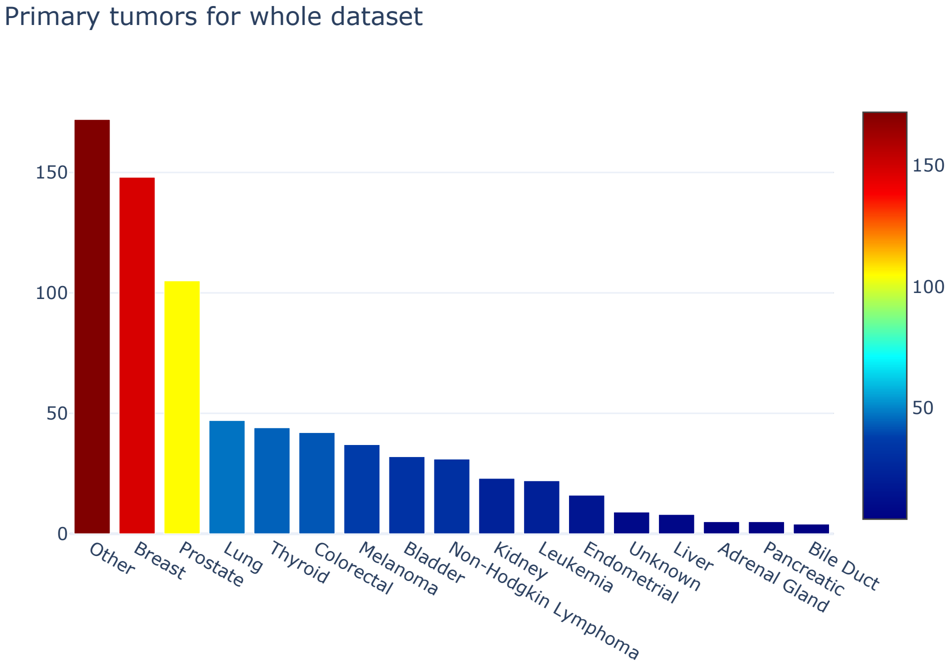


Figure 3.1. Cancer type of participants.

We implemented extremely randomized trees algorithm for radiotherapy and chemotherapy treatments for target variables *usefulness(1)* and *usefulness(2)*, which are respectively the responses to the following questions: “Usefulness of free/discounted rides to cancer patients - I could have a better life if there was free or discounted rides during my treatments between my house and work?”, “I could have a better life if there was free or discounted rides during my treatments between my house and health care providers?” and options were re-coded as Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5. We treated this problem as a multi-class classification problem. We found optimal parameters by hyper-parameter tuning with a grid search. For *usefulness(1)* most important predictors are in Figure 3.2. This figure shows the importance of each features that calculated by gini measure.

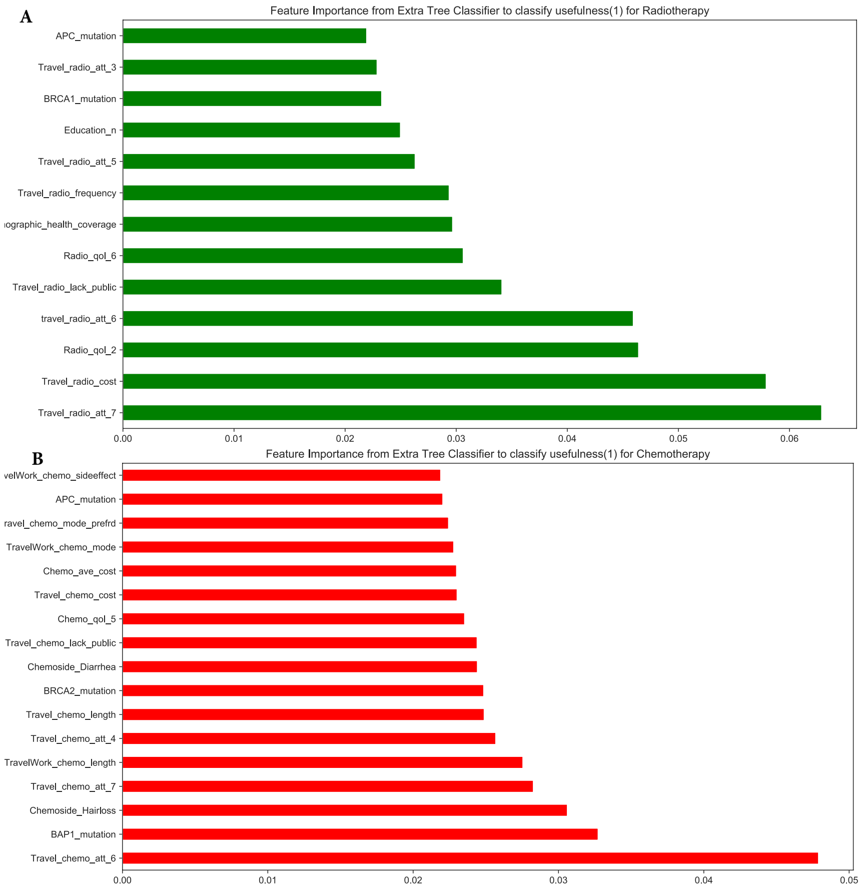


Figure 3.2. *usefulness(1)* (**A**) Important features of the model by using data of respondent who had radiotherapy. (**B**) Important features of the model by using data of respondent who had chemotherapy.

For *usefulness(2)*, most important predictors that corresponds to each treatment are in Figure 3.3.

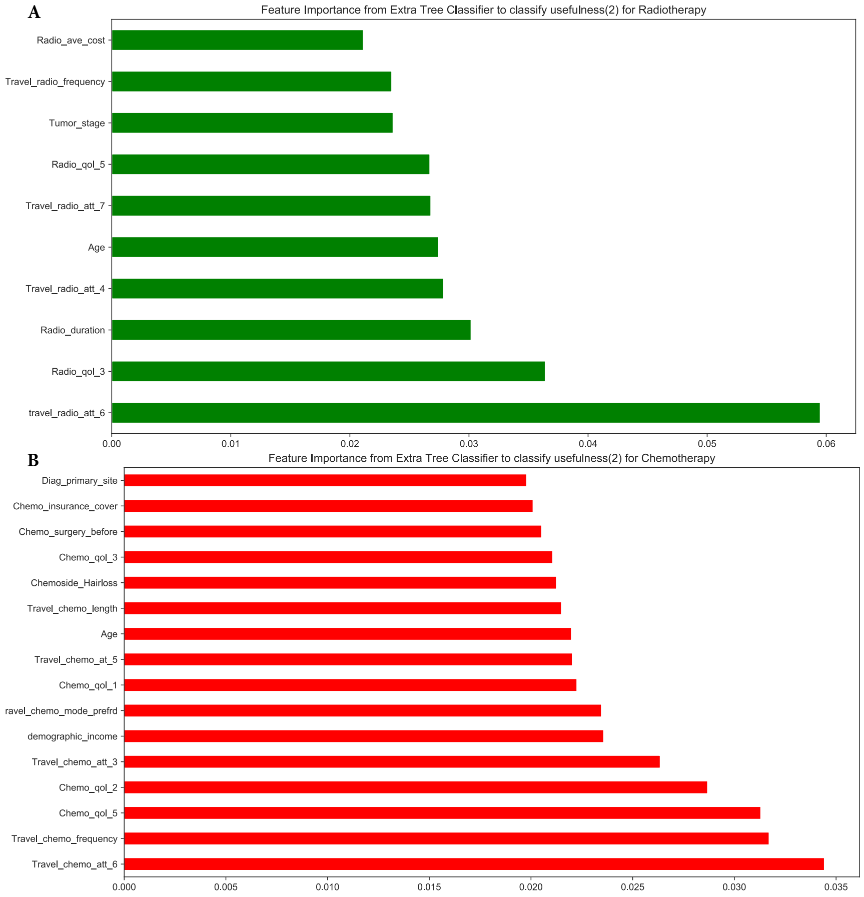


Figure 3.3. *usefulness(2)* (**A**) Important features of the model by using data of respondent who had radiotherapy. (**B**) Important features of the model by using data of respondent who had chemotherapy.

We further analyzed these important features by exploratory data analysis. We label them as follows: Level 1 corresponds to intersection between most important features from classification model and significant features that we observed from exploratory data analysis. Level 2 corresponds to important features from model but not significant from our exploratory analysis. Level 3 corresponds to features that we observed significant relation to target variable but not in the most important features from the model.

