ADVANCED ANALYTICS TO PREDICT SURVIVABILITY OF BREAST CANCER PATIENTS

by

Sonal Bajaj

B.Tech, Uttar Pradesh Technical University, UP, India, 2014

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER SCIENCE

UNIVERSITY OF NORTHERN BRITISH COLUMBIA

November 2018

Abstract

Cancer is a significant burden of disease worldwide. Amongst women, breast cancer is the most common cancer and the primary cause of death of women followed by heart diseases. With increasing breast cancer cases and technological improvements, large volumes of data related to breast cancer are collected every year around the globe. This historical data is a vast source of knowledge, and when extracted, this knowledge could be used in making decisions in the future. Descriptive analytics uncovers hidden patterns and trends and provides insights into the past to answer "What has happened?". Predictive analytics uses different modeling techniques on historical data to predict future medical outcomes and answer "What could happen?".

Cancer care institutions and registries have collected large volumes of cancer data in various formats. Unfortunately, these repositories are not easily accessible, and the stored formats are difficult to analyze. The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program is a premier source for cancer statistics in the United States. Although the data is accessible, it lacks consistency; generating reports from such data is a labour-intensive process. An end-to-end process is proposed through which such data can be cleansed, integrated and presented in the form of interactive dashboards with drill-down and drill-through reporting capabilities. This provides a comprehensive view of over forty years of data consisting of over one million records with provisions to slice this data along several dimensions. The underlying patterns and trends could be utilized in improving treatment plans, data-driven resource allocation, and better patient care. The dashboard would be extensible, scalable, and update in real-time with new data.

Additionally, developing a breast cancer predictive model that predicts survival months for diagnosed patients is proposed. The cleansed and pre-processed data from the analysis is used for creating data subsets, which in turn, trains the predictive model. The outcomes of different modeling techniques along with assessing the impacts of retraining the predictive model are observed in the experimentations conducted for this research.

Acknowledgement

I would first like to thank my supervisor Dr. Waqar Haque for his constant support, guidance and motivation. The door to Dr. Haque's office was always open whenever I ran into trouble or had a question about my research. It would never have been possible for me to take this work to completion without his constant support and generosity.

Next, I would like to thank my supervisory committee Dr. Alex Aravind and Dr. Pranesh Kumar for their support and direction in my research. I would also like to thank Dr. Robert Olson for his suggestions towards this research and my colleagues at Northern Health for supporting and motivating me at all times.

Lastly, I would like to express my gratitude to my family and friends for providing me with unfailing support and continuous encouragement throughout my years of study. A huge shout out to my elder brother Sameer Bajaj for being my pillar of strength and source of inspiration throughout this journey. This accomplishment would not have been possible without them. Thank you.

Sonal Bajaj

Contents

List of Figures

List of Tables

List of Equations

Acronyms

ADP	Automated Data Preparation
ANN	Artificial Neural Network
AUC	Area Under Curve
BI	Business Intelligence
BN	Bayesian Network
BOSOM	Breast Cancer Outcome-Survival Online Measurement
C&RT	Classification and Regression Tree
CDC	Centers for Disease Control and Prevention
CHAID	Chi-squared Automatic Interaction Detection
CI5	Cancer Incidence in Five Continents
COD	Cause of Death
COPE	Corporate Oncology Program for Employees
CSU	Section of Cancer Surveillance
DAVE	Data Analysis, Visualization, and Exploration
DCCPS	Division of Cancer Control and Population Sciences
EMR	Electronic Medical Records
GBD	Global Burden Disease
GCO	Global Cancer Observatory
GDC	Genomic Data Commons
IARC	International Agency for Research on Cancer
IHME	Institute of Health Metrics and Evaluation
IICC	International Incidence of Childhood Cancer
KDD	Knowledge Data Discovery
LCOC	Lung Cancer Outcome Calculator
MLP	Multilayer Perceptron

NCI	National Cancer Institute
NPCR	National Program of Cancer Registries
QUEST	Quick, Unbiased, Efficient Statistical Tree
RBF	Radial Basis Function
SEER	Surveillance, Epidemiology, and End Results
SPR	Surveillance Research Program
STR	Survival Time Recode
SVM	Support Vector Machine
VSR	Vital Status Recode
WDBC	Wisconsin Diagnostic Breast Cancer Database
WHO	World Health Organization

1. Introduction

Cancer is generally referred to as a large group of diseases that can affect any part of the human body. It is the uncontrolled growth of cells, which can invade one site or spreads to many sites within the body [1]. A constant increase in the number of cancer cases has been reported over the past decade globally. According to the American Cancer Society, there are over 1.6 million new cases, and over half a million cancer-related deaths were estimated to have been reported in the United States in 2016 alone [2]. Hence it is of no surprise that cancer is the second most leading cause of deaths in the United States. Cancer is a leading cause of deaths in Canada as well, accounting for 30% of all deaths according to the Canadian Cancer Society [3].

There exist more than 100 types of cancer with different symptoms and treatment. Breast cancer is one of the most common cancer among women. However, due to technological advancements and increasing cancer-related research, many new early detection methods and treatments have been developed which have helped to decrease cancer-related deaths [4]. However, cancer in general and breast cancer, in particular, is still a significant cause of concern [5]. Breast cancer accounted for about 25% of all the new cancer cases among women in Canada in 2017, with an estimated total count over 26,000 for the year [3]. The numbers are on the rise globally, for both incidence and mortality due to breast cancer.

“Breast cancer starts when cells in the breast begin to grow out of control. These cells usually form a tumor that can often be seen on an x-ray or felt as a lump” [6]. Breast cancer is most common in women, but it is also possible in men. The tumor is considered malignant if the cancer cells multiply and start affecting the surrounding cells or metastasize to other parts of the body such as the liver, lung, bone or brain. The tumor is benign or non-malignant when there are abnormal growths, but do not invade other cells outside of the breast. Although non-malignant tumors are considered not life-threatening, some benign breast tumors can increase a woman's risk of getting a malignant tumor in breast [6].

Even though breast cancer research is clinical or biological [5], the data-driven research outcomes can be valuable and a significant step forward in cancer treatment. Dr. Robert Stein, UCL breast cancer consultant, states that the one-size-fits-all basis is the common practice in all areas of cancer treatment [7]. However, tools like IBM Watson for Oncology [8] and SAP’s Corporate Oncology Program for Employees (COPE) [9] help physicians provide better cancer care. SAP’s COPE is corporate oncology program run by SAP for its employees in Germany, U.S and Canada. The program offers treatment cost coverage of tumor analysis and helps physicians find safe and effective treatment for each patient. IBM Watson [8] for oncology is trained to design cancer treatments. Over time, Watson has performed very well at recommending a treatment plan for different types of cancer. The clinicians continue to add to the Watson cancer assessing repository [10]. By selecting treatment plan based on genetic changes and evidence by understanding the data points, Watson, COPE and other tools have brought a paradigm shift in cancer care in today’s world.

With such new tools and research aid, breast cancer care can be improved further. Predicting survival of breast cancer patient can provide physicians help in developing the treatment plan specific to each patient. Different modeling techniques that can be used to develop such predictive model. One hypothesis here is that Ensemble modeling technique can be more accurate than other modeling techniques when designing the breast cancer predictive model. Ensemble modeling is the process of combining two or more modeling techniques and score the combined results by using voting or averaging techniques.

1.1 Background

One of the most frequently asked questions by cancer patients post-diagnosis is the lifespan they are left with. To predict how long a cancer patient will live is a tough question for oncologists to answer. Oncologists answer to such questions are based on past records of cancer patients with similar prognosis or by consulting other physicians and researchers working on comparable cases. Although careful prognosis is vital, it is difficult to find accurate survival time of patients because survivability is based on many factors [11]. Also, these predictions may not be absolute as the past records are not entirely reliable and the prognosis from different oncologists are generally inconsistent [12].

1.1.1 Common Breast Cancer Research Methods

The common breast cancer research methods include experimental, observational and clinical trials. The experimental method involves new medication, treatment plan or new treatment aid introduced to a new set of people. The results of a new intervention are then compared with another set of people (control group) who are not exposed to the new intervention being tested. The control group and other group's members are selected randomly by the researchers. Experimental research methods help in testing the new techniques and learning how cancer starts or metastasizes. The second common research method is observational, which involves observing a set of people in a natural environment to determine factors associated with a specific outcome [13]. As a result, observational studies can establish the association of the variables to the outcome. Another common breast cancer research method is clinical trials.

Clinical trials are medical experiments performed on humans. The development of new procedures and drugs are usually developed based on clinical trials [4, 14].

In this research, observational research methodology is used. The historical data records of cancer patients are used to understand the patterns, trends (data analysis) and factors involved in the disease outcomes (data mining) and predict the outcome for new patients (predictive analytics). The purpose of this research is twofold: to develop a breast cancer dashboard and build a predictive model by utilizing the relationship between a set of independent variables and the survival months.

1.1.2 Medical Prognosis and Survival Analysis

Predicting the outcome of a disease is one of the most challenging tasks for researchers [5]. In cancer treatment survival is considered as the most important outcome [15]. “Survival analysis is a study of time between entry into observation and a subsequent event” [16]. In today’s world, scientists use survival analysis for not just commencement time of disease but also for the time before the stock market crash, time until weather changes or equipment failure, the time before natural calamities like flood, and earthquake [16]. For cancer, the essential event of interest is “death”. Other events include relapse of disease, recovery from disease or death. Survival analysis tools are also used for leukemia patient readmission time, time for an average person to develop a heart disease, the time before death for elderly population, and so on [17].

In medical prognosis “Survival analysis” is referred to a field which uses different methods and techniques on collected historical data, to predict the patient’s survival from a disease over a specific period of time [5]. With Electronic Medical Records (EMR) systems development, storing patient’s history, test results, diagnoses and other relevant facts has become easy and manageable. Moreover, such open source data source acts a tremendous resource for researchers who want to develop survivability prediction models. Data analysis and knowledge discovery research techniques are used by researchers to predict the outcome of disease by identifying the patterns and relationship between different variables of historical data [5].

1.1.3 Knowledge Data Discovery and Data Mining

Historical data from cancer patients' medical records are a powerful source of information. It helps oncologists and researchers find the grounds for inter-relationships of present to historical cases [18]. Using historical data to predict outcomes in breast cancer could be dated back to 1992 where neural network analysis was used to predict the recurrence of breast cancer [19]. However, with no specific global standard to record the patient data, a vast inconsistency is often observed across the data available globally. Despite this inconsistency, these records remain invaluable medical literature.

Knowledge Data Discovery (KDD) is a significant process of extracting knowledge from raw data. KDD is defined as a step-by-step process of understanding the realm, preparation of data, followed by the collection and formulation of knowledge from extracted patterns. The ''post-processing of the knowledge'' then can be applied to capture the knowledge from a large amount of recorded data [20]. In KDD, data mining is the step of collecting and formulating knowledge from data using different pattern extraction methods. The knowledge discovery process is an essential part of medical data mining [21].

Data mining has influenced many fields such as medicine, media, astronomy, business, marketing, investment, manufacturing, and telecommunications. In today's digital world, large volumes of data are produced every day, and manual data analysis is an impractical approach.

Thus an automated data analysis is the need of the hour. Data mining is the attempt to address a problem of the digital era, i.e. data overload [22]. Data mining uses different algorithms to extract useful patterns from data [23] and then use the patterns to build predictive models. There are two major data mining tasks, grouped as descriptive data mining and predictive data mining. The descriptive data mining includes association, clustering and summarization tasks.

On the other hand, predictive data mining tasks include classification, prediction, and time-series analysis tasks [24]. Descriptive data mining tasks describe the significant properties of the actual data and predictive data mining tasks aim to do predictions based on existing data [25]. In this research, descriptive and predictive data mining tasks are used.

1.1.4 Data Visualization

A powerful option to make complex data usable and relevant is through data visualization. The goal of interactive data visualization is to display data in forms of visuals to help the user understand the data quickly, identify the areas of improvement and take decisions accurately based on historical events or data [26]. On a broader scale, there are two goals of data visualization – explanatory i.e. to explore data to solve a particular problem, or exploratory, i.e. to explore large sets of data for enhancing understanding of data and finding crucial missing information. Converting structured data into meaningful charts and graphical depictions enable the users to gain insights into all the captured data [27]. As the healthcare sector is rapidly moving towards data analytics and has started relying on digital information to improve health care and reduce the costs, data analysis and visualization of data has become a core component. Many health organizations also have online visualization tools available for public to explore the trends of incidence, mortality, demographics and other statistics. Institute of Health Metrics and Evaluation's "GBD (Global Burden Disease) Compare" is an interactive analytics and visualization tool available online [28]. The tools allow users to visualize and compare disease causes and risks in the form of treemaps, maps, diagrams, charts within the region (inter-country or intra-country) or worldwide [28]. There are some online visualization tools and dashboards available for cancer data also. The National Program of Cancer Registries (NPCR) has the United States Cancer Data Visualization online tool [29] available which fetches data from National Cancer Institute (NCI) and Centres for Disease Control and Prevention's CDC for all cancer types for the year 2010 - 2014. The visualization tools have U.S Cancer demographics represented graphically by cancer incidence rate, mortality rate and type of cancer. World Health Organization's Global Cancer Observatory (GCO) is a web-based platform which provides global cancer statistics [30].

1.1.5 Past Research

Many former researchers have established the capacity of data mining by the medium of application to medical records [18]. For instance, Agrawal et al.'s Lung Cancer Outcome Calculator [5] is a survival prediction model for lung cancer patients using different data mining techniques and SEER data [31]. It can predict survival of lung cancer patients from 6 months, 9 months, 1 year, 2 years and 5 years of diagnosis [5]. Zorluoglu et al. used Wisconsin Diagnostic Breast Cancer Database (WDBC) for similar work to predict whether a breast cancer tumor is malignant or benign [32]. Several other prognostic applications, using different tools and algorithms, also exist that can predict breast cancer survivability, for example, Adjuvant [33], PREDICT [34].

1.2 Problem Statement

Cancer care institutions and registries have collected large volumes of cancer data in various formats. Unfortunately, these repositories are not easily accessible, and the stored formats are difficult to analyze. The National Cancer Institute's SEER Program is one of the organizations which maintains cancer statistics of the US population. Though the data is accessible, it lacks consistency and generating reports from such data is a labour intensive process.

An end-to-end process is proposed through which this data is cleansed, integrated and presented in the form of interactive dashboards with drill-down and drill-through reporting capabilities. This provides a coherent view of the data and allows users to observe hidden patterns and trends which could be utilized towards improving treatment plans, data-driven resource allocation and better patient care. The top-level dashboard presents the main KPIs and is supported by a sequence of visualizations to convey information which can be sliced and diced along several dimensions. This real-time dashboard can be updated as soon as new data is uploaded to the database.

Further by using the pre-processed and cleansed data, a breast cancer predictive model is proposed that predicts survival months of a breast cancer patient from the time of diagnosis. The predictive model is trained, tested and validated with different subsets. The predictor's selection is based on main KPIs identified by analysis along with an expert's opinion from a total of 134 variables available in SEER. Different data mining is used, and the predictive models will be designed based on data mining technique which performs best amongst all, along with the ensemble of selected techniques.

1.3 Motivation of the Research

"Time is shortening. But every day that I challenge this cancer and survive is victory for me" – Ingrid Bergman (Cancer patient)

The traditional cancer survival prediction methods such as research on past experiences using spreadsheets require an unacceptably large amount of time and effort to study the data. The classic ways of identifying survival time of cancer patients include a comparison study between the patient's health situation and symptoms with previously recorded cancer patient's medical records, statistical computation of survival rates based on historical records, or consulting another breast cancer expert.

"Cancer prognosis is the doctor's best estimate of how cancer will affect person" [3].

Many factors that can affect a person's prognosis. Survival statistics is one of the methodologies that physicians use to develop a prognosis for a person with cancer. Researchers, when developing a prognosis, often look at studies that measure survival for one specific cancer type, stage or risk group. The survival rate is the percentage of people with cancer who are alive at some point in time (i.e. 1, 3, 5 or 10 years) after their diagnosis [3]. Survival prediction is the process of finding the time left for a patient to live. It is generally associated with diseases, which have high mortality rate such as cancer. Survival prediction is part of physician's prognostic investigation, where a result takes the form of a numerical percentage of survival over a period that depends on a factor such as tumor size, time after diagnosis and stage of cancer [18]. Table 1 shows global survivability rates of breast cancer for the year 2016.

Time since Diagnosis	Survival Rate
5-year	89%
10-year	83%
15-year	78%

Table 1. Breast Cancer Survival Statistics [2]

All cancer types have high mortality rates [35] and cancer patients are always frenzied to know how much time they are left with. Despite many treatment options available for cancer patients, there is no assurance that the patient will be cured after treatment. Each patient responds to the treatment differently. Physicians estimate the prognosis of cancer by using statistics collected by researchers over many years. Various statistics are used to estimate cancer prognosis [36], most commonly: Cancer-specific survival, Relative survival, Overall survival and Disease-free survival. These statistical survival prediction methodologies are time-consuming and lack accuracy. By applying data mining techniques to breast cancer data, a breast cancer survival prediction model is built. Data mining techniques help rank and link cancer attributes to survival outcome [5]. The outcome of the breast cancer survival predictive model will help both physicians and patients to determine accurate survivability, serving as a reference for patients and provide them with a second opinion [18]. It can also assist physicians in decision-making to determine the best treatment for breast cancer patient. The modeling technique can predict outcome depending on patient-specific attributes instead of relying on personal experiences or time-consuming statistical evaluation. The combination of breast cancer effects across the different age range, the promising results and benefits of data-driven research in healthcare, and the desire to contribute towards improving healthcare and treatment of breast cancer have together motivated this research.

1.4 Research Methodology

The main steps involved in this research are:

- Literature review
- Determining relevant data source
- Consultation with medical personnel/oncologist to shortlist the relevant variables
- Design data analysis dashboard
- Determine relevant tools and modeling techniques
- Develop a predictive model
- Training, testing and validation of the predictive model

1.5 Contributions

This research has two primary contributions. The first contribution focuses on data visualization for breast cancer data of over 40 years. This is achieved by developing a dashboard built from breast cancer patients' data to uncover hidden patterns as well as provide easy-to-understand metrics for users of all backgrounds. The dashboard includes breast cancer data for patient population demographics, patient volumes, diagnosis and treatment.

The second contribution focuses on building a predictive model to predict breast cancer patient survivability. This model is built from the preprocessed data extracted from the SEER database. The model could be used by doctors to determine their predicted survival time ranging in months for their patients diagnosed with breast cancer. The model is trained with the existing data and processes the given breast cancer-related attributes and predict survival months [18]. This research will contribute towards the healthcare field in following ways:

1. **Increase accuracy of the diagnoses:** Predictive algorithms help physicians input the patient's clinical symptoms and get more accurate diagnosis thereby assisting their judgements [37]. The treatment plan of patients can then be enriched with predictive analytics results such as survival months.

2. **Provide physicians with answers they are seeking for individual patients:** There are possibilities where a treatment plan works best for a set of patients, but may or may not work for another individual patient. Predictive analytics can help physicians plan the treatment specific to the patient. The breast cancer treatment plan can include a combination of different treatments such as surgery, radiation therapy, chemotherapy, hormonal therapy and targeted therapies. An ideal treatment plan should work against all things inside the cells that caused cancer to develop, grow, and possibly spread to other parts of the body [38].

3. **Increase the patients' understanding and participation:** The outcomes of predictive analytics like survivability and possible health risk indicators can help the patients' educate themselves about their disease. An educated patient can take more responsibility of their treatment plan. The patients can then equally participate with their physicians to help make decisions on their treatment plan.

4. **Analyze massive healthcare data:** In current scenarios, the massive amount of data generated by the healthcare industry is digitized for ease. It is a tedious work to make sense from such massive digital data. The analysis of healthcare data will help by providing actionable insights to both physicians and healthcare industries regarding their

planning, administration and assessment [37]. It thereby augments the decision-making ability of the administration, by evaluating the critical opportunities, such as quality care improvement or patient injury prevention, and allocating resources to the fundamental processes [39].

1.6 Organization of Thesis

This thesis consists of five chapters. In this chapter, the background, problem statement, the motivation and contributions of this research has been provided. In Chapter 2, the related work, challenges and different research approaches are discussed. Chapter 3, extensively discusses the implementation steps involved in developing the breast cancer dashboard and the predictive model. Also, data mining and modeling techniques are formally introduced in this chapter. Chapter 4, provides experimentation and analysis of results. Chapter 5 provides the conclusion and the future directions to extend the work done in this research.

2. Literature Review

To better understand the research questions raised and addressed in the problem statement, an extensive literature review is conducted. For this purposes, traditional/narrative literature review methodology is used. It involves critiquing and reviewing existing work and deduce a conclusion about the research questions raised. This review type comes handy in collecting a volume of related work in a specific research area and then summarize it by highlighting the current techniques and approaches used. By finding gaps or disparity in the literature, researchers can determine or define a new research approach or hypotheses.

This process is utilized to understand the nuances of the approaches used in the literature specifically concerning two areas – data analysis and visualization of existing cancer data, and various predictive modeling techniques to determine the survivability of cancer patients. In this chapter, the observations in the two corresponding sections are presented.

2.1 Cancer Data Analysis and Visualization

Over the past few years, several studies have been conducted focusing on the analysis of data related to multiple types of diseases including breast cancer [28, 30, 40, 41]. More recently, several health organizations have developed online analytics and visualization tools for cancer and other diseases from their data repositories. Some of these tools available are discussed in the following sections:

2.1.1 Global Burden of Disease Compare

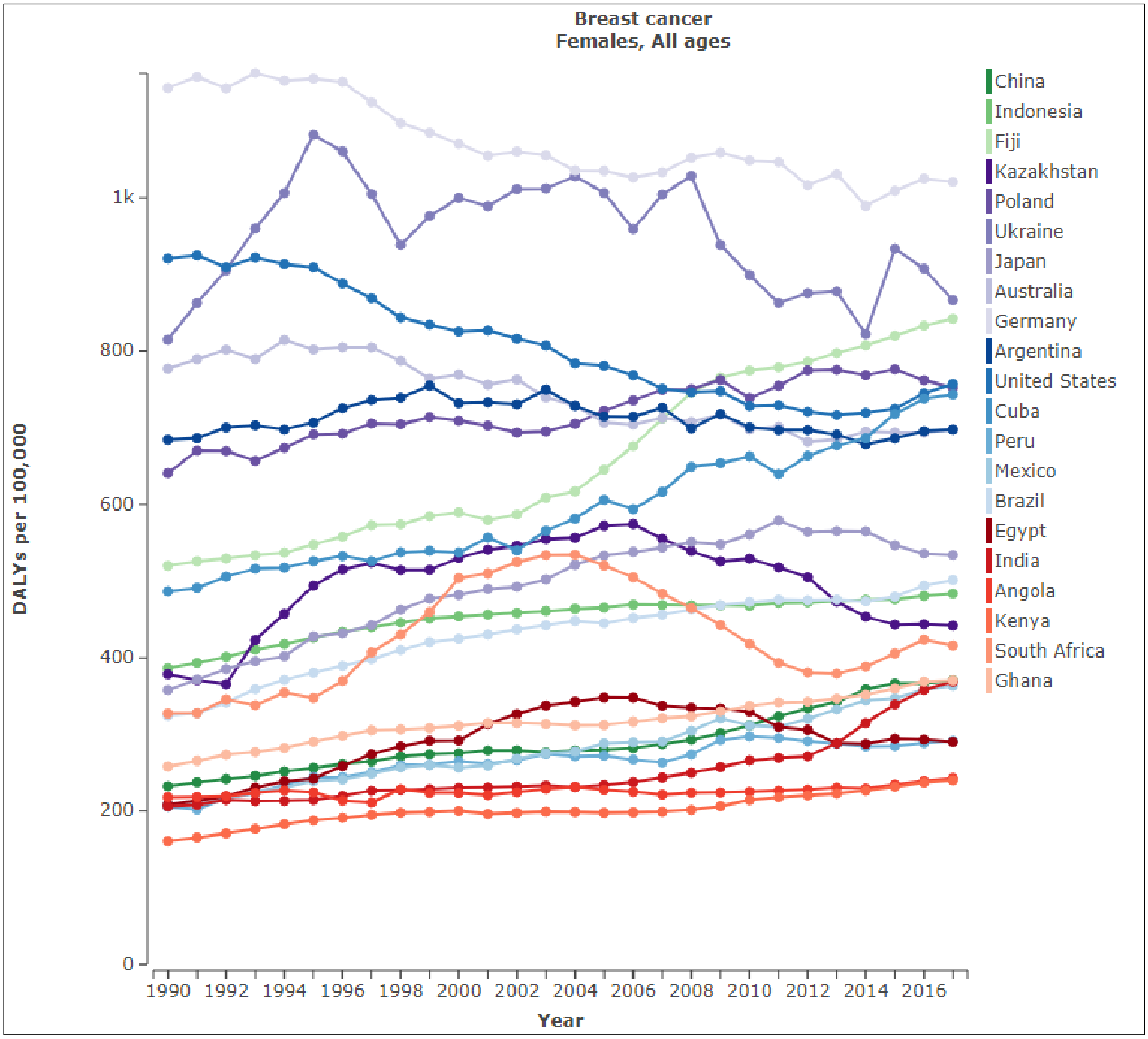


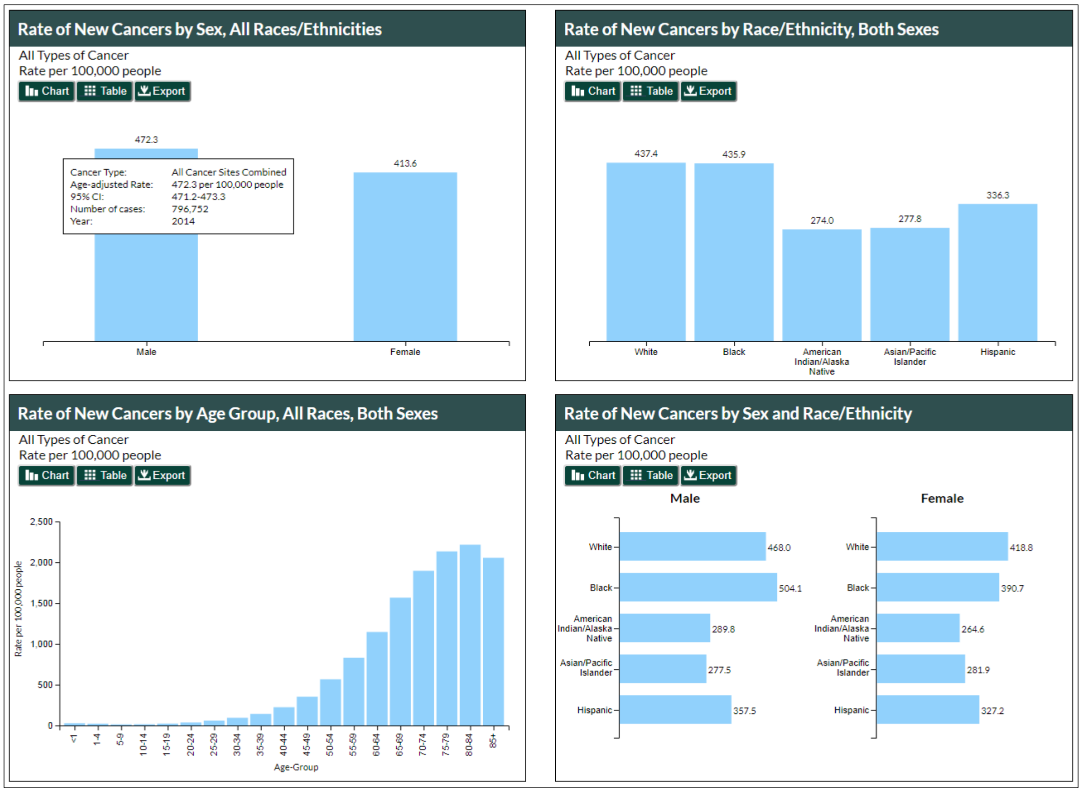
Figure 1. Screenshot of GBD Compare Breast Cancer Incidence Visualization [28]

The Institute of Health Metrics and Evaluation (IHME) [42] at the University of Washington measures, compares and evaluates strategies for various health issues, diseases, injuries and risk factors, around the globe. IHME's tool, GBD [28] evaluates global health challenges and risk factors so that health systems can be aligned with the disease trends. The tool analyzes data from 1990-2016 and provides a comparison of the effects of different diseases on a set of population. The policymakers can thus make more informed decisions with respect to the allocation of resources for better health care. The data related to premature deaths, disabilities, and injury is collected from over 130 countries and can be visualized along several dimensions including demographics, mortality, disease causes and risk factors [28]. The visualization is available in different formats such as map, treemap, line chart, patterns bar chart, pyramid chart, arrow chart and heat map. The dashboard can be drilled down to specific countries and states. The tool has three main tabs- single (single chart type), explore (map and one additional chart) and compare (two of the same chart by year, age, sex, cause of disease, risk and location).

Figure 1 is a screenshot of line graph visualization from 'compare by cause' tab where the cause is selected as breast cancer. The colour-coded lines show breast cancer rate and trend for selected countries. The same website also provides links to other visualization projects such as Mortality, Cause of Death (COD), Epidemiological (Epi), and Financing Global Health. [43].

2.1.2 U.S. Cancer Statistics Data Visualizations Tool

The Centers for Disease Control and Prevention (CDC) [44] and the National Cancer Institute (NCI) [45] collects cancer data from hospitals, physicians, clinics and health labs all over the U.S. and have made this data available through a visualization tool [40]. CDC recommends professionals like planners, policymakers, health advisors, researchers, and journalists to use this information to view and report cancer statistics [44]. The dashboard has multiple tabs and dropdowns for cancer types, historical trends, incidence, mortality rate, gender, age and demographics. While the data includes cases registered in the year 2010-2014, nation-wide changes in rates are available for the period 2006-2014.



Cancer Type: All Cancer Sites Combined
Age-adjusted Rate: 472.3 per 100,000 people
95% CI: 471.2-473.3
Number of cases: 796,752
Year: 2014

Figure 2. Screenshot of CDC Visualization tool [40]

Additional functionality includes geographical distribution displayed on an interactive map together with comparative numbers for all states [40]. Figure 2 is a screenshot of CDC visualization from demographics tab showing the rate of new cancer by sex, age group, race/ethnicity for all types of cancer in the United States (2015).

2.1.3 Global Cancer Observatory

World Health Organization's (WHO) [46] GCO [30] is a web-based visualization tool for global cancer statistics. The data presented is gathered from different projects of International Agency for Research on Cancer (IARC) Section of Cancer Surveillance (CSU) [47] including GLOBOCAN [48], Cancer Incidence in Five Continents (CI5) [49], International

Incidence of Childhood Cancer (IICC) [50], and Cancer Survival in Africa, Asia, the Caribbean and Central America [51]. The dashboard has four main tabs – Cancer Today, Cancer over Time, Cancer Tomorrow and Cancer Causes. The 'Cancer Today' tab presents incidence, mortality, and types of cancer estimates for 184 countries, broken down by age group and gender. The 'Cancer over Time' tab shows trends of cancer incidence and mortality for the last 50 years for 40 countries. The 'Cancer Tomorrow' tab provides visualization of cancer prediction up to the year 2035 by country and cancer type. Finally, 'Cancer Causes' tab highlights the causes of cancer, and the vital contributing risk factors [30]. Figure 3 is a screenshot of GCO visualization from 'Cancer Today' tab displaying estimated top 10 cancer incidence (cases) in the year 2018.

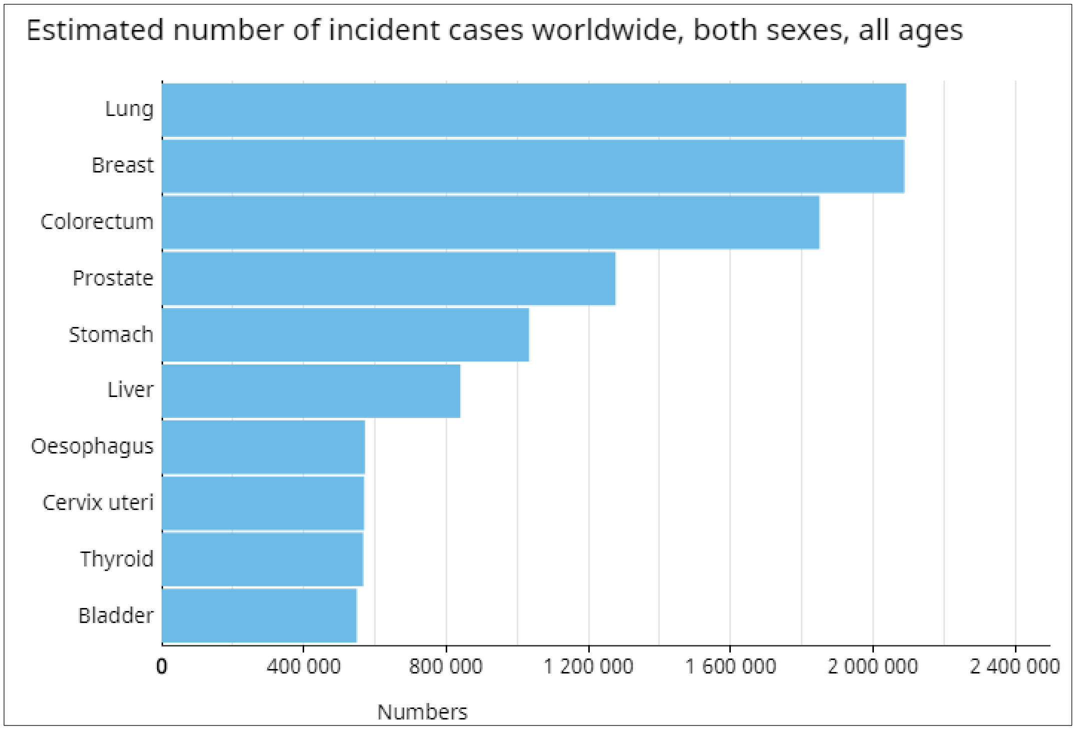


Figure 3. Screenshot of GCO Visualization [30]

2.1.4 Genomic Data Commons DAVE Tools

The National Cancer Institute's [45] Genomic Data Commons (GDC) [52] provides access to standardized clinical and genomic data. GDC also includes data from, The Cancer Genome Atlas (TCGA) [53] and Therapeutically Applicable Research to Generate Effective Therapies (TARGET) [54]. The GDC Data Analysis, Visualization, and Exploration (GDC DAVE) Tools [41] provides cancer supporting gene and variant level analysis of GDC data. These tools provide researchers' ability to visualize gene data with high impact mutations, most frequently mutated genes, survival analysis of different cases, and graphical visualization of cancer gene mutations. The data from each analysis can be visualized in bar charts, graph plots, trend lines and tabular format, along with download functionality. Figure 4 is a screenshot from GDC DAVE Tool visualization, showing a distribution of most frequently mutated genes. This tool can be accessed by using the GDC Data Portal.

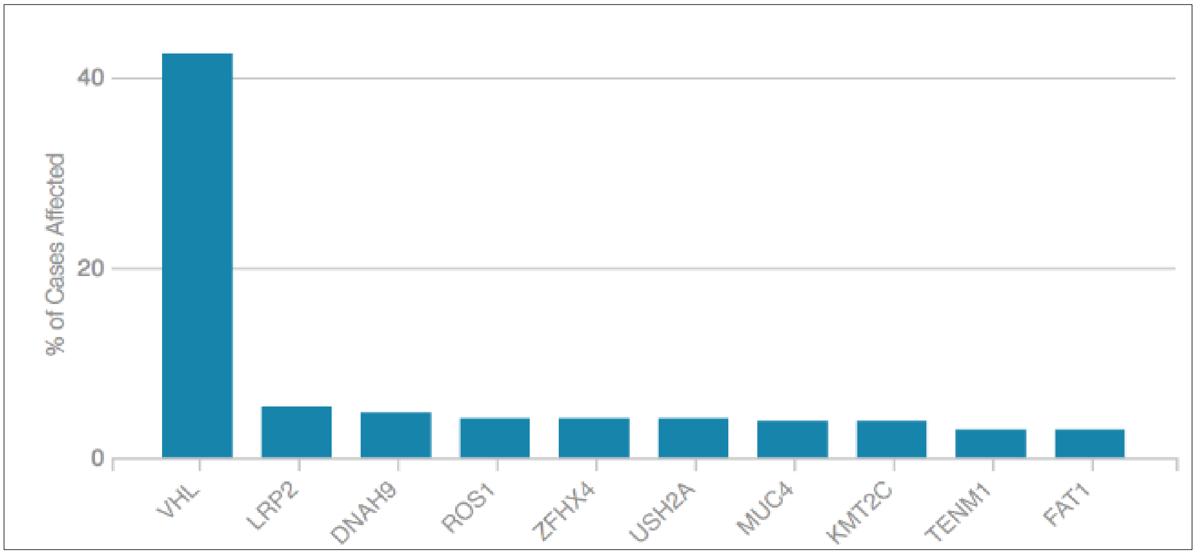


Figure 4. Screenshot of GDC DAVE Tool Visualization [41]

GDC provides data sharing, data submission across different cancer genomic studies and research thereby supports the development of precision medicine for cancer. A secure GDC API is also developed to provide batch data submissions [52].