Level 1 - usefulness(1) - radiotherapy	**Level 2 - usefulness(1) - radiotherapy**	**Level 3 - usefulness(1) - radiotherapy**
Travel_radio_att_7	Radio_qol_2	Radio_stop_side_effect
Travel_radio_att_6	Education	Tumor_stage
APC_mutation	Travel_radio_att_5	Radioside_Hairloss
Travel_radio_cost	Travel_radio_frequency	Travel_radio_att_2
Travel_radio_lack_public	Radio_qol_6	Travel_radio_att_4
Travel_radio_att_3	demographic_health_coverage	HER2_mutation
BRCA1_mutation		TravelWork_radio_sideeffect
		TravelWork_radio_reliable
		TravelWork_rdio_losejob
		TravelWork_radio_publictr_impact
		demographic_race
		BAP1_mutation
		Radio_qol_5
		demographic_car
		Diag_primary_site
Level 1 - usefulness(1) - chemotherapy	**Level 2 - usefulness(1) - chemotherapy**	**Level 3 - usefulness(1) - chemotherapy**
Travel_chemo_att_6	Travel_chemo_att_4	Gender
BAP1 mutation	Chemoside_Diarrhea	Chemo_Nights_inhospital
Chemoside_Hairloss		Chemo_qol_6
Travel_chemo_att_7		Travel_chemo_att_2
TravelWork_chemo_length		Travel_chemo_at_5
Travel_chemo_length		Chemoside_Headache
BRCA2_mutation		Chemo_stop_side_effect
Travel_chemo_lack_public		TravelWork_chemo_reliable
Chemo_qol_5		TravelWork_chemo_losejob
Travel_chemo_cost		BRCA1_mutation
TravelWork_chemo_mode		Chemoside_Dizziness
Travel_chemo_mode_prefrd		demographic_race
APC_mutation		HER2_mutation
TravelWork_chemo_sideeffect		

Table 3.1. Levels of features from Figure 3.2 and exploratory data analysis

Patient travel attitudes, cost of travel during treatment and having mutations showed significance effect on patients' decision making during radiotherapy treatment. Similarly and additionally, side effects of chemotherapy and quality of life during treatments seem to be important for patients. We plotted some of the features from Level 1 categories to show that the algorithm was able to catch important connections. For Figure 3.4A, respondents who agreed to *usefulness(1)*, they mostly thought that app-based services was more reliable than car. For Figure 3.4B, if their travel for treatment cost was higher, they were more willing to have discounted/free ride. Figure 3.4C shows that most of patients with BAP1 mutations would like to have discounted/free rides.

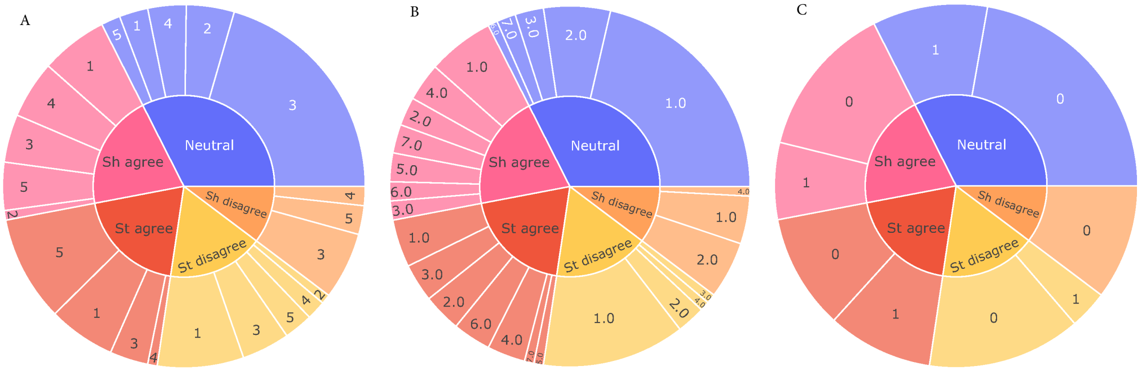


Figure 3.4. *usefulness(1)* (**A**) The relation between Travel_radio_att_7 and target variable. Travel_radio_att_7 corresponds to the question: "During the radiotherapy - Traveling to health care providers by Uber, Lyft or similar app-based services was more reliable than car" and options were Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5. (**B**) The relation between Travel_radio_cost and target variable. Travel_radio_cost corresponds to the question: "How much did it cost you to travel for per Radiation Treatment session? Please select an approximate amount" and options were Less than \$10: 1, \$11-20: 2, \$21-30: 3, \$31-40: 4, \$41-50: 5, \$51-60: 6, More than \$60: 7. (**C**) The relation between BRCA1_mutation and target variable. BRCA1_mutation corresponds to the question: "Please indicate all mutations that you are aware of? - Selected Choice" and options were Yes: 1, No: 0.

Patients who did not feel comfortable when traveling with others by public transit during chemotherapy treatment, we observed most of them would like to have discounted/free rides between their house and work (Figure 3.5A). Interestingly, if patients thought that discounted/free ride would be useful, they were more among those who had BAP1 mutation or APC mutation (Figure 3.5B and 3.5C, respectively).

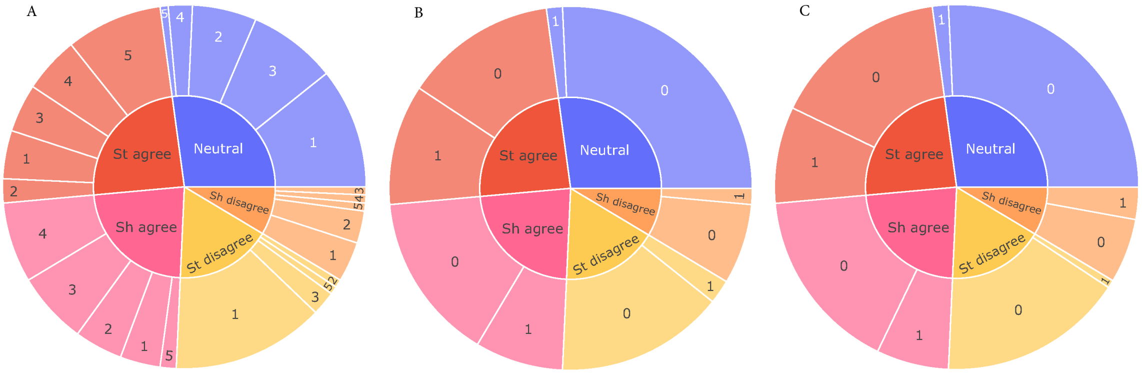


Figure 3.5. *usefulness(1)* (**A**) The relation between Travel_chemo_att_6 and target variable. Travel_chemo_att_6 corresponds to the question: “During your Chemotherapy to what extent you agree with the following statements: - I did not feel comfortable if I traveled with others by public transit” and options were Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5. (**B**) The relation between BAP1_mutation and target variable. BAP1_mutation corresponds to the question: “Please indicate all mutations that you are aware of? - Selected Choice” and options were BAP1 Yes: 1, No: 0. (**C**) The relation between APC_mutation and target variable. APC_mutation corresponds to the question: “Please indicate all mutations that you are aware of? - Selected Choice” and options were APC Yes: 1, No: 0.

We also look at some of the level 3 features from Table 3.1. More cancer patients with higher tumor stages agreed to have discounted rides especially if patient stopped radiotherapy due to side effects. Although we do not have too many black (6.2%) and Hispanic or Latino (5.2%) patients, they were more willing to have discounted/free rides (Figure 3.6A). Furthermore, most of female Hispanic or Latino respondents responded that discounted/free ride could help them to have better life during chemotherapy (Figure 3.9B).

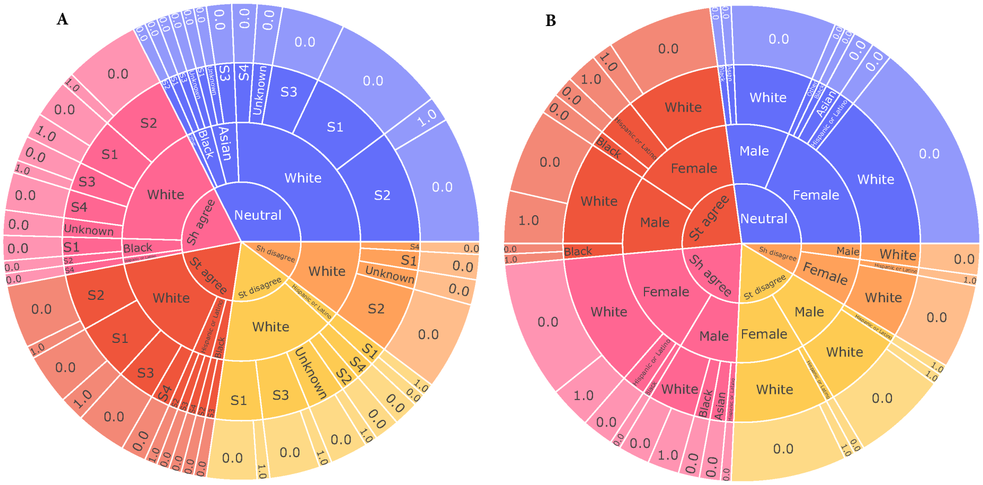


Figure 3.6. *usefulness(1)* (**A**-radiotherapy) The relation between variables race, tumor stage, Radio_stop_sideeffect and target variable. (**B**-chemotherapy) The relation between variables gender, race, Chemo_stop_side_effect and target variable.

Table 3.2 level 1 features indicates the most important factors for patients who benefit from free/discounted rides are travel attitudes, quality of life of patients during radiotherapy and age. The length of travel to health care center, the percentage of chemotherapy treatment cost covered by insurance also seem to be important key factors to classify patients.

Level 1 - usefulness(2) - radiotherapy	**Level 2 - usefulness(2) - radiotherapy**	**Level 3 - usefulness(2) - radiotherapy**
Travel_radio_att_6	Radio_duration	Radio_insurance_cover
Travel_radio_att_4	Travel_radio_frequency	Travel_radio_mode
Radio_qol_3	Radio_ave_cost	Travel_radio_lack_car
Travel_radio_att_7	Tumor_stage	TravelWork_radio_losejob
Radio_qol_5		TravelWork_radio_publictr_impact
Age		APC_mutation
		BAP1_mutation
		BRCA1_mutation
		HER2_mutation
		Radioside_Dizziness
		Radioside_Headache
		demographic_household
		Radio_qol_2
		Radio_surgery_before
		demographic_employment
		Radio_qol_4
		Radioside_Hairloss
		Radio_qol_1
Level 1 - usefulness(2) - chemotherapy	**Level 2 - usefulness(2) - chemotherapy**	**Level 3 - usefulness(2) - chemotherapy**
Travel_chemo_att_6	Travel_chemo_frequency	Chemo_tumor_free_years
Chemo_qol_2	Chemo_qol_5	BRCA2_mutation
Travel_chemo_att_3	Travel_chemo_att_5	Travel_chemo_lack_public
Travel_chemo_mode_preferred	Diag_primary_site	TravelWork_chemo_losejob
Chemo_surgery_before	demographic_income	TravelWork_publictr_impact
Chemoside_Hairloss		TravelWork_chemo_length
Chemo_qol_1		Travel_chemo_workathome
Age		demographic_licence
Travel_chemo_length		demographic_race
Chemo_qol_3		APC_mutation
Chemo_insurance_cover		BRCA1_mutation
		HER2_mutation
		Chemoside_Dizziness
		Chemoside_Headache
		BAP1_mutation
		Travel_chemo_att_7
		Travel_chemo_att$_4$

Table 3.2. Levels of features from Figure 3.3 and exploratory data analysis

We observed that if radiotherapy affected patients' ability to work which was one of the quality of life related questions in the survey, they were more willing to have discounted/free ride (Figure 3.7B). Additionally, we saw more people responded as agree to *usefulness(2)* if they felt more exhausted to travel by car than public transit and/or other services to health care provide (Figure 3.7C).

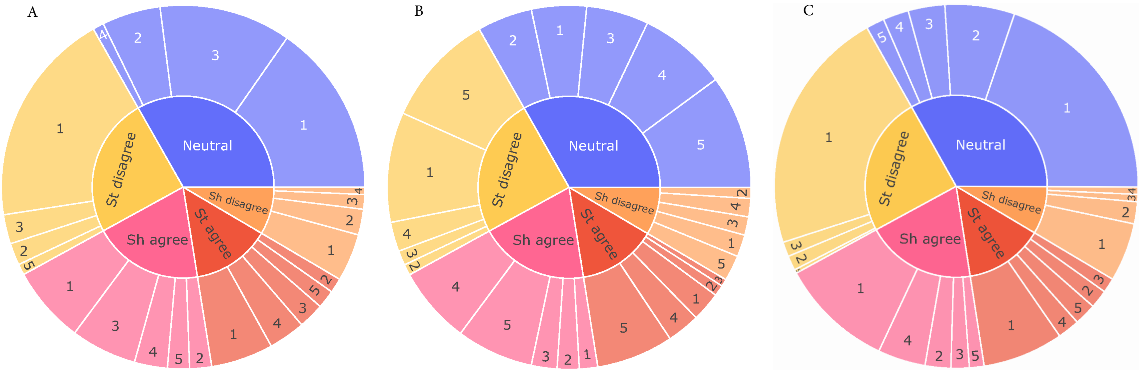


Figure 3.7. *usefulness(2)* (**A**) The relation between Travel_radio_att_6 and target variable. Travel_radio_att_6 corresponds to the question: "During your Radiotherapy to what extent you agree with the following statements: - I did not feel comfortable if I traveled with others by public transit" (**B**) The relation between Radio_qol_3 and target variable. Radio_qol_3 corresponds to the question: "To what extent do you agree with these statements During Radiation Treatments? - Treatment affected my ability to work" (**C**) The relation between Travel_radio_att_4 and target variable. Travel_radio_att_4 corresponds to the question: "During your Radiation Treatments to what extent do you agree with the following statements: - Traveling by car to health care provider made me more exhausted compared to riding public transit, Uber/Lyft or other services" For all above-mentioned questions, options were Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5.

Moreover, participants who had difficulties to pay chemotherapy treatment cost and who had surgery before chemotherapy, they believed to have discounted/free ride could be useful (Figure 3.8A and 3.8C, respectively).

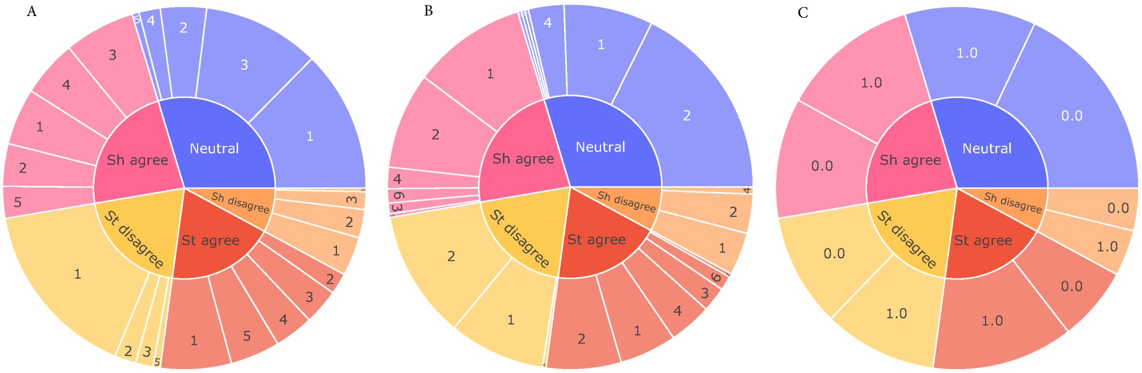


Figure 3.8. *usefulness(2)* (**A**) The relation between Chemo_qol_2 and target variable. Chemo_qol_2 corresponds to the question: "To what extent do you agree with these statements During Chemotherapy Treatments? - I had difficulties in paying treatment costs" and options were Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5. (**B**) The relation between Travel_chemo_mode_preferred and target variable. Travel_chemo_mode_preferred corresponds to the question: "During Chemotherapy Treatment-If you have had choice to choose your transportation mode to health care provider, what would you prefer?" and options were Car, alone: 1, Car, with others: 2, Bus/Rail: 3, Free transportation services for cancer patients: 4, Taxi/cab: 5, Uber/Lyft or similar services: 6. (**C**) The relation between Chemo_surgery_before and target variable. Chemo_surgery_before corresponds to the question: "For your First Chemotherapy Period: - Have you had a surgery before chemotherapy treatment to remove the tumor(s)?" and options were No: 0, Yes: 1.

Figure 3.9 shows some level 3 variables from Table 3.2. We observed a connection between lower insurance coverage, having side effect of radiotherapy and being agree to ride usefulness from home to health care center (Figure 3.9A). Another interesting finding was that if people had less tumor free years after chemotherapy and if they had dizziness due to treatment, they made their choices towards to usefulness of rides (Figure 3.9B).

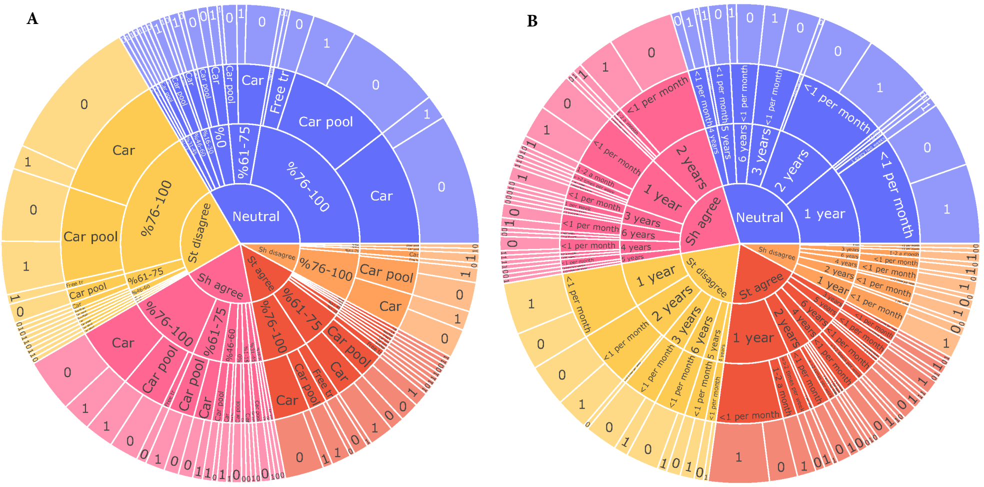


Figure 3.9. *usefulness(2)* (**A**-radiotherapy) The relation between variables Radio_insurance_cover, Travel_radio_mode, Radioside_Dizziness and target variable. (**B**-chemotherapy) The relation between variables Chemo_tumor_free_years, Travel_chemo_lack_public, Chemoside_Dizziness and target variable.