2.1.5 Summary of Cancer Data Analysis and Visualization

The tools discussed above are new and recently available platforms to analyze and visualize cancer data. While each of them is simple and easy to use, they are all tied to the databases which are not publically available. With no scope of adding a custom database at the backend, it leaves these tools as standalone projects that cannot be integrated into other projects or tools. More importantly, these tools haven't been available for public use for long which diminishes the scope for a comprehensive evaluation or comparison among them.

Since all these tools are deployed over the web, there are no distributions available to use them offline or on desktop modes. The wide variety of options to visualize data, prove handy to understand the trends better and inspires the researchers to design a similar tool. It serves as a motivation to design a flexible, scalable, easy to integrate platform, with visualization features comparable to these tools that could accommodate varying databases.

2.2 Predicting Breast Cancer Survival using Data Modeling Techniques

Predicting the outcome of a disease is one of the most challenging tasks for researchers and medical personnel. In cancer treatment, survival is considered the most critical outcome [15]. Cancer survival prediction using data mining on historical records is possible. The existing predictive models have used data mining techniques such as artificial neural networks, decision trees and statistical methods to predict cancer survival. The different approaches used for cancer survival prediction are grouped into three categories – (i) comparison of different modeling techniques [11, 55, 57, 59, 60, 61, 62] to identify the most accurate prediction model, (ii) hybrid prediction model [65, 66] and (iii) ensemble of different modeling techniques [5, 18, 32, 68].

2.2.1 Comparison of different Modeling Techniques

In 2005, Delen et al. [55] developed breast cancer prediction model and compared different modeling techniques. Two data mining techniques, i.e. artificial neural networks, decision trees (C5) and one statistical technique, logistic regression were compared. The study used the SEER public-use database [31] for the year 1973-2000. The software packages used in the research for exploring data were MS Access database, SPSS statistical analysis tool, STATISTICA data miner and Clementine data mining toolkit. The data source was pre-processed, and the final dataset consisted of 202,932 records. The modification and removing of records were completed to predict survivability exclusive to breast cancer. Data cleansing and preparation strategies followed are as described below:

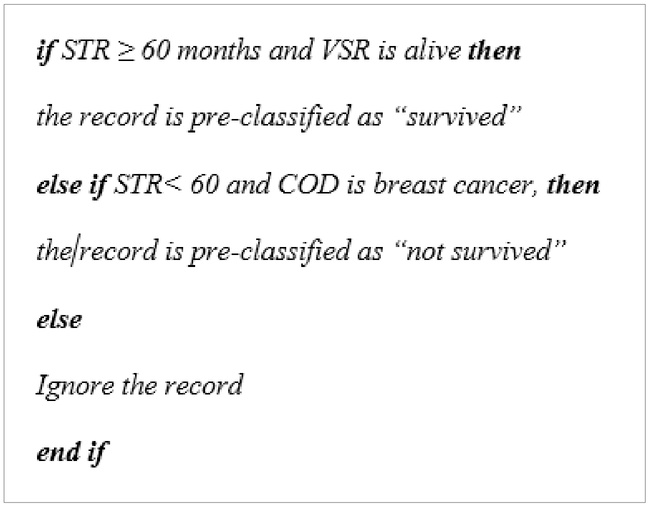
- The records in which the patient didn’t survive for sixty months post-diagnosis were removed.

- The records with “Cause of Death” other than breast cancer were removed.
- The records of those that were not followed up for sixty months were removed.
- The records with missing values were removed.
- The records with unusual “Tumor Size” variable values were also eliminated.

Only 17 out of 72 variables were selected; these included 16 predictor variables and 1 dependent variable. Some of the key variables used were: race, age, grade, marital status, primary site code, histology, behaviour, extension of disease, lymph node involvement, radiation, stage of cancer and tumor size. The binary ‘dependent variable’ was assigned values of 0 and 1, where **0** denoted ‘did not survive’ and **1** denoted ‘survived’. The comparative performance of three data mining methods was evaluated by accuracy, sensitivity, specificity and k-fold cross-validation. The results showed that decision tree (C5) was the best predictor with the highest accuracy of 93%; followed by artificial neural networks with an accuracy of 91.2%, and logistic regression with an accuracy of 89.2% [55]. Some of the shortcomings of this study are:

1) The pre-classification method used in the study for determining the records of ‘died/not survived’ category was incorrect.
2) The study is based on the assumption that all patients died due to breast cancer only, which is not always the case [56].
3) The study did not use Vital Status Recode (VSR) and Cause of Death variables. VSR marks whether the patient is dead or alive as of study cut-off date and Cause of Death provides the reason of cause of death of the patient. These two variables have been shown as important variables for cancer survival prediction [57] and other related studies.

Several spin-offs of their work followed through the years, and the SEER public-use data was observed to be used as the primary data source for these studies. In 2006, Bellaachia and Guven [57] implemented data mining techniques on breast cancer data to enhance Delen et al.'s study. The study included two more variables in addition to the 17 variables selected by [55], i.e. Cause of Death and Vital Status Recode. A new dependent variable Survivability was derived using Survival Time Recode (STR) and VSR. For a 60-month threshold, the 'Survivability' variable was calculated using the logic shown in Figure 5.



```
if STR $\geq$ 60 months and VSR is alive then
    the record is pre-classified as "survived"
else if STR< 60 and COD is breast cancer, then
    the record is pre-classified as "not survived"
else
    Ignore the record
end if
```

Figure 5. Survivability calculation by Bellaachia and Guven [57]

SEER public-use database [31] was used for the period 1973-2002. The study compares three data mining techniques: Naïve Bayes, back-propagated neural networks, and C4.5 decision tree algorithm. WEKA [58] toolkit software package was used for developing the prediction model. Accuracy, precision, and recall performance measures were used to evaluate the data mining techniques. The experimentation ranked Naïve Bayes technique as best with 84.5% accuracy, followed by artificial neural networks and C4.5 algorithms with 86.5% and 86.7% accuracy, respectively. With respect to Delen et al.'s study, the variation in the accuracy of the two studies is due to different SEER datasets, pre-processing and data mining techniques

[57]. One limitation of this study, as stated by the authors, is the exclusion of records with missing data (Extent of Disease and Site Specific Surgery).

Endo et al. [59] compared seven algorithms to predict breast cancer survival using the SEER public-use database [31] from the year 1992 to 1997. Logistic Regression model, Artificial Neural Network, Naïve Bayes, Bayes Net, Decision Trees with Naïve Bayes, Decision Trees (ID3), Decision Trees (J48) were used to develop the prediction models. Among these methods, the Logistic Regression model showed the highest accuracy with 85±0.2%, Decision tree (J48) showed the highest sensitivity and ANN displayed the highest specificity. The study used accuracy to evaluate the model performance, but the authors also state that sensitivity is a comparatively better parameter for survival based prediction models.

A study by Wang et al. [11] predicts 5-year breast cancer patient survivability by using two data mining techniques, i.e. Logistic Regression model and a Decision Tree model. The study is performed on the the SEER public-use database [31] for the year 2010. The study concludes that the Logistic Regression model is better than the Decision Tree model [11]. The study uses the same data preparation method as used by [55]. The dataset used in both the studies are different. The incidence and mortality trends in the datasets used by both these studies are significantly different. This study concludes with higher accuracy than [55]. The former study shows 91.19% and 91.34% accuracy while the latter shows 85.8% and 86.0% accuracy of decision trees and logistic regression respectively.

Few studies [60, 61, 62] as discussed next, have developed a breast cancer detection model which predicts whether the cancer is present or not. These studies have also used data mining techniques and performed a comparison of these techniques.

Chaurasia and Pal [60] developed a diagnosis system for breast cancer detection. The model uses RepTree, RBF Network and Simple Logistic modeling techniques. The study uses

University Medical Centre Institute of Oncology, Ljubljana, Yugoslavia database. The extracted breast cancer data had 286 rows and 10 variables for each row. The results of the study state that Simple Logistic modeling technique has higher accuracy (74.47%) as compared to RepTree (71.32%) and RBF Network (73.77%).

Senturck and Kara [61] performed a breast cancer diagnosis study using data mining on the UCI Machine Learning database from the University of Wisconsin Hospitals, Madison. The study aimed to analyze the performance of seven different algorithms. RapidMiner 5.0 [63] tool was used for data mining and prediction. The study concluded that a Support Vector Machines algorithm is best for breast cancer diagnosis prediction with an accuracy of 96.0% [61]. The study predicts if the tumor is benign or malignant. The prediction model is trained with 699 cases only (records with missing information were removed).

Chaurasia and Pal [62] developed a breast cancer detection model using WEKA [58] software for data mining. This study also used UCI Machine Learning database from the University of Wisconsin Hospitals, Madison. The study aims to compare the performance of three classification techniques: Sequential Minimal Optimization (SMO), IBK and BF Tree. The study concludes that SMO classification techniques have the highest prediction accuracy (96.2%) amongst all three techniques. This study is different from their earlier work [60] in which that Simple Logistic, RepTree and RBF Network modeling techniques were used. The databases used in both studies are from different demographics.

2.2.2 Hybrid modeling for Predicting Breast Cancer Survival

Hybrid modeling is an approach when two or more modeling techniques are combined, for example, clustering and classification techniques combined or clustering and association modeling techniques used together [64].

In 2008, Khan et al. [65] investigated a hybrid scheme based on fuzzy decision trees as an alternative to breast cancer prognosis. The data source used for the study was the SEER public-use database [31] for the period 1973-2003. An essential aspect of the research was to use a hybrid modeling technique based on Fuzzy Decision trees. Data pre-processing method removed the records having missing data and included only records with a Cause of Death (COD). The final dataset of 162,500 records with 16 variables and a binary target variable (**0** denoted 'did not survive' and **1** denoted 'survived') was used for experimentation. The performance measure evaluation stated that hybrid fuzzy decision tree classification technique (accuracy 85%) is more powerful and fair than independently applied decision tree classification technique (accuracy 82%) [65]. The study is also based on an assumption similar to Delen et al. [55] that all patients died due to breast cancer only, which is not always the case. A 2012 survey states that external causes, heart failure, suicide and gastrointestinal diseases are other reasons for breast cancer patient death [56].

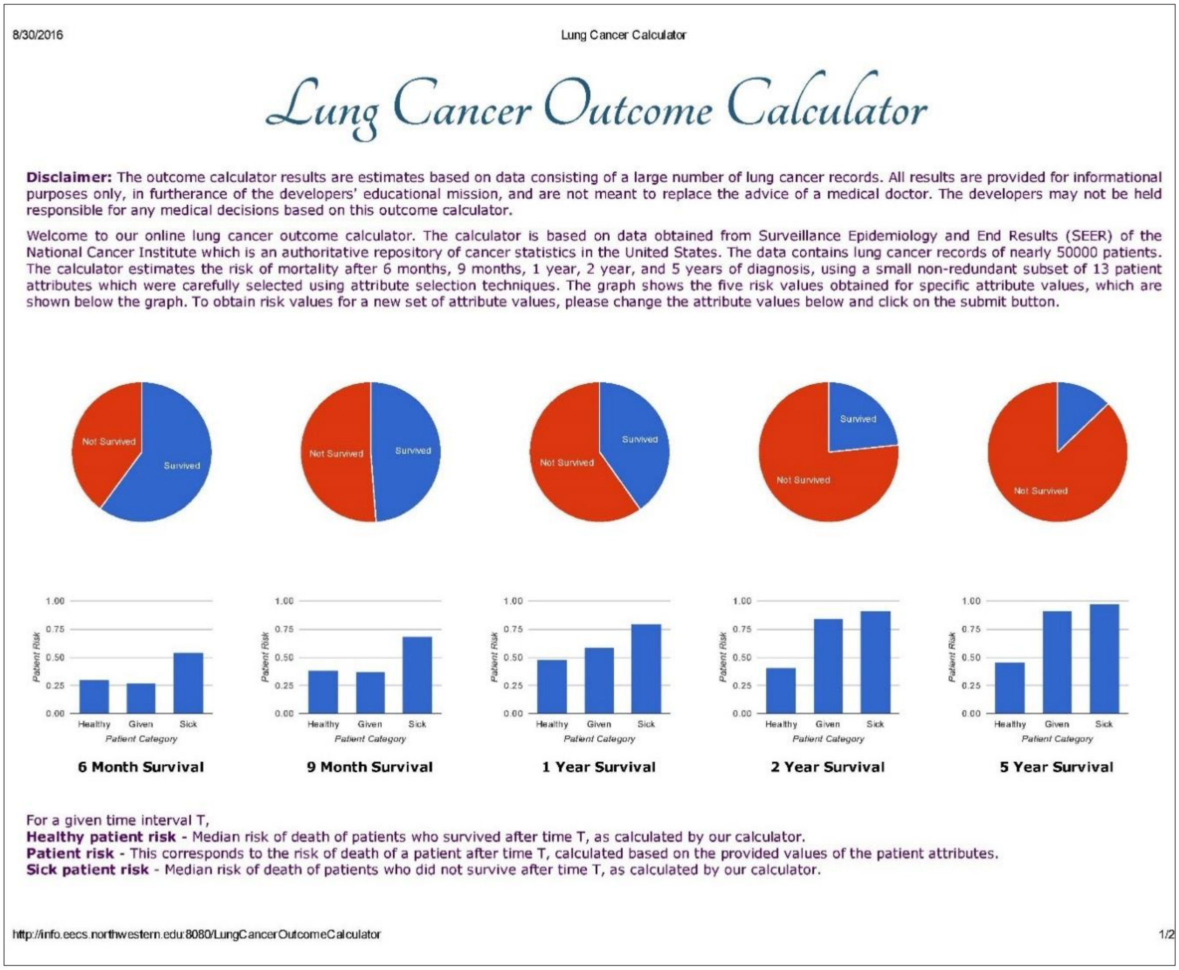
According to Choi et al. [66], a hybrid Bayesian model for predicting breast cancer prognosis can outperform other models. Three different model for cancer prognosis were examined: Bayesian Network (BN) model, Artificial Neural Network (ANN) model and hybrid BN model. The hybrid model developed was a combination of an ANN model and the BN model. The SEER public-use database [31] for the time period 1973-2003 was used to build the model with 294,275 records and 9 input variables. For a threshold of 60 months, the proposed hybrid BN model performed better than the Bayesian network [66]. The study states that the proposed hybrid BN model and the ANN models outperformed the BN model. The goal of this study was to explain the power of BN models over ANN models. However, the study results showed that the hybrid BN model's performance was mainly due to ANN model instead of the BN model. The Area Under Curve (AUC) of ANN (0.930) difference from

Hybrid BN (0.935) is minimal (0.005) as compared to BN (0.813). Authors stated that the better performance of the hybrid BN was originated from ANN instead of BN.

2.2.3 Ensemble Modeling Technique for Predicting Cancer Survival

Ensemble modeling techniques are used to improve the performance of individual classification techniques such as decision trees, regression, neural networks, support vector machines and Bayesian networks. Ensembles combine predictions of multiple classification techniques to achieve better prediction accuracy [67]. Common ensemble techniques used are bagging, boosting, voting and stacking. Ensembles modeling techniques only combine classification techniques, unlike hybrid modeling technique which can combine classification and clustering, or clustering (example- K-Means and two-step clustering models) and association techniques (example- Apriori and Carma models).

In 2010, Agrawal et al. [5] developed an online lung cancer outcome calculator, using data mining and predictive modeling. The research aimed at developing an accurate survival prediction model by using the SEER public-use database [31]. The study used 1998-2001 data for five-year prediction. Records earlier than 1998 were eliminated because some variables were added only after 1998. Few variables were modified and merged to form new variables for the study. Only records with “Cause of Death” as lung cancer were used. The WEKA [58] software tool was used to evaluate the data mining techniques used. An ensemble of the data mining algorithms- J48 Decision Tree, Alternating Decision Tree, Logit Boost, Random Subspace and the Random Forest was used in the study. The predictive model was built with 64 variables, and the online calculator was built by using 13 of 64 variables.



8/30/2016 Lung Cancer Calculator

Lung Cancer Outcome Calculator

Disclaimer: The outcome calculator results are estimates based on data consisting of a large number of lung cancer records. All results are provided for informational purposes only, in furtherance of the developers' educational mission, and are not meant to replace the advice of a medical doctor. The developers may not be held responsible for any medical decisions based on this outcome calculator.

Welcome to our online lung cancer outcome calculator. The calculator is based on data obtained from Surveillance Epidemiology and End Results (SEER) of the National Cancer Institute which is an authoritative repository of cancer statistics in the United States. The data contains lung cancer records of nearly 50000 patients. The calculator estimates the risk of mortality after 6 months, 9 months, 1 year, 2 year, and 5 years of diagnosis, using a small non-redundant subset of 13 patient attributes which were carefully selected using attribute selection techniques. The graph shows the five risk values obtained for specific attribute values, which are shown below the graph. To obtain risk values for a new set of attribute values, please change the attribute values below and click on the submit button.

For a given time interval T,
Healthy patient risk - Median risk of death of patients who survived after time T, as calculated by our calculator.
Patient risk - This corresponds to the risk of death of a patient after time T, calculated based on the provided values of the patient attributes.
Sick patient risk - Median risk of death of patients who did not survive after time T, as calculated by our calculator.

http://info.eecs.northwestern.edu:8080/LungCancerOutcomeCalculator 1/2

Figure 6. Lung Cancer Outcome Calculator screenshot [5]

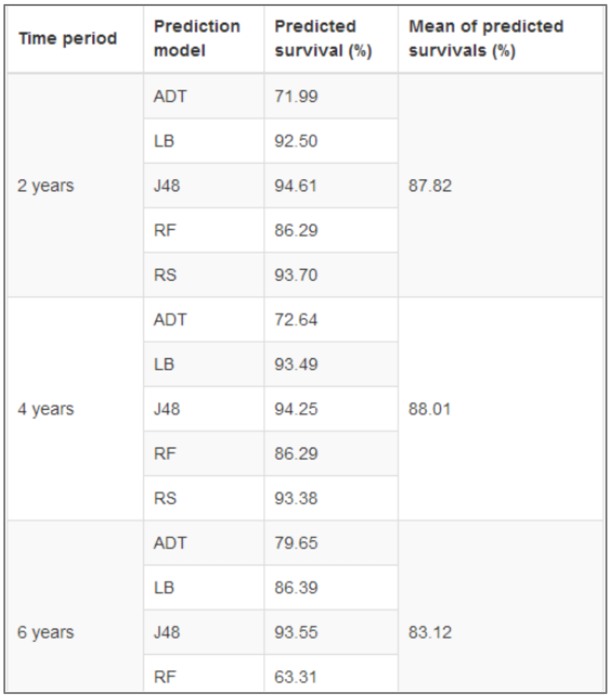
The 13 variables were selected based on the predictive power[1] by using a feature selection[2] method. Some of the key variables used were: age, birthplace, cancer grade, farthest extension of tumor, lymph node involvement and total regional lymph nodes examined. Overall, the ensemble voting classification technique performed best with the highest

[1] Predictive power: The predictive power of an attribute here refers to the ranking or inter relatability ability of that attribute with other attributes to form patterns [101].

[2] Feature selection: It is used to identify the fields that are most important for a given analysis.

prediction accuracy (91.4%) and AUC (94%) [5]. Figure 6 shows a snapshot of the online lung cancer calculator result window.

In 2014, GilTroy Paular Meren [18] developed a Breast Cancer Outcome-Survival Online Measurement (BOSOM) calculator. An online survival measurement calculator using data mining and predictive modeling on the SEER public-use database [31] (1973-2010). The study uses the framework and data mining techniques as Agrawal et al.'s 'Online Lung Cancer Outcome Calculator' to establish the prediction calculator. The time interval used for prediction includes 2 years, 4 years, 6 years, 8 years and 10 years. The study concluded with the average accuracies of the calculator and completed dataset as 88.27% and 91.71%, respectively. Figure 7 is a snapshot of the BOSOM calculator output screen. The classifiers used in this study are same as used in [5]. The calculator of this study is a replica of LCOC [5] using the same methodology for breast cancer survivability.



Time period	Prediction model	Predicted survival (%)	Mean of predicted survivals (%)
2 years	ADT	71.99	87.82
	LB	92.50	
	J48	94.61	
	RF	86.29	
	RS	93.70	
4 years	ADT	72.64	88.01
	LB	93.49	
	J48	94.25	
	RF	86.29	
	RS	93.38	
6 years	ADT	79.65	83.12
	LB	86.39	
	J48	93.55	
	RF	63.31	

Figure 7. BOSOM Calculator - Table for Predicted Survival [19]

Gokhan and Mustafa [32] used an ensemble of three data mining techniques to diagnose breast cancer. Clementine software was used for data mining on the Wisconsin Diagnostic Breast Cancer Database. Amongst Decision Trees, Support Vector Machines, Artificial Neural Network and an ensemble of all three, the ensemble model proved to be better than individual models. The dataset used in this study has 569 instances or records. The limited data used in this study was one of the major shortcomings of this study. The study only determines whether the case is malignant or benign.

Lastly, in 2017, Huang et al. [68] compared Support Vector Machine (SVM) modeling technique with SVM ensemble technique for breast cancer prediction. The datasets used in the study are Wisconsin Diagnostic Breast Cancer Database, UCI machine learning database and ACM SIGKDD Cup 2008. WEKA [58] data mining software is used to construct SVM classifiers. It is concluded that for smaller datasets, SVM ensembles performed better than individual SVM classification technique. For large datasets, SVM ensembles using the boosting method performs better than the other classification techniques. The results are presented in Table 2 [68].

Dataset	**Modeling Technique**		**Accuracy (%)**
Small-scale dataset	Individual SVM	GA+linear SVM	96.57%
	SVM Ensemble	GA+RBF SVM (boosting)	98.28%
		GA+linear SVM (boosting/bagging)	96.57%
Large-scale dataset	SVM Ensemble	Poly SVM (boosting)	99.51%
		GA+poly SVM (bagging)	99.50%
		RBF SVM (boosting)	99.52%

Table 2. Accuracy Comparison of SVM and SVM Ensembles [68]

2.2.4 Summary of Breast Cancer Prediction using Data Modeling Techniques

Table 3 summarizes the data modelling techniques used for cancer survivability prediction in literature.

Type	Author	Year	Dataset	Technique
Comparison of different modeling techniques	Delen et al.	2005	SEER public-use database (1973 – 2000)	Decision Tree (C5), Artificial Neural Network, Logistic regression
	Bellaachia and Guven	2006	SEER public-use database (1973 – 2000)	Naïve Bayes, Back-propagated Neural Network, C4.5 Decision Tree
	Endo et al.	2008	SEER public-use database (1992 – 1997)	Logistic Regression, Artificial Neural Network, Naïve Bayes, Bayes Net, Decision Trees with naïve Bayes, Decision Trees (ID3), Decision Trees (J48)
	Wang et al.	2013	SEER public-use database (1973 – 2007)	Logistic Regression, Decision Tree (J48)
	Senturk and Kara	2014	UCI Machine Learning Repository	Artificial Neural Network, Decision Trees, Logistic Regression, Support Vector Machines, Naïve Bayes, K-Nearest Neighborhood
	Chaurasia and Pal	2017	University Medical Centre Institute of Oncology, Ljubljana, Yugoslavia database	RepTree, RBF Network, Simple Logistic
		2017	UCI Machine Learning Repository	SequentialMinimalOptimization, IBK, BF Tree
Hybrid	Khan et al.	2008	SEER public-use database (1973 – 2003)	Decision Trees, Fuzzy Decision Trees
	Choi et al.	2009	SEER public-use database (1973-2003)	Artificial Neural Network, Logistic Regression, Bayesian Network

Ensemble	Agrawal et al.	2010	SEER public-use database (1988-2001)	J48 decision tree, Random forest, LogitBoost, Random subspace Alternating Decision Tree
	GilTroy Paular Meren	2014	SEER public-use database (1973-2010)	ZeroR, Random forest, LogitBoost, Random subspace, J48 Decision Tree, Alternating decision Tree
	Gokhan and Mustafa	2015	Wisconsin Diagnostic Breast Cancer Database	Decision Trees, Support Vector Machines, Artificial Neural Network
	Huang et al	2017	Wisconsin Diagnostic Breast Cancer Database, UCI machine learning database, ACM SIGKDD Cup 2008	SVM, SVM ensemble

Table 3. Data Modeling Techniques used in Cancer Survivability Prediction

2.3 Summary

In this chapter, an overview of research done on breast cancer survivability has been provided. The models presented in the literature are based on several data mining techniques and different datasets. Most of these works focus on comparing data mining techniques for building predictive models, and only a few have developed specific tools to predict the outcome (survival) based on the patient-specific input. Predicting survivability of breast cancer patients can greatly assist physicians in developing the treatment plan specific to each patient. In the next chapter, the methodology used for this research along with implementation steps are discussed.

3. Methodology

As mentioned in Chapter 1, the two main contributions of this thesis are the prediction of breast cancer survivability and analysis of breast cancer data. This chapter breaks down the overarching contributions into the sets of smaller tasks and explains each of the tasks along with presenting the rationale behind finalizing the routes adopted and choices made to accomplish them. In precision, the chapter presents various approaches that could be used for predicting survivability and analyzing breast cancer data, the predictive model along with implementation details of the different modeling techniques used for developing the predictive model to predict survival months, and the dashboard to visualize 40 years of historical breast cancer data.

3.1 Proposed Approach

Figure 8. Methodology gives an overview of the method used for this research. It is primarily divided into five tasks: data extraction (from raw data), data analysis (including pre-processing), data visualization, predictive modeling, and evaluation of the predictive model.

3.1 Data Extraction

The SEER database maintains cancer statistics for the US and monitors annual cancer incidence progression of various types of cancer. The breast cancer data occurring in different population subgroups is available for the period from 1973-2013. Among other variables, the data includes patient records, race/ethnicity, primary site, the first course of treatment, and follow-up vital status [69]. The “SEER limited-use” data is defined by demographics, treatment

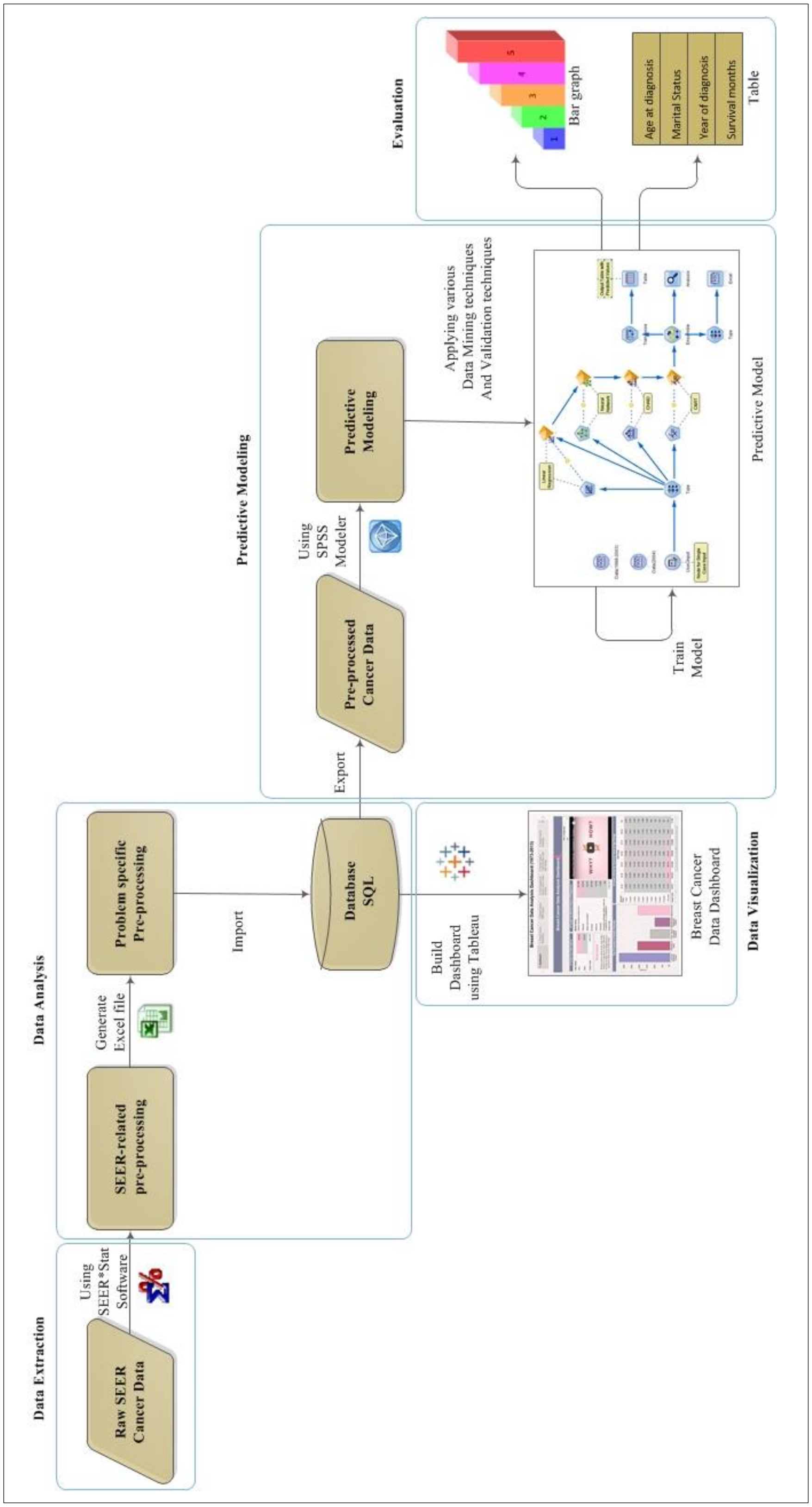


Figure 8. Methodology

(Icons copyright: SEER*Stat, Microsoft, tableau and IBM) (e.g. surgery, radiation therapy), diagnosis (e.g. primary site, tumor size), and an outcome characteristic (e.g. survival time, cause of death), which makes SEER an excellent source for outcome analysis and prediction-based studies. The SEER dataset used for this research is a collection of data from 18 registries. SEER*stat statistical software [70] is used to extract raw data from the SEER database. This software allows viewing of patient record and production of different sessions such as Frequency, Rate, Survival, and Case Listing. After consultation with a radiation oncologist, 30 variables were selected (from a total of 134 variables available in SEER) to prepare the relevant dataset. The dataset is filtered to only include cases, which died due to cancer, i.e. 'Dead' and 'N/A not first tumor' are selected for the SEER cause-specific death classification (the detailed definition of SEER variables are presented in Appendix B).

3.2 Data Analysis

Data analysis is the process of transforming raw data into usable form. Once the raw records are extracted, data preprocessing is performed to produce a relevant subset. The pre-processed data is imported into the SQL Server database [71], followed by analysis leading to building the dashboard and reports using Tableau.

a) Data preprocessing

The data preprocessing is done at two levels:

- **SEER-related preprocessing**: Normalization of data, such as converting text values to numeric representation is performed as a part of preprocessing. SEER*stat software is used to accomplish this task. The derived data is cleansed for eliminating redundant content. Male breast cancer cases are also eliminated as a part of data cleansing.

- **Problem-specific preprocessing**: This includes selecting data records for a distinct time of significance and eliminating attributes which do not hold any considerable predictive power. One of the steps is the removal of records which represent deaths due to a reason other than breast cancer.

b) Data Modeling

The pre-processed data is imported into MS SQL Server to create a database consisting of relevant dimensions and measures. Tableau [72] is connected to the database, and the tables are joined to create a view, extending horizontally by adding columns of data, as needed. The data is further cleansed (such as by changing data types, renaming & resetting fields) and prepared for analysis. Calculated fields, formulas, grouping and sets are added via SQL queries. The data is sliced and diced by using filters and parameters. The dissected data is then visualized using workbook, dashboards, and stories.

3.3 Data Visualization

The dashboard and dynamic reports contained in them use views and tables created in the SQL Server database that help convert massive amounts of data into meaningful and actionable information. This is accomplished by building dashboards which contain visually appealing and interactive components including charts, graphs, tooltips, and drill-down/ drill-through reports.

Dashboards provide interactive access to informative data and help understand the enormous data generated by every cancer incidence. In addition to tracking of KPIs, the reports also allow discovery of hidden patterns in the data. All of this, in turn, can improve the quality of cancer care. The policymakers, health professionals, advisors, and planners could use this data to view and report breast cancer statistics which provide a better understanding of the incidence and mortality trends. Physicians can identify treatment options, wellness programs, and patient engagement. It can also empower patients to choose the right care through interactive visualization of treatment cost, quality and effectiveness.

3.4 Predictive Modeling

By using the pre-processed and cleansed data, a breast cancer predictive model has been designed to predict survival months of a breast cancer patient from the year of diagnosis. The predictive model has been trained, tested and validated with SEER data. The predictors shortlisting is based on main KPIs identified by the analysis. Seeking an expert's opinion in choosing the relevant predictors from a total of 134 variables available in SEER is an inevitable process for this task. By using various data modeling techniques, the predictive model has been developed with modeling techniques of highest accuracy, along with their ensemble. There are three crucial steps in this stage: selection of predictors, selection of training and testing datasets, and developing the predictive model using the training dataset.

3.5 Evaluation

The predictive model has been evaluated on the testing dataset:

a) By comparing the actual average survival month with predicted survival months.

b) By calculating the performance metrics such as accuracy.

3.2 Breast Cancer Analysis and Visualization

With increasing cases of cancer, the amount of data associated with cancer has also increased proportionally. Analyzing such huge datasets is difficult. Dashboards provide a visual mechanism to track KPIs and other metrics relevant to specific processes [73]. The purpose of the dashboard is to capture, process, and distribute information in an intelligible format to enable users to understand the data better [74]. This, in turn, can improve the quality of cancer care.

For this research, a combined dataset of over 40-year historical breast cancer data has been used. The dataset, which holds data in the raw form, is analyzed and then set for visualization by means of interactive reports and dashboard. The dashboard helps uncover hidden patterns as well as provide easy-to-understand metrics for users of all backgrounds. The dashboard includes breast cancer data for patient population demographics, patient volumes, diagnosis and treatment.

There are several data visualization tools available such as SSRS, Tableau, Power View and Qlik View, which can be used to build dashboards and visualize data. In this research, Tableau is chosen because of its versatility, flexible interface and other capabilities described in the following section.

3.2.1 Tableau

Tableau [75] is a commercially available tool, which allows building dashboards by transforming data into visually appealing and interactive visualizations. It can be connected to a variety of data sources such as Access, Excel, and data warehouse or web-based data [76]. It is an easy-to-use tool for data analysis and building dashboards together with drill-down and drill-through reports. Such visualization can help users (physicians, researchers & patients) to find the right path forward. For example, physicians can identify treatment options, wellness programs, and patient engagement. It can even empower patients to choose the right care through interactive visualizations of treatment cost, quality and effectiveness [72].

Tableau is a useful tool for organizations which have massive amounts of data (e.g. the healthcare industry) to be converted into meaningful and actionable information. Also, the reports generated by Tableau can be published to a shareable URL [77]. Tableau can extract useful information out of large data, which is otherwise difficult to examine manually. The advantages of using Tableau as a Business Intelligence (BI) tool in this research are:

- Tableau can discover hidden patterns in the data.
- Tableau is an easy to navigate tool with simple drag and drop features for creating dashboards.
- Tableau can analyze millions of rows of data in seconds.
- Gartner's BI Magic Quadrant [78] report has ranked Tableau as a leader for four consecutive years [79].
- Tableau can link multiple data sources (such as databases, flat files, web services) for quick and accurate analysis.
- Tableau can provide real-time dashboards.
- Tableau provides an option of extracting data from the data source or have a live connection with the data source such as healthcare environments.