3.4 Discussion

Machine learning models can find the complex linear/non-linear relationship between variables and outcome in many areas and recently including survey research. In particular, tree-based models have been shown to be useful to model these relations due to their flexibility, non-parametric nature and computational effectiveness [103, 104, 105]. Since the single decision tree approach often causes high variance in prediction, this limitation seem to be addressed by ensemble tree methods such as random forest and extremely randomized trees [102, 106]. In this project, our major focus was on discovering the key factors in cancer patients' quality of life. We wanted to know the effect of socioeconomic, demographic factors, tumor characteristics, treatment related work travel burden during and after treatments in their life. To explore these, we designed our survey without hypothesizing any assumptions on respondents' behaviour. Our data driven approach gave us flexibility to handle diverse data of participants with many cancer types. We found that in order to help cancer patients, especially those who had APC, BAP1, BRCA1 and BRCA2 mutations, discounted/free rides from their home to work during treatments can be beneficial.

It is worth noting that we observed treatment burdens such as cost of treatments and percentage of insurance coverage of treatment influences cancer patients' life. Travel attitudes during treatment have a significant impact on life of cancer patients regardless of type of treatments. People who thought that Uber, Lyft or similar app-based services was more reliable than car, most of them believed that discounted/free rides from home to work may improve their well-being.

Additionally, our analysis suggests that providing free rides from home to work is likely to be helpful for patients who suffer from cost of travel to health care providers during radiotherapy treatments.

Another important practical recommendation of our study is that providing discounted/free rides from home to treatment centers seem to be a reasonable approach to tackle problems of patients especially who had difficulties to pay chemotherapy treatment costs and whose work ability got affected while getting radiotherapy treatment.

Side effects of chemotherapy such as having problem to get to work during treatment, losing hair were among the important features in our model (Table 3.1). These may result a diminished quality of life [95].

Majority of Hispanic/Latino patients cited discounted/free ride could improve their life. This work contributes to existing knowledge of transportation difficulties for minorities in the U.S [107].

Some study limitations should be noted. Although we did not focus on classification algorithm performance on a validation set, we will focus on this as a future study. Moreover, the number of participant was 750 due to the cost of conducting a nationwide survey. This number can be increased which may likely to help improve model performance.

Our research is different than most of the literature in a manner that our questions did not only cover traveling from home to health care provider but also traveling from home to work. We hope that our research will drive researchers attention to daily life of cancer patients as well. More study is needed to further develop the methods and investigate features of usefulness of rides and their association with quality of life of cancer patients while considering treatments, daily life activities and work related problems.

APPENDIX A

Appendix

Codebook for Survey: The Role of Transportation on Cancer Patient Decisions-Making Through Machine Learning Techniques

Survey year : 2019

"What is your age?" : **Age**

Numeric

"What is your biological sex?" : **Gender**

Female: 0, Male: 1

"Are you" : **RecentDiag_remission**

In remission: 1, Not recently diagnosed, but seeking treatment: 2, Recently diagnosed: 3

"At the first time of diagnosis, what was - stage of your tumor": **Diag_tumor_stage**

Stage I: 1, Stage II: 2, Stage III: 3, Stage IV: 4, Unknown: 99

"At the first time of diagnosis, what was - grade of your tumor" : **Diag_tumor_grade**

Grade 1: 1, Grade 2: 2, Grade 3: 3, Grade 4: 4, Unknown: 99

"At the first time of diagnosis, what was - number of metastatic tumors": **Diag_no_metastatic**

0: 0, 1: 1, 2: 2, 3 or more: 3

"At the first time of diagnosis, what was - primary site of your tumor": **Diag_primary_site**

Multiple choice question. 16 types of cancer

"Please first write the location of your primary tumor, then the locations of tumors in the order that they appeared over time and treatments. Example: colon, surgery and 3 months chemo, 10 yrs tumor free, colon and liver, surgery, 3 months chemo, and 2 weeks radiotherapy, now tumor free for 5yrs": **Primary_tumorlocation**

Only comment, no re-coding is needed

"At the first time of diagnosis, what was - Your age (at the time of diagnosis)": **Diag_age**

Numeric

"At the first time of diagnosis, what was - HPV status": **HPV_status**

HPV negative: 0, HPV positive: 1, I am not sure: 99

"At the first time of diagnosis, what was - HIV status": **HIV_status**

HIV negative: 0, HIV positive: 1, I am not sure: 99

"When was your cancer diagnosed? Please specify date of diagnosis. - Month": **Diag_month**

January: 1, February: 2, March: 3, April: 4, May: 5, June: 6, July: 7, August: 8, September: 9, October: 10, November: 11, December: 12

"When was your cancer diagnosed? Please specify date of diagnosis. - Year": **Diag_year**

Numeric

"Please indicate all mutations that you are aware of? - Selected Choice": **mutations**

Multiple choice question***** APC,BAP1,BRCA1, BRCA2, HER2, P53.

Other,please specify: **mutation_other**

Only comment, no re-coding is needed

"Please indicate all treatments other than surgery that you have received? - Selected Choice": **treatments**

Multiple choice question***** Radiotherapy, Chemotherapy, Other

"Please indicate all treatments other than surgery that you have received? - Other, please specify": **Treatments_other**

Only comment, no re-coding needed

"How many times your tumor reoccurred?": **Reoccurence**

Numeric

"On average, how long have you been tumor free between reoccurrences?" : **Tumor_free_reoccurence**

No re-occurrences: 0, Less than 1 year: 1, 1-3 years: 2, 4-6 years: 3, 7-10 years: 4, 11-15 years: 5, more than 15 years: 6

"How many surgeries you had for removing tumor(s)?": **remove_tumor_surgery**

Numeric

"If you had surgery before initial treatment, what kind of surgery you had - Selected Choice": **surgery_before**

Complete removing: 0, Partial removing: 1, Other, please specify: 2, Missing: 98

"If you had surgery before initial treatment, what kind of surgery you had - Other, please specify" : **surgery_before_other**

Only comment, no re-coding is needed

"Have you ever changed your house (living location) because of your disease?": **location_change**

No: 0, Yes: 1, Missing: 98

"Have you ever changed your house (living location) because of your disease? If Yes, - From Zipcode": **location_change_fromzip**

Numeric

"Have you ever changed your house (living location) because of your disease? If Yes, - To Zipcode": **location_change_topzip**

Numeric

"Have you ever changed your house (living location) because of your disease? If Yes, - The main reason": **location_change_reason**

Only comment, no re-coding is needed

Radiation Therapy

“For your Radiation Treatments: - Year of first radiotherapy”: **Radio_firstyear_of_radiotherapy**

Numeric

“For your Radiation Treatments: - Year of last radiation therapy”: **Radio_lastyear_of_radiotherapy**

Numeric

“For your Radiation Treatments: - Number of separate radiotherapy periods”: **Radio_no_separate_radiotherapy**

Numeric

“For your Radiation Treatments: - Number of nights you spent in hospital during your radiation treatments”: **Radio_Nonights_inhospital**

Numeric

“For your Radiation Treatments: - Number of days you needed an assistant to do your day to day activities during radiotherapies and their recovery periods”: **Radio_Need_assistant**

Numeric

“Please answer following questions about your FIRST Radiation Treatment - Radiotherapy duration in weeks”: **Radio_duration**

Missing: 98, Less than 2 weeks: 1, 2-3 weeks: 2, 4-5 weeks: 3, 6-7 weeks: 4, 8-9 weeks: 5, more than 9 weeks: 6

“Please answer following questions about your FIRST Radiation Treatment - Frequency of treatments”: **Radio_frequency**

Missing: 98, daily: 1, 2 days a week: 2, 3 days a week: 3, 4 days a week: 4, 5 days a week: 5, more than 5 days weeks: 6

“Tumor stage before radiotherapy”: **Radio_before_tumor_stage**

Missing: 98, Stage I: 1, Stage II: 2, Stage III: 3, Stage IV: 4, Tumor free: 0, Unknown: 99

“Tumor stage right after radiotherapy”: **Radio_after_tumor_stage**

Missing: 98, Stage I: 1, Stage II: 2, Stage III: 3, Stage IV: 4, Tumor free: 0, Unknown: 99

“Years being tumor free after radiotherapy”: **Radio_tumor_free_years**

Missing: 98, less than 1 year: 1, 1-3 years: 2, 4-5 years: 3, 6-7 years: 4, 8-10 years: 5, more than 10 years: 6

“After radiation, stage of reoccurred cancer”: **Radio_stageof_reoccured_cancer**

Missing: 98, Stage I: 1, Stage II: 2, Stage III: 3, Stage IV: 4, Tumor free: 0, Unknown: 99

“Have you experienced any of the following side effects During Radiation Treatments? - Selected Choice” : **Radio_side_effect**

Multiple choice question***** Dizziness, Diarrhea, Headache, Nausea, Hair loss, Poor Appetite.

“If other, please specify”: **Radio_side_effect_other**

Only comment, no re-coding is needed

“For your First Radiation Treatment: - Are you currently having your first radiation treatment?”: **Radio_current_first**

No: 0, Yes: 1, Missing: 98

“For your First Radiation Treatment: - Did you stop the radiation treatment because of the side effects?”: **Radio_stop_side_effect**

No: 0, Yes: 1, Missing: 98

“For your First Radiation Treatment: - Have you had a surgery before radiation treatment to remove tumors?”: **Radio_surgery_before**

No: 0, Yes: 1, Missing: 98

"For your First Radiation Treatment: - Have you had surgery during radiation to remove tumors?": **Radio_surgery_during**

No: 0, Yes: 1, Missing: 98,

"For your First Radiation Treatment: - Have you had radiation and chemotherapy at the same time?": **Radio_chemo_and_radio**

No: 0, Yes: 1, Missing: 98

"For your First Radiation Treatment: - Have you used Anti-Inflammatory Medications (NSAIDs) during your radiotherapy?": **Radio_NSAID**

No: 0, Yes: 1, Missing: 98

"Which of the following best describes your condition after your LAST Radiation Treatment? - Selected Choice" **Radio_condition_after**

Multiple choice question*****

"Tumor free, for how long: - Text": **Radio_condition_after_2**

Only comment, no re-coding is needed

"New tumors appeared in what locations: - Text": **Radio_condition_after_3**

Only comment, no re-coding is needed

"What was the location of facility providing Radiation Treatment? - Health care provider name": **Radio_provider_name**

Only comment, no re-coding is needed

"What was the location of facility providing Radiation Treatment? - Zipcode": **Radio_provider_zip**

Numeric

"To what extent do you agree with these statements During Radiation Treatments? - I needed pain-killers to do my day-to-day activities": **Radio_qol_1**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

'"To what extent do you agree with these statements During Radiation Treatments? - I had difficuties in paying treatment costs": **Radio_qol_2**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

"To what extent do you agree with these statements During Radiation Treatments? - Treatment affected my ability to work": **Radio_qol_3**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

"To what extent do you agree with these statements During Radiation Treatments? - Treatment affected my ability to drive": **Radio_qol_4**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

"To what extent do you agree with these statements During Radiation Treatments - My employer let me work flexible schedule to meet my treatment needs": **Radio_qol_5**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

"To what extent do you agree with these statements During Radiation Treatments? - I was satisfied with my overall quality of life": **Radio_qol_6**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

"Cost of your Radiation Treatment: - What percentage of your cost was covered by your insurance?": **Radio_insurance_cover**

Missing: 98, 0: 0, %1-15: 1, %16-30: 2, %31-45: 3, %46-60: 4, %61-75: 5, %76-100: 6

“Cost of your Radiation Treatment: - How much on average you spent for your radiation treatment?”: **Radio_ave_cost**

Missing: 98, Less than $50: 1, $51-$100: 2, $101-$250: 3, $251-$500: 4, $501-$1000: 5, $1001-$2000: 6, More than $2000: 7

Transportation for radiation

“Please tell us about your trip to health care providers During Radiation Treatments. - How often did you make a trip to your health care provider?”: **Travel_radio_frequency**

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

“Please tell us about your trip to health care providers During Radiation Treatments. - How long did it usually take to get to your health care provider?”: **Travel_radio_length**

Missing: 0, Less than 15 minutes: 1, 15-30 minutes: 2, 30-45 minutes: 3, 45-60 minutes: 4, More than 60 minutes: 5

“Please tell us about your trip to health care providers During Radiation Treatments. - What was your main transportation mode to get to your health care provider?”: **Travel_radio_mode**

Missing: 98, Car, alone: 1, Car, with others: 2, Bus/Rail: 3, Free transportation services for cancer patients: 4, Taxi/cab: 5, Uber/Lyft or similar services: 6, Uber pool or similar services: 7

“Please tell us about your trip to health care providers During Radiation Treatments. - If you have had choice to choose your transportation mode to health care provider, what would you prefer? ”: **Travel_radio_mode_prefrd**

Missing: 98, Car, alone: 1, Car, with others: 2, Bus/Rail: 3, Free transportation services for cancer patients: 4, Taxi/cab: 5, Uber/Lyft or similar services: 6, Uber pool or similar services: 7

“During your Radiation Treatments to what extent you agree with the following statements: - I preferred to ride a car by a family member/friend rather than ride public transit to get to health care providers”: **Travel_radio_att_1**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

“During your Radiation Treatments to what extent you agree with the following statements: - Traveling by car was more time-saving and safer for me comparing to riding public transit”: **Travel_radio_att_2**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

“During your Radiation Treatments to what extent you agree with the following statements: - Traveling by car to health care provider made me more exhausted compared to riding public transit, Uber/Lyft or other services”: **Travel_radio_att_3**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

“During your Radiation Treatments to what extent you agree with the following statements: - I preferred to ride public transit rather than asking other people to take me a ride to health care provider”: **Travel_radio_att_4**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

“During your Radiation Treatments to what extent you agree with the following statements: - I did not feel comfortable if I traveled with others by public transit”: **Travel_radio_att_5**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

“During your Radiation Treatments to what extent you agree with the following statements: - Traveling to health care providers by Uber, Lyft or similar app-based services was more reliable than car”: **Travel_radio_att_6**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

“During your Radiation Treatments to what extent you agree with the following statements: - Traveling to health care providers by traditional public transport was cheaper than Uber, Lyft or similar app-based services”: **Travel_radio_att_7**

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

“During your Radiation Treatment how often did you miss your appointments due to following reasons? - Lack of access to private car”: **Travel_radio_lack_car**

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

“During your Radiation Treatment how often did you miss your appointments due to following reasons? - Lack of access to traditional public transit (bus and rail)”: **Travel_radio_lack_public**

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

“During your Radiation Treatment how often did you miss your appointments due to following reasons? - Lack of access to app-based mobility such as Uber, Lyft, Uber pool or similar services”: **Travel_radio_lack_appbased**

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

“During your Radiation Treatment how often did you miss your appointments due to following reasons? - Lack of access to free transportation services for cancer patients”: **Travel_radio_lack_freetr**

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

“How much did it cost you to travel for per Radiation Treatment session? Please select an approximate amount”: **Travel_radio_cost**

Missing: 98, Less than $10: 1, $11-20: 2, $21-30: 3, $31-40: 4, $41-50: 5, $51-60: 6, More than $60: 7

“Please answer following questions for specifically During Radiation Treatments : Did you have a job?”: **Radio_job**

Missing: 98, No: 0, Yes: 1

“Please answer following questions for specifically During Radiation Treatments : - Did you have problem to get to work because of the side effects of treatment?”: **TravelWork_radio_sideeffect**

Missing: 98, No: 0, Yes: 1

“Please answer following questions for specifically During Radiation Treatments : - Was there a reliable public transportation between your home and work?”: **TravelWork_radio_reliable**

Missing: 98, No: 0, Yes: 1

“Please answer following questions for specifically During Radiation Treatments : Did you lose your job or did not accept a job offer because of lack of public transportation?”: **TravelWork_radio_losejob**

Missing: 98, No: 0, Yes: 1

“Please answer following questions for specifically During Radiation Treatments : - Did the existence or lack of public transportation have any impact on your decision regarding choosing radiation treatment?”: **TravelWork_radio_publictr_impact**

Missing: 98, No: 0, Yes: 1

“Please tell us about your work/school trips During your Radiation Treatment. - How long did it usually take to get to your primary place of work/school (by car)?”: **TravelWork_radio_length**

Missing: 98, Less than 15 minutes: 1, 15-30 minutes: 2, 30-45 minutes: 3, 45-60 minutes: 4, More than 60 minutes: 5

“Please tell us about your work/school trips During your Radiation Treatment. - How often did you work at home instead of making the trip to work/school?”: **Travel_radio_workathome**

Missing: 98, Never: 0, Less than per month: 1, Once or twice a month: 2, About once every two week: 3, About once per week: 4, 2-3 times per week: 5, More than 4 times per week: 6

“Please tell us about your work/school trips During your Radiation Treatment. - What was your main transportation mode to get to your primary place of work/school?”: **TravelWork_radio_mode**

Missing: 98, Car, alone: 1, Car, with others: 2, Bus/Rail: 3, Free transportation services for cancer patients: 4, Taxi/cab: 5, Uber/Lyft or similar services: 6, Uber pool or similar services: 7

Treatment related questions are same for all treatments, so we will not repeat questions here. Variables are re-coded as follows.