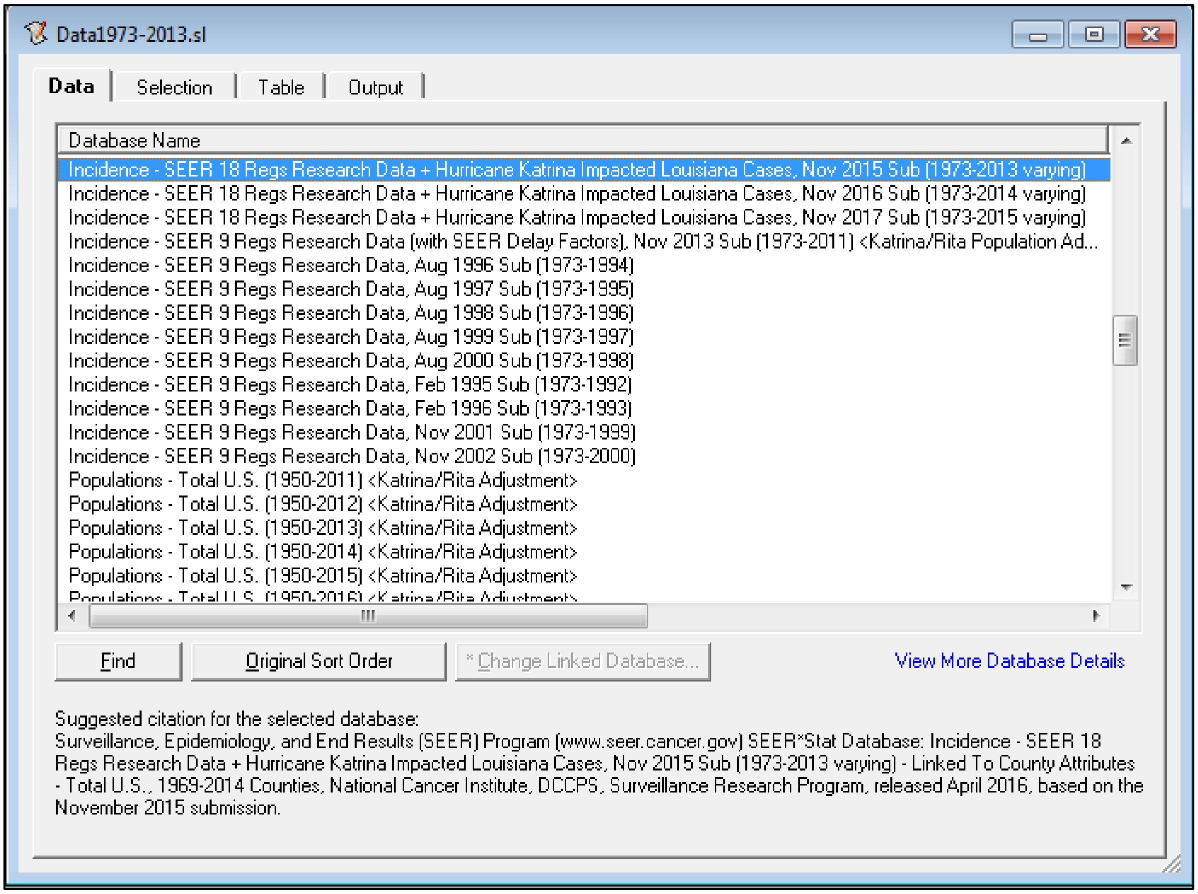
3.2.2 Accessing SEER Database and SEER*Stat

The SEER Program [80] is a database of cancer statistics in the United States. SEER is supported by the Surveillance Research Program (SPR) [81] in the National Cancer Institute's Division of Cancer Control and Population Sciences (DCCPS) [82]. The SEER database monitors the annual cancer incidence progression of each type of cancer.

The SEER public-use data is available from the SEER web site on submitting a SEER limited-use data agreement form [31]. The data agreement form is available in Appendix A. The data can be accessed by two different options, SEER*Stat's client-server mode or by downloading compressed files. The first method requires the download and installation of SEER*Stat software [70] on the machine. It requires an internet connection to extract data from the SEER database. The variables selected by the user are transferred from the user's machine to the SEER*Stat server to retrieve data as per the ad-hoc requests submitted. The second method requires the user to download compressed files of the data in two formats, i.e. binary and ASCII version of data. These files can then be accessed using the SEER*Stat software.

3.2.3 Data Understanding, Preparation and Extraction

The SEER*Stat statistical software version 8.3.5 is used to obtain breast cancer data. SEER*Stat software is associated with the SEER research data - available either directly from SEER's server or a local file. For this research, the first method is used, i.e. extracted data using SEER*Stat software and SEER*Stat server. SEER*Stat software provides different types of sessions, designed to calculate specific statistics. The software helps to view a given record of a cancer patient and can produce different sessions such as frequency, rate, survival, and case listing sessions. In this research, a case listing session is used to get data at the individual case or patient level. Case listing session also allows accessing the data in ASCII text format and generates a dictionary for each variable selected.



Database Name

Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2015 Sub (1973-2013 varying)
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2016 Sub (1973-2014 varying)
Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2017 Sub (1973-2015 varying)
Incidence - SEER 9 Regs Research Data (with SEER Delay Factors), Nov 2013 Sub (1973-2011) <Katrina/Rita Population Ad...
Incidence - SEER 9 Regs Research Data, Aug 1996 Sub (1973-1994)
Incidence - SEER 9 Regs Research Data, Aug 1997 Sub (1973-1995)
Incidence - SEER 9 Regs Research Data, Aug 1998 Sub (1973-1996)
Incidence - SEER 9 Regs Research Data, Aug 1999 Sub (1973-1997)
Incidence - SEER 9 Regs Research Data, Aug 2000 Sub (1973-1998)
Incidence - SEER 9 Regs Research Data, Feb 1995 Sub (1973-1992)
Incidence - SEER 9 Regs Research Data, Feb 1996 Sub (1973-1993)
Incidence - SEER 9 Regs Research Data, Nov 2001 Sub (1973-1999)
Incidence - SEER 9 Regs Research Data, Nov 2002 Sub (1973-2000)
Populations - Total U.S. (1950-2011) <Katrina/Rita Adjustment>
Populations - Total U.S. (1950-2012) <Katrina/Rita Adjustment>
Populations - Total U.S. (1950-2013) <Katrina/Rita Adjustment>
Populations - Total U.S. (1950-2014) <Katrina/Rita Adjustment>
Populations - Total U.S. (1950-2015) <Katrina/Rita Adjustment>
Populations - Total U.S. (1950-2016) <Katrina/Rita Adjustment>

Suggested citation for the selected database:
Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2015 Sub (1973-2013 varying) - Linked To County Attributes - Total U.S., 1969-2014 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2016, based on the November 2015 submission.

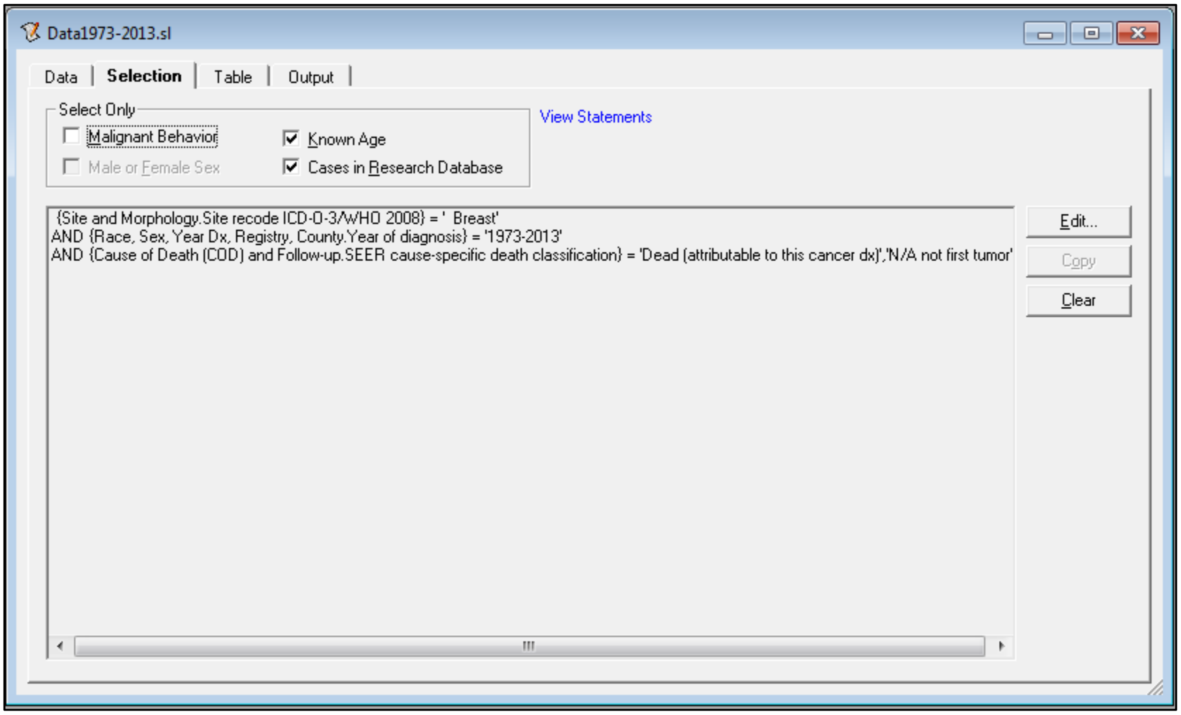
Figure 9. Selecting Database in SEER*Stat

The database selection could be made as shown in Figure 9. For this research database used is “Incidence-SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2015 Sub (1973-2013 varying)”. This dataset is a collection of data from 18 different SEER registries i.e. Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, Seattle-Puget Sound, Utah, Lost Angeles, San Jose-Monterey, Rural

Georgia, Alaska Native Tumor Registry, Greater California, Greater Georgia, Kentucky, Louisiana and New Jersey [45]. The selected dataset has data for all types of cancer and 134 variables to use.

The 'Selection' tab allows filtering the database by site and morphology, year of diagnosis and other factors to customize the data extraction with respect to the research. The filters used on the selected database are shown in Figure 10. Following are filters are applied:

- Only known age cases in research database are included
- Site and Morphology is selected as 'Breast'
- Year of diagnosis is selected from 1973-2013
- Only 'Dead' and 'N/A not first tumor' is selected for SEER cause-specific death classification.



Select Only
☐ Malignant Behavior
☐ Male or Female Sex
☑ Known Age
☑ Cases in Research Database

```
 {Site and Morphology.Site recode ICD-O-3/WHO 2008} = ' Breast'
AND {Race, Sex, Year Dx, Registry, County.Year of diagnosis} = '1973-2013'
AND {Cause of Death (COD) and Follow-up.SEER cause-specific death classification} = 'Dead (attributable to this cancer dx)','N/A not first tumor'
```

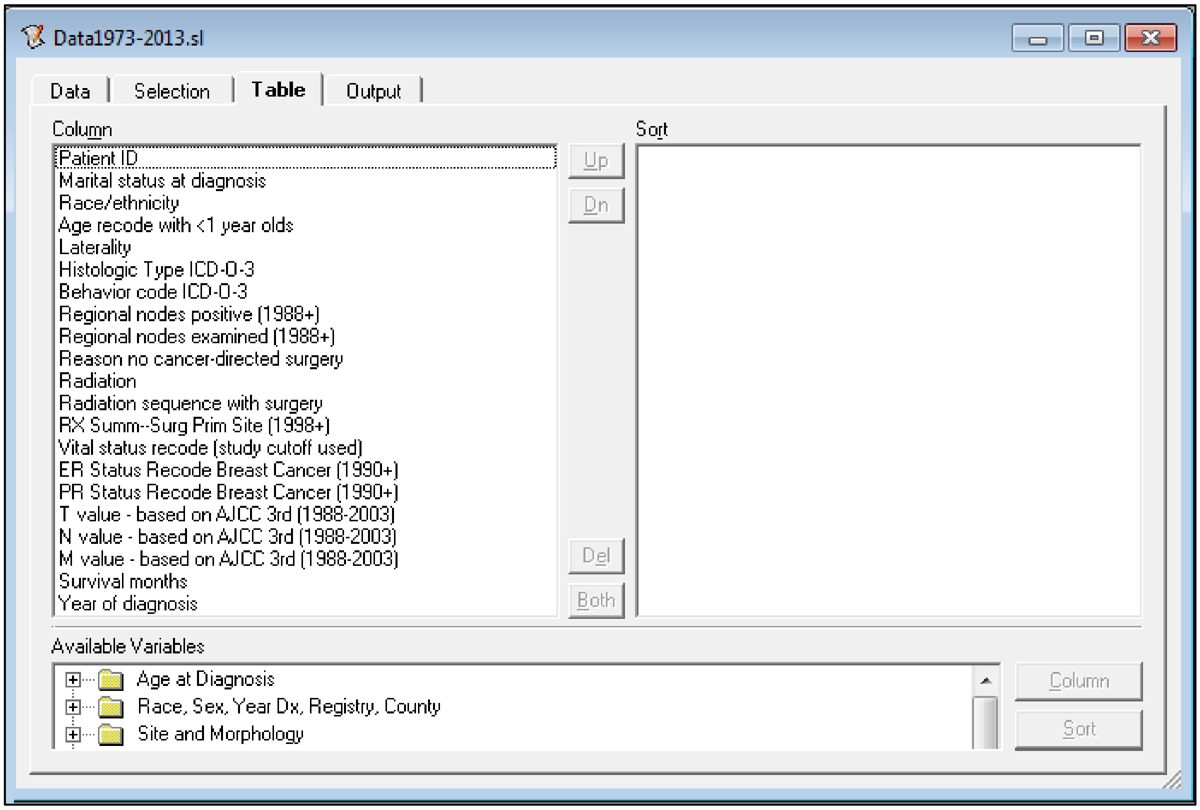
Figure 10. Filters used in SEER*Stat

List of short-listed variables
1. CS mets[3] at Diagnosis-bone
2. CS mets at Diagnosis-lung
3. CS mets at Diagnosis-liver
4. CS mets at Diagnosis-brain
5. Breast Subtype
6. Vital status recode (study cut-off used)
7. Age recode
8. Radiation
9. Radiation sequence with surgery
10. T value[4]
11. N value[5]
12. M value[6]
13. Regional nodes positive
14. CS lymph nodes
15. CS mets at Diagnosis
16. CS extension
17. CS tumor size
18. Marital status at diagnosis
19. Regional nodes examined
20. Estrogen Receptor Status
21. Progesterone Receptor Status
22. Survival months
23. Laterality
24. Histologic Type ICD-O- 3
25. Race/ethnicity
26. Year/Month of Diagnosis
27. Behavior code ICD-O-3
28. Surgery of Primary Site
29. Reason no cancer-directed surgery
30. SEER cause-specific death classification

Table 4. List of short-listed Variables

[3] Mets - Metastasis
[4] T value - Size of the original tumor
[5] N value - Degree of nearby lymph nodes involved
[6] M value - Presence of distant metastasis

Out of 134 variables available in SEER, it is crucial to use variables relevant to the prediction of breast cancer survivability. Seeking an expert's opinion for shortlisting the variables is crucial. Hence, Dr. Robert Olson, who is a Radiation Oncologist at the BC Cancer Agency Centre for the North [83] and Regional Director of Faculty Development, Affiliate Assistant Professor, Northern Medical Program, UNBC [84] was consulted. Dr. Robert Olson suggested a list of 30 relevant variables out of the 134 variables available (Table 4). These variables along with their definitions are listed in Appendix B and were selected in the case listing session window (Figure 11).



Patient ID
Marital status at diagnosis
Race/ethnicity
Age recode with <1 year olds
Laterality
Histologic Type ICD-O-3
Behavior code ICD-O-3
Regional nodes positive (1988+)
Regional nodes examined (1988+)
Reason no cancer-directed surgery
Radiation
Radiation sequence with surgery
RX Summ--Surg Prim Site (1998+)
Vital status recode (study cutoff used)
ER Status Recode Breast Cancer (1990+)
PR Status Recode Breast Cancer (1990+)
T value - based on AJCC 3rd (1988-2003)
N value - based on AJCC 3rd (1988-2003)
M value - based on AJCC 3rd (1988-2003)
Survival months
Year of diagnosis

Age at Diagnosis
Race, Sex, Year Dx, Registry, County
Site and Morphology

Figure 11. Selecting Variables in SEER*Stat

The 'Output' tab allows naming the dataset and the session is executed which produces an output table or matrix. The resulting SEER*Stat matrix window could be exported in the

CSV format. Before extraction, the data is converted into ASCII text format. A dictionary file is auto-generated along with CSV file which provides codes for the variables formatted in the matrix. The SEER*Stat case listing session can be saved and reused to run and extract data.

3.2.4 Database

The extracted CSV file is imported into the SQL Server. The database is created based on star schema. The dictionary file is used to create dimension tables. Each variable now has primary and foreign keys. The master CSV file serves as the fact table using the foreign keys of each variable (dimension tables).

3.2.5 Visualization Dashboard

There are several data visualization tools available such as SSRS, Tableau, Power View and Qlik View, which can be used to build dashboard and visualize data. As stated earlier, Tableau is used primarily because of its versatility, and flexible interface. Tableau is connected to SQL Server breast cancer database. The tables are joined to create a view, extending horizontally by adding columns of data, as needed. Tableau is further used to cleanse the data such as – changing data types, renaming and resetting fields suitable for analysis. Calculated fields, formulas, grouping and sets are added via SQL queries. The grouping is done based on the level at which analysis is to be performed. Some of the groupings done are as shown in Table 5.

Grouping	Dimension Used	Level	Data elements used
Age-Range (group)	Age-Range	10-year level expect 80-84 and 85+ (01-09, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-84, 85+)	Ages 01-04, Ages 05-09, Ages 10-14, Ages 15-19, Ages 20-24, Ages 25-29, Ages 30-34, Ages 35-39, Ages 40-44, Ages 45-49, Ages 50-54, Ages 55-59, Ages 60-64, Ages 65-69, Ages 70-74, Ages 75-79, Ages 80-84, Ages 85+
Race/Ethnicity (group)	Race/Ethnicity	White, Black, Others, Unknown	Others include: American Indian, Aleutian, Alaskan Native or Eskimo (includes all indigenous populations of the Western hemisphere), Chinese, Japanese, Filipino, Hawaiian, Korean, Vietnamese, Laotian, Hmong, Kampuchean (including Khmer and Cambodian), Thai, Asian Indian or Pakistani, Asian Indian, Pakistani, Micronesian, Chamorran, Guamanian, Polynesian, Tahitian, Samoan, Tongan, Melanesian, Fiji Islander, New Guinean, Other Asian, including Asian, Pacific Islander, Other
Regional nodes examined (group)	Regional nodes examined	Exact number (01-89) of nodes examined, 90 or more nodes were examined, No nodes were examined, No regional nodes were removed, Regional node removal documented as dissection, Regional lymph node removal as sampling, Regional nodes were surgically removed, Unknown	Exact number (01-89) of nodes examined: Exact 1 nodes examined to Exact 89 nodes examined Unknown: Unknown, Unknown whether nodes were examined, Not Applicable or negative, Not stated in patient record

Regional nodes positive (group)	Regional nodes positive	Exact number (01-89) of nodes positive, 90 or more nodes were examined, All nodes examined are negative, No nodes were examined, Positive aspiration of lymph nodes(s) was performed, Positive nodes are documented, but not specified, Unknown	Exact number (01-89) of nodes examined: Exact 1 nodes examined and tagged positive to Exact 89 nodes examined and tagged positive Unknown: Unknown, Unknown whether nodes were examined, Not Applicable or negative, Not stated in patient record

Table 5. Grouping of Variables for Data Preprocessing

The data can be extracted in the Tableau workbook which is a compressed snapshot of actual data. It is used to make the data engine work faster and provides faster analytical and query performance. The extracted data could be further sliced and diced by using filters and parameters such as year of diagnosis, race/ethnicity and state. The dissected data is then visualized using workbook, dashboards, and stories. The screenshots of the dashboard are presented in section 4.1.

3.3 Breast Cancer Predictive Model

Identifying patterns from the historical data and using them to make predictions forms the basis of predictive analysis [86]. Predictive analysis deals with developing models using wide varieties of data modeling techniques [87]. Decision trees (C&RT, QUEST, CHAID), Neural Networks, Linear Regression, and Support Vector Machines are some of the popular data modeling techniques.

The breast cancer predictive model is developed from the preprocessed data extracted from the SEER database. This model is trained with the existing data of breast cancer patients and could be used by doctors to determine the patient's survival time ranging in months. The step by step process of developing a predictive model is discussed in the following subsections. However, before developing the predictive model, it is essential to finalize on the technology to create the model. For this research IBM SPSS Modeler 18.1 was used. The underlying rationale for selecting IBM SPSS Modeler is discussed next.

3.3.1 SPSS Modeler

Many of the studies discussed in the literature [5, 18], have used WEKA [58] as the underlying software to design their predictive models. However, with the growing popularity of IBM's SPSS Modeler [88] and SAS Enterprise Miner [89], an inclination towards the commercially available tools over the open source ones used earlier was imminent. Amongst the two, IBM's SPSS Modeler is selected for this research. The rationale for this decision was two-fold:

- Base this research by designing a predictive model of a different, reliable platform to increase the probability of any differences one could observe with the techniques used to design the predictive models, and
- The reliability of the brand name of IBM, and the assurance of adequate technical support, along with the detailed and publicly available documentation which helped address many of the primary concerns.

IBM SPSS Modeler is a software package used for building predictive models using advanced algorithms and data mining techniques, such as decision trees (C&RT, QUEST, CHAID), neural networks, linear regression, and support vector machines. SPSS Modeler 18.1 is used for this research and has the following features:

- It has a highly interactive and user-friendly interface.
- SPSS Modeler can do data preparation for the user and has the capability of automated preparation of raw data via the ‘Automated Data Preparation’ (ADP) node.
- Modeling nodes such as ‘Auto Classifier’, ‘Auto Numeric’ and Auto Cluster’ are powerful techniques that can compare several modelling methods and rank them in effective order.
- It helps extract values from a variety of data including structured data, unstructured data, and that from other sources such as a database, variable file, statistics file, IBM Cognos BI [90] and SAS file.

Once IBM SPSS Modeler was finalized as the platform to create the predictive model, the next task was to put it to use and commence designing the predictive model. Although the IBM documentation guides creating custom models, the content on a few occasions was ambiguous and the documentation, in general, was verbose. Broadly, the following essential tasks were identified to accomplish the predictive model:

- Preparing the data for modeling
- Determining input and target variables of breast cancer predictive model
- Selection of modeling techniques for developing the predictive model
- Training the predictive model
- Testing the predictive model

These tasks are described in the following sections.

3.3.2 Data Preparation

The SEER data is available from 1973-2013 which was divided into two datasets, one for training and the other for validating/testing. The 1988-2003 dataset is selected for training the model due to the following reasons:

- The model needs to be trained on one particular dataset and then tested and validated on different datasets, i.e. data outside trained dataset.
- The training dataset should have data for all shortlisted variable.
- The range of the number of years to predict survivability in this research is arbitrarily set at 10 years. Since the follow-up cut-off date for selected SEER data is December 31, 2013, the cases registered in 2003 or before are considered.

Thus, the dataset from 1988-2003 is selected for training the predictive model and the dataset for the year 2004 is used for testing and determining the predictive model's accuracy.

3.3.3 Determining Input Variables

Post data preparation, the subsequent task involved shortlisting relevant variables which have predictive power. The independent or target variable's relation with the input variables determines the power of the predictive model. The first screening process of narrowing down 30 relevant variables out of 134 total variables played an instrumental role in this process. The shortlisted variables were relevant if they are the best fit or have predictive power. The target or outcome variable is 'survival months' which is the dependent variable. The remaining 29 variables are independent variables and would be checked if they have a relationship with the dependent variable, i.e. 'survival months'. Another important factor required to take into consideration is to select input variables which are available for the selected period (1988-2003).

Feature selection modeling technique is used for shortlisting relevant input variables. This technique is used during the preliminary stages of analysis to locate variables that are most likely to be of interest. The feature selection consists of three steps: Screening, Ranking and Selecting. The ‘Feature Selection’ node (Figure 12) is configured to find the rankings of all input variables, i.e. important, marginal and unimportant. Variables not available for training period are removed from the list of input variables because there is no data for those variables to train the model.

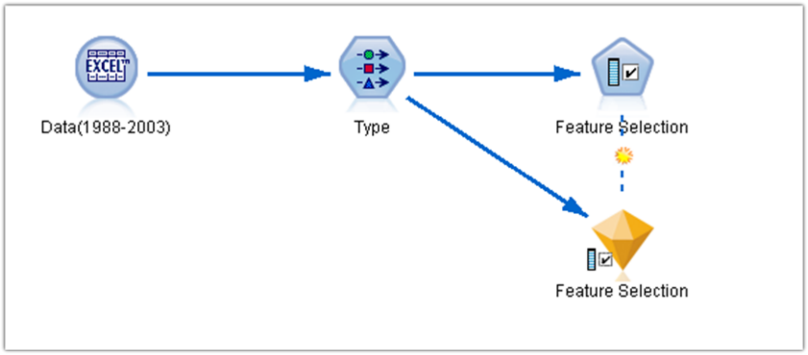


Figure 12. Feature Selection Model Snapshot

The ‘Feature Selection’ model nugget filters out nine variables, of the 30 total variables, as unimportant. The remaining 21 variables are ranked by their importance. By cutting down the number of fields in the model, scoring time and amount of data collected in future iterations can be reduced [88]. The list of variables according to importance ranking are shown in Table 6. The detailed definitions and coding of these variables are listed in Appendix B.

Input variables:
1. Marital Status
2. Race/ethnicity
3. Age recode
4. Laterality
5. Histologic Type ICD-O-3
6. Behavior code ICD-O-3
7. Regional nodes positive
8. Regional nodes examined
9. Reason no cancer-directed surgery
10. Radiation
11. Radiation sequence with surgery
12. Surgery of Primary Site
13. Vital Status recode
14. Estrogen Receptor Status
15. Progesterone Receptor Status
16. T value
17. N value
18. M value
19. Year/Month of diagnosis
Target variables:
20. Survival months
Record ID (unique identifier):
21. Patient ID

Table 6. List of Variables (Input & Target) for Predictive Modeling

3.3.4 Selecting Modeling Techniques

The predictive model is developed using modeling technique(s) which are based on the use of algorithms. There are three modeling technique classes in SPSS Modeler, namely Classification, Association and Segmentation. Some of the examples of modeling techniques in these classes are described in Table 7. Classification models take one or more input fields and can predict one or more target variables. Association models find patterns in the data, where one or more entities are associated with one or more entities. These models allow a variable to act as input and target both. On the other hand, Segmentation models divide the data into clusters that have similar patterns of input variables.

The goal of this research is to predict survival months of patients. Survival months is of continuous (numeric) data type thus the modeling techniques selection is based on the models which allow continuous numeric range target. The classification techniques which support continuous numeric range target include Neural Networks, C&R Tree, CHAID, Linear Regression, Generalized Linear Regression and Support Vector Machines.

Classification	Association	Segmentation (clustering)
Decision Trees: C&R Tree, Quest, CHAID, C5.0	Apriori model Carma model Sequential detection model	Kohnen Networks K-Means clustering Two-step clustering Anomaly detection
Regression: Linear, Logistic, Generalized linear, Cox regression		
Neural Networks		
Support Vector Machines		
Bayesian Networks		

Table 7. Modeling Technique Classes

The predictive model is built using the modeling techniques classes described above. The modeling techniques which complete execution in a reasonable time and have a high correlation of variables were selected. The top three classification techniques selected are Neural Network, CHAID and C&R Tree.. These techniques are discussed in detail in the following sections.

3.3.4.1 Neural Network

Neural Networks work by finding unknown and intricate patterns in the data. They resemble the human brain as it gains knowledge from the learning process. The basic units called neurons are organized in layers. The neurons are connected with different weights, and the network learns from training. There are three layers in the neural network, namely, the input layer, hidden layers and output layer (Figure 13). The input variables are presented to the input layer. The values are propagated from each unit in the hidden layers. The predicted outcome is delivered from the output layer. The network learns by examining each record and predicting target and making the adjustment to the weights if the prediction is incorrect. This is a recursive process and is only stopped if there is a stopping criterion defined before training. The Neural Network models are recommended to use if interpretability is not a priority.

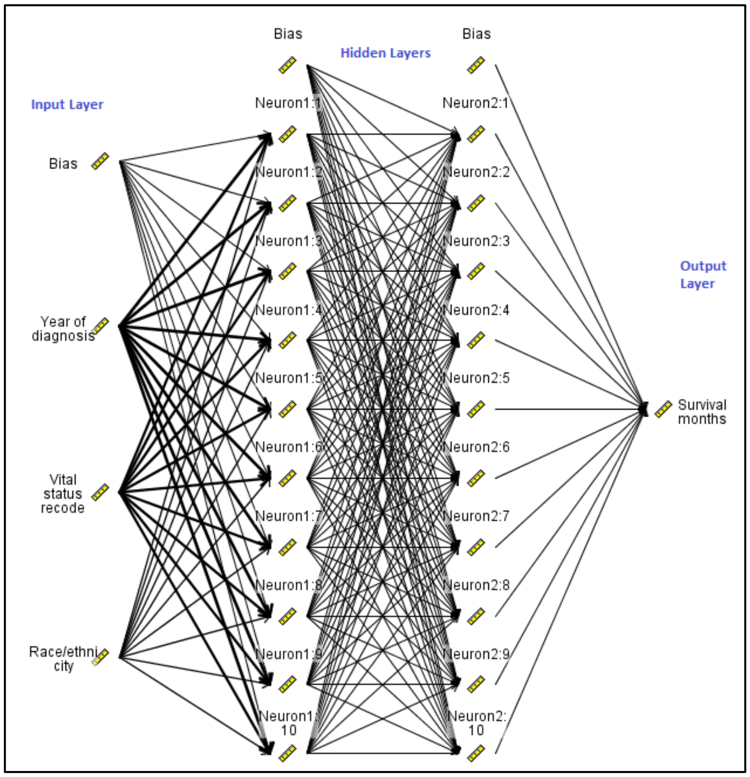


Figure 13. Structure of Neural Network

It is not easy to understand the underlying process of creating a relationship between the target and input variables. There are two types of Neural Network model available in IBM SPSS Modeler: Multilayer Perceptron (MLP) and Radial Basis Function (RBF). MLP is made up of two or more hidden layers. It is a feed-forward, supervised learning network. It is a function of multiple input variables that minimize the prediction error of one or more targets [87]. MLP has higher training and scoring time compared to RBF. Further, RBF has low predictive power as compared to MLP. Training the Neural Network model is a critical step. The model's accuracy is dependent on the training process. Finding suitable model settings for training the Neural Network models is an iterative process. Post data preparation, next step is identifying factors such as the type of Neural Network model (MLP or RBF), structure (number of hidden layers and number of neurons in each layer), stopping rules, training time, training cycles, and ensemble voting if applicable. After training the model, the results are validated and the process is repeated if required. The field requirements are simple as there only must be at least one target value and one input field [88]. Neural Networks deal with the missing values with these two options:

- Records with missing values are excluded.
- Missing values are imputed – Continuous fields impute the average value of minimum, and maximum values observed.

3.3.4.2 Chi-squared Automatic Interaction Detection

Chi-squared Automatic Interaction Detection (CHAID) is a classification scheme for building a decision tree by using chi-square statistics. Decision tree models predict future outcomes based on a set of decision rules. Decision trees work best with categorical data elements such as surgery versus non-surgery treatment, married versus unmarried patients, and types of nodes involved. The decision trees also relate predictions to the values of continuous variables. CHAID uses statistical tests as criteria for evaluating the predictors (input variables). The unique feature of CHAID is that it groups variables that are statistically similar to the target variable. It also maintains a group of variables that are statistically dissimilar. Taking the similar ones into consideration, CHAID finds the best predictor to create first branch of the tree. This tree has child nodes which also fall under the statistically similar variables group. The tree is completed by continuing this process. The statistical test for grouping similar variables is done by using F-test for the continuous target and chi-squared test for the categorical target. CHAID not being a binary tree method produces two or more groupings at all levels of the tree. The CHAID decision trees are wider than binary tree methods. The advantages of using CHAID are:

- It works for all types of variables.
- It accepts both frequency and case-weighted variables.
- It can handle missing values by treating them as a single category.

The predictions in CHAID are made by following the terminal node of the tree which has a specific predicted value associated with it. For numeric target, the terminal node's predicted category is calculated as the weighted mean of the target values for records in the node. CHAID works for all types of inputs. Target and inputs can be continuous or categorical [88].

3.3.4.3 Classification and Regression Tree

The Classification and Regression Tree (C&RT) model is used when there are multiple inputs and one target variable. C&RT groups data into two subsets and repeats the process until homogeneous records are grouped together. The subsets split again and process repeats and stops when stopping rule is applied or homogeneity is achieved totally. C&RT might use the same predictor field at different tree levels. All the splits are binary; thus C&RT is strictly a binary tree. It provides the option first to grow the tree and then prune based on cost-complexity criteria. This technique requires one or more input variables and exactly one target variable. The advantage of using the C&RT modeling node are:

- It does not require long training time.
- It is quite adaptive if data has missing values or has a large number of fields.

The predictions in C&RT are also done by following the tree splits to a terminal node of the tree. The terminal node has a specific predicted value associated with it. For numeric target, the terminal node's predicted category is calculated as the weighted mean of the target values for records in the node. Both target and input variables can be continuous or categorical [88].

3.3.4.4 Ensemble

Ensemble modeling technique combines multiple individual models to provide better prediction accuracy. In literature, it is observed that the researchers use an ensemble of different modeling techniques and predict better outcomes as compared to individual models [5, 18]. By combining the predictions from multiple models, limitations of individual models can be avoided and thereby result in high accuracy overall. Models when combined in this manner perform at least as well as the best of individual models and even better [88].

Some of the difficulties faced when developing the ensemble model are finding out the combination of models to be used. Individual models selected for ensemble should have high prediction accuracy and should not overfit [64]. There are different rules for combining predicted values from individual models to compute ensemble score value. For categorical targets, combining rules available are voting, the highest probability and highest mean probability. For continuous targets, averaging is the only combining rule. Since the target variable is a continuous variable, an ensemble score is computed by averaging the values of individual models. The ensemble prediction is calculated by using the following formula (Equation 1):

$$\hat{y}_{i,M} = \frac{1}{M}\sum_{m=1}^{M} \hat{y}_{i,m}$$

Equation 1. Ensemble Prediction Equation

where, $\hat{y}_{i,M}$ is the final predicted value of case *i*

$\hat{y}_{i,m}$ is the mth base model's predicted value for case *i*

A hypothesis is proposed that Ensemble model outperforms the individual models, i.e. Neural Network, C&RT and CHAID models. After selecting the modeling techniques, the predictive model is built. The following sections present the three phases of building a prediction model in detail.

3.3.5 Training Predictive Model

Figure 14 is a snapshot of the training phase of the predictive model built in IBM SPSS Modeler [88]. The model description is as follows:

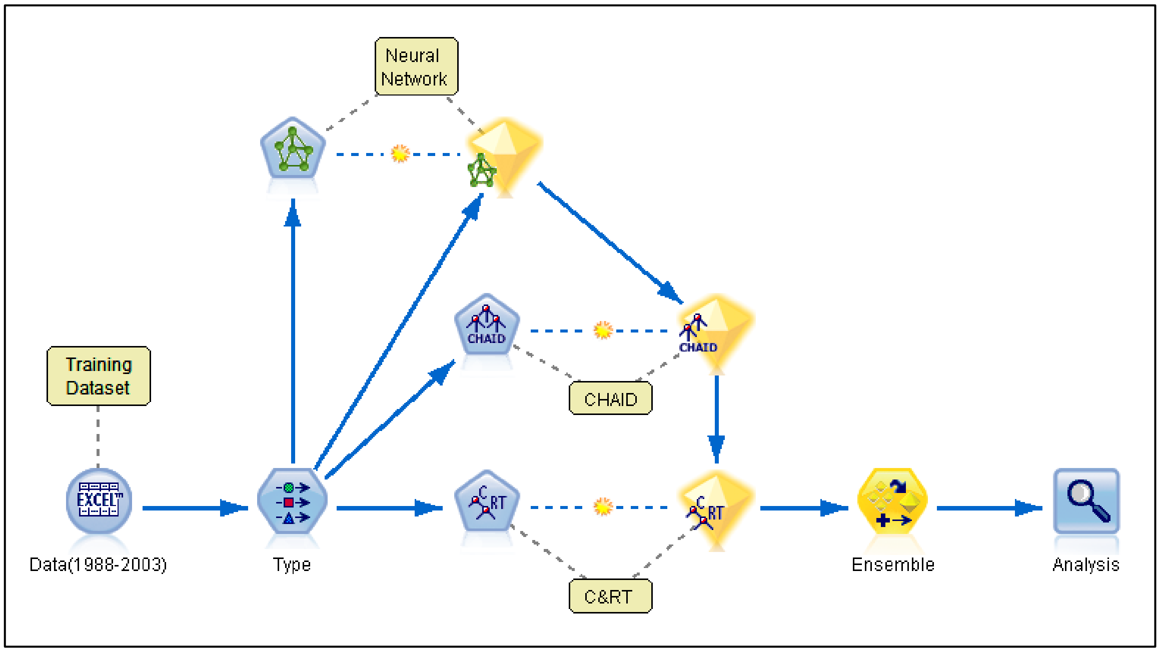
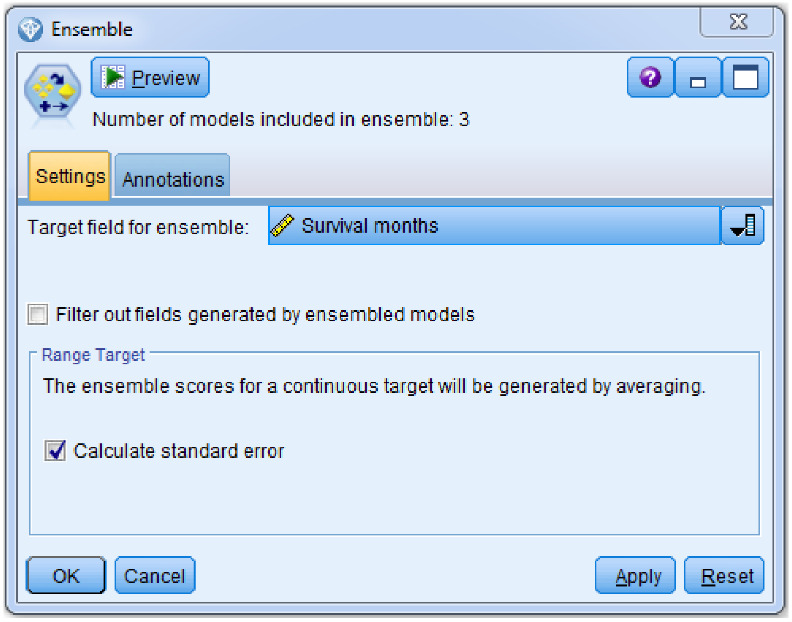


Figure 14. Predictive Model Training snapshot

- **Data (1988-2003)** node is the Excel Source node. Excel Source node allows importing any Excel workbook available on the local machine. The import can be customized by selecting a specific range of cells defined in the Excel worksheet or by selecting a specific worksheet from the entire workbook. The variable names can be changed in Source node settings. The Excel worksheet imported for training the model has SEER extracted breast cancer data from 1988-2003.
- **Type node:** The Source node is connected to a Type node which defines the measurement level for each variable such as Nominal, Ordinal, Continous, Categorical, Flag or Typeless.

Type node also defines the role of each input field such as Input, Target, Both (Input & Target), None, Partition, Split, Frequency, and Record ID. Input fields are the predictors and Target is the field that the model has to predict. In the model, there exist a total 21 fields, out of which there are 19 predictors (input), 1 Target (Survival months) and 1 Record ID (Patient ID).

- **Modeling nodes:** The Type node is connected to three modeling nodes - Neural Network, CHAID and C&R Tree. The selected modeling nodes are classification models which use one or more predictors to predict the target. Each modeling node has field option where variables are specified as input and target. However, since Type node is used, there is no need to address the field specification in modeling nodes.
- When the model completes its execution, the resulting nuggets are added to stream for each modeling node (Figure 14). The nuggets contain complete information of the model (rules and equations developed) and accuracy of the independent model formulated by IBM SPSS Modeler. The model summary can be browsed by double-clicking the generated nuggets.
- **Ensemble node** is added to the stream to create an ensemble of these techniques. For this purpose, the modeling nuggets are connected to the ensemble node as shown in Figure 14. The ensemble model provides minimal options such as selecting a target field for ensemble, filtering out the fields generated by ensemble models and calculating standard error. Figure 15 shows Ensemble node settings, ‘Survival Months’ is selected as the target and the option ‘Filter out fields generated by ensemble models’ is unselected to get individual models’ prediction along with ensemble. An option for calculating the standard error was also chosen.



Number of models included in ensemble: 3

Target field for ensemble: Survival months

☐ Filter out fields generated by ensembled models

Range Target

The ensemble scores for a continuous target will be generated by averaging.

☑ Calculate standard error

Figure 15. Ensemble Node Setting

On completing execution, each modeling technique node creates nuggets (diamond-shaped) except for the Ensemble which is a combining rule model. Each nugget has the model summary which displays a number of variables used, predictor importance, stopping rules, and number of layers. The nugget summary also has the training accuracy of each modeling technique. The following training accuracies were observed for selected modeling techniques of the predictive model:

	Modeling Techniques		
	Neural Network	CHAID	C&RT
Training Accuracy	82.9%	82.1%	82.0%

Table 8. Training Accuracy

In the Analysis node, training and actual outcomes are analyzed for the individual model as well as Ensemble model. The statistical measure used to compare is mean, minimum, maximum, mean absolute error and standard deviation. Once, the model has trained it is now ready to be tested and validated. The next section discusses testing and validation methods used in this research.

3.3.6 Testing and Validation of Predictive Model

Figure 16 is a snapshot of the testing phase of the developed predictive model. The model is now tested for the cases outside its training range, that is, cases registered in 2004. The source node now represents the testing dataset, i.e. Data (2004). The output of individual models and ensemble model can be captured in different ways:

- Table - a static matrix window generated in IBM SPSS Modeler
- Analysis node - this node allows testing the measured (predicted) values against the known result
- Excel node - it generates an Excel sheet with predicted values
- The additional nodes used here are Transpose node to transpose columns into rows, and Type node to extract data into Excel sheet.

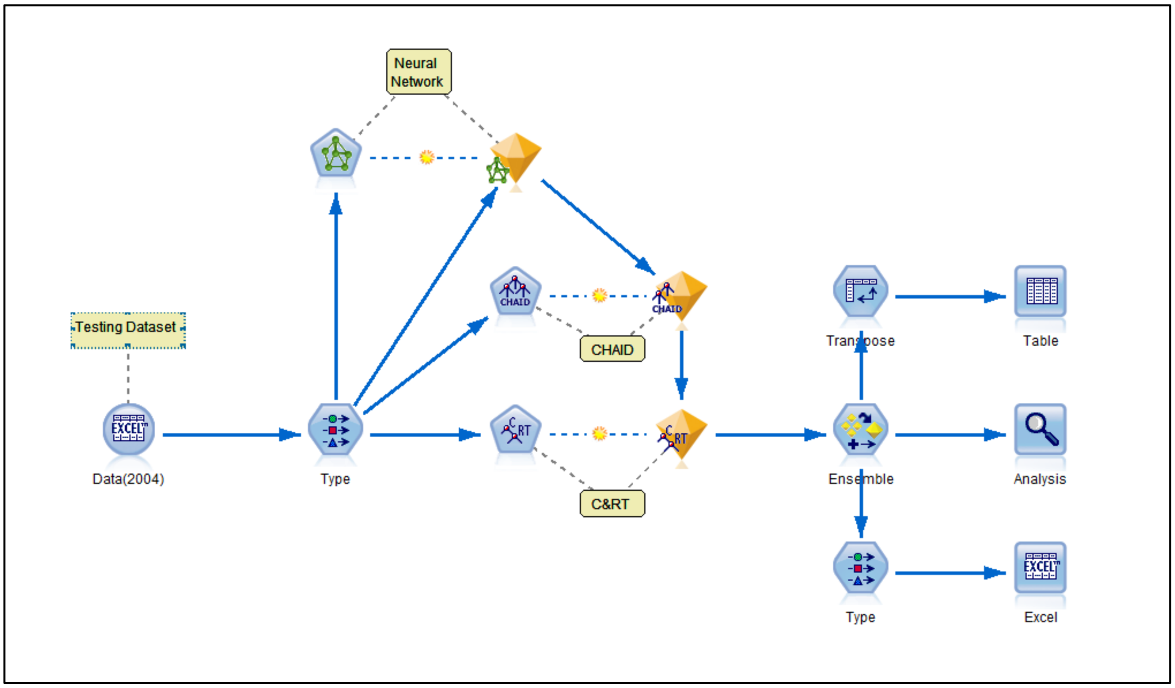


Figure 16. Predictive Model Testing Snapshot

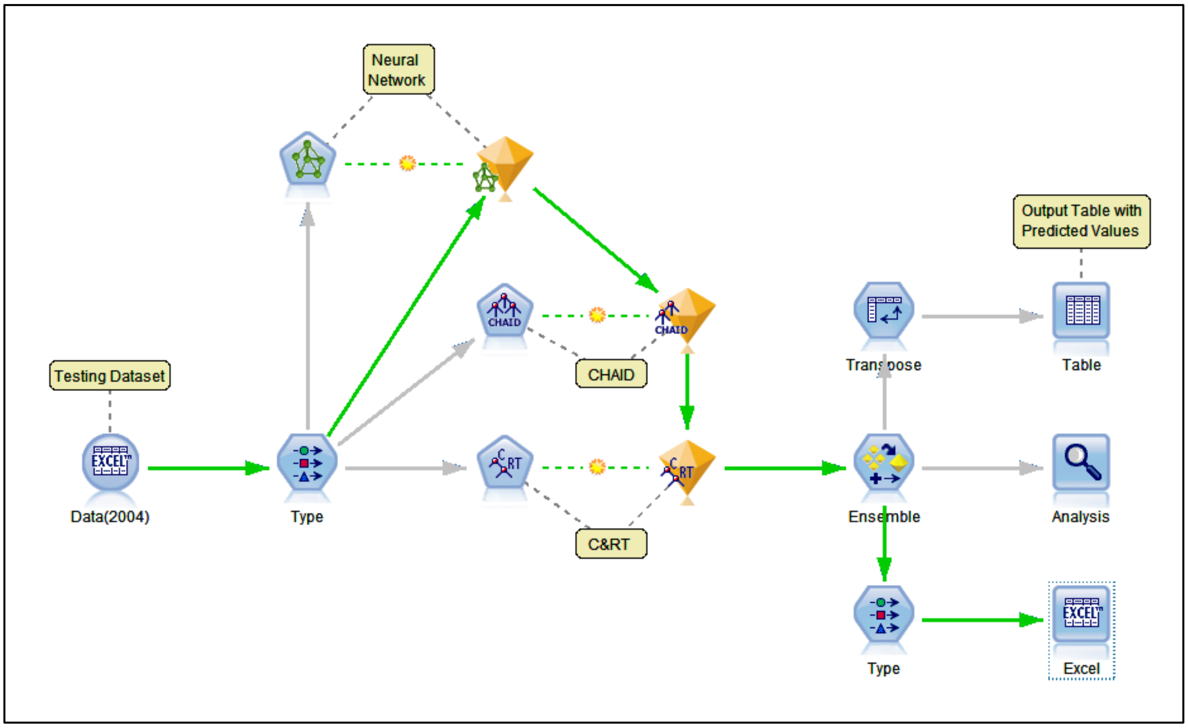


Figure 17. Execution of Excel Output Node

For testing the model with all cases diagnosed in 2004 together, Excel output node is executed (as shown in Figure 17) which generates an Excel sheet with predicted outcomes for each record. This newly generated Excel sheet is then used to compare measured values with actual values (survival months) to validate the accuracy of the predictive model. The results of the validation of the model are discussed in the next chapter.

The method presented above works well for testing a large number of cases together. However, if the user wants to predict survival months for one individual case, they could use the User Input to enter values of all variables and use the same predictive model as a calculator. Figure 18 is a snapshot of using the predictive model as a calculator with the User Input node instead of the source node. Figure 19 shows the User Input node window. The definitions and coding of each variable are attached in Appendix B.

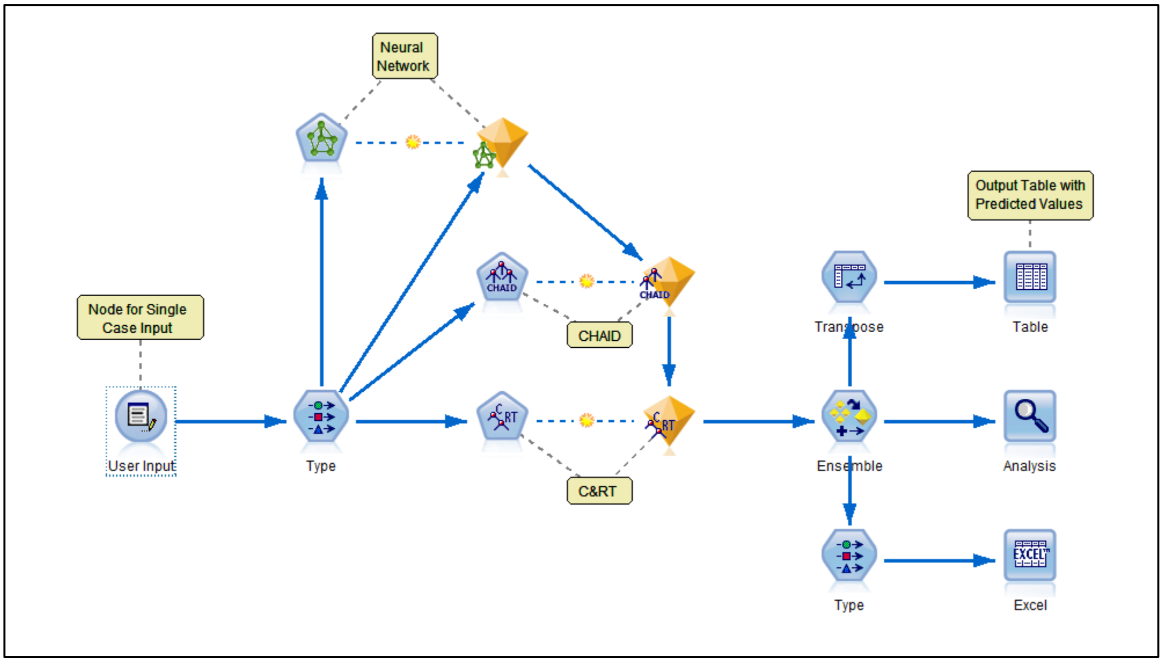
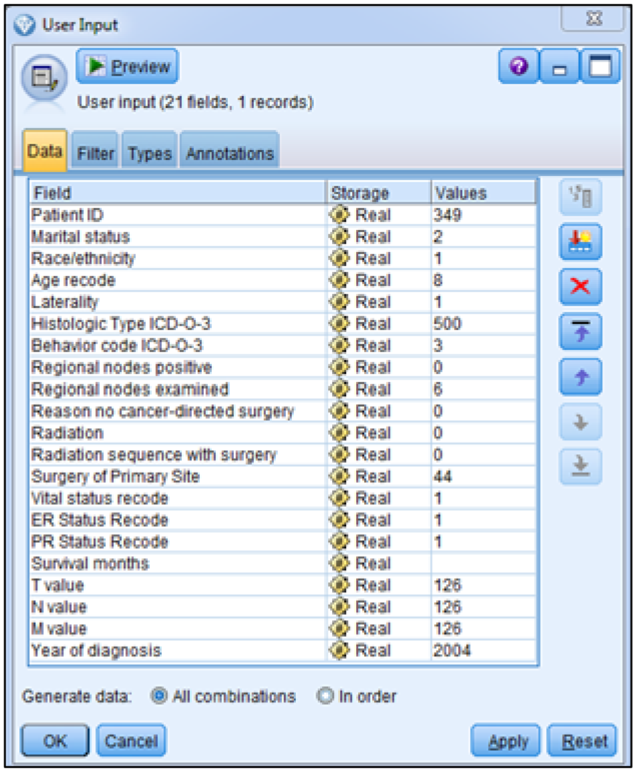


Figure 18. Predictive Model as calculator for individual case



User input (21 fields, 1 records)

Field	Storage	Values
Patient ID	Real	349
Marital status	Real	2
Race/ethnicity	Real	1
Age recode	Real	8
Laterality	Real	1
Histologic Type ICD-O-3	Real	500
Behavior code ICD-O-3	Real	3
Regional nodes positive	Real	0
Regional nodes examined	Real	6
Reason no cancer-directed surgery	Real	0
Radiation	Real	0
Radiation sequence with surgery	Real	0
Surgery of Primary Site	Real	44
Vital status recode	Real	1
ER Status Recode	Real	1
PR Status Recode	Real	1
Survival months	Real	
T value	Real	126
N value	Real	126
M value	Real	126
Year of diagnosis	Real	2004

Generate data: ● All combinations ○ In order

Figure 19. User Input Node Snapshot

3.4 Summary

In this chapter, a detailed description of the five main tasks, namely, data extraction, data analysis, data visualization, predictive modeling, and evaluation of predictive model has been presented. These tasks collectively contribute to accomplishing the implementation of the predictive model along with visualization dashboard. A comprehensive understanding of several predictive modeling techniques along with the rationale behind choosing the right tools, technologies, and approaches to accomplish the primary contributions have also been presented. The experiments and analysis of results of both dashboard and predictive model are presented in the next chapter.

4. Experiments and Results

In this chapter, the experimental results of data analysis and predictive techniques are presented. The first section demonstrates the data analysis and visualization reports and dashboard representing forty years of breast cancer data. The second section demonstrates the accuracy of the predictive model and comparison study of predictive model's accuracy on different datasets.

4.1 Data Analysis and Visualization

For visual analytics, the pre-processed SEER data is first imported to SQL Server to create 'Cancer' database in a format suitable for analysis. Tableau is connected to the Cancer database, and the tables are joined to create a virtual table, extending horizontally by adding columns of data needed. The data is further cleaned up (by changing data types, renaming & resetting fields, wherever needed) to prepare data for analysis. Data is analyzed by adding new calculated fields, formulas, grouping and sets by writing SQL queries. The data is sliced and diced by using filters on measures and dimensions and creating parameters. The data is then visualized using workbook, dashboards and stories. The dashboard and dynamic reports use the views and tables created in the SQL Server database.

The subsequent sections present each of the reports generated using Tableau [75]. Tableau story [91] and story points are used for visual analytics of breast cancer data. Tableau story is a sequence of visualizations that work together to convey the findings. Each sub-report is also called a story-point. The top-level dashboard provides an overview of the KPIs and includes navigation controls including a tab panel (Figure 20) which allows switching between breast cancer metastasis, TNM system, cases by geo-mapping, geo-mapping by race/ethnicity, cases by race and age range, lymph node involvement, incidence/mortality, survival/mortality and anatomy dashboards.

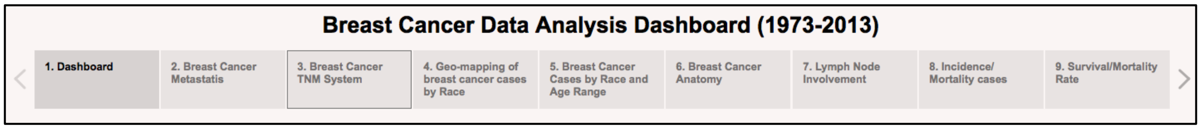
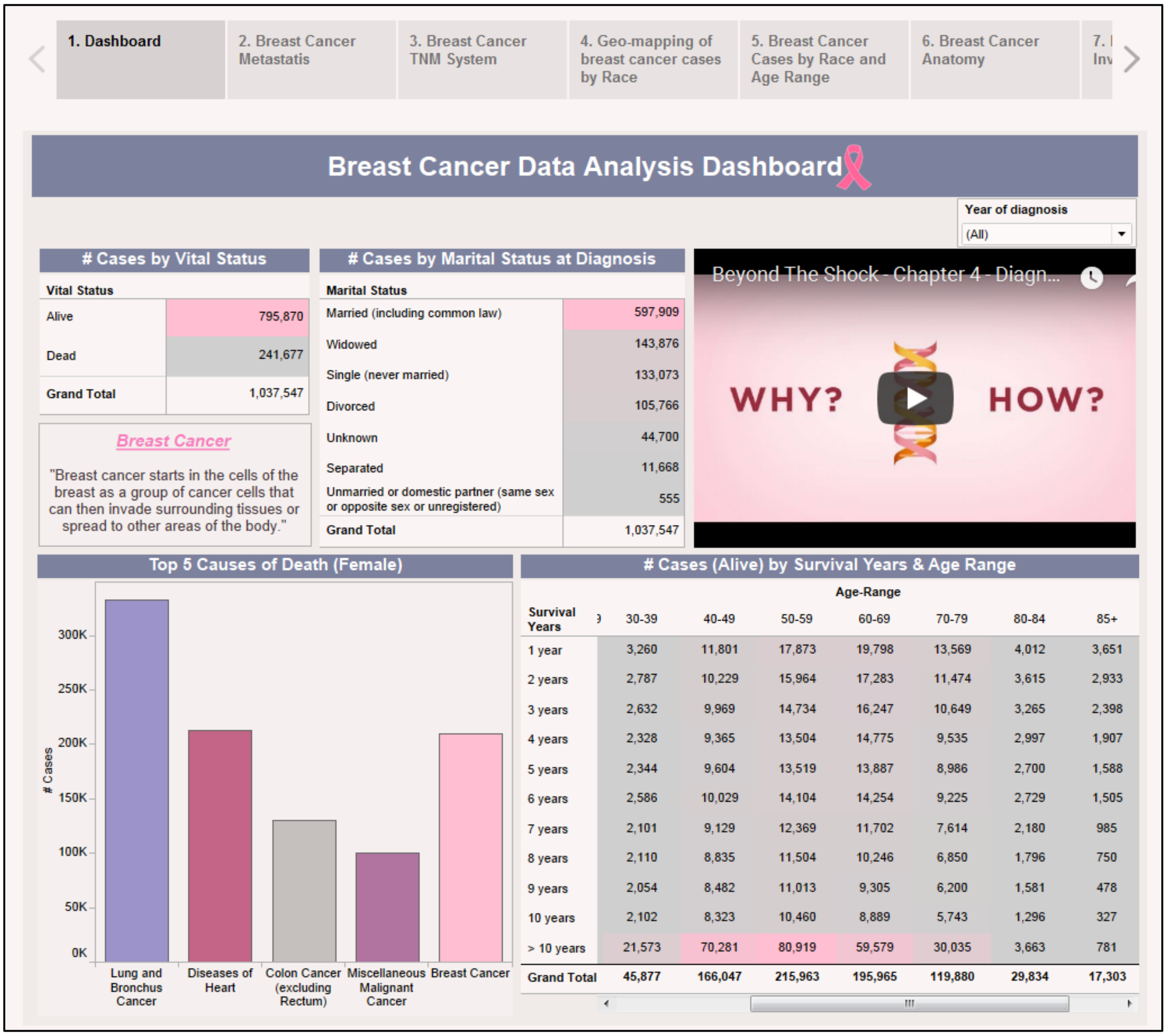

Breast Cancer Data Analysis Dashboard (1973-2013)
1. Dashboard
2. Breast Cancer Metastatis
3. Breast Cancer TNM System
4. Geo-mapping of breast cancer cases by Race
5. Breast Cancer Cases by Race and Age Range
6. Breast Cancer Anatomy
7. Lymph Node Involvement
8. Incidence/ Mortality cases
9. Survival/Mortality Rate

Figure 20. Breast Cancer Dashboard Story Point Panel

4.1.1 Breast Cancer Survivability Dashboard


1. Dashboard
2. Breast Cancer Metastatis
3. Breast Cancer TNM System
4. Geo-mapping of breast cancer cases by Race
5. Breast Cancer Cases by Race and Age Range
6. Breast Cancer Anatomy
7. I Inv
Breast Cancer Data Analysis Dashboard
Year of diagnosis
(All)
Beyond The Shock - Chapter 4 - Diagn...
WHY?
HOW?

Breast Cancer Data Analysis Dashboard

Cases by Vital Status

Vital Status	
Alive	795,870
Dead	241,677
Grand Total	1,037,547

Breast Cancer

"Breast cancer starts in the cells of the breast as a group of cancer cells that can then invade surrounding tissues or spread to other areas of the body."

Cases by Marital Status at Diagnosis

Marital Status	
Married (including common law)	597,909
Widowed	143,876
Single (never married)	133,073
Divorced	105,766
Unknown	44,700
Separated	11,668
Unmarried or domestic partner (same sex or opposite sex or unregistered)	555
Grand Total	1,037,547

Top 5 Causes of Death (Female)

Cases (Alive) by Survival Years & Age Range

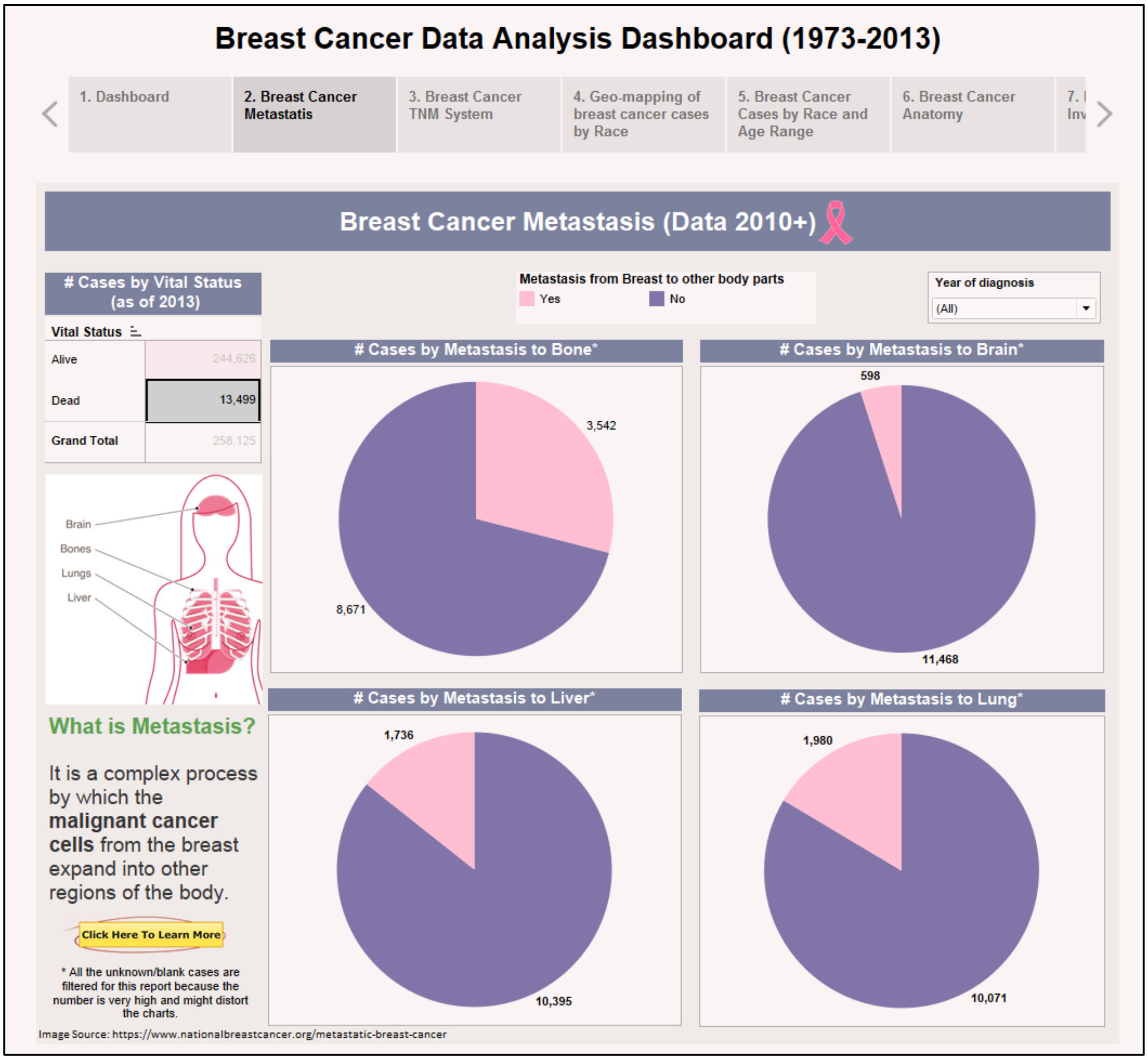
	Age-Range						
Survival Years	30-39	40-49	50-59	60-69	70-79	80-84	85+
1 year	3,260	11,801	17,873	19,798	13,569	4,012	3,651
2 years	2,787	10,229	15,964	17,283	11,474	3,615	2,933
3 years	2,632	9,969	14,734	16,247	10,649	3,265	2,398
4 years	2,328	9,365	13,504	14,775	9,535	2,997	1,907
5 years	2,344	9,604	13,519	13,887	8,986	2,700	1,588
6 years	2,586	10,029	14,104	14,254	9,225	2,729	1,505
7 years	2,101	9,129	12,369	11,702	7,614	2,180	985
8 years	2,110	8,835	11,504	10,246	6,850	1,796	750
9 years	2,054	8,482	11,013	9,305	6,200	1,581	478
10 years	2,102	8,323	10,460	8,889	5,743	1,296	327
> 10 years	21,573	70,281	80,919	59,579	30,035	3,663	781
Grand Total	45,877	166,047	215,963	195,965	119,880	29,834	17,303

Figure 21. Breast Cancer Survivability Dashboard

The main dashboard (Figure 21) contains five sub-reports/story points which can be filtered by year of diagnosis (1973-2013). The first table shows the total number of cases by vital stats (i.e. Alive or Dead). Out of the total of 1,037,457 cases registered in SEER over the forty-year period, 241,677 cases died due to breast cancer and rest were tagged Alive at the study cut-off date (i.e. December 31, 2013). Amongst all the causes of death in females (including cancer and non-cancer related deaths), breast cancer is the third highest cause of death after lung-bronchus cancer and diseases of the heart. For registered breast cancer cases, a similar kind of categorization is represented by marital status at diagnosis. The analysis shows that the majority of women (597,909 cases) were married at the time of breast cancer diagnosis. This is followed by widowed, single (never married), divorced, separated and unmarried in descending order. There are about 44,700 cases whose marital status was unknown. All the cases are further categorized by age-range and survival years. For this table, only cases that are tagged Alive as of cut-off date are used. Out of total ‘Alive’ tagged cases, a majority of the patients who survived more than 10 years are in age-range of 50-59, followed by 40-49 and 60-69 years. Brandt et al. [92] also concluded that women aged under 40 and above 80 years at the time of diagnosis have a poor survival rate independent of any factors. This is consistent with observation made by the dashboard.

4.1.2 Breast Cancer Metastasis

The Breast Cancer Metastasis (Figure 22) dashboard consists of five sub-reports which can be filtered by year of diagnosis and alive cases or cases who died. The data for metastasis is only available from 2010 onwards. The vital stats table shows that out of a total of 258,125 cases registered with metastasis, 13,499 patients had died due to breast cancer and rest were alive at the study cut-off date. The four pie-charts display information regarding metastasis to bone, brain, liver and lung. Cases are categorized based on whether the breast cancer metastasized, no metastasis and unknown. The number of unknown cases is low and hence is not included in the pie-charts. It is noted that the majority of the cases who died had their breast cancer metastasized to the brain, followed by lung, liver and brain, respectively. Similarly, 'Alive' cases can be selected to observe the impact of metastasis.



Breast Cancer Data Analysis Dashboard (1973-2013)

1. Dashboard
2. Breast Cancer Metastatis
3. Breast Cancer TNM System
4. Geo-mapping of breast cancer cases by Race
5. Breast Cancer Cases by Race and Age Range
6. Breast Cancer Anatomy
7. I Inv

Breast Cancer Metastasis (Data 2010+)

Cases by Vital Status (as of 2013)

Vital Status	
Alive	244,626
Dead	13,499
Grand Total	258,125

Metastasis from Breast to other body parts
Yes No

Year of diagnosis: (All)

# Cases by Metastasis to Bone*
3,542
8,671
# Cases by Metastasis to Brain*
598
11,468
# Cases by Metastasis to Liver*
1,736
10,395
# Cases by Metastasis to Lung*
1,980
10,071
Brain
Bones
Lungs
Liver
Click Here To Learn More

What is Metastasis?

It is a complex process by which the **malignant cancer cells** from the breast expand into other regions of the body.

* All the unknown/blank cases are filtered for this report because the number is very high and might distort the charts.

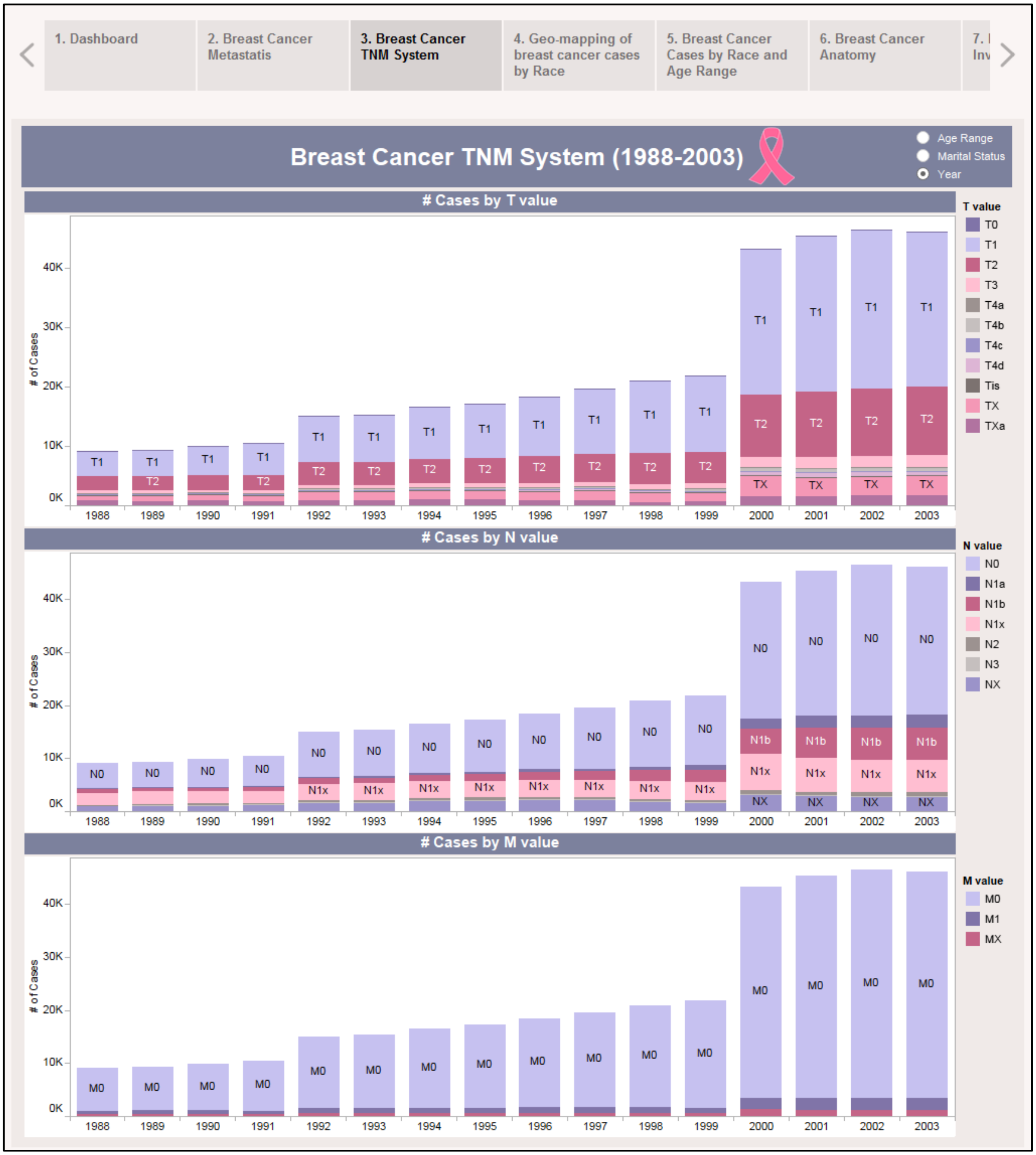
Image Source: https://www.nationalbreastcancer.org/metastatic-breast-cancer

Figure 22. Breast Cancer Metastasis (Data 2010+)

4.1.3 Breast Cancer TNM System

The Breast Cancer TNM System dashboard (Figure 23) consists of three sub-reports categorized by T, N, M values for cases registered from 1988 to 2003. The bar graph views can be switched by age-range, marital status and year of diagnosis by using the radio buttons on the top-right of the dashboard. A massive jump in the number of cases registered is observed in 1992 and 2000. This is because new SEER registries were added to the SEER database in these two years [93]. Figure 23 consists of three sub reports which present cases by T value (size of the original tumor), N value (degree of nearby lymph nodes involved) and M value (presence of distant metastasis). Each value type has different categories such as T0, T1, TX, N0, N1, Nx, M0, and M1. Each point in the charts has tooltips to display more information and can be clicked to open relevant web pages.