Chemotherapy

Chemo_firstyear_of_chemotherapy

Numeric

Chemo_lastyear_of_chemotherapy

Numeric

Chemo_no_separate_chemotherapy

Numeric

Chemo_Nights_inhospital

Numeric

Chemo_Need_assistant

Multiple choice question

Chemo_duration

Missing: 98, Less than 3 months: 1, 3-6 months: 2, 7-9 months: 3, 10-12 months: 4, 13-18 months: 5, More than 18 months: 6

“Please answer following questions about your FIRST Chemotherapy. - The length of each cycle (in average)”: **Chemo_cycle_length**

Missing: 98, <= 2 weeks: 1, 3 weeks: 2, 4 weeks: 3, 5 weeks: 4, 6 weeks: 5, more than 6 weeks: 6

Chemo_before_tumor_stage

Missing: 98, Stage I: 1, Stage II: 2, Stage III: 3, Stage IV: 4, Tumor free: 0, Unknown: 99

Chemo_after_tumor_stage

Missing: 98, Stage I: 1, Stage II: 2, Stage III: 3, Stage IV: 4, Tumor free: 0, Unknown: 99

Chemo_tumor_free_years

Missing: 98, less than 1 year: 1, 1-3 years: 2, 4-5 years: 3, 6-7 years: 4, 8-10 years: 5, more than 10 years: 6

Chemo_stageof_reoccured_cancer

Missing: 98, Stage I: 1, Stage II: 2, Stage III: 3, Stage IV: 4, Tumor free: 0, Unknown: 99

Chemo_drug

Only comment, no re-coding is needed

Chemo_side_effect

Left, multiple choice question

Chemo_side_effect_other

Only comment, no re-coding is needed

Chemo_current_first

Missing: 98, No: 0, Yes: 1

Chemo_stop_side_effect

Missing: 98, No: 0, Yes: 1

Chemo_surgery_before

Missing: 98, No: 0, Yes: 1

Chemo_surgery_during

Missing: 98, No: 0, Yes: 1

Chemo_NSAID

Missing: 98, No: 0, Yes: 1

Chemo_condition_after

Multiple choice question*****

Chemo_condition_after_2

Only comment, no re-coding is needed

Chemo_condition_after_3

Only comment, no re-coding is needed

Chemo_provider_name

Only comment, no re-coding is needed

Chemo_provider_zip

Numeric

Chemo_qol_1

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Chemo_qol_2

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Chemo_qol_3

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Chemo_qol_4

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Chemo_qol_5

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Chemo_qol_6

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Chemo_insurance_cover

Missing: 98, %0: 0, %1-15: 1, %16-30: 2, %31-45: 3, %46-60: 4, %61-75: 5, %76-100: 6

Chemo_ave_cost

Missing: 98, Less than $50: 1, $51-$100: 2, $101-$250: 3, $251-$500: 4, $501-$1000: 5, $1001-$2000: 6, More than $2000: 7

Transportation for chemotherapy

Travel_chemo_frequency

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

Travel_chemo_length

Missing: 98, Less than 15 minutes: 1, 15-30 minutes: 2, 30-45 minutes: 3, 45-60 minutes: 4, More than 60 minutes: 5

Travel_chemo_mode

Missing: 98, Car, alone: 1, Car, with others: 2, Bus/Rail: 3, Free transportation services for cancer patients: 4, Taxi/cab: 5, Uber/Lyft or similar services: 6, Uber pool or similar services: 7

Travel_chemo_mode_prefrd

Missing: 98, Car, alone: 1, Car, with others: 2, Bus/Rail: 3, Free transportation services for cancer patients: 4, Taxi/cab: 5, Uber/Lyft or similar services: 6

Travel_chemo_att_1

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Travel_chemo_att_2

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Travel_chemo_att_3

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Travel_chemo_att_4

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Travel_chemo_att_5

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Travel_chemo_att_6

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Travel_chemo_att_7

Missing: 98, Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

Travel_chemo_lack_car

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

Travel_chemo_lack_public

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

Travel_chemo_lack_appbased

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

Travel_chemo_lack_freetr

Missing: 98, Two or more times per week: 5, About once per week: 4, About once every two weeks: 3, Once or twice a month: 2, Less than once per month: 1

Travel_chemo_cost

Missing: 98, Less than $10: 1, $11-20: 2, $21-30: 3, $31-40: 4, $41-50: 5, $51-60: 6, More than $60: 7

Chemo_job

Missing: 98, No: 0, Yes: 1

TravelWork_chemo_sideeffect

Missing: 98, No: 0, Yes: 1

TravelWork_chemo_reliable

Missing: 98, No: 0, Yes: 1

TravelWork_chemo_losejob

Missing: 98, No: 0, Yes: 1

TravelWork_chemo_publictr_impact

Missing: 98, No: 0, Yes: 1

TravelWork_chemo_length

Missing: 98, Less than 15 minutes: 1, 15-30 minutes: 2, 30-45 minutes: 3, 45-60 minutes: 4, More than 60 minutes: 5

Travel_chemo_workathome

Missing: 98, Never: 0, Less than per month: 1, Once or twice a month: 2, About once every two week: 3, About once per week: 4, 2-3 times per week: 5, More than 4 times per week: 6

TravelWork_chemo_mode

Missing: 98, Car, alone: 1, Car, with others: 2, Bus/Rail: 3, Free transportation services for cancer patients: 4, Taxi/cab: 5, Uber/Lyft or similar services: 6, Uber pool or similar services: 7,

"Please tell us about yourself - Your race": **demographic_race**

Missing: 98, White: 1, Black or African American: 2, Hispanic or Latino: 3, Asian: 4, Other: 5

"Please tell us about yourself - Your marital status": **demographic_marital_status**

Missing: 98, Married: 1, Divorced: 2, Separated: 3, Widowed: 4, Never married: 5

"Please tell us about yourself - Your educational background": **demographic_education**

Missing: 98, Some school but no degree: 0, Less than high school degree: 1, High school degree or equivalent(e.g.,GED): 2, Associate degree: 3, Bachelor degree: 4, Graduate degree: 5

“Which statement best describes your current employment status? - Selected Choice”: **demographic_employment**

Working (paid employee): 7, Working (self-employed): 6, Not working (temporary layoff from a job): 5, Not working (looking for work): 4, Not working (disabled): 3, Not working (retired): 2, Other,please specify: 1, Prefer not to answer: 0

“Which statement best describes your current employment status? - Other,please specify - Text”: **demographic_employment_other**

Only comment, no re-coding is needed

“Please answer the following questions - Your annual income”: **demographic_income**

Missing: 98, Less than $20,000: 1, $20,000-$34,999: 2, $35,000-$49,999: 3, $50,000-$74,999: 4, $75,000-$99,999: 5, $100,000 or more: 6

“Please answer the following questions - Your current residency (own or rent)”: **demographic_residency**

Missing: 98, Own: 1, Rent: 2

“Please answer the following questions - Do you have a driver’s licence?”: **demographic_licence**

Missing: 98, No: 0, Yes: 1

“What is your household size (by counting yourself) ? - Household size is”: **demographic_household**

Numeric

“How many cars do you have access in your household?”: **demographic_car**

None: 0, 1 car: 1, 2 cars: 2, 3 or more cars: 3

“Which health coverage are you currently enrolled with? - Selected Choice”: **demographic_health_coverage**

Missing: 98, Medicaid: 7, Medicare: 6, Affordable Care Act: 5, Employer-paid insurance: 4, Private health insurance: 3, Uninsured: 2, Other,please specify: 1

“Which health coverage are you currently enrolled with? - Other,please specify - Text”: **demographic_health_coverage_other**

Only comment, no re-coding needed

“Your current overall quality of life: - How would you rate your overall quality of life after treatments?”: **Qol_1**

Terrible: 1, Poor: 2, Average: 3, Good: 4, Excellent: 5

“Your current overall quality of life: - How would you rate your overall physical condition after treatments?”: **Qol_2**

Terrible: 1, Poor: 2, Average: 3 , Good: 4, Excellent: 5

“Usefulness of free/discounted rides to cancer patients and any further comments - I could have a better life if there was free or discounted rides during my treatments between my house and work?”: **Usefulness_1_n**

Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5,

“Usefulness of free/discounted rides to cancer patients and any further comments - I could have a better life if there was free or discounted rides during my treatments between my house and health care providers?”: **Usefulness_2_n**

Strongly disagree: 1, Somewhat disagree: 2, Neither agree nor disagree: 3, Somewhat agree: 4, Strongly agree: 5

“If you have any other thoughts/comments/feedback that you would like to share with us, please write them below”: **other_comments**

Only comment, no re-coding is needed

“What of the following statements would best describe your response to this survey”: **honest_2**

I answered all questions thoughtfully and honestly. : 5, I answered most of the questions thoughtfully and honestly. : 4, I answered some of the questions thoughtfully and honestly. : 3, I answered small number of the questions thoughtfully and honestly. : 2, I answered none of the questions thoughtfully and honestly. : 1

REFERENCES

[1] Wagner, E. et al. The quality of cancer patient experience: Perspectives of patients, family members, providers and experts. *BMJ Quality and Safety*, **19,** 484-489 (2010).

[2] Epstein, R., Fiscella, K., Lesser, C., Stange, K. Why the nation needs a policy push on patient-centered health care. *Health Affairs*, **29,** 1489-1495 (2010).

[3] David Crawford, E. et al. Comparison of perspectives on prostate cancer: analyses of survey data. *Urology*, **50,** 366-372 (1997).