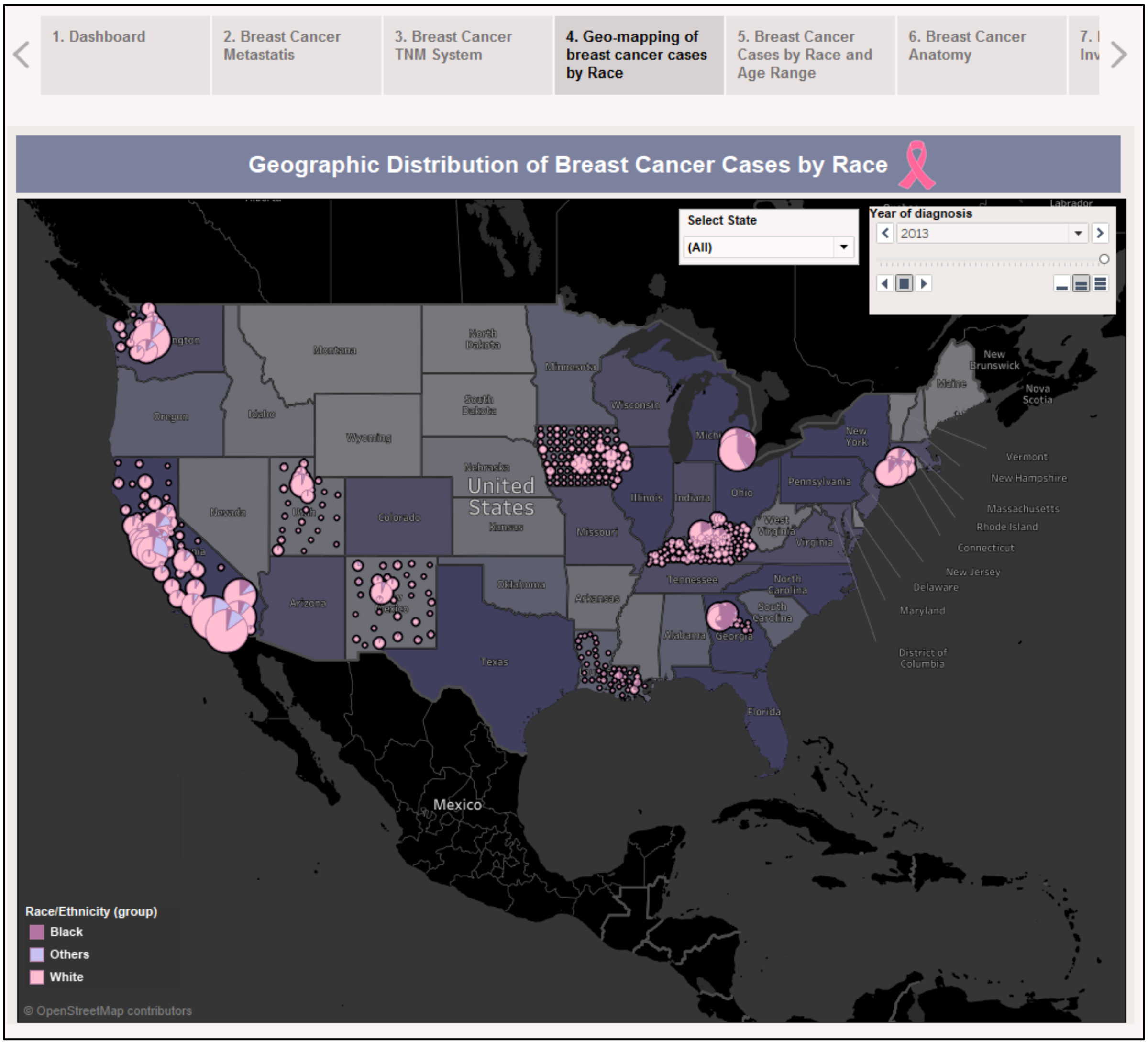
The dashboard shows that for all cases diagnosed in 1988-2003, the maximum number of cases have T1 value (i.e. one primary tumor) followed by the T2 value (i.e. two primary tumors). On the other hand, the maximum number of cases did not have any nearby lymph nodes containing cancer (N0 value) and no distant cancer metastasis was found (M0 value). When the view is switched, and cases are categorized by age range, the dashboard shows that 50-59 age range, with maximum number of cases, 58.28% of cases have T1 value, followed by T2 (24.04%), TX, Txa, T3, T4d, T4b, T4a, T0, Tis, and T4c. The N value charts show that for the 50-59 age range, 60.46% cases have N0 value followed by N1x (16.20%), N1b, NX, N1a and N2. The M value chart shows 93.29% of cases in the 50-59 age range have M0 value, i.e. no distant metastasis. Similar trends are observed when the view is switched to marital status.



○ Age Range
○ Marital Status
● Year

Figure 23. Breast Cancer TNM System

4.1.4 Geographic Distribution of Breast Cancer Cases by Race



Geographic Distribution of Breast Cancer Cases by Race

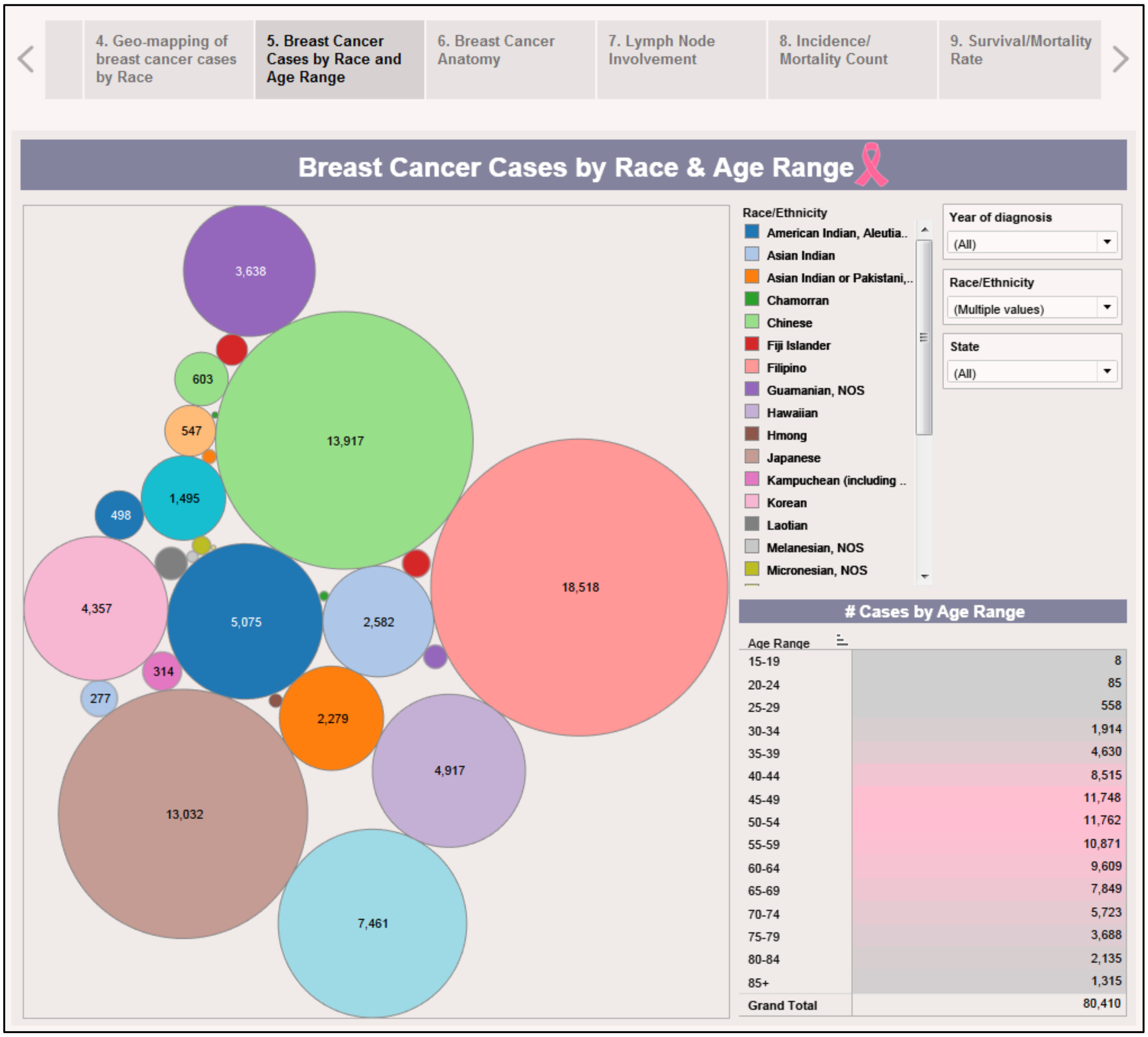
Select State: (All)

Year of diagnosis: 2013

Figure 24. Geographic Distribution of Breast Cancer Cases by Race

The Geo-mapping by Race/Ethnicity dashboard (Figure 24) is a dynamic map showing the case concentration and changes over the years. The pie-charts categorize the cases by Race/Ethnicity: White, Black, and Others, the latter being a clickable option which drills down to display the distribution of cases by all races in data. A "play/pause" feature allows a dynamic display of the changes over the years (1973-2013). The State parameter can be used to filter and select one or more States and observe the trend for selected regions.

4.1.5 Breast Cancer Cases by Race and Age Range


4. Geo-mapping of breast cancer cases by Race | 5. Breast Cancer Cases by Race and Age Range | 6. Breast Cancer Anatomy | 7. Lymph Node Involvement | 8. Incidence/ Mortality Count | 9. Survival/Mortality Rate

Breast Cancer Cases by Race & Age Range

Race/Ethnicity
- 🟦 American Indian, Aleutia..
- 🟦 Asian Indian
- 🟧 Asian Indian or Pakistani,..
- 🟩 Chamorran
- 🟩 Chinese
- 🟥 Fiji Islander
- 🟥 Filipino
- 🟪 Guamanian, NOS
- 🟪 Hawaiian
- 🟫 Hmong
- 🟫 Japanese
- 🟪 Kampuchean (including ..
- 🟥 Korean
- ⬛ Laotian
- ⬜ Melanesian, NOS
- 🟨 Micronesian, NOS

Year of diagnosis: (All)

Race/Ethnicity: (Multiple values)

State: (All)

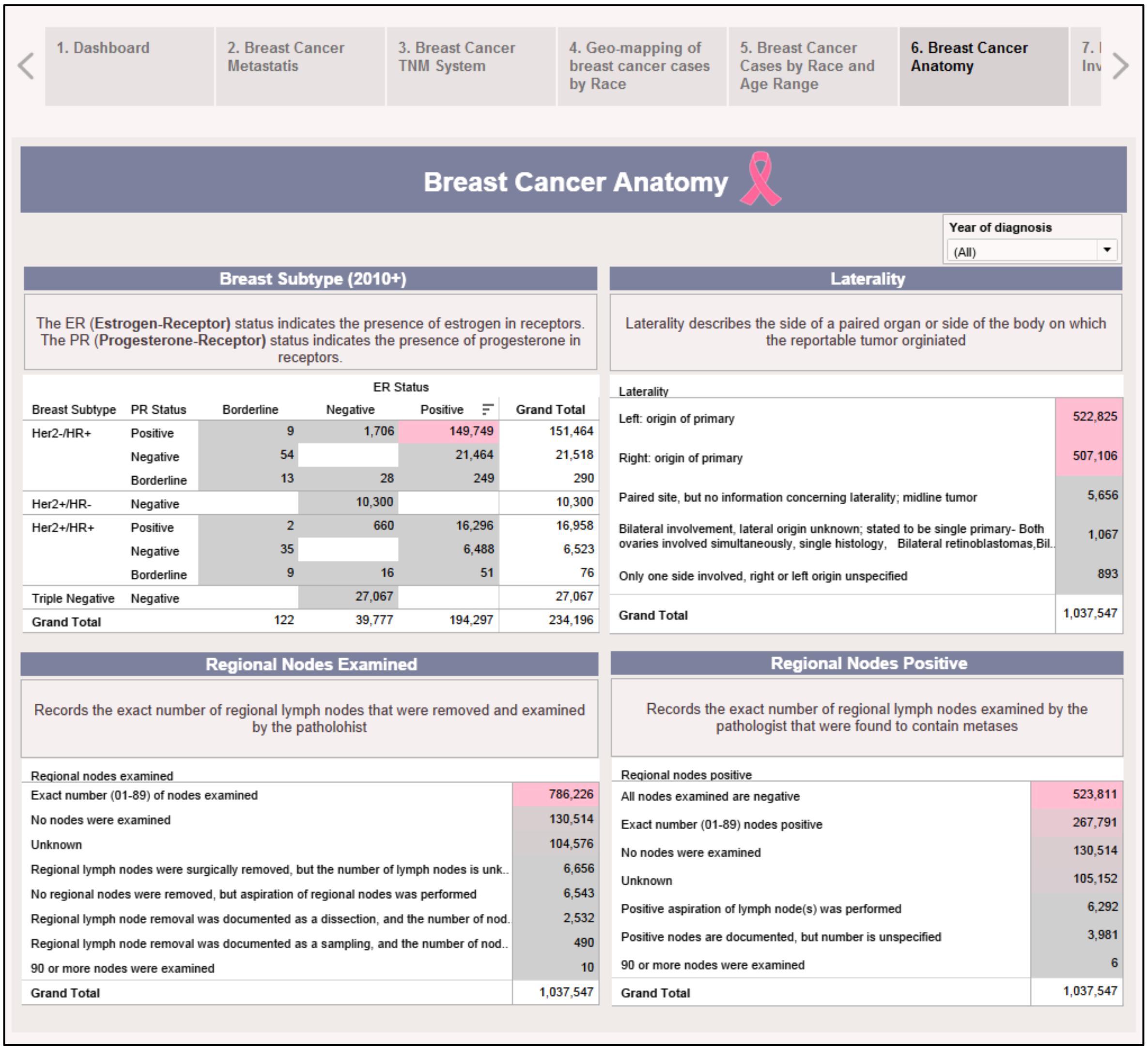
# Cases by Age Range

Age Range	
15-19	8
20-24	85
25-29	558
30-34	1,914
35-39	4,630
40-44	8,515
45-49	11,748
50-54	11,762
55-59	10,871
60-64	9,609
65-69	7,849
70-74	5,723
75-79	3,688
80-84	2,135
85+	1,315
Grand Total	80,410

Figure 25. Breast Cancer Cases by Race and Age Range

Breast cancer cases by race and age-range are shown in Figure 25. The bubble chart shows the cases categorized by the races other than white and black race which are filtered from the Geo-mapping dashboard. A tabular breakdown of cases by age range shows that the 50-54 age range has the highest number of cases followed by 45-49 and 55-59. This dashboard can be further filtered by year of diagnosis, race and state parameters.

4.1.6 Breast Cancer Anatomy Dashboard


1. Dashboard
2. Breast Cancer Metastatis
3. Breast Cancer TNM System
4. Geo-mapping of breast cancer cases by Race
5. Breast Cancer Cases by Race and Age Range
6. Breast Cancer Anatomy
7. I Inv

Breast Cancer Anatomy

Year of diagnosis
(All)

Breast Subtype (2010+)

The ER (**Estrogen-Receptor)** status indicates the presence of estrogen in receptors. The PR (**Progesterone-Receptor)** status indicates the presence of progesterone in receptors.

		ER Status			
Breast Subtype	PR Status	Borderline	Negative	Positive	Grand Total
Her2-/HR+	Positive	9	1,706	149,749	151,464
	Negative	54		21,464	21,518
	Borderline	13	28	249	290
Her2+/HR-	Negative		10,300		10,300
Her2+/HR+	Positive	2	660	16,296	16,958
	Negative	35		6,488	6,523
	Borderline	9	16	51	76
Triple Negative	Negative		27,067		27,067
Grand Total		122	39,777	194,297	234,196

Laterality

Laterality describes the side of a paired organ or side of the body on which the reportable tumor orginiated

Laterality	
Left: origin of primary	522,825
Right: origin of primary	507,106
Paired site, but no information concerning laterality; midline tumor	5,656
Bilateral involvement, lateral origin unknown; stated to be single primary- Both ovaries involved simultaneously, single histology, Bilateral retinoblastomas,Bil..	1,067
Only one side involved, right or left origin unspecified	893
Grand Total	1,037,547

Regional Nodes Examined

Records the exact number of regional lymph nodes that were removed and examined by the patholohist

Regional nodes examined	
Exact number (01-89) of nodes examined	786,226
No nodes were examined	130,514
Unknown	104,576
Regional lymph nodes were surgically removed, but the number of lymph nodes is unk..	6,656
No regional nodes were removed, but aspiration of regional nodes was performed	6,543
Regional lymph node removal was documented as a dissection, and the number of nod.	2,532
Regional lymph node removal was documented as a sampling, and the number of nod..	490
90 or more nodes were examined	10
Grand Total	1,037,547

Regional Nodes Positive

Records the exact number of regional lymph nodes examined by the pathologist that were found to contain metases

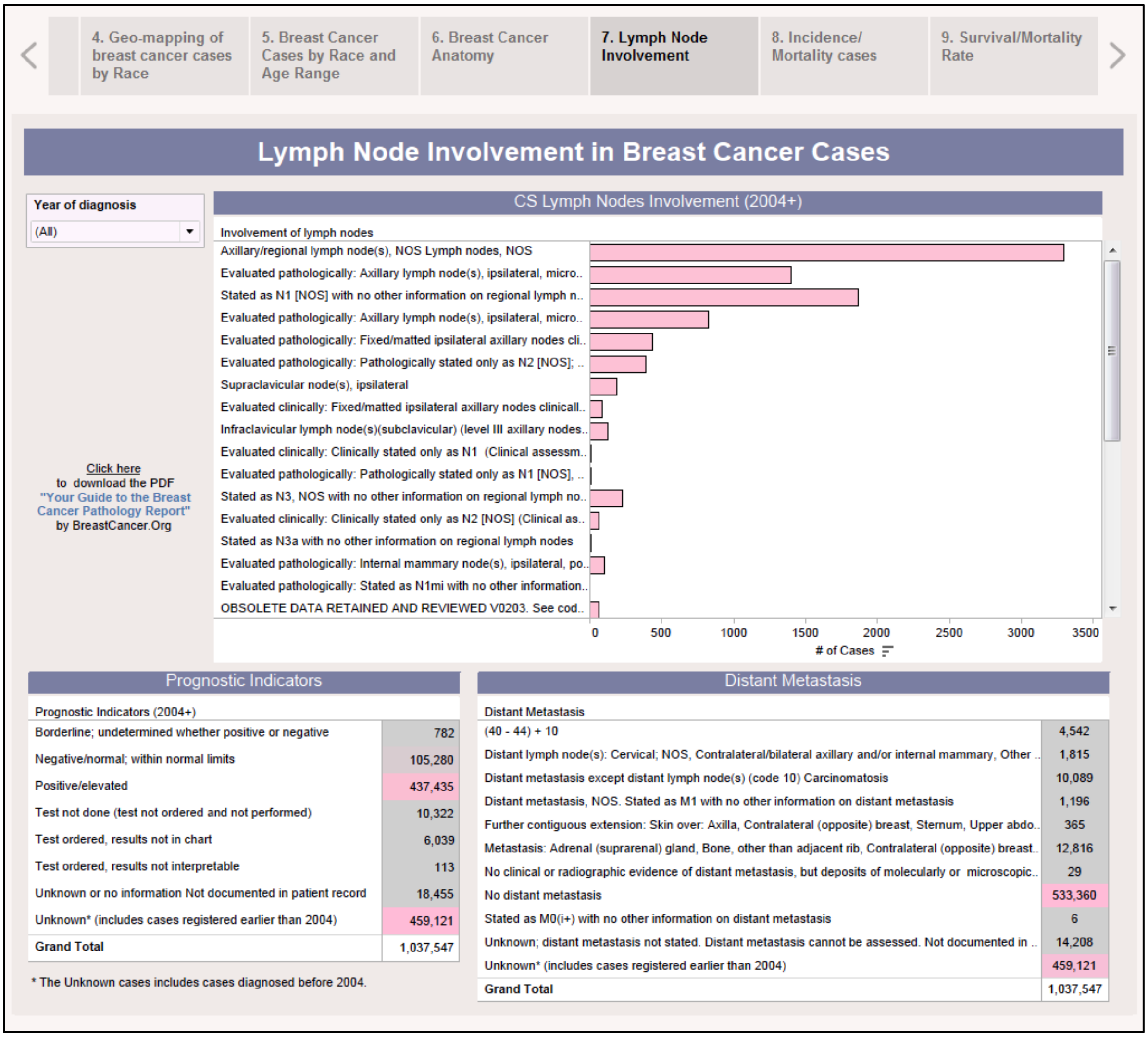
Regional nodes positive	
All nodes examined are negative	523,811
Exact number (01-89) nodes positive	267,791
No nodes were examined	130,514
Unknown	105,152
Positive aspiration of lymph node(s) was performed	6,292
Positive nodes are documented, but number is unspecified	3,981
90 or more nodes were examined	6
Grand Total	1,037,547

Figure 26. Breast Cancer Anatomy

The Breast Cancer Anatomy dashboard (Figure 26) consists of four sub-reports/story points. The data used is for the period 1973-2013 except for breast subtype for which data is only available from 2010 onwards. The tables represent breast cancer cases categorized by laterality, examined and positive regional nodes, and breast subtype. The highlighted cells in each table display the highest number of cases in that category. For example, in most cases, the tumour originated on the left side of the body/organ. The regional nodes are the lymph nodes present in the armpits [94]. For a maximum number of cases, regional nodes were examined (1-89 in number) to test cancer cells involvement; 50% of these were negative (absence of cancer cells). The breast subtype is further categorized by ER (Estrogen-Receptor-positive) status and PR (Progesterone-Receptor-positive) status. The analysis shows that a maximum number of cases have positive ER and PR status and the breast subtype is Her2-/HR+ (*Human epidermal growth factor receptor 2 negative/*Hormone Receptor-positive).

4.1.7 Lymph Node Involvement Dashboard

The lymph node involvement in breast cancer is shown in Figure 27. The presence of cancer cells in a lymph node under arms acts as an indicator of increased risk of cancer spreading in the body [94]. Higher the number of lymph nodes containing cancer cells, higher is the seriousness of cancer. Physicians often use the count of lymph nodes involved to design treatment plans. The data for this dashboard was only available from 2004 onwards. The cases registered before 2004 are categorized as Unknown. The axillary/regional lymph nodes involvement is observed in a maximum number of cases. Axillary lymph nodes dissection is performed on axillary nodes which are suspected to have cancer cells in them. Since, axillary nodes constitute around 75% of the lymph nodes drain from breasts, the analysis indicated that from over 88 thousand cases of breast cancer reported, axillary/regional lymph nodes involvement accounted for almost one third of them, thus making axillary/regional lymph node involvement as the top category in the number of cases. No distant metastasis was found in a maximum number of cases, i.e. the cancer is not spread to other parts for such cases. The prognosis factors table is displayed in the prognostic indicators table. Positive/elevated topped the list of prognostic indicators.


4. Geo-mapping of breast cancer cases by Race
5. Breast Cancer Cases by Race and Age Range
6. Breast Cancer Anatomy
7. Lymph Node Involvement
8. Incidence/ Mortality cases
9. Survival/Mortality Rate
CS Lymph Nodes Involvement (2004+)
Involvement of lymph nodes
Axillary/regional lymph node(s), NOS Lymph nodes, NOS
Evaluated pathologically: Axillary lymph node(s), ipsilateral, micro..
Stated as N1 [NOS] with no other information on regional lymph n..
Evaluated pathologically: Axillary lymph node(s), ipsilateral, micro..
Evaluated pathologically: Fixed/matted ipsilateral axillary nodes cli..
Evaluated pathologically: Pathologically stated only as N2 [NOS]; ..
Supraclavicular node(s), ipsilateral
Evaluated clinically: Fixed/matted ipsilateral axillary nodes clinicall..
Infraclavicular lymph node(s)(subclavicular) (level III axillary nodes..
Evaluated clinically: Clinically stated only as N1 (Clinical assessm..
Evaluated pathologically: Pathologically stated only as N1 [NOS], ..
Stated as N3, NOS with no other information on regional lymph no..
Evaluated clinically: Clinically stated only as N2 [NOS] (Clinical as..
Stated as N3a with no other information on regional lymph nodes
Evaluated pathologically: Internal mammary node(s), ipsilateral, po..
Evaluated pathologically: Stated as N1mi with no other information..
OBSOLETE DATA RETAINED AND REVIEWED V0203. See cod..
0
500
1000
1500
2000
2500
3000
3500
# of Cases

Lymph Node Involvement in Breast Cancer Cases

Year of diagnosis

(All)

Click here
to download the PDF
"Your Guide to the Breast Cancer Pathology Report"
by BreastCancer.Org

Prognostic Indicators

Prognostic Indicators (2004+)	
Borderline; undetermined whether positive or negative	782
Negative/normal; within normal limits	105,280
Positive/elevated	437,435
Test not done (test not ordered and not performed)	10,322
Test ordered, results not in chart	6,039
Test ordered, results not interpretable	113
Unknown or no information Not documented in patient record	18,455
Unknown* (includes cases registered earlier than 2004)	459,121
Grand Total	1,037,547

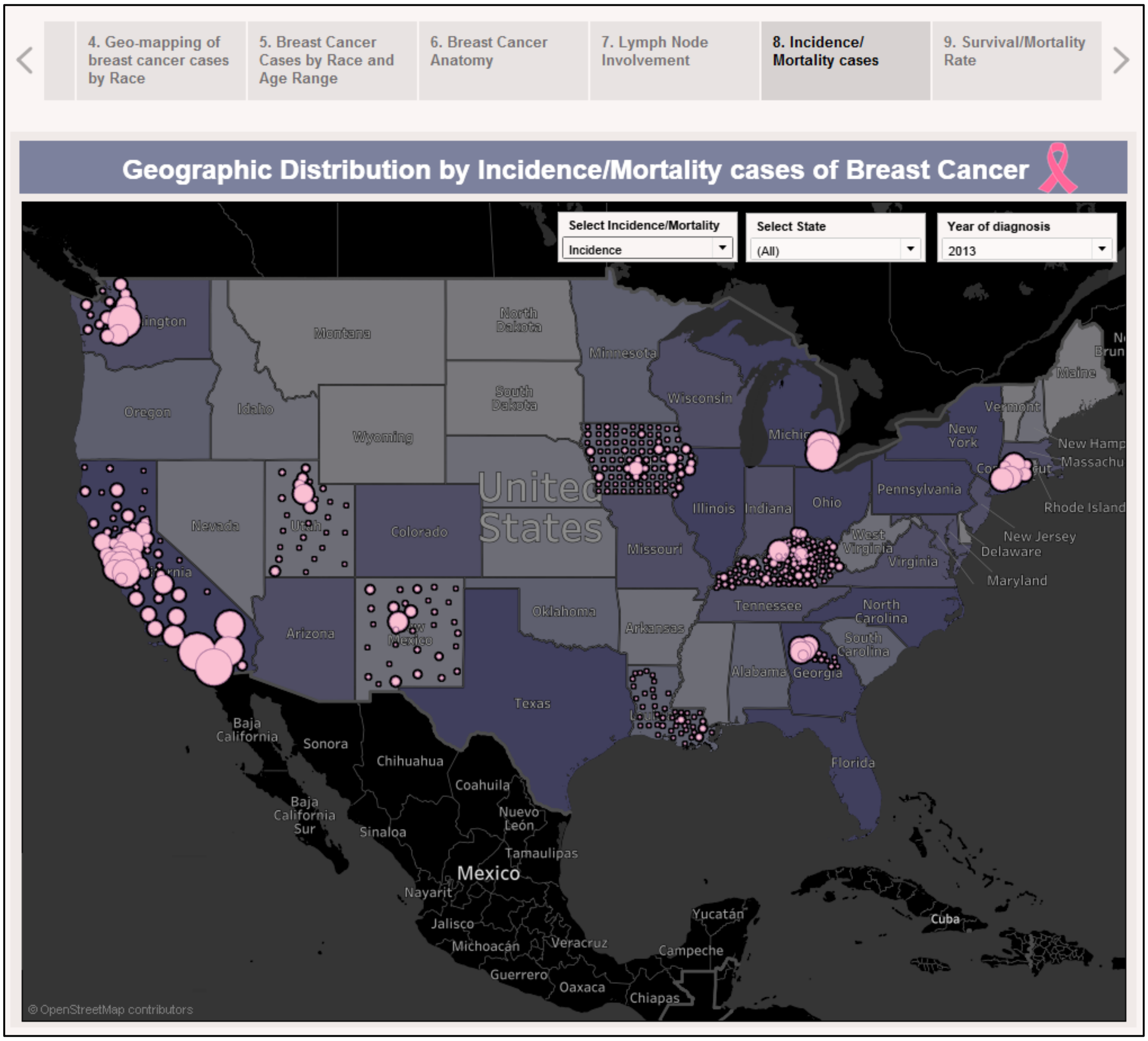
* The Unknown cases includes cases diagnosed before 2004.

Distant Metastasis

Distant Metastasis	
(40 - 44) + 10	4,542
Distant lymph node(s): Cervical; NOS, Contralateral/bilateral axillary and/or internal mammary, Other ..	1,815
Distant metastasis except distant lymph node(s) (code 10) Carcinomatosis	10,089
Distant metastasis, NOS. Stated as M1 with no other information on distant metastasis	1,196
Further contiguous extension: Skin over: Axilla, Contralateral (opposite) breast, Sternum, Upper abdo..	365
Metastasis: Adrenal (suprarenal) gland, Bone, other than adjacent rib, Contralateral (opposite) breast..	12,816
No clinical or radiographic evidence of distant metastasis, but deposits of molecularly or microscopic..	29
No distant metastasis	533,360
Stated as M0(i+) with no other information on distant metastasis	6
Unknown; distant metastasis not stated. Distant metastasis cannot be assessed. Not documented in ..	14,208
Unknown* (includes cases registered earlier than 2004)	459,121
Grand Total	1,037,547

Figure 27. Lymph Node Involvement in Breast Cancer Cases

4.1.8 Geographic Distribution by Incidence/Mortality cases of Breast Cancer Cases


4. Geo-mapping of breast cancer cases by Race | 5. Breast Cancer Cases by Race and Age Range | 6. Breast Cancer Anatomy | 7. Lymph Node Involvement | **8. Incidence/ Mortality cases** | 9. Survival/Mortality Rate

Geographic Distribution by Incidence/Mortality cases of Breast Cancer

Select Incidence/Mortality: Incidence
Select State: (All)
Year of diagnosis: 2013

Figure 28. Geographic Distribution by Breast Cancer Incidence Cases

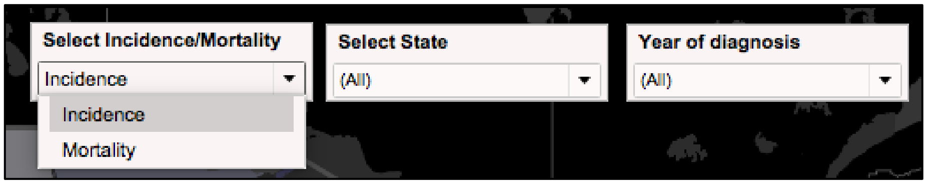
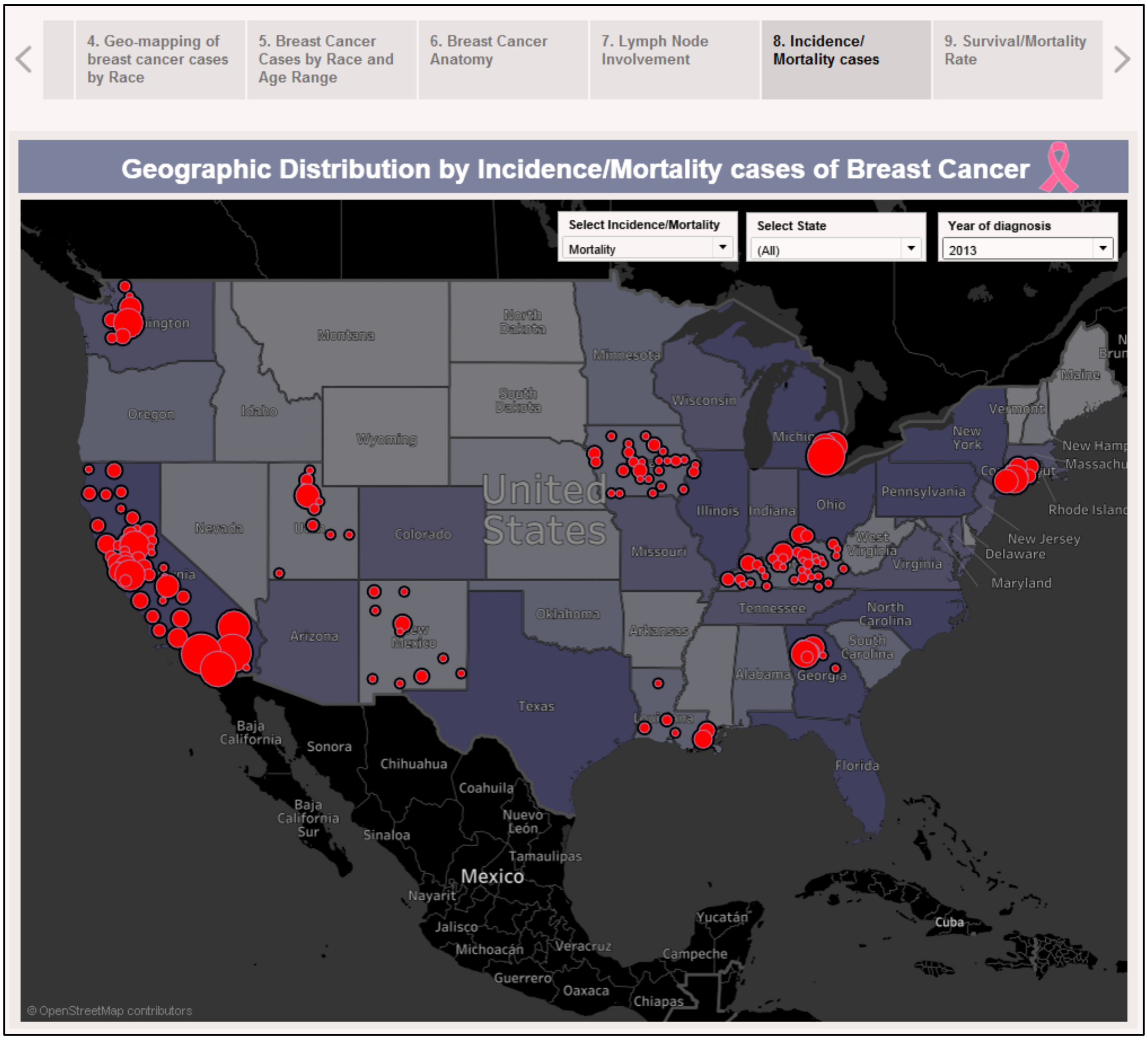

Select Incidence/Mortality: Incidence
Select State: (All)
Year of diagnosis: (All)

Figure 29. Selection panel - Breast Cancer Incidence/Mortality Cases

Geographic distribution by Incidence/Mortality provides Incidence and Mortality cases mapping by State and County as two separate dashboards (Figure 28 and Figure 30) either of which can be accessed by switching the views (Figure 29). Both dashboards can be filtered by the year of diagnosis and the State. Figure 28 and Figure 30 display breast cancer incidence and mortality cases for the year 2013 and shows that California had the highest number of breast cancer cases, followed by Washington, Michigan, Kentucky, and Connecticut. The pink plots denote the number of cases only and don't take into consideration the population of the states while determining the rankings by state. The dashboard helps identify that incidence count of breast cancer is not directly proportional to population. This is evident from the fact that the population of Michigan in 2013 was higher than the population of Washington, yet Washington had higher incidence count.

The mortality (i.e. cases died due to breast cancer) trends for the year 2013 are not coherent with the incidence trends observed for the same year. The state of California had the highest mortality count followed by Kentucky, Michigan and Connecticut. Alaska had the lowest mortality count. The mortality count, however, has some similarities, especially with respect to their dependency on the total population count for the states for that given year. This is evident from the fact that, Kentucky, like Washington, had a total population lower than that of Michigan, yet had higher mortality count than Michigan.



Geographic Distribution by Incidence/Mortality cases of Breast Cancer

Select Incidence/Mortality: Mortality

Select State: (All)

Year of diagnosis: 2013

Figure 30. Geographic Distribution by Breast Cancer Mortality Cases

4.1.9 Breast Cancer Survival/Mortality rate by Age Range

A box and whisker visual as shown in Figure 31 is a chart type which displays data distribution by quartiles. The box represents the value between first and third quartiles. The whiskers (A – lower whisker and B – upper whisker) represent the distance between the lowest value to the first quartile and the fourth quartile to the highest value. At the median (E) the box colour changes and becomes lighter showing the upper and lower quartiles (D and C respectively).

The lower (C) and upper hinge (D) are medians of the lower and upper half of the data [95]. Each box and whisker is for specific age range showing the distribution of cases by incidence or mortality rates.

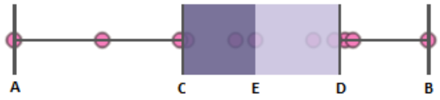
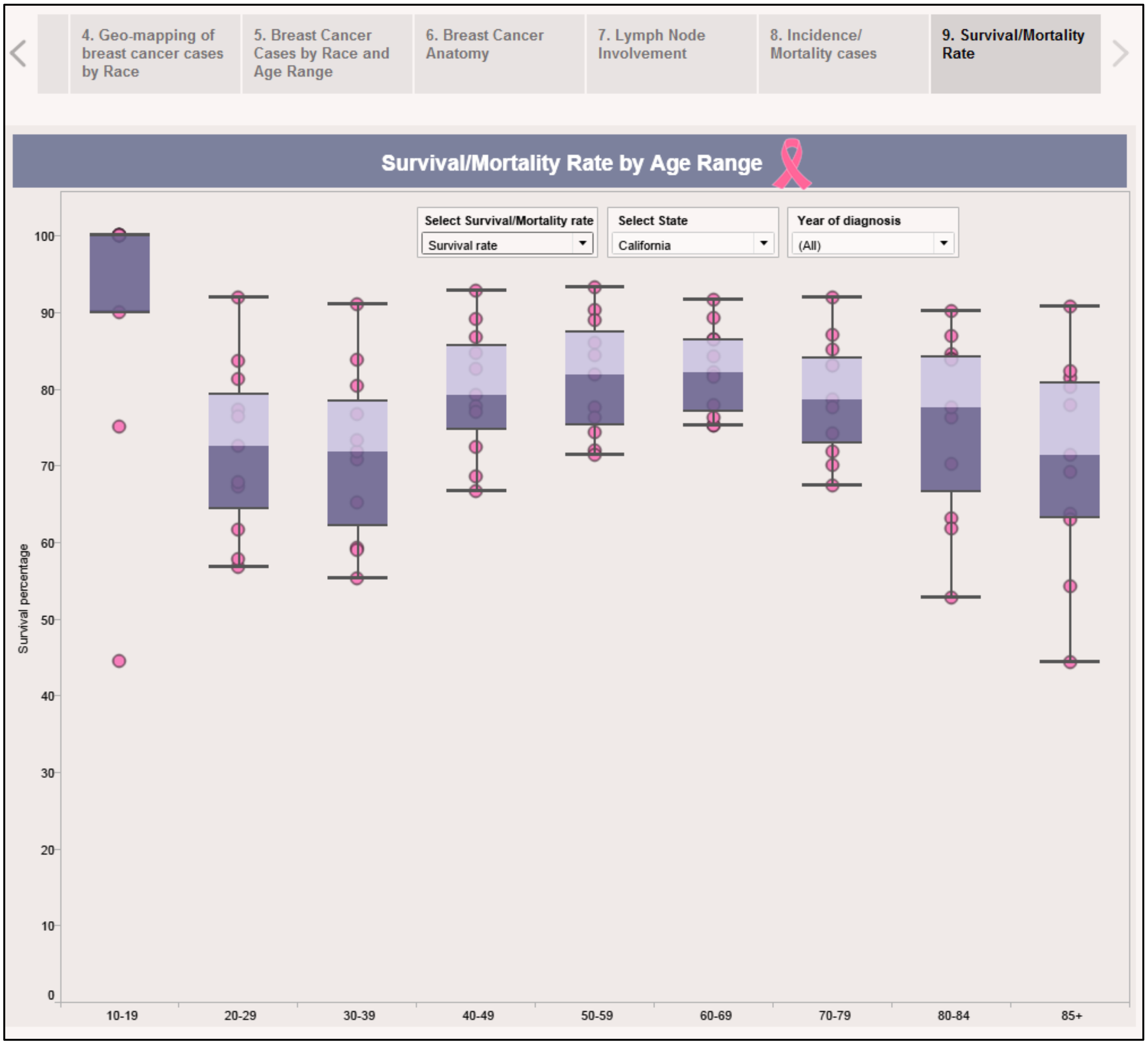


Figure 31. Box and Whisker Visual


4. Geo-mapping of breast cancer cases by Race
5. Breast Cancer Cases by Race and Age Range
6. Breast Cancer Anatomy
7. Lymph Node Involvement
8. Incidence/ Mortality cases
9. Survival/Mortality Rate
Survival percentage
100
90
80
70
60
50
40
30
20
10
0
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-84
85+

Survival/Mortality Rate by Age Range

Select Survival/Mortality rate: Survival rate

Select State: California

Year of diagnosis: (All)

Figure 32. Breast Cancer Survival Rate by Age Range

Figure 32 shows a box-whisker chart of the survival rate of by age-range for California (1973-2013). The survival rate is the ratio of cases tagged Alive to the total number of cases registered. The survivability (by years) is plotted against age-range and survival rate. The median survival rate is highest for age-range 60-69 (82.2%) followed by the age-ranges 50-59, 40-49, and 70-79. There is an exception, and age range 10-19 has a 100% survival rate because of the low number of cases (12).

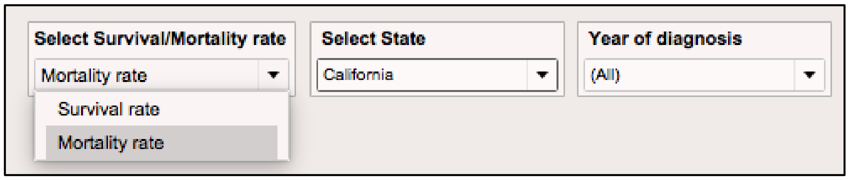
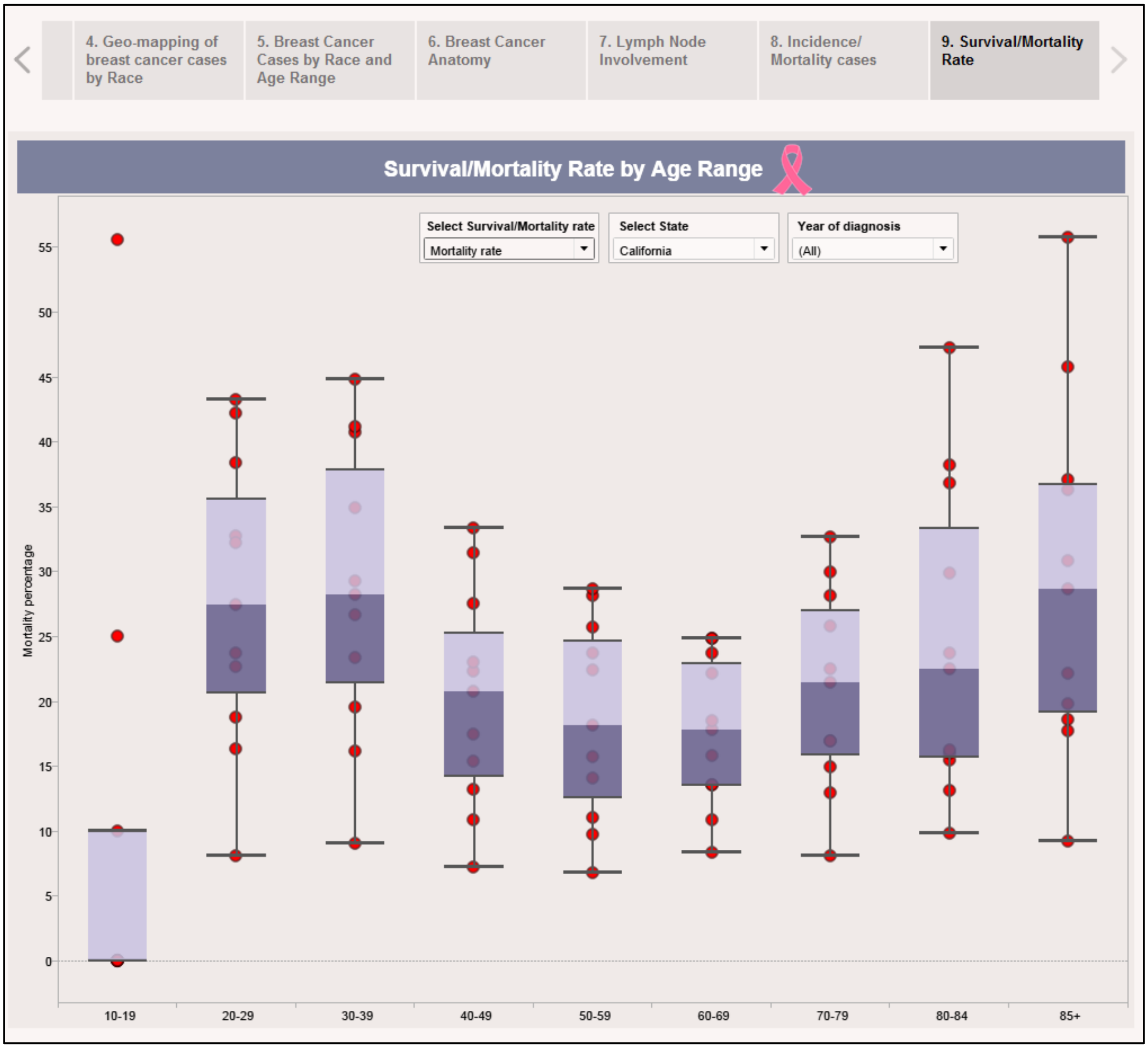


Figure 33. Selection panel - Breast Cancer Survival/Mortality Rate

A similar dashboard is shown for the mortality rate (Figure 34) for the same geographical area. The incidence and mortality rate are two separate dashboards either of which can be accessed by switching the views (Figure 33). The mortality rate is the ratio of cases tagged ‘Dead’ to the total number of cases registered. The median mortality rate is highest for age-ranges 85+ (28.6%) and 30-39 (28.2%). The age-range 60-69 has the lowest mortality rate of 17.7%. Each plot of the box-whisker chart shows the total number of cases, Alive/Dead cases, and survival/mortality rate for selected survivability period. The survivability period ranges from 1 to >10 years. The bottom point on each box-plot displays the lowest survivability period and increases in the top to bottom fashion (Figure 32). However, the pattern reverses in mortality rate dashboard (Figure 33), the top point of each box-plot starts with the lowest survivability period and increases in the top to bottom fashion.


4. Geo-mapping of breast cancer cases by Race
5. Breast Cancer Cases by Race and Age Range
6. Breast Cancer Anatomy
7. Lymph Node Involvement
8. Incidence/ Mortality cases
9. Survival/Mortality Rate
Survival/Mortality Rate by Age Range
Mortality percentage
55
50
45
40
35
30
25
20
15
10
5
0
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-84
85+

Select Survival/Mortality rate: Mortality rate

Select State: California

Year of diagnosis: (All)

Figure 34. Breast Cancer Mortality Rate by Age Range

4.1.10 Summary of Data Analysis and Visualization Results

An interactive, end-to-end process to cleanse, integrate, analyze, and visualize enormous amount of data is developed. The purpose is to enable healthcare professionals, patients and policymakers with a better understanding of the hidden patterns in data, which in turn, could be useful to to improve the quality of healthcare collectively. Forty years of breast cancer data (over one million records) extracted from the SEER database was used to demonstrate the analytical power of data visualization. The research approach involved using several data preprocessing techniques on the raw data followed by a selection of the relevant 30 variables, out of a total of 134 variables, before feeding the processed data to the SQL Server for analysis. Tableau was used for understanding and interpreting the data along several dimensions. Dynamic reports generated using the drill-down capabilities of the dashboard provide insights at a finer granularity. The dashboard shows incidence and mortality trends and highlights the underlying patterns observed in breast cancer patients which could be used to support clinical decisions made by physicians in formulating treatment plans. The dashboard is scalable and capable of integrating new data in real-time. Although SEER data from the US currently power the dashboard, it can be configured to use data from other sources as well. A better understanding of the incidence and mortality trends could potentially guide data-driven resource allocation. Physicians could also use this information to educate patients and create more awareness about the disease.

4.2 Predictive Modeling

The predictive model is trained with the breast cancer dataset, consisting of 400,000 patients registered from 1983-2003. The dataset is obtained by performing pre-processing and transformation of SEER dataset. The variables used to train the model are selected by using the feature selection algorithm. Twenty-one variables are selected out of 134 total variables available from the SEER. The model uses CHAID, C&RT, Neural Network, and Ensemble modeling techniques. Equal weights are assigned to selected models and Ensemble score is generated by averaging. The outcome variable 'Survivability' refers to the survival time (in months) of each patient. The performance metrics used are average and the accuracy. The metrics and graphs are computed by using Tableau.

In the following sections, graphs showing a comparison of actual and measured (predicted) average survival months are plotted by age range, marital status, positive to examined regional nodes ratio (percentage), radiation sequence surgery, ER status, PR status and Behavior code ICD-O-3. Similarly, a comparison of the accuracy of individual modeling techniques and ensemble modeling technique is plotted for the same variables. The results are presented for cases tagged as 'Dead' on the cut-off date. According to SEER variable description cases tagged as 'Alive' are the cases who died after the follow-up cut-off date, i.e. December 31, 2013. The noted survival months of 'Alive' cases is not their actual survival months as they were not followed up after cut-off. Thus, only cases tagged as 'Dead' at cut-off date are used to validate the predictive model's accuracy.

The developed predictive model is tested for cases registered in the year 2004 which is outside trained period (1988-2003) of the model. This dataset is selected for testing and validating the predictive model due to the following reasons:

- The cut-off date of study is 2013 and cases diagnosed in 2004 are followed up till 2013 which gives survivability range from 1 to >10 years.
- The majority of cases fall under 10 and >10 years survivability period as shown in the visualization dashboard (Figure 21).

Thus the dataset for the year 2004 has a maximum number of cases in any calendar year outside the training period. Additionally, the same dataset yields all survivability ranges which makes it apt to select it for testing and experimentation purposes. There are a total of 55,268 cases registered in 2004 (only cases with the exact month of diagnosis are included). A total of 46,365 cases are tagged 'Alive' at the cut-off date, and 8,903 cases are marked 'Dead' at the cut-off date.

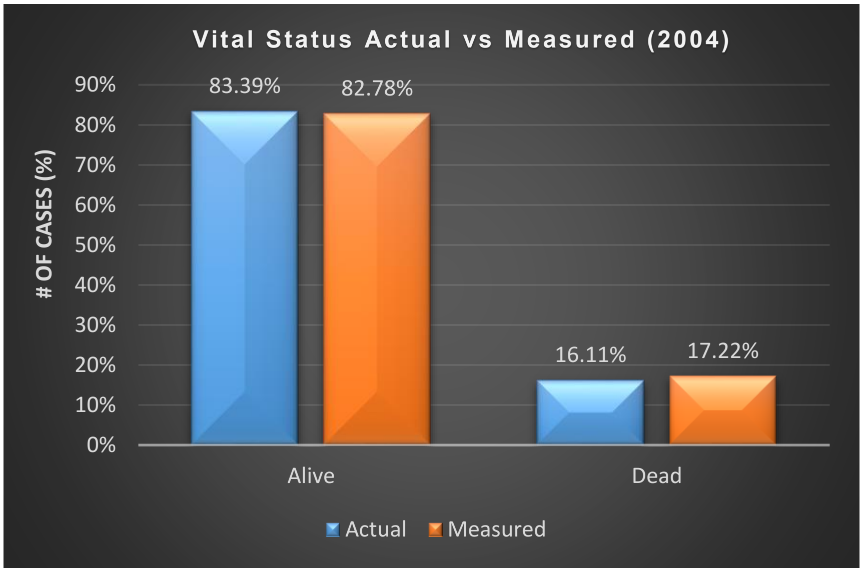


Figure 35. Vital Status comparison (2004)

The vital status statistics of actual and predicted model's output (i.e. measured[7]) are compared in Figure 35. For cases diagnosed in 2004, 83.39% of cases are tagged 'Alive' and

[7] Measured survival months are the survival months predicted by predictive model

16.61% are tagged 'Dead' at the cut-off date. The predictive model developed predicts 82.78% of cases as 'Alive' and 17.22% as 'Dead' as of the cut-off date (December 31, 2013) thus demonstrating accuracy of 99.26% and 96.45%, respectively.

4.2.1 Comparison of Average Survival Months (Actual vs Measured)

The average survival months for the following experiments is calculated by using the formula (Equation 2):

$$\text{Average Survival Months} = \frac{\text{Sum of measured Survival Months}}{\text{Total number of cases}}$$

Equation 2. Average Survival Months

Figure 36 shows measured (predicted) survival months by each selected modeling technique and their ensemble. Actual survival months (average) of cases registered in the year 2004 and tagged 'Dead' at cut-off date, as shown in the yellow bar, is 42 months. Both Ensemble and CHAID measured survival months (average) as 45 months which is closest to actual survival months. C&RT and Neural Network predicts average survival months of 33 and 56 respectively. Overall, Ensemble is performing best along with CHAID. C&RT and Neural Network modeling predictions are off from actual average survival months.

In the following sections, graphs of actual vs measured survival months are plotted as trend-lines. Each line in the graphs show the trend of average predicted survival months of each model (C&RT, CHAID and Neural Network), and their Ensemble along with actual survival months (denoted by Survival Months). The bars in each graph display the number of cases falling under the specific category, i.e. such as age-range, marital status, and lymph node involvement. Next, the performance of individual modeling techniques is observed.

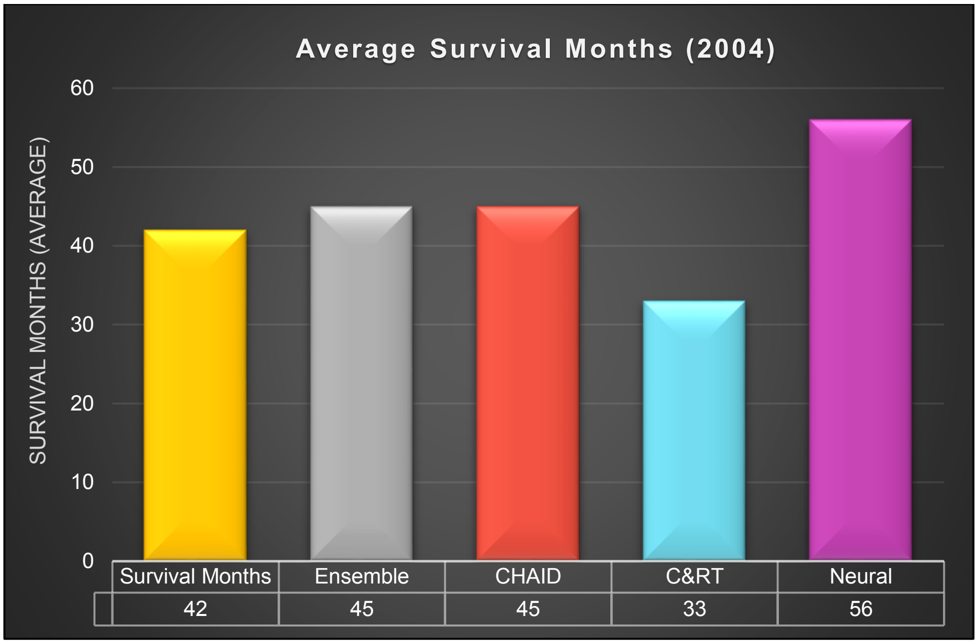


Figure 36. Average Survival Months (2004)

From Figure 36, it is evident that Neural Networks do not demonstrate higher accuracy (i.e. 66%) in comparison to the other modeling techniques used. This is likely due to the following:

The training dataset used for training (from the years 1988-2003) had values for T, N, and M variables. On the contrary, for the testing dataset (2004 and onwards) the values of these variables are missing. Neural Networks tend to learn more and extract better knowledge by observing trends in the datasets [88]. Neural Networks relative predictor importance[8] is not uniform in its training. Hence, an inconsistent data across training and testing dataset resulted in lower accuracy. Other modeling techniques perform better even with the missing

[8] "Predictor importance is determined by computing the reduction in variance of the target attributable to each predictor, via a sensitivity analysis" [88].

value of T, N and M variables because they give equal preference to all the variables (i.e. equal predictor importance to all predictors).

To validate the above hypothesis, a new experimentation was performed by eliminating the T, N and M variables from the training dataset for all the modeling techniques. The results from this experiment (Figure 37) confirm the hypothesis stated above. Neural Networks produced higher accuracy in comparison to the other modeling techniques (Table 9).

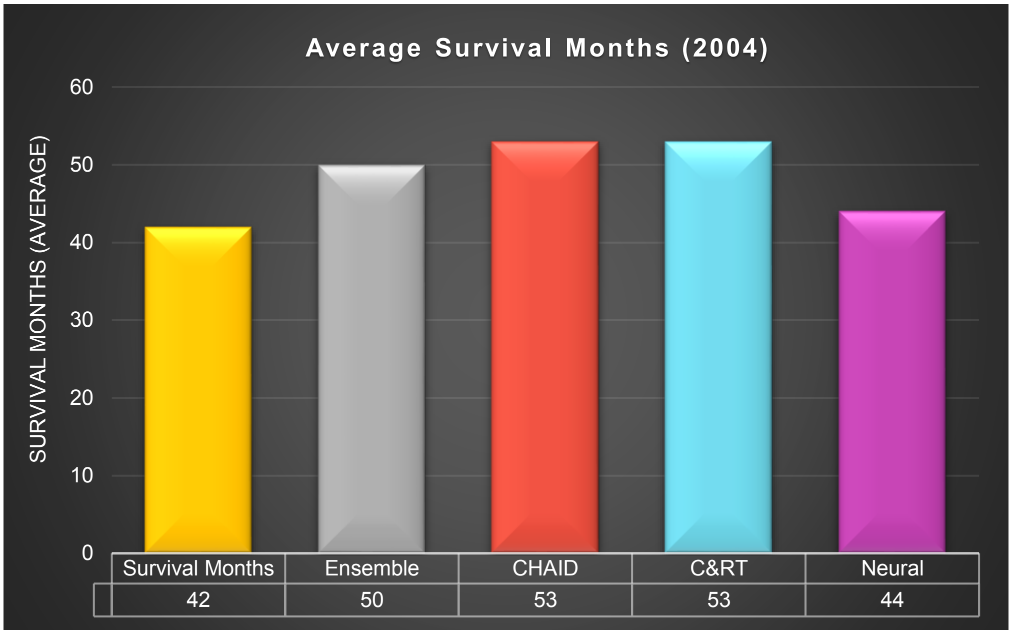
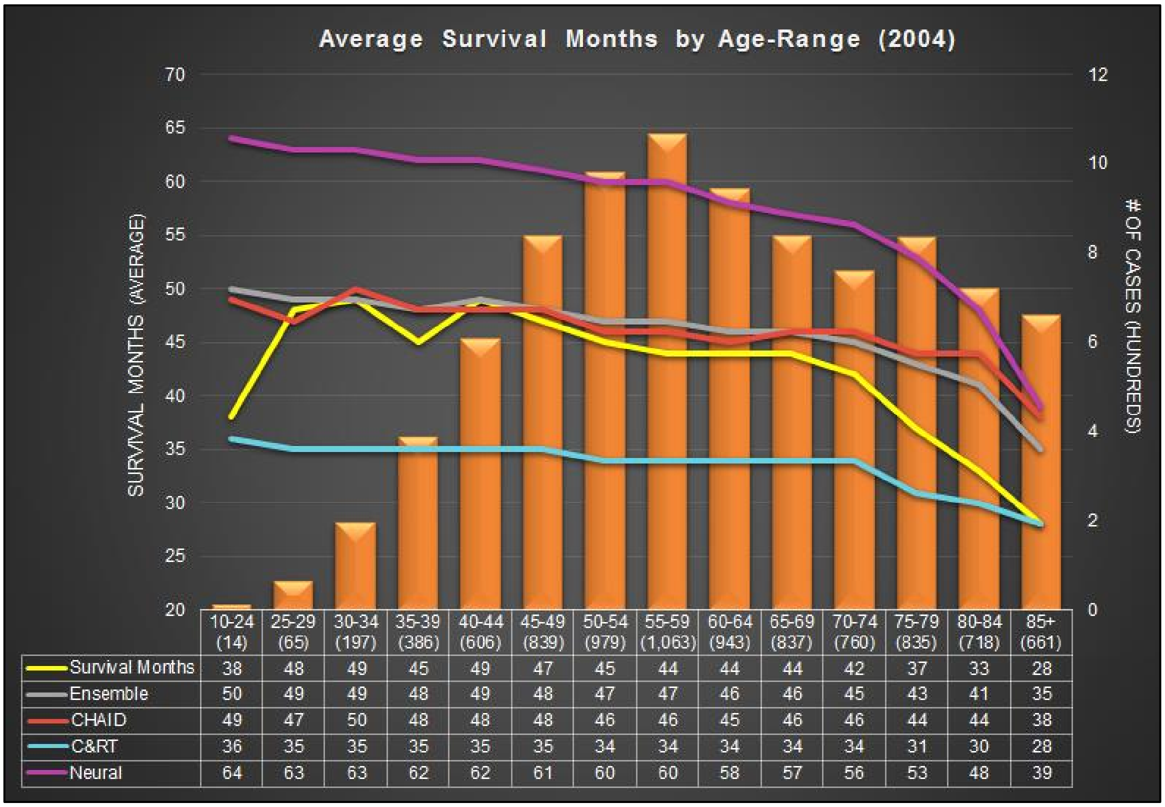


Figure 37. Average Survival Months (2004 - excluding TNM variables)

Accuracy (%)	Modeling Techniques			
	Ensemble	CHAID	C&RT	Neural Network
	81	74	74	95

Table 9. Accuracy of Modeling Techniques

4.2.1.1 Average Survival Months by Age Range

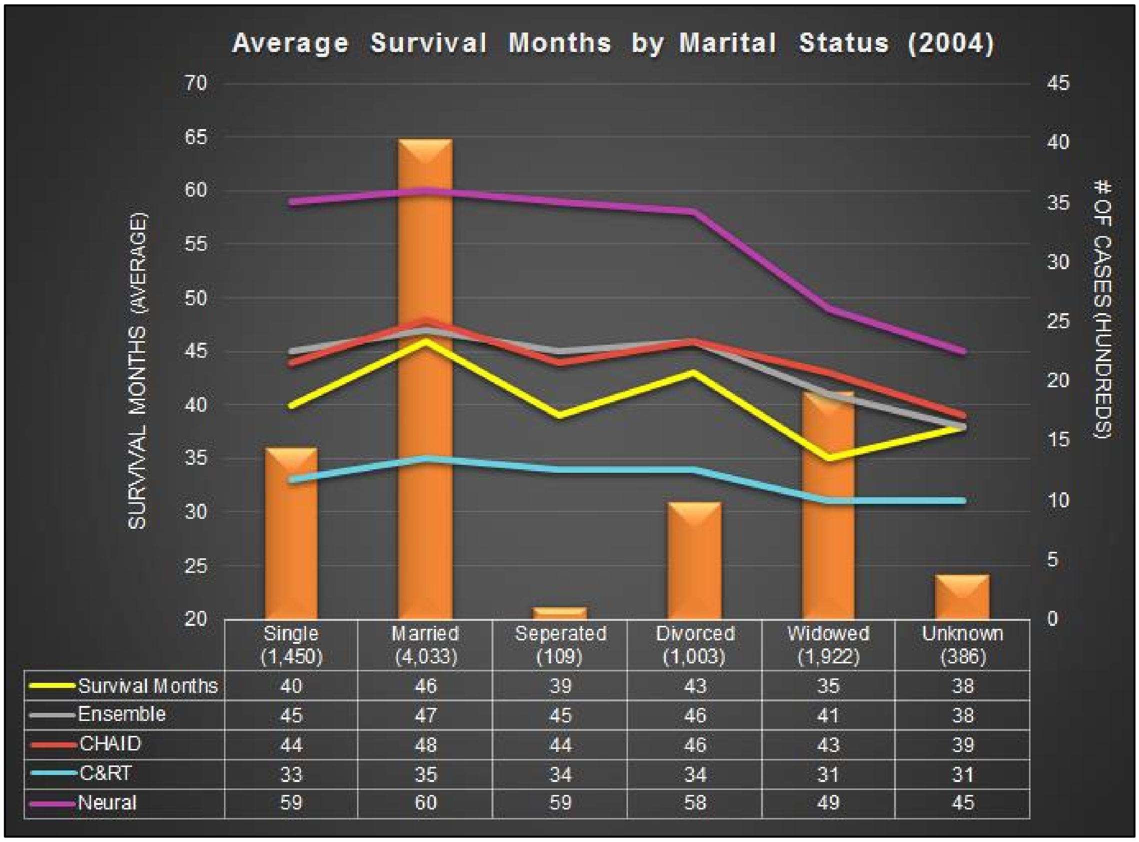


	10-24 (14)	25-29 (65)	30-34 (197)	35-39 (386)	40-44 (606)	45-49 (839)	50-54 (979)	55-59 (1,063)	60-64 (943)	65-69 (837)	70-74 (760)	75-79 (835)	80-84 (718)	85+ (661)
Survival Months	38	48	49	45	49	47	45	44	44	44	42	37	33	28
Ensemble	50	49	49	48	49	48	47	47	46	46	45	43	41	35
CHAID	49	47	50	48	48	48	46	46	45	46	46	44	44	38
C&RT	36	35	35	35	35	35	34	34	34	34	34	31	30	28
Neural	64	63	63	62	62	61	60	60	58	57	56	53	48	39

Figure 38. Average Survival Months by Age Range (2004)

Figure 38 compares actual and measured average survival months by age range 10 to 85+ for cases registered in 2004 which are tagged ‘Dead’, at the cut-off date. The bars in the chart display the number of cases for each age range. Approximately 43% of cases tagged ‘Dead’ at cut-off date, fall under the 45-64 age range. The actual survival months varies over different age ranges. Cases with age 85 and above have lowest survival months, i.e. 28 months. The graph shows both Ensemble and CHAID performing closest to the actual survival months and also overlap at few data points. However, Ensemble performs better than C&RT for age 70 and onwards. C&RT, on the other hand, performs best for lowest and highest age range categories. Neural Network predicts high survival month as compared to actual overall age ranges.

4.2.1.2 Average Survival Months by Marital Status of patient

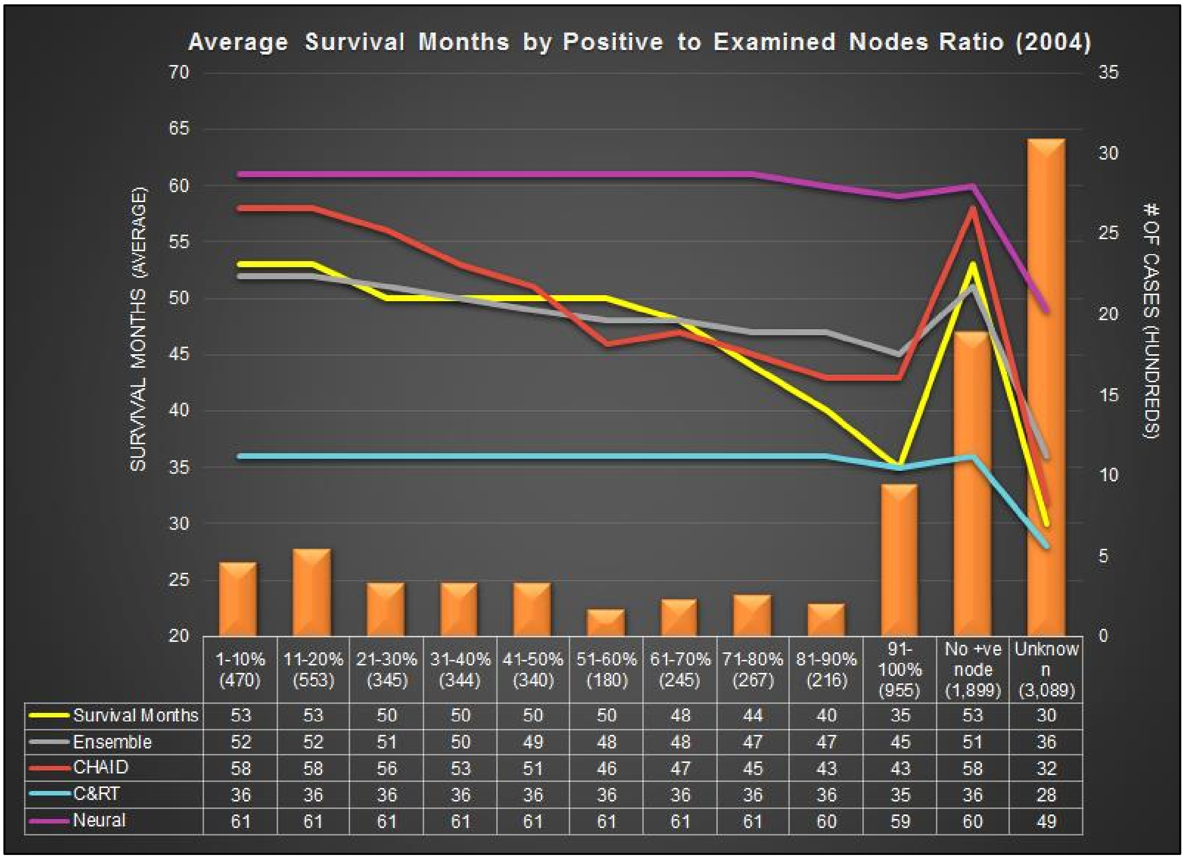


	Single (1,450)	Married (4,033)	Seperated (109)	Divorced (1,003)	Widowed (1,922)	Unknown (386)
Survival Months	40	46	39	43	35	38
Ensemble	45	47	45	46	41	38
CHAID	44	48	44	46	43	39
C&RT	33	35	34	34	31	31
Neural	59	60	59	58	49	45

Figure 39. Average Survival Months by Marital Status (2004)

Figure 39 compares actual and measured average survival months by marital status (married, single, widowed, divorced, separated and unknown) for cases registered in 2004 which are tagged ‘Dead’, at the cut-off date. The bars in the chart displays a number of cases for each marital status. About 45% of cases tagged ‘Dead’, at cut-off date, are married at diagnosis. The actual survival months vary by marital status. Widowed cases have the lowest survival months, i.e. 35 months. The graph shows both, Ensemble and CHAID predict closest to the actual survival months. The trend lines overlap for divorced cases. C&RT predicts low survival months as compared to other techniques. However, Neural Network predicts in range of 45-60 months when actual survival months range from 35-46 survival months. Overall, CHAID and Ensemble perform closely.

4.2.1.3 Average Survival Months by Positive to Examined Nodes Ratio



	1-10% (470)	11-20% (553)	21-30% (345)	31-40% (344)	41-50% (340)	51-60% (180)	61-70% (245)	71-80% (267)	81-90% (216)	91-100% (955)	No +ve node (1,899)	Unknown (3,089)
Survival Months	53	53	50	50	50	50	48	44	40	35	53	30
Ensemble	52	52	51	50	49	48	48	47	47	45	51	36
CHAID	58	58	56	53	51	46	47	45	43	43	58	32
C&RT	36	36	36	36	36	36	36	36	36	35	36	28
Neural	61	61	61	61	61	61	61	61	60	59	60	49

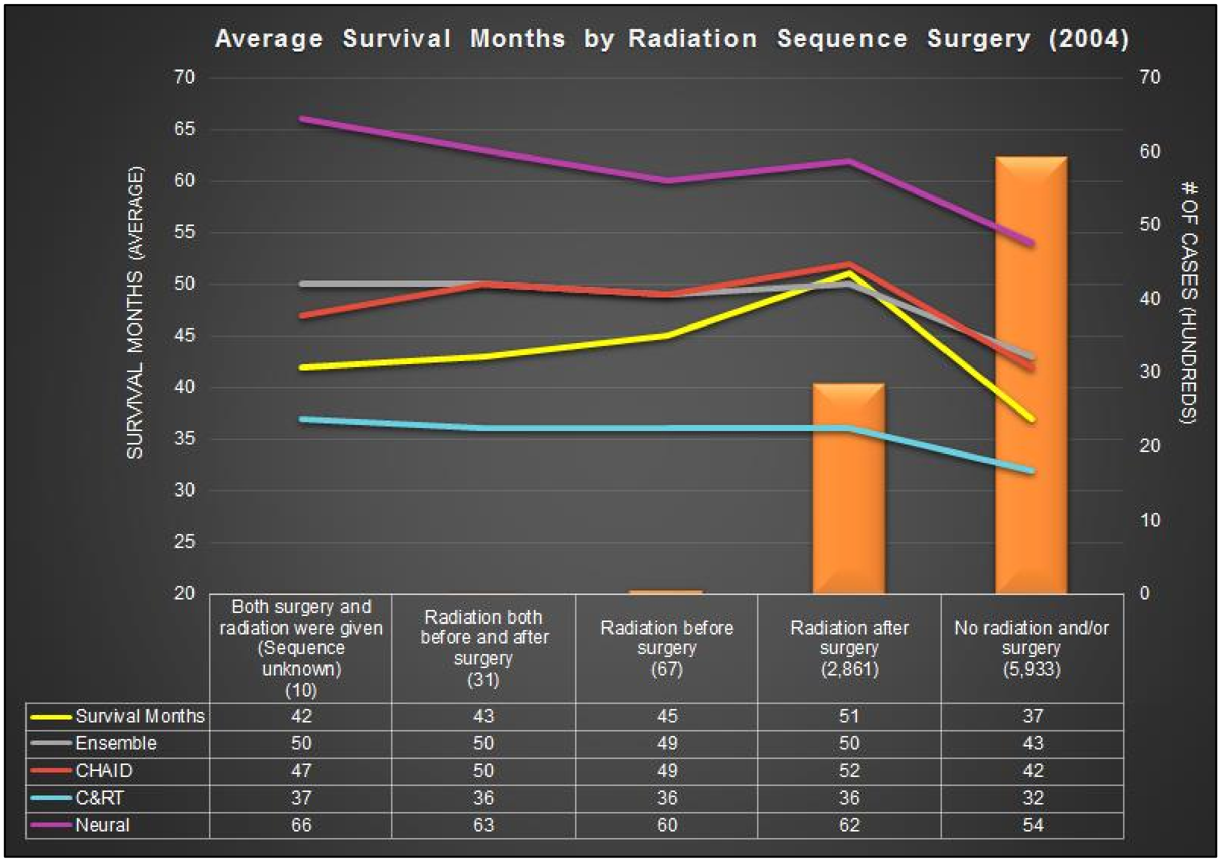
Figure 40. Average Survival Months by Positive to Examined Regional Nodes Ratio (2004)

Figure 40 compares actual and measured average survival months by the ratio of positive to examined regional nodes for cases registered in 2004 which are tagged ‘Dead’ at the cut-off date. The bars in the chart displays the number of cases with the positive to examined regional nodes ratio. Amongst the cases with examined nodes, 33% cases have no positive regional nodes. The actual survival months is lowest for cases with unknown nodes examined i.e. 30 months. The graph shows that the Ensemble performs better than other models for cases having

70% or less positive to regional nodes ratio and cases having no positive node at all. CHAID performs better for cases having 81-90% and unknown positive to regional nodes ratio. Neural Network predicts survival months in a range of 49-61 months when actual survival months ranges from 30-53 survival months.

4.2.1.4 Average Survival Months by Radiation and Surgery Sequence

Figure 41 shows the case counts and survival months' distribution by radiation and surgery sequence of cases registered in 2004 which are tagged 'Dead', at the cut-off date. The bars show a number of cases which had radiation and surgery performed categorized as – intraoperative radiation therapy, intraoperative radiation with other radiation given before or after, sequence unknown yet both surgery and radiation are given, radiation both before and after surgery, radiation before surgery, radiation after surgery and no radiation and/or surgery. A maximum (67%) number of cases tagged 'Dead' did not have radiation and/or surgery performed followed by 32% of cases who received radiation after surgery. The actual survival months is lowest for cases who died without radiation or surgery, i.e. 37 months. Both CHAID and Neural Network predict 42 and 43 months for such cases, respectively. Next, for cases which had radiation after surgery have 51 survival months, highest survival months amongst all other categories. Ensemble and CHAID predict 50 and 52 survival months, respectively. C&RT and Neural Network predicted survival months are off the actual range. Both are ranging between 32-37 and 54-60 months, respectively.

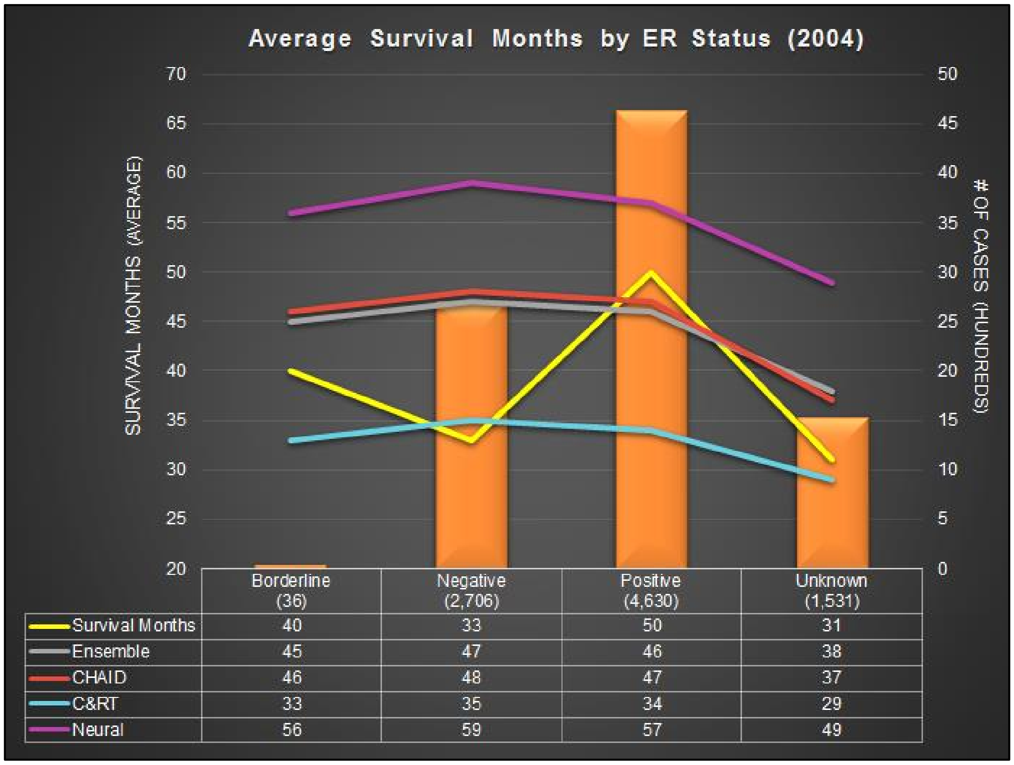


	Both surgery and radiation were given (Sequence unknown) (10)	Radiation both before and after surgery (31)	Radiation before surgery (67)	Radiation after surgery (2,861)	No radiation and/or surgery (5,933)
Survival Months	42	43	45	51	37
Ensemble	50	50	49	50	43
CHAID	47	50	49	52	42
C&RT	37	36	36	36	32
Neural	66	63	60	62	54

Figure 41. Average Survival Months by Positive to Radiation and Sequence Surgery (2004)

4.2.1.5 Average Survival Months by Estrogen and Progesterone Receptor Status

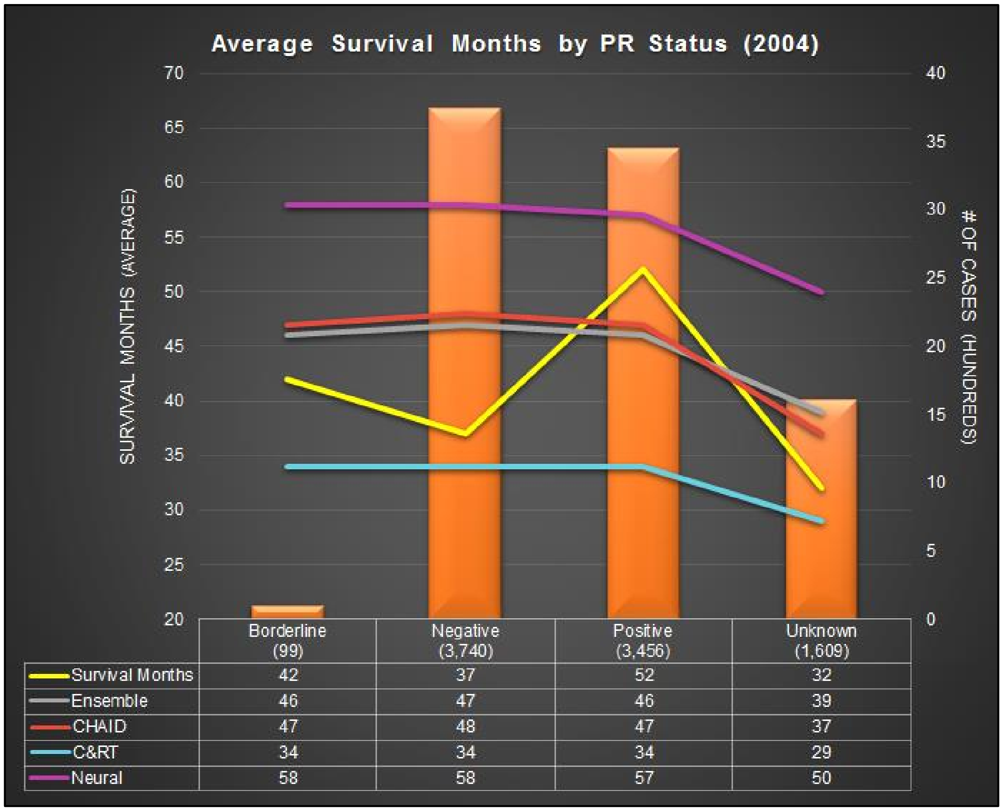
Figure 42 shows actual and measured average survival months are plotted by ER status for cases registered in 2004 which are tagged ‘Dead’ at the cut-off date. The status is recorded as – positive, negative, borderline and unknown. 52% of cases tagged ‘Dead’, at cut-off date, have positive ER status and 30% cases have negative ER status. The actual survival months recorded varies from 32-50 months. Cases with unknown and negative ER status have the lowest survival months, i.e. 31 and 33 months respectively. Cases with positive ER status have the highest survival months, i.e. 50 months. CHAID (47 months) performs better than Ensemble (46 months) and other models for such case. For cases having negative ER status, C&RT predicts 35 months compared to actual survival months (33 months). Neural Network predicts survival months ranging from 49-56 months over all ER status available.



	Borderline (36)	Negative (2,706)	Positive (4,630)	Unknown (1,531)
Survival Months	40	33	50	31
Ensemble	45	47	46	38
CHAID	46	48	47	37
C&RT	33	35	34	29
Neural	56	59	57	49

Figure 42. Average Survival Months by Estrogen Receptor Status (2004)

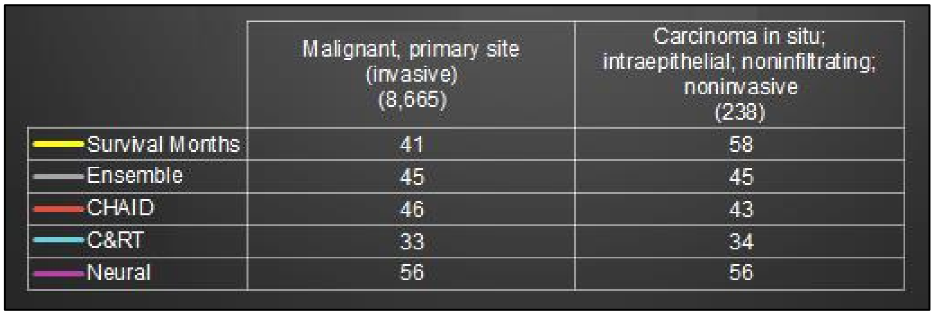
Next, Figure 43 compares the actual and measured average survival months by PR status which is recorded as – positive, negative, borderline and unknown. 42% of ‘Dead’ cases have negative PR status, and 38% of cases have a positive PR status. The actual survival months recorded varies from 33-52 months. Cases with unknown and negative PR status have the lowest survival months, i.e. 32 and 37 months, respectively. The graph shows that for the maximum number of cases (negative PR status), C&RT performs best with 34 survival months as compared to 37 actual survival months. The highest survival months recorded is 52 months. CHAID and Neural Network perform best with 47 and 57 survival months.



	Borderline (99)	Negative (3,740)	Positive (3,456)	Unknown (1,609)
Survival Months	42	37	52	32
Ensemble	46	47	46	39
CHAID	47	48	47	37
C&RT	34	34	34	29
Neural	58	58	57	50

Figure 43. Average Survival Months by Progesterone Receptor Status (2004)

4.2.1.6 Average Survivability by Behavior Code ICD-O-3



	Malignant, primary site (invasive) (8,665)	Carcinoma in situ; intraepithelial; noninfiltrating; noninvasive (238)
Survival Months	41	58
Ensemble	45	45
CHAID	46	43
C&RT	33	34
Neural	56	56

Figure 44. Average Survival Months by Behavior Code ICD-O-3 (2004)

In Figure 44, actual and measured average survival months are compared by behavior type for cases registered in 2004 which are tagged ‘Dead’ at the cut-off date. The cases are categorized as malignant (invasive) and carcinoma in situ (non-invasive). 97% of cases tagged ‘Dead’, at cut-off date, have the malignant tumor and invasive primary breast cancer. The Ensemble and

CHAID models perform best with 45 and 46 average predicted survival months, respectively, compared to 41 actual average survival months. Neural Network model performs best for cases which have carcinoma in situ behavior type.

4.2.2 Comparison of Accuracy of Modeling Techniques

The accuracy of the experiments is calculated by using the formula (Equation 3)-

$$\text{Accuracy} = \left\{1 - \text{abs}\,\{\frac{(\text{Actual Survival Months} - \text{ Measured Survival Months})}{\text{Actual Survival Months}}\}\right\} * 100$$

Equation 3. Accuracy

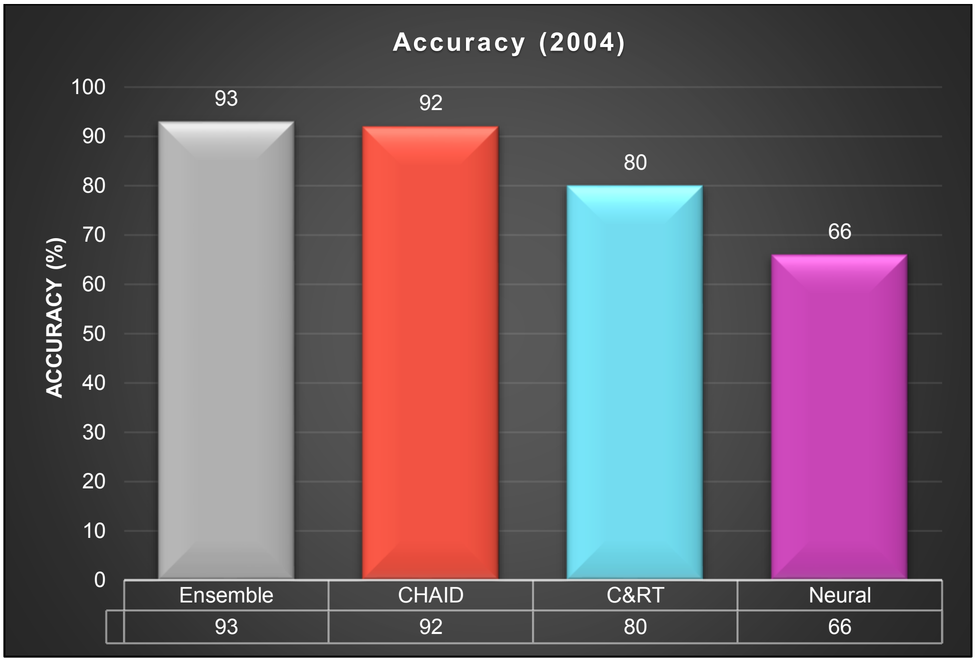
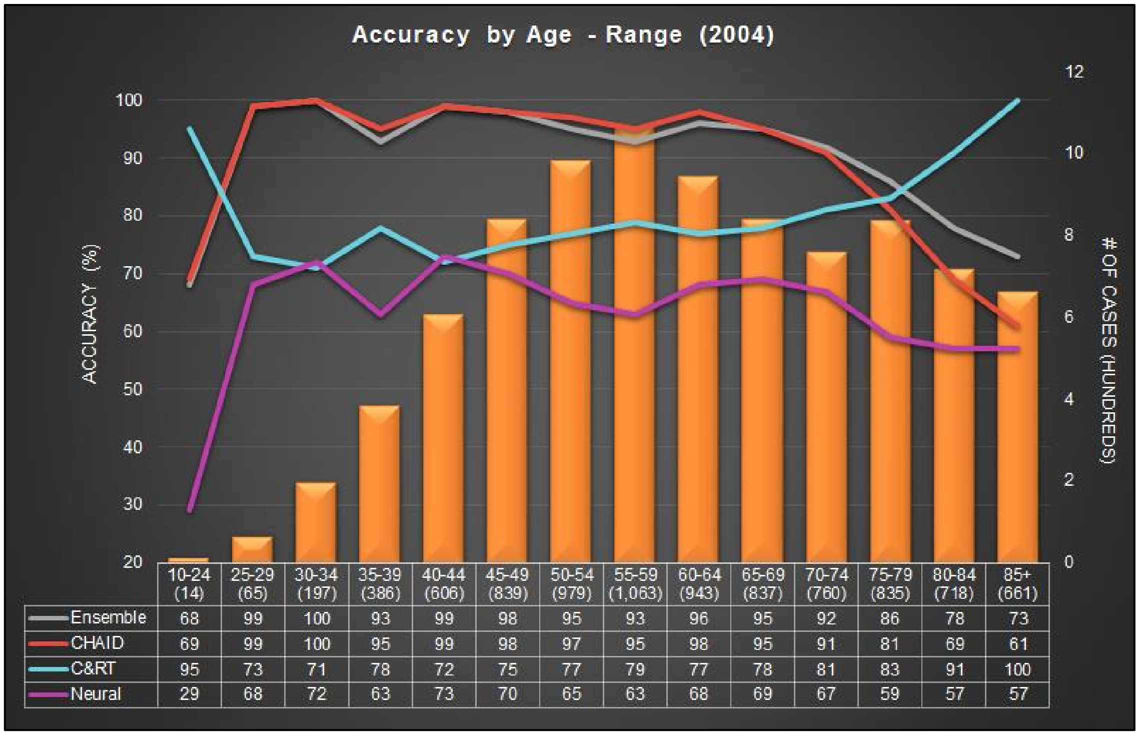


Figure 45. Accuracy of different modeling techniques (2004)

Figure 45 compares accuracies of CHAID, C&RT, Neural Network and Ensemble modeling techniques. Ensemble has the highest accuracy when compared at aggregated year level, i.e. 2004. This is followed by CHAID and C&RT with 92% and 80%, respectively. Neural Network has the lowest accuracy, i.e. 66%.

In the following sections, the accuracy of the predictive model is plotted as trend-lines where each line shows the accuracy of the individual modeling technique (C&RT, CHAID and Neural Network), and their Ensemble. The bars in each graph display the number of cases following under the specific category i.e. such as age-range, marital status, etc.

4.2.2.1 Accuracy by Age Range

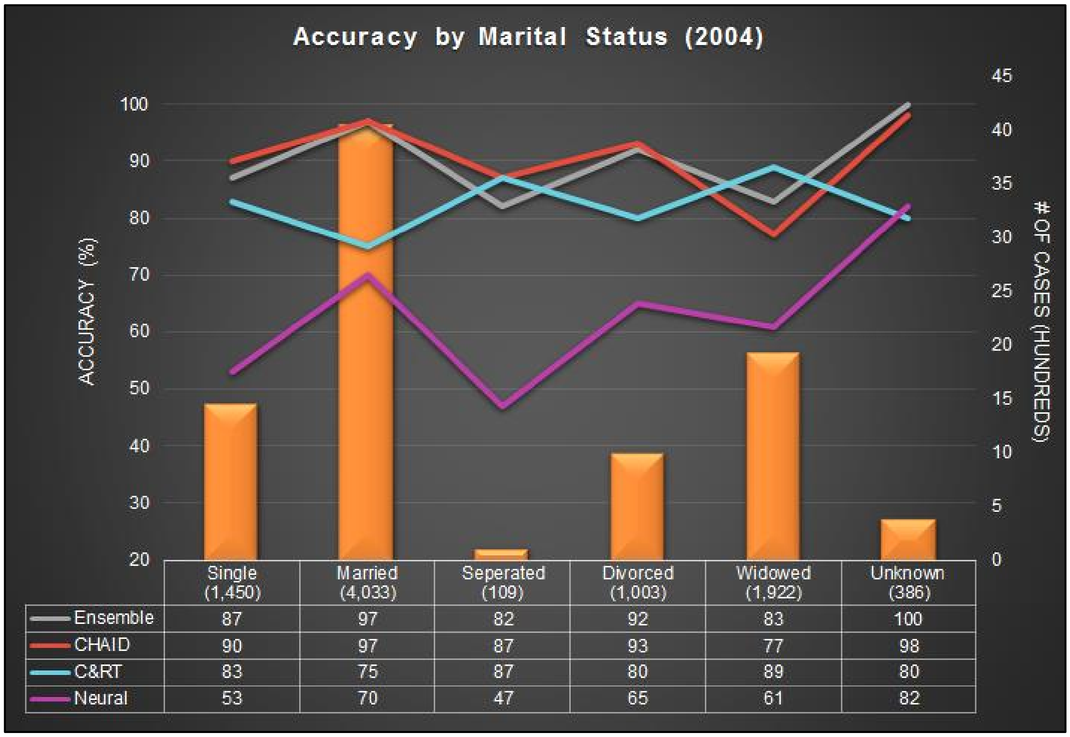


	10-24 (14)	25-29 (65)	30-34 (197)	35-39 (386)	40-44 (606)	45-49 (839)	50-54 (979)	55-59 (1,063)	60-64 (943)	65-69 (837)	70-74 (760)	75-79 (835)	80-84 (718)	85+ (661)
Ensemble	68	99	100	93	99	98	95	93	96	95	92	86	78	73
CHAID	69	99	100	95	99	98	97	95	98	95	91	81	69	61
C&RT	95	73	71	78	72	75	77	79	77	78	81	83	91	100
Neural	29	68	72	63	73	70	65	63	68	69	67	59	57	57

Figure 46. Accuracy by Age Range (2004)

Figure 46 shows the comparison of the accuracy of predictive models with respect to age for cases registered in 2004 which are tagged ‘Dead’ at the cut-off date. The age ranges from 10 to 85+. The bars in the chart displays a number of cases for each age range. 15% of total cases tagged ‘Dead’ fall under 80-85+ age range. The graph shows interesting trends, CHAID and Ensemble model has the highest accuracy overall but drops down for age 75 and onwards. C&RT model on the other hand has the highest accuracy of 95%, 91% and 100% for the age range 10-24, 80-84 and 85+, respectively. Neural Network prediction ranges from 29-79% which is lowest compared to other models. Ensemble outperforms the CHAID model for some age-ranges.

4.2.2.2 Accuracy by Marital Status

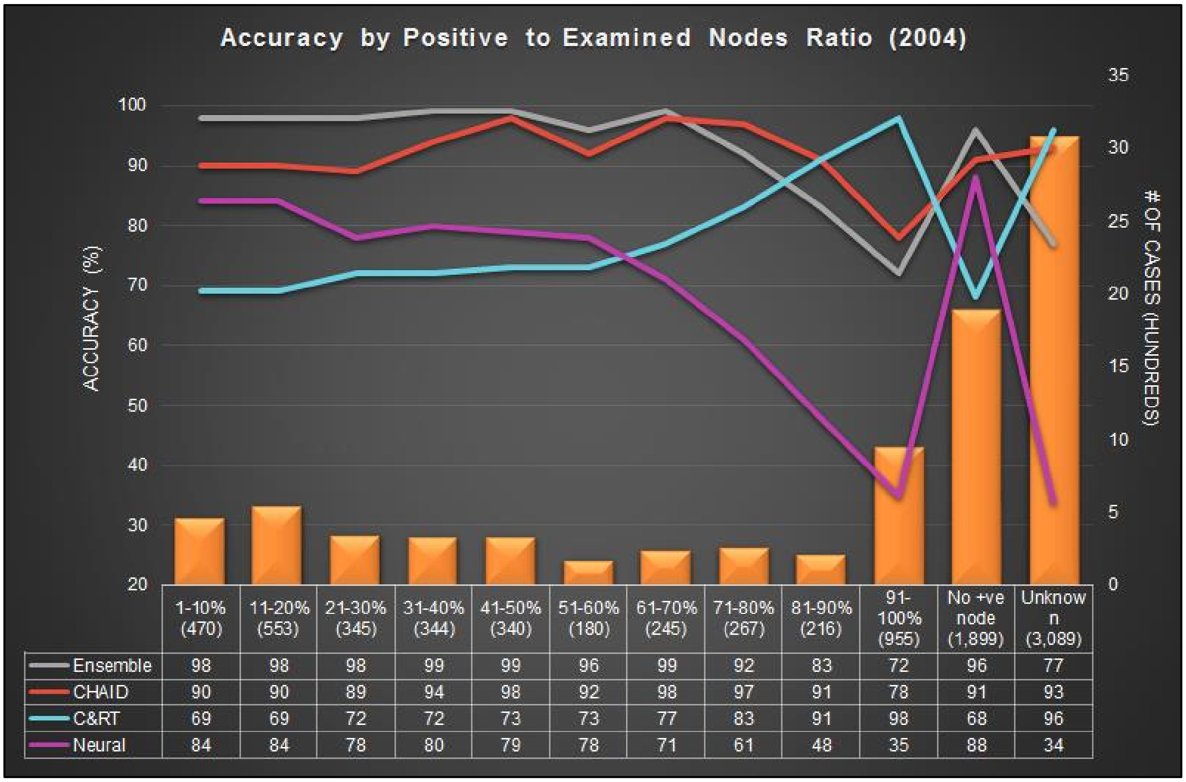


	Single (1,450)	Married (4,033)	Seperated (109)	Divorced (1,003)	Widowed (1,922)	Unknown (386)
Ensemble	87	97	82	92	83	100
CHAID	90	97	87	93	77	98
C&RT	83	75	87	80	89	80
Neural	53	70	47	65	61	82

Figure 47. Accuracy by Marital Status (2004)

Figure 47 shows the comparison of the accuracy of predictive models by marital status (married, single, widowed, divorced, separated and unknown) for cases registered in 2004 which are tagged ‘Dead’ at the cut-off date. The bars in the chart displays a number of cases for each marital status. The graph shows that 45% of died cases are married, 22% cases are widowed, and 11% are divorced. Ensemble and CHAID have the highest accuracy for married cases. CHAID has the highest prediction accuracy of divorced cases, i.e. 93%. C&RT has the highest prediction accuracy for widowed cases. Neural Network tends to have low prediction accuracy ranging from 47-82%.

4.2.2.3 Accuracy by Positive to Examined Nodes Ratio

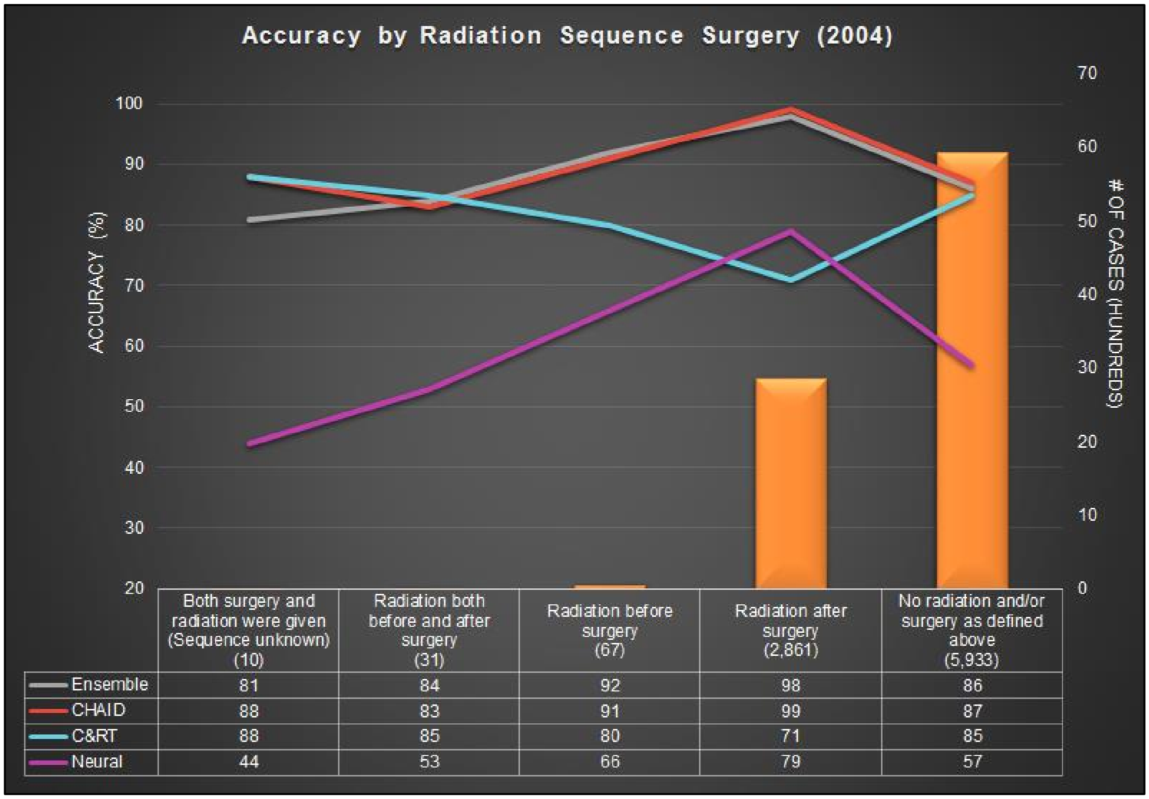


	1-10% (470)	11-20% (553)	21-30% (345)	31-40% (344)	41-50% (340)	51-60% (180)	61-70% (245)	71-80% (267)	81-90% (216)	91-100% (955)	No +ve node (1,899)	Unknown (3,089)
Ensemble	98	98	98	99	99	96	99	92	83	72	96	77
CHAID	90	90	89	94	98	92	98	97	91	78	91	93
C&RT	69	69	72	72	73	73	77	83	91	98	68	96
Neural	84	84	78	80	79	78	71	61	48	35	88	34

Figure 48. Accuracy by Positive to Examined Ratio (2004)

Figure 48 shows the comparison of the accuracy of predictive models by the ratio of positive to examined regional nodes for cases registered in 2004 which are tagged ‘Dead’ at the cut-off date. The bars in the chart displays a number of cases with positive to examined regional nodes ratio. Amongst the cases with examined nodes, 67% cases have positive to examined regional nodes (ranging 1-100%). The Ensemble has the highest accuracy for cases having 70% or less positive to regional nodes ratio and cases having no positive nodes. CHAID and C&RT have the highest accuracy for cases having 81-90% and unknown positive to regional nodes ratio. Neural Network prediction accuracy ranges from 34-84%.

4.2.2.4 Accuracy by Radiation and Surgery Sequence

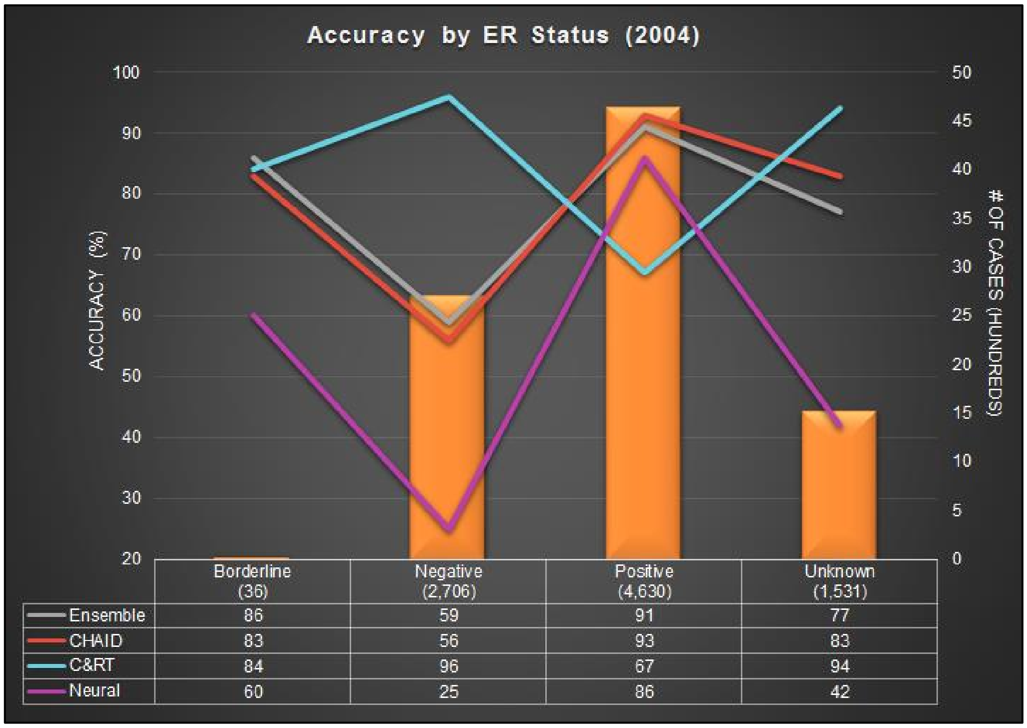


	Both surgery and radiation were given (Sequence unknown) (10)	Radiation both before and after surgery (31)	Radiation before surgery (67)	Radiation after surgery (2,861)	No radiation and/or surgery as defined above (5,933)
Ensemble	81	84	92	98	86
CHAID	88	83	91	99	87
C&RT	88	85	80	71	85
Neural	44	53	66	79	57

Figure 49. Accuracy by Radiation Sequence Surgery (2004)

Figure 49 shows the case counts and survival months' distribution by the radiation and surgery sequence of cases registered in 2004 which are tagged 'Dead' at the cut-off date. The bars show a number of cases which had radiation and surgery performed categorized as – intraoperative radiation therapy, intraoperative radiation with other radiation given before or after, sequence unknown yet both surgery and radiation are given, radiation both before and after surgery, radiation before surgery, radiation after surgery and no radiation and/or surgery. The graph shows that 66% of cases tagged 'Dead' did not have radiation and/or surgery performed followed by 32% of cases who got radiation after surgery is performed. Ensemble and CHAID have the highest accuracies ranging from 81-98% and 83-99% respectively. Neural Network has the lowest accuracy overall, except for cases categorized as radiation after surgery. C&RT has the lowest accuracy for cases categorized as radiation after surgery.

4.2.2.5 Accuracy by Estrogen and Progesterone Receptor Status

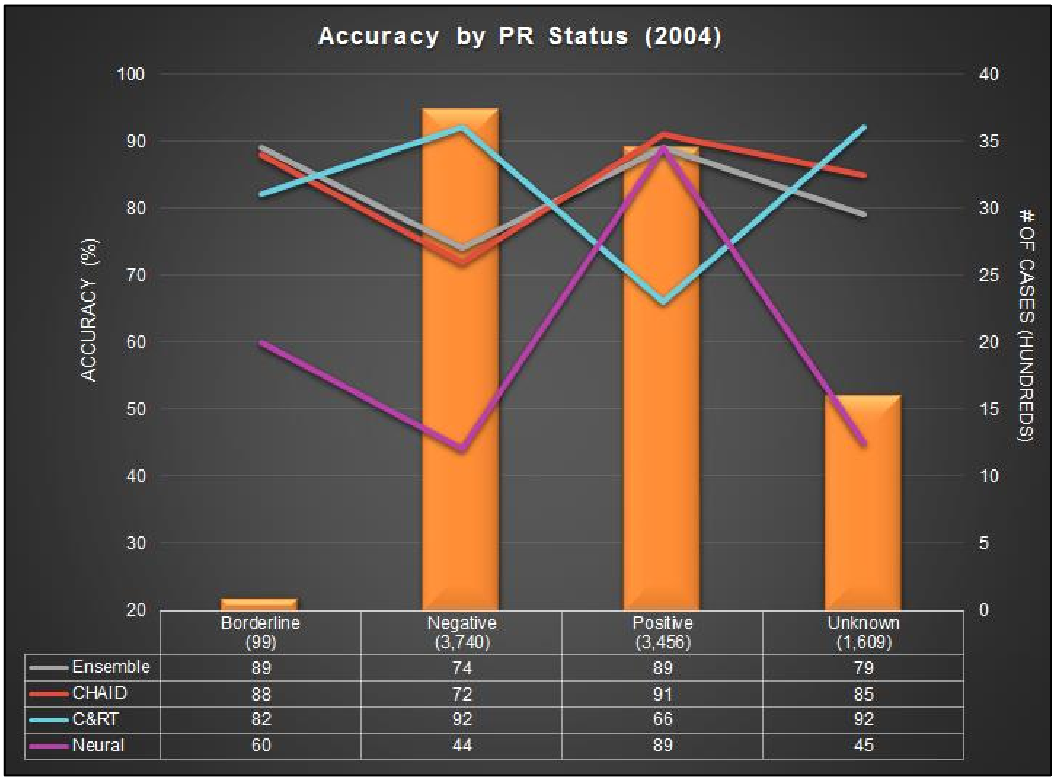


	Borderline (36)	Negative (2,706)	Positive (4,630)	Unknown (1,531)
Ensemble	86	59	91	77
CHAID	83	56	93	83
C&RT	84	96	67	94
Neural	60	25	86	42

Figure 50. Accuracy by Estrogen Receptor Status (2004)

Figure 50 compares the accuracy of each modeling technique by ER status for cases registered in 2004 which are tagged 'Dead' at the cut-off date. The status is recorded as – positive, negative, borderline and unknown. 52% of cases tagged 'Dead', at cut-off date, have positive ER status and 30% cases have negative ER status. The highest number of cases have positive

ER status, CHAID has the highest accuracy for such cases, i.e. 93% followed by Ensemble with 91% accuracy. C&RT has the lowest accuracy for such cases. However, C&RT has the highest accuracy for cases having negative ER status. Neural Network overall has low accuracy as compared to other modeling techniques.