[4] Feldman-Stewart, D., Brundage, M., Manen, L. Svenson, O. Patient-focussed decision-making in early-stage prostate cancer: insights from a cognitively based decision aid. *Health Expectations*, **7,** 126-141 (2004).

[5] Volk, R. et al. Preferences of husbands and wives for outcomes of prostate cancer screening and treatment. *J. Of General Internal Medicine*, **19,** 339-348 (2004).

[6] Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA. *Cancer J. Clin.* **68,** 394-424 (2018).

[7] Cancer of the Breast (Female): Cancer Stat Facts. SEER (2019). Available at: https://seer.cancer.gov/statfacts/html/breast.html. (Accessed: 11th July 2019)

[8] Wan, D., Villa, D., Woods, R., Yerushalmi, R. Gelmon, K. Breast cancer subtype variation by race and ethnicity in a diverse population in British Columbia. *Clin. Breast Cancer*. **16,** e49-e55 (2016).

[9] Troester, M. A. et al. Racial differences in PAM50 subtypes in the Carolina breast cancer study. *J. Natl. Cancer Inst.* **110,** 176-182 (2017).

[10] Keegan, T. H. M., DeRouen, M. C., Press, D. J., Kurian, A. W. Clarke, C. A. Occurrence of breast cancer subtypes in adolescent and young adult women. *Breast Cancer Res.* **14,** (2012).

[11] Parada, H. et al. Race-associated biological differences among luminal A and basal-like breast cancers in the Carolina Breast Cancer Study. *Breast Cancer Res.* **19,** 1-9 (2017).

[12] Sørlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci.* **100,** 8418-8423 (2003).

[13] Brenton, J., Carey, L., Ahmed, A. Caldas, C. Molecular classification and molecular forecasting of breast cancer: Ready for clinical application? *J. of Clin. Oncol.*. **23,** 7350-7360 (2005).

[14] O'Brien, K. M. et al. Intrinsic breast tumor subtypes, race, and long-term survival in the carolina breast cancer study. *Clin. Cancer Res.* **16,** 6100-6110 (2010).

[15] Clarke, C. A. et al. Age-specific incidence of breast cancer subtypes: Understanding the black-white crossover. *J. Natl. Cancer Inst.* **104,** 1094-1101 (2012).

[16] Engstrøm, M. J. et al. Molecular subtypes, histopathological grade and survival in a historic cohort of breast cancer patients. *Breast Cancer Res. Treat.* **140,** 463-473 (2013).

[17] Parise, C. A., Bauer, K. R. Caggiano, V. Variation in breast cancer subtypes with age and race/ethnicity. *Crit. Rev. Oncol. Hematol.* **76,** 44-52 (2010).

[18] Pilewskie, M. King, T. A. Age and molecular subtypes: Impact on surgical decisions. *J. Surg. Oncol.* **110,** 8-14 (2014).

[19] Petkov, V. I. et al. Breast-cancer-specific mortality in patients treated based on the 21-gene assay: A SEER population-based study. *npj Breast Cancer* **2,** (2016).

[20] Zhu, W., Perez, E. A., Hong, R., Li, Q. Xu, B. Age-related disparity in immediate prognosis of patients with triple-negative breast cancer: A population-based study from SEER cancer registries. *PLoS One* **10,** 1-15 (2015).

[21] Chen, H. L., Zhou, M. Q., Tian, W., Meng, K. X. He, H. F. Effect of age on breast cancer patient prognoses: A population-based study using the SEER 18 database. *PLoS One* **11,** 1-11 (2016).

[22] Ho-Yen, C., Bowen, R. L. Jones, J. L. Characterization of basal-like breast cancer: An update. *Diagnostic Histopathol.* **18,** 104-111 (2012).

[23] Stead, L. A. et al. Triple-negative breast cancers are increased in black women regardless of age or body mass index. *Breast Cancer Res.* **11,** 1-10 (2009).

[24] Bauer, K. R., Brown, M., Cress, R. D., Parise, C. A. Caggiano, V. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: A population-based study from the California Cancer Registry. *Cancer* **109,** 1721-1728 (2007).

[25] Warner, E. T. et al. Racial and ethnic differences in breast cancer survival: Mediating effect of tumor characteristics and sociodemographic and treatment factors. *J. Clin. Oncol.* **33,** 2254-2261 (2015).

[26] Bellavance, E. C. Kesmodel, S. B. Decision-Making in the surgical treatment of breast cancer: Factors influencing women?s choices for mastectomy and breast conserving surgery. *Front. Oncol.* **6,** 1-7 (2016).

[27] McCrate, F. et al. Surgical treatment choices for breast cancer in Newfoundland and Labrador: A retrospective cohort study. *Can. J. Surg.* **61,** 377-384 (2018).

[28] Ann Gilligan, M., Kneusel, R., Hoffmann, R., Greer, A. Nattinger, A. Persistent differences in sociodemographic determinants of breast conserving treatment despite overall increased adoption. *Med. Care* **40,** 181-189 (2002).

[29] Molenaar, S. et al. Predictors of patients' choices for breast-conserving therapy or mastectomy: A prospective study. *Br. J. Cancer* **90,** 2123-2130 (2004).

[30] Blichert-Toft M, Rose C, Andersen JA, et al. Danish randomized trial comparing breast conservation therapy with mastectomy: six years of life-table analysis. *J. Natl. Cancer Inst. Monogr.* **11,** 19-25 (1992).

[31] Fisher B. et al. Eight-year results of a randomized clinical trial comparing total mastectomy and lumpectomy with or without irradiation in the treatment of breast cancer. *N. Engl. J. Med.* **320,** 822-828 (1989).

[32] Sarrazin D. et al. Ten-year results of a randomized trial comparing a conservative treatment to mastectomy in early breast cancer. *Radiother Oncol..* **14,** 177-184 (1989).

[33] NIH consensus conference. Treatment of early-stage breast cancer. *The J. of the Amer. Med. Assoc.* **265,** 391-395 (1991).

[34] Whelan, T. J. et al. Long-Term results of hypofractionated radiation therapy for breast cancer. *N. Engl. J. Med.* **362,** 513-520 (2010).

[35] Fisher, B. et al. Lumpectomy compared with lumpectomy and radiation therapy for the treatment of intraductal breast cancer. *N. Engl. J. Med.* **328,** 1581-1586 (1993).

[36] Dubey, A., Gupta, U., Jain, S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. *Int J. CARS.* **11,** 2033-2047 (2016).

[37] Wang, C., Machiraju, R., Huang, K. Breast cancer patient stratification using a molecular regularized consensus clustering method. *Methods* **67,** 304-312 (2014).

[38] Bradley, P.S., Mangasarian, O.L. K-Plane clustering. *J. of Global Opt.* **16,** 23-32 (2000).

[39] Mann, H.B., Whitney D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18,** 50-60 (1947).

[40] Wilcoxon, F., Individual comparisons by ranking methods. *Biomet. Bull.* **1,** 80-3 (1945).

[41] Jolliffe, I.T. Principal component analysis (Springer, New York, 2002).

[42] Steinhaus, H., Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci.* **4,** 801-804 (1956).

[43] Lloyd, S., Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28,** 129-137 (1982).

[44] MacQueen, J., Some methods for classification and analysis of multivariate observations. *Fifth Berkeley Symposium on Mathematics, Statistics and Probability, University of California Press* **28,** 281-297 (1967).

[45] Jain, A.K., Data clustering: 50 years beyond k-means. *Pattern Recog. Letters* **31,** 651-666 (2010).

[46] James, G., Witten, D., Hastie, T. Tibshirani, R., An introduction to statistical learning. 388 (Springer, 2013).

[47] Riely, G.J., Marks, J., Pao, W., KRAS mutations in non-small cell lung cancer. *Proc. Am. Thorac. Soc.* **2,** 201-205 (2009).

[48] Raman, J.P. et al. A comparison of survival analysis methods for cancer gene expression RNA-Sequencing data.*Cancer Genetics* **235-236,** 1-12 (2019).

[49] Kaplan, E.L., Meier P. Non-parametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53,** 457-481 (1958).

[50] Mootha, V. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34,** 267-273 (2003).

[51] Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545-15550 (2005).

[52] Luo, W., Friedman, M., Shedden, K., Hankenson, K. Woolf, P. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10,** 161 (2009).

[53] Saxena, V., Orgill, D. Kohane, I. Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic Acids Research* **34,** e151-e151 (2006).

[54] Shahriyari, L. Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Brief. Bioinform.* **20,** 985-994 (2017).

[55] Pedregosa et al. Scikit-learn: Machine learning in Python. *J. of Machine Learning Research* **12,** 2825-2830 (2011).

[56] Cox, D., Regression models and life-tables. *J. Royal Stat. Soc. Ser B (Method)* **34,** 187-220 (1972).

[57] Valenzuela, S. et al. Molecular cloning and expression of a chloride ion channel of cell nuclei. *J. of Biological Chemistry*, **272,** 12575-12582 (1997).