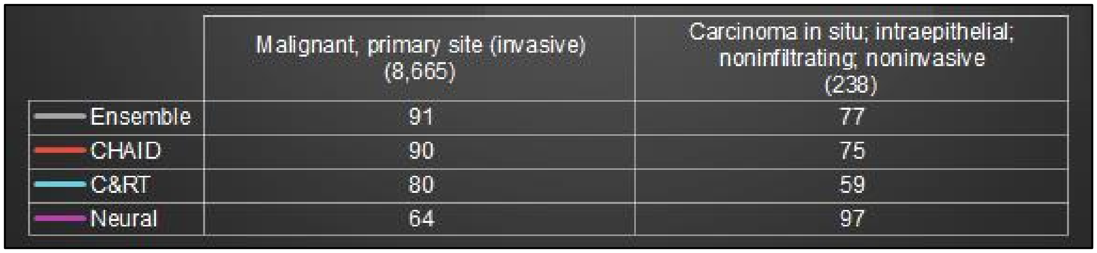
	Borderline (99)	Negative (3,740)	Positive (3,456)	Unknown (1,609)
Ensemble	89	74	89	79
CHAID	88	72	91	85
C&RT	82	92	66	92
Neural	60	44	89	45

Figure 51. Accuracy by Progesterone Receptor Status (2004)

Figure 51 compared the actual and measured average survival months by PR status for cases registered in 2004. The PR status is recorded as – positive, negative, borderline and unknown. 42% of 'Dead' cases have negative PR status, and 38% of cases have a positive PR status. The graph shows exciting trends. C&RT has highest accuracy for cases with negative PR status and unknown PR status. However, CHAID has the highest accuracy for cases with positive PR status. On the other hand, Neural Network and Ensemble have second highest accuracy, i.e.

89% for cases with positive PR status. The Ensemble has the highest accuracy for cases with borderline PR status.

4.2.2.6 Accuracy by Behavior Code ICD-O-3



	Malignant, primary site (invasive) (8,665)	Carcinoma in situ; intraepithelial; noninfiltrating; noninvasive (238)
Ensemble	91	77
CHAID	90	75
C&RT	80	59
Neural	64	97

Figure 52. Accuracy by Behavior ICD-O-3 (2004)

In Figure 52, the accuracy of modeling techniques is compared by behavior type for cases registered in 2004 which are tagged ‘Dead’ at the cut-off date. The cases are categorized as malignant (invasive) and carcinoma in situ (non-invasive). The Ensemble has the highest accuracy of 91% followed by CHAID and C&RT respectively. However, Neural Network has the lowest accuracy for malignant cases. On the other hand, for the cases having carcinoma in situ behavior type, Neural Network has the highest accuracy. However, the Neural Network also highest accuracy for carcinomic cases (97%) followed by Ensemble, CHAID and C&RT respectively.

4.2.3 Impact of Retraining Breast Cancer Predictive Model

Predictive model deployment is an iterative process. The data with which the prediction model is trained should have the same data distribution as the data on which it is tested [96]. With new technological improvements, early detection and screening techniques, the breast cancer data distribution has changed over time [97, 98]. To get more accurate results, retraining the existing model with new data is a good practice. Retraining the predictive model in IBM SPSS Modeler is possible. The existing model is trained with new data, reproducing/refreshing the modeling nuggets. The impact of retraining the model is presented in the next subsections.

4.2.3.1 Case Study of Cases Diagnosed in 2008

In the year 2008, a total of 67,017 breast cancer cases were registered in the SEER database. The predictive model is tested for cases registered in the year 2008 (which is outside training range). The vital status of actual and predicted model's outcome (i.e. measured) are compared.

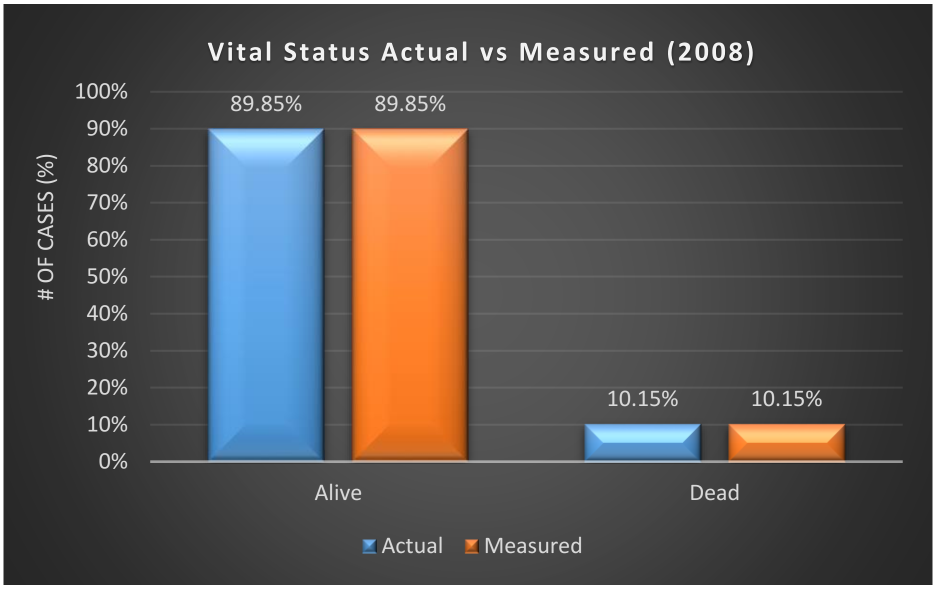


Figure 53. Vital Status comparison (2008)

As shown in Figure 53, out of total cases diagnosed in 2008, 83.39% of cases are tagged 'Alive' and 16.11% are tagged 'Dead' at the cut-off date. The predictive model also predicts 83.39% of cases as 'Alive' and 16.11% as 'Dead' as of the cut-off date (December 31, 2013). This shows that the predictive model is predicting the exact number of mortalities for the year 2008. To check the predictive model's accuracy in predicting survival months, measured survival months are compared with actual survival months. The results are presented by average and accuracy.

Further, the existing model is retrained with more recent data, i.e. cases diagnosed from 2003 to 2007. The same testing dataset (of cases diagnosed in 2008) is used to test the measured average and accuracy of the retrained model. The results from both the experiments are compared by plotting actual vs measured survival months. As long as the model is trained with the recent data possible, it produces a more accurate outcome.

4.2.3.1.1 Comparing outcomes of Predictive Model with Retrained Model by Average Survival Months

Figure 54 compares the actual and measured average survival months. A total of 6,774 cases are tagged 'Dead' at the cut-off date. The average survival months of 'Dead' cases is 28.37 months. CHAID predicts 32 survival months followed by C&RT (39 months), Ensemble (42 months) and Neural Network (53 months). This graph shows that the Neural Network, C&RT and Ensemble are not performing very well.

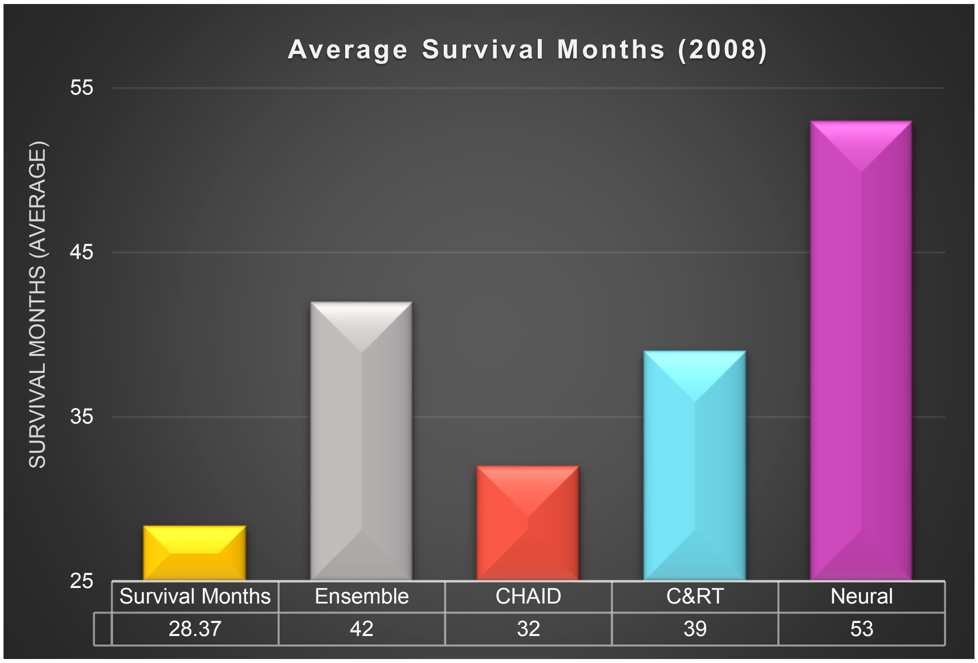


Figure 54. Average Survival Months Actual vs Measured (2008)

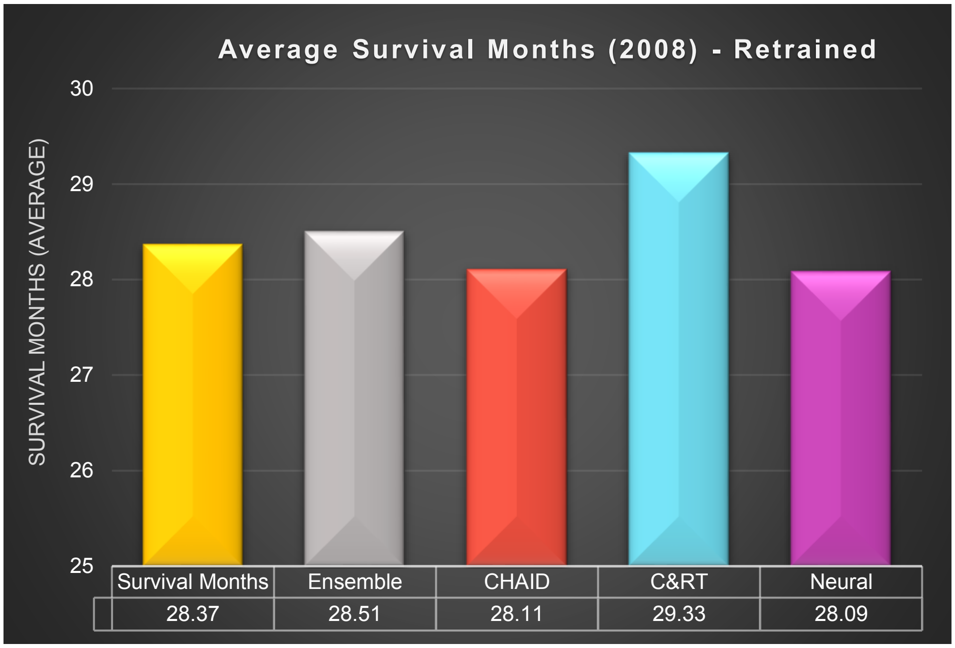


Figure 55. Average Survival Months Actual vs Measured (2008) (Retrained model)

On the other hand, when the predictive model is retrained with 2004-2007 data and tested with the same dataset (2008), the results change drastically. The measured survival months of retrained model ranges from 28.09 to 28.51 survival months when the actual survival months is 28.37 (Figure 55). This shows the impact of retraining the model with the prediction outcomes improving exceptionally.

4.2.3.1.2 Comparing outcomes of Predictive Model with Retrained Model by Accuracy

Figure 56 compares the accuracy of the predictive model when trained with 1988-2003 data. Only cases tagged 'Dead' at the cut-off date are taken into consideration. The graph shows the CHAID has the highest accuracy, i.e. 86%, and C&RT and Ensemble technique's accuracy is comparatively lower, i.e. 61% and 53%, respectively. Neural Network has the lowest accuracy i.e. 14%. When the model is retrained with 2004-2007 historical data, the accuracy improves drastically.

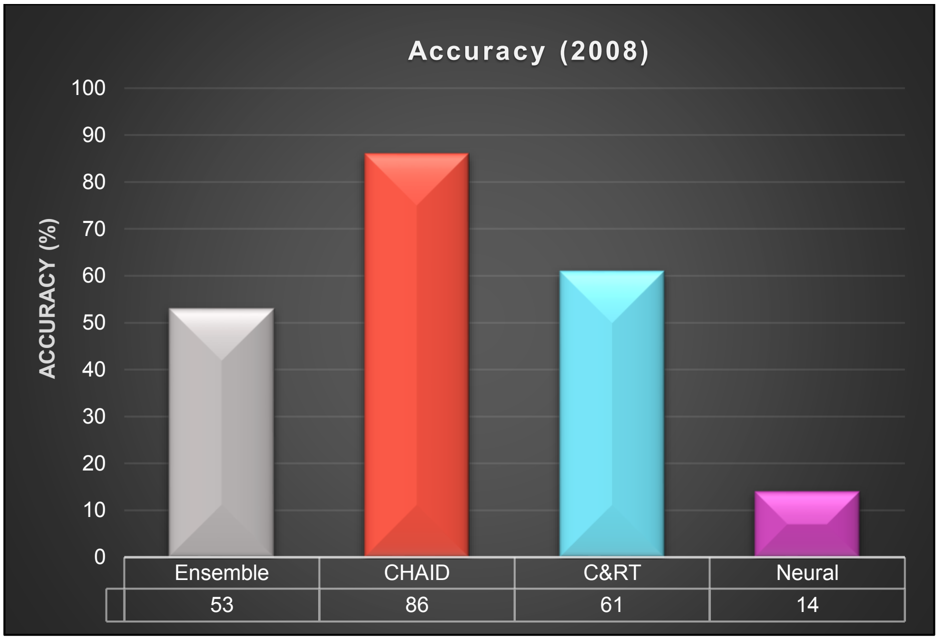


Figure 56. Accuracy of Predictive Model (2008)

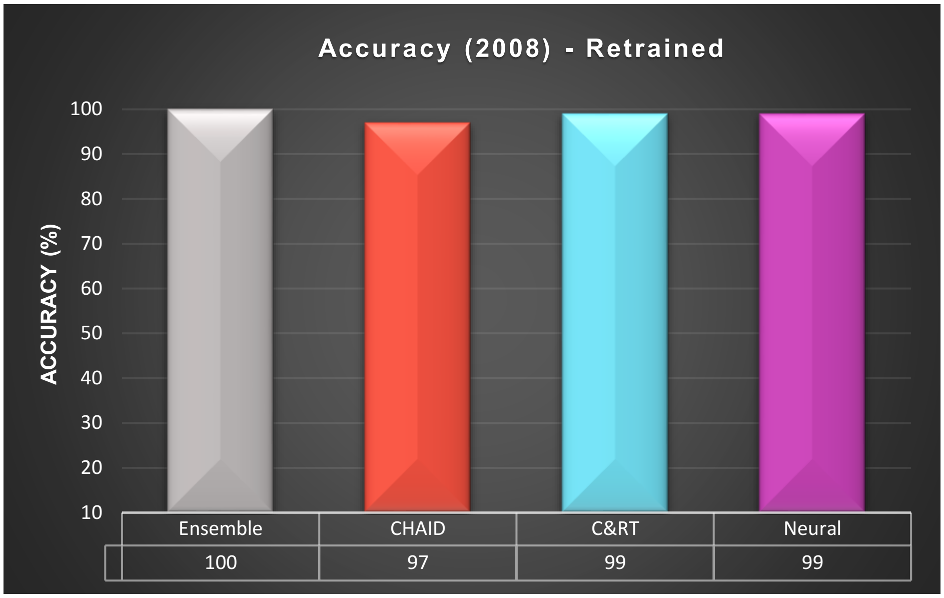


Figure 57. Accuracy of Retrained Predictive Model (2008)

Figure 57 shows Ensemble has 100% accuracy, C&RT and Neural Network have 99% accuracy followed by CHAID. The summary of the impact of retraining the model is presented in Table 10. The 'Trained' columns represent results generated from 1988-2003 trained predictive model, and 'Retrained' columns represent the results generated by the retrained model (2004-2007).

Actual Survival Months	**Average Survival Months**		**Accuracy**	
28.37	**Trained**	**Retrained**	**Trained**	**Retrained**
Ensemble	42	28.51	53	100
CHAID	32	28.11	86	97
C&RT	39	29.33	61	99
Neural	53	28.09	14	99

Table 10. Impact of Retraining

Apart from 2004 and 2008 test datasets as shown in above subsections, similar experiments were conducted for other datasets which are outside the training range, and similar results are achieved.

4.2.4 Summary of Predictive Modeling

The developed predictive model is tested with different datasets outside the training data. Predicted average survival months and accuracy of each modeling technique is observed for cases diagnosed in 2004 and 2008. The test results are validated by comparing the actual survival months of cases tagged ‘Dead’ on the cut-off date. Following key observations are made:

- The Ensemble of selected modeling techniques performs better than other models with 93% accuracy.
- CHAID with 92% accuracy ranks second following Ensemble technique.
- C&RT and Neural Network modeling techniques have 80% and 66% accuracy, respectively.
- Results are consistent across various attributes (age, marital status, regional nodes).
- Ensemble and CHAID perform best amongst all other modeling techniques used in this study.
- Neural Network performs poor consistently across all the experiments. Some exceptions were observed when the number of cases was low.
- Testing the predictive model with 2008 dataset gives low accuracies for all the models
- On retraining the predictive model with more recent data, the accuracy of the predictive model improved significantly resulting in Ensemble performing best with 100% accuracy.
- The accuracy of Neural Networks improves upon retraining. The inconsistency of data, i.e. missing values of T, N and M variables in testing and training datasets does not exist anymore. The 2004-2007 retraining dataset has over 200,000 cases. The Neural

Networks is retrained with missing values of these variables and thus predicts with very high accuracy.

All experimental results presented are for cases tagged “Dead” at the study cut-off date due to uncertainty of survival months beyond this date. The noted survival months of “Alive” cases is not their survival months as patients could have died one month, one year or 5 years beyond the study cut-off date. It is not possible to validate the results of “Alive” cases and hence they are not included in the results. However, if both “Alive” and “Dead” cases are taken into consideration, Neural Networks also perform well with higher accuracy across all the experiments.

On testing with various other datasets outside training ranges, similar patterns of results are observed. By analysis of various data sets, it is concluded that the developed predictive model is a powerful application when retrained periodically.

4.3 Summary

In this chapter, the experimental results and analysis of the predictive model’s outcome are presented. The results are presented by comparing actual and measured survival months and validated by computing accuracy of each modeling technique. The visualization dashboard snapshots are also presented along with an analysis of each report. In the next chapter, the conclusion of the thesis is presented along with the possible future work.

5. Conclusion and Future Work

Breast cancer is the most common cancer in women across the globe. With an increasing number of breast cancer incidences an early detection and treatment is the ideal way to decrease breast cancer mortalities. It is also important to note that not all breast cancer patients die due to breast cancer, many deaths happen due to other diseases which are a consequence of breast cancer and metastasized cancers. Breast cancer remains the second most common cause of death in women after heart diseases. Despite technological advancements and early cancer detection techniques, one-size-fits-all is a common practice used for developing a cancer treatment. Data-driven research outcomes are steps towards advancements in cancer treatment. These outcomes include cancer prognosis, survival outcome and side effects of therapies. This research focuses on the survival outcome of breast cancer patient and uses historical data to develop a visualization dashboard and a predictive model. The survival outcome of the patient not only helps physicians in designing a custom treatment plan for each patient, but it also helps keeps the patient informed and involved in the process of cancer treatment.

In this research, an interactive, end-to-end process to cleanse, integrate, analyze, and visualize enormous amount of data has been presented. The purpose is to enable healthcare professionals, patients and policymakers with a better understanding of the hidden patterns in data, which in turn, could be useful to collectively improve the quality of healthcare. Forty years of breast cancer data (over one million records) extracted from the SEER database was used to demonstrate the analytical power of data visualization. The research approach involved using several data preprocessing techniques on the raw data followed by a selection of the relevant variables, before feeding the processed data to the SQL Server for analysis. Tableau was used for understanding and interpreting the data along several dimensions. Dynamic reports generated using the drill-down capabilities of the dashboard provide insights at a finer granularity. The dashboard shows incidence and mortality trends and highlights the underlying patterns observed in breast cancer patients which could be used to support clinical decisions made by physicians in formulating treatment plans. The dashboard is scalable and capable of integrating new data in real-time. Although SEER data from the US currently power the dashboard, it can be configured to use data from other sources as well. A better understanding of the incidence and mortality trends could potentially guide data-driven resource allocation. Physicians could also use this information to educate patients and create more awareness about the disease.

The pre-processed data is used to develop a predictive model to predict survival months of individual breast cancer patient. For predictions, Neural Network, CHAID, C&RT modeling techniques along with their Ensemble are used, which predict survival months of each patient. The predictive model calculator allows user to enter values of variables such as *Marital Status; Race/ethnicity; Age; Laterality; Histologic Type; Behavior Code; Regional nodes positive and examined; Cancer-directed surgery; Radiation; Surgery of Primary Site; ER and PR status; T, N, M value; Year of diagnosis.* The predictive model is then run and produces survival months predicted by each modeling technique and the Ensemble modeling technique. The predictive model is developed in SPSS Modeler 18.1 [88] (explained in Chapter 3).

5.1 Future Work

This research can be extended in several ways as described below:

- The predictive model can be deployed on cloud and re-trained and tested online.
- The calculator can be developed with a web-interface and hosted on available web services such as Amazon Web Services, IBM Bluemix.
- A similar predictive model can be developed for other cancer types and diseases.
- The developed predictive model can be trained and tested with different demographics data.
- Automating the process of re-training the developed predictive model with the most recent data.
- The Visualization dashboard can be extended to other diseases and cancer types.

Bibliography

[1] WHO, "Cancer," [Online]. Available: http://www.who.int/cancer/en/.

[2] American Cancer Society, "Cancer Facts & Figures 2016," [Online]. Available: http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079.pdf.

[3] Cancer Canadian Society, "Cancer statistics at a glance," Cancer Canadian Society, 2015. [Online]. Available: http://www.cancer.ca/en/cancer-information/cancer-101/cancer-statistics-at-a-glance/?region=on#.

[4] Dursun Delen, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artifiical Intelligence in Medicine,* vol. 34, no. 2, pp. 113-127, 2005.

[5] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on SEER data," *Scientific Programming - Biological Knowledge Discovery and Data Mining,* vol. 20, no. 1, pp. 29-42, August 2012.

[6] American Cancer Society, "About Breast Cancer," [Online]. Available: https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html.

[7] SAP Voice, "Four-Time Cancer Survivor on the Importance of Data-Driven Health Decisions," BrandVoice - Forbes, 2016.

[8] IBM, "IBM Watson Health," IBM, [Online]. Available: https://www.ibm.com/watson/health/oncology-and-genomics/oncology/.

[9] SAP, "The SAP Corporate Oncology Program," SAP, [Online]. Available: https://www.sap.com/canada/about/careers/joining/benefits.html.

[10] L. Mallory, "IBM's Watson is really good at creating cancer treatment plans," Engadget, 2017. [Online]. Available: https://www.engadget.com/2017/06/01/ibm-watson-cancer-treatment-plans/.

[11] K.-M. Wang, M. Bunjira, W.-L. Wu and Y. Lin, "Optimal Data Mining Method or Predicting Breast Cancer Survivability," *International Journal of Innovative Management, Information and Production (ISME International),* vol. 3, no. 2, pp. 28-33, 2013.

[12] N. A. Christakis, J. L. Smith, C. M. Parkes and E. B. Lamont, "Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort studyCommentary: Why do doctors overestimate? Commentary: Prognoses should be based on proved indices not intuition," *British Medical Journal,* vol. 320, no. 7233, pp. 469-472, 2000.

[13] ASCO Cancer.Net, "Understanding Cancer Research Study design and How to Evaluate Results," Cancer.Net, [Online]. Available: https://www.cancer.net/research-and-advocacy/introduction-cancer-research/understanding-cancer-research-study-design-and-how-evaluate-results.

[14] Susan G. Komen, "Komen Perspectives - The Importance of Clinical trials in Breast Cancer Treatment," [Online]. Available: https://ww5.komen.org/KomenPerspectives/Komen-Perspectives---The-Importance-of-Clinical-Trials-in-Breast-Cancer-Treatment-(July-2012).html.

[15] American Society of Clinical Oncology, "Outcomes of cancer treatment for technology assessment and cancer treatment guidelines," *Journal of Clinical Oncology,* vol. 14, no. 2, pp. 671-679, 1996.

[16] T. Smith and B. Smith, "Survival Analysis And The Application Of Cox's Proportional Hazards modeling using SAS," in *Proceedings of the twenty-sixth annual SAS user's group international conference*, 2001, pp. 224-246.

[17] D. G. Kleinbaum and M. Klein, "Survival Analysis," in *Survival Analysis*, Springer, 2012, pp. 1-54.

[18] G. P. Meren, "BOSOM Calculator: A Breast Cancer Outcome- Survival Online Measurement Calculator using Data Mining and Predictive Modeling on SEER data," University of the Philippines, 2014.

[19] P. M. Ravdin, G. M. Clark, S. G. Hilsenbeck, M. A. Owens, P. Vendely, M. R. Pandian and W. L. McGuire, "A demonstration that breast cancer recurrence can be predicted by Neural Network analysis," *Breast Cancer Research and Treatment,* vol. 21, no. 1, pp. 44-53, 1992.

[20] N. Lavrac, "Selected techniques for data mining in medicine," *Artificial Intelligence in Medicine,* vol. 16, no. 1, pp. 3-23, 1999.

[21] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine,* vol. 26, no. 1-2, pp. 1-24, 2002.

[22] U. Fayyad, G. Piatestsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine,* vol. 17, no. 3, p. 37, 1996.

[23] A. G. Eapen, "Application of Data mining in Medical Applications," University of Waterloo, 2004.

[24] WIDESKILLS, "Introduction to Data Mining Tasks," [Online]. Available: http://www.wideskills.com/data-mining-tutorial/05-data-mining-tasks.

[25] S. Gupta, D. Kumar and A. Sharma, "Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis," *Indian Journal of Computer Science and Engineering (IJCSE),* vol. 2, no. 2, pp. 188-195, 2011.

[26] J. G. Stadler, D. Kipp, J. D. Siewert, F. Tessa and N. E. Lewis, "Improving the Efficiency and Ease of Healthcare Analysis Through Use of Data Visualization Dashboards," *Big Data,* vol. 4, no. 2, pp. 129-135, 2016.

[27] Einforchips, "White Paper - Revolutionizing The Healthcare Industry with Big Data, Analytics and Visualization".

[28] Institute of Health Metrics and Evaluation (IHME), "GBD Compare Data Visualization," IHME, University of Washington, 2017. [Online]. Available: https://vizhub.healthdata.org/gbd-compare/.

[29] Centres for Disease Control and Prevention, "United States Cancer Statistics: Data Visualizations," Centres for Disease Control and Prevention, 2017. [Online]. Available: https://nccd.cdc.gov/USCSDataViz/rdPage.aspx.

[30] WHO, "Global Cancer Observatory," World Health Organization, [Online]. Available: https://gco.iarc.fr/.

[31] SEER, "Data & Software for Researchers," National Cancer Institute, [Online]. Available: seer.cancer.gov/resources/.

[32] G. Zorluoglu and M. Agaoglu, "Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods," *International Journal of Bioinformatics and Biomedical Engineering,* vol. 1, no. 3, pp. 318-322, 2015.

[33] P. M. Ravdin, L. A. Siminoff, G. J. Davis, M. B. Mercer, J. Hewlett, N. Gerson and H. L. Parker, "Computer Program to Assist in Making Decisions about Adjuvant Therapy for Women With Early Breast Cancer," *Journal of Clinical Oncology,* vol. 19, no. 4, pp. 980-991, 2001.

[34] G. C. Wishart, E. M. Azzato, D. C. Greenberg, J. Rashbass, O. Kearins, G. Lawrence, C. Caldas and P. D. Pharoah, "PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer," *Breast Cancer Research,* vol. 12, no. 1, p. 401, 2010.

[35] Statistics Canada, "The 10 leading causes of death, 2011," [Online]. Available: http://www.statcan.gc.ca/pub/82-625-x/2014001/article/11896-eng.htm.

[36] National Cancer Institute, "Diagnosis and Staging," [Online]. Available: www.cancer.gov/about-cancer/diagnosis-staging/prognosis.

[37] L. A. Winters-Miner, "Seven ways predictive analytics can improve healthcare," *Elsevier's Daily stories for the science, Technology and health communities,* 2014.

[38] Breast Cancer Organisation, "Why So Many Types of Breast Cancer Treatment?," [Online]. Available: http://www.breastcancer.org/treatment/planning/types_treatment.

[39] Health Catalyst, "Predictive Analytics Solutions," [Online]. Available: https://www.healthcatalyst.com/predictive-analytics.

[40] U.S. Cancer Statistics Working Group., "U.S. Cancer Statistics Data Visualizations Tool," .S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute, June 2018. [Online]. Available: www.cdc.gov/cancer/dataviz.

[41] Genomic Data Commons, "GDC DAVE Tools," National Cancer Institute, [Online]. Available: https://portal.gdc.cancer.gov/.

[42] IHME, "About IHME," IHME, [Online]. Available: http://www.healthdata.org/about.

[43] Institute of Health Metrics and Evaluation (IHME), "Data Visualizations," Institute of Health Metrics and Evaluation, 2017. [Online]. Available: http://www.healthdata.org/results/data-visualizations.

[44] CDC, "Centers for Disease Control and Prevention," Centers for Disease Control and Prevention, 2017. [Online]. Available: https://www.cdc.gov/.

[45] National Cancer Institute (NCI), "About NCI," National Cancer Institute, [Online]. Available: https://www.cancer.gov/about-nci.

[46] WHO, "World Health Organization Home," World Health Organization , [Online]. Available: http://www.who.int/en/.

[47] WHO, "Section of Cancer Surveillance," World Health Organization, [Online]. Available: http://www.iarc.fr/en/research-groups/sec1/index.php.

[48] GLOBOCAN, "Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012," World Health Organization, 2012. [Online]. Available: http://globocan.iarc.fr/Pages/fact_sheets_population.aspx.

[49] WHO, "CI5 Cancer Incidence in Five Continents," World Health Organization, [Online]. Available: http://ci5.iarc.fr/.

[50] WHO, "International Incidence of Childhood Cancer 3," World Health Organization, [Online]. Available: http://iicc.iarc.fr/.

[51] WHO, "Cancer survival in Africa, Asia, the Caribbean and Central America," World Health Organization, [Online]. Available: http://survcan.iarc.fr/.

[52] Genomic Data Commons, "About the GDC," National Cancer Institute, [Online]. Available: https://gdc.cancer.gov/about-gdc.

[53] National Cancer Institute (NCI), "About TCGA," National Human Genome Research Institute, [Online]. Available: https://cancergenome.nih.gov/abouttcga.

[54] National Cancer Institute (NCI), "TARGET: Therapeutically Applicable Research To Generate Effective Treatments," National Cancer Institute Office of Cancer Genomics, [Online]. Available: https://ocg.cancer.gov/programs/target.

[55] D. Delen, G. Walker and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine,* vol. 34, no. 2, pp. 113-127, 2005.

[56] M. Riihimäki, H. Thomsen, A. Brandt, J. Sundquist and K. Hemminki, "Death causes in breast cancer patients," *Annals of Oncology,* vol. 23, no. 3, pp. 604-610, 2012.

[57] A. Bellaachia and E. Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques," *Age,* vol. 58, no. 13, pp. 10-110, 2006.

[58] WEKA, "Home," The University of Waikato, [Online]. Available: https://weka.wikispaces.com/.

[59] A. Endo, S. Takeo and H. Tanaka, "Comparison of Seven Algorithms to Predict Breast Cancer Survival," *Biomedical Soft Computing and Human Sciences,* vol. 13, no. 2, pp. 11-16, 2008.

[60] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability," *International Journal of Computer Science and Mobile Computing,* vol. 3, no. 1, pp. 10-22, 2017.

[61] Z. K. Senturk and R. Kara, "Breast Cancer Diagnosis via Data Mining: Performance Analysis of Seven Different Algorithms," *Computer Science & Engineering: An International Journal (CSEIJ),* vol. 4, no. 1, p. 35, 2014.

[62] V. Chaurasia and S. Pal, "A Novel Approach for Breast Cancer Detection Using Data Mining Techniques," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 2, no. 1, 2017.

[63] RapidMiner, "Data Science Behind Every Decision," [Online]. Available: https://rapidminer.com/.

[64] B. V. Sumana and T. Santhanam, "An empirical comparison of ensemble and hybrid classification," *Proc Processing and VLSI,* pp. 463-470, 2014.

[65] M. U. Khan, J. P. Choi, H. Shin and M. Kim, "Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare," in *30th Annual International IEEE EMBS Conference*, Vancouver, British Columbia, 2008.

[66] J. P. Choi, T. H. Han and R. W. Park, "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis," *Journal of Korean Society of Medical Informatics,* vol. 15, no. 1, pp. 49-57, 2009.

[67] E. Alpaydin, Introduction to Machine Learning Second Edition, MIT Press, 2004.

[68] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke and C.-F. Tsai, "SVM and SVM Ensembles in Breast Cancer Prediction," *PloS one,* vol. 12, no. 1, p. e0161501, 2017.

[69] B. F. Hankey, L. A. Ries and B. K. Edwards, "The surveillance, epidemiology, and end results program: A National Resource," *Cancer Epidemiology & Prevention Biomarkers,* vol. 8, no. 12, p. 1117, 1999.

[70] NCI Surveillance, Epidermiology, and End Results Program (SEER), "SEER*Stat Software," National Cancer Institute, [Online]. Available: https://seer.cancer.gov/seerstat/.

[71] Microsoft, "SQL Server 2016," Microsoft, [Online]. Available: https://www.microsoft.com/en-ca/sql-server/sql-server-2016.

[72] Tableau, "Healthcare," Tableau, [Online]. Available: https://www.tableau.com/stories/topic/healthcare.

[73] Klipfolio, "Operational dashboards vs Analytical Dashboards," [Online]. Available: http://www.klipfolio.com/resources/articles/operational-analytical-bi-dashboards.

[74] X. Zhang, K. Gallagher and S. Goh, "BI Application: Dashboards for Healthcare.," in *AMCIS*, Detroit, 2011.

[75] Tableau, "Tableau Home," [Online]. Available: http://www.tableau.com/.

[76] Interworks, "Why Tableau," Interworks, [Online]. Available: https://interworks.co.uk/business-intelligence/why-tableau/.

[77] D. Howlett, "Why Tableau is crushing it for 21st century analysis," 2014. [Online]. Available: http://diginomica.com/2014/09/11/tableau-crushing/.

[78] Gartner, "Magic Quadrant for Business Intelligence and Analytics Platforms," Gartner, [Online]. Available: https://www.gartner.com/doc/reprints?ct=160204&id=1-2XXET8P.

[79] A. Chandramouly., "For fourth year, Gartner names Tableau a 'leader' in Magic Quadrant," February 2016. [Online]. Available:

https://www.tableau.com/about/blog/2016/2/fourth-year-gartner-names-tableau-leader-magic-quadrant-49719.

[80] SEER, "Overview of SEER program," Surveillance Epidemiology and End Results, [Online]. Available: http://www.seer.cancer.gov/about/overview.html.

[81] National Cancer Institute, "Surveillance Research Program," National cancer Institute, [Online]. Available: https://surveillance.cancer.gov/.

[82] National Cancer Institute, "Divison of Cancer Control and Population Sciences," National Cancer Institute, [Online]. Available: https://cancercontrol.cancer.gov/.

[83] NCI's Surveillance, Epidermiology, and End Results Program (SEER), "SEER Registry Grouping for Analyses," National Cancer Institure Surveillance, Epidermiology, and End Results Program, [Online]. Available: http://seer.cancer.gov/registries/terms.html.

[84] BC Cancer Agency, "Research at BC Cancer Agency- Centre for the North," [Online]. Available: http://www.bccancer.bc.ca/our-services/centres-clinics