[58] Lim, S. et al. Identification of the kinase STK25 as an upstream activator of LATS signaling. *Nature Communications*, **10,** (2019).

[59] Park, J. et al. Prolactin regulatory element-binding (PREB) protein regulates hepatic glucose homeostasis. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, **1864,** 2097-2107 (2018).

[60] Yan, Y. et al. Loss of polycomb group protein Pcgf1 severely compromises proper differentiation of embryonic stem cells. *Scientific Reports*, **7,** (2017).

[61] Webb, K., Lipson, R., Al-Hadid, Q., Whitelegge, J. Clarke, S. Identification of protein N-Terminal methyltransferases in yeast and humans. *Biochemistry*, **49,** 5225-5235 (2010).

[62] Brenner, V., Nyakatura, G., Rosenthal, A. Platzer, M. Genomic organization of two novel genes on human Xq28: Compact head to head arrangement of IDHγ and TRAPδ is conserved in rat and mouse. *Genomics*, **44,** 8-14 (1997).

[63] Li, A. et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Research* **69,** 2091-2099 (2009).

[64] Yu, J. et al. A Transcriptional fingerprint of estrogen in human breast cancer predicts patient survival. *Neoplasia* **10,** 79-88 (2008).

[65] Thongkam, J., Xu, G., Zhang, Y. Huang, F. Proceedings of the second Australasian workshop on health data and knowledge management - Volume 80. 55-64 (Australian Computer Society, Inc., 2008). Available at https://dl.acm.org/doi/pdf/10.5555/1385089.1385098?download=true (Accessed: 17th February 2020)

[66] Griffith, O. et al. A robust prognostic signature for hormone-positive node-negative breast cancer. *Genome Medicine* **5,** 92 (2013).

[67] Jones, C. et al. Expression profiling of purified normal human luminal and myoepithelial breast cells. *Cancer Research* **64,** 3037-3045 (2004).

[68] Agarwal, S. et al. Effect of breast conservation therapy vs mastectomy on disease-specific survival for early-stage breast cancer. *JAMA Surgery* **149,** 267 (2014).

[69] Poggi, M. M. et al. Eighteen-year results in the treatment of early breast carcinoma with mastectomy versus breast conservation therapy: The National Cancer Institute randomized trial. *Cancer* **98,** 697-702 (2003).

[70] Mersin, H. et al. Prognostic factors affecting postmastectomy locoregional recurrence in patients with early breast cancer: Are intrinsic subtypes effective? *World J. Surg.* **35,** 2196-2202 (2011).

[71] Laurberg, T. et al. Impact of age, intrinsic subtype and local treatment on long-term local-regional recurrence and breast cancer mortality among low-risk breast cancer patients. *Acta Oncol.* (Madr). **56,** 59-67 (2017).

[72] Arvold, N. D. et al. Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy. *J. Clin. Oncol.* **29,** 3885-3891 (2011).

[73] Kim, S. H. et al. Role of secreted type I collagen derived from stromal cells in two breast cancer cell lines. *Oncol. Lett.* **8,** 507-512 (2014).

[74] Lin, J., Goldstein, L., Nesbit, A. Chen, M. Y. Influence of hormone receptor status on spinal metastatic lesions in patients with breast cancer. *World Neurosurg.* **85,** 42-48 (2016).

[75] Sengupta, P.K. et al. DNA hypermethylation near the transcription start site of collagen alpha2(I) gene occurs in both cancer cell lines and primary colorectal cancers. *Cancer research* **63**, 1789–1797 (2003).

[76] Guo, C., Liu, S., Wang, J., Sun, M. Z. Greenaway, F. T. ACTB in cancer. *Clin. Chim. Acta* **417,** 39-44 (2013).

[77] Ferreira, E. Cronjé, M. J. Selection of suitable reference genes for quantitative real-time PCR in apoptosis-induced MCF-7 breast cancer cells. *Mol. Biotechnol.* **50,** 121-128 (2012).

[78] Lin, C., Beattie, A., Baradaran, B., Dray, E. Duijf, P. Contradictory mRNA and protein misexpression of EEF1A1 in ductal breast carcinoma due to cell cycle regulation and cellular stress. *Scientific Reports* **8,** (2018).

[79] MeSH Browser (2020). Available at: https://meshb.nlm.nih.gov/record/ui?ui=D009243 (Accessed: 11th April 2020)

[80] Li, L. et al. Down-Regulation of NDUFB9 promotes breast cancer cell proliferation, metastasis by mediating mitochondrial metabolism. *PLOS ONE* **10,** e0144441 (2015).

[81] Ko, J. et al. Expression profiling of ion channel genes predicts clinical outcome in breast cancer. *Molecular Cancer* **12,** 106 (2013).

[82] Weber, L. et al. Olfactory receptors as biomarkers in human breast carcinoma tissues. *Frontiers in Oncol.* **8,** (2018).

[83] Grambsch, P. Therneau, T. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81,** 515-526 (1994).

[84] Heitman, S., Au, F., Manns, B., Pattison, P. Hilsden, R. Nonmedical costs of colorectal cancer screening using CT colonography. *J. Of The American College Of Radiology*, **7,** 943-948 (2010).

[85] Peipins, L. et al. Time and distance barriers to mammography facilities in the Atlanta Metropolitan Area. *J. Of Community Health*, **36,**, 675-683 (2011).

[86] Chihuri, S. et al. Driving cessation and health outcomes in older adults. *J. Of The American Geriatrics Society*, **64,** 332-341 (2016).

[87] Dickerson, A. et al. Transportation and aging: An updated research agenda for advancing safe mobility. *J. Of Applied Gerontology*, **38,** 1643-1660 (2017).

[88] Syed, S., Gerber, B. Sharp, L. Traveling towards disease: Transportation barriers to health care access. *J. Of Community Health*, **38,** 976-993 (2013).

[89] Heckman, T. et al. Barriers to care among persons living with HIV/AIDS in urban and rural areas. *AIDS Care*, **10,** 365-375 (1998).

[90] Martin, M. et al. What do cancer patients worry about when making decisions about treatment? Variation across racial/ethnic groups. *Supportive Care In Cancer*, **22,** 233-244 (2014).

[91] Ramondetta, L. et al. Advanced cervical cancer treatment in Harris County: Pilot evaluation of factors that prevent optimal therapy. *Gynecologic Oncol.*, **103,** 547-553 (2006).

[92] Massa, S. et al. An assessment of patient burdens from head and neck cancer survivorship care. *Oral Oncol.*, **82,** 115-121 (2018).

[93] Ngoc Thi Dang, D., Ngoc Thi Nguyen, L., Thi Dang, N., Quang Dang, H. Ta, T. Quality of life in Vietnamese gastric cancer patients. *Biomed Research International*, **2019,** 1-9 (2019).

[94] Dehkordi, A., Heydarnejad, M. Fatehi, D. Quality of life in cancer patients undergoing chemotherapy. *Oman Medical J.*, **24,** 204–207 (2011).

[95] Richardson, L., Wang, W., Hartzema, A. Wagner, S. The role of health-related quality of life in early discontinuation of chemotherapy for breast cancer. *The Breast J.*, **13,** 581-587 (2007).

[96] Detmar, S., Muller, M., Schornagel, J., Wever, L. Aaronson, N. Health-related quality-of-life assessments and patient-physician communication. *JAMA*, **288,**, 3027 (2002).

[97] Silvestri, G., Pritchard, R. Welch, H. Preferences for chemotherapy in patients with advanced non-small cell lung cancer: descriptive study based on scripted interviews. *BMJ*, **317,** 771-775 (1998)

[98] Gordon, A., Breiman, L., Friedman, J., Olshen, R. Stone, C. Classification and regression trees. *Biometrics*, **40,** 874 (1984).

[99] Hothorn, T., Hornik, K. Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. *J. of Computational and Graphical statistics*, **15,** 651–674 (2006).

[100] Plotly Python Graphing Library. Plotly.com. (2020). Available at: https://plotly.com/python/. (Accessed: 24th April 2020)

[101] Spearman, C. The proof and measurement of association between two things. The American Journal of Psychology *Am. J. Psychol.*, **15,** 72–101 (1904).

[102] Geurts, P., Ernst, D. Wehenkel, L. Extremely randomized trees. *Machine Learning*, **63,** 3-42 (2006).

[103] Kern, C., Klausch, T., Kreuter, F. Tree-based machine learning methods for survey research. *Survey Research Methods,* **13,** 73-93 (2019).

[104] Lohr, S., Hsu, V. Montaquila, J. Using classification and regression trees to model survey nonresponse (2015). Available at: https://www.semanticscholar.org/paper/Using-Classification-and-Regression-Trees-to-Model-Lohr-Hsu/e3d8bbeaf70d292c367ba31ab9e5199ef02c57df. (Accessed: 10th April 2020)

[105] Buskirk, T. Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Survey Practice*, **18,** (2018).

[106] Breiman, L. Random forests. *Machine Learning*, **45,** 5–32 (2001).

[107] Flores, G., Abreu, M., Olivar, M. Kastner, B. Access barriers to health care for Latino children. *Archives of Pediatrics Adolescent Medicine*, **152,** (1998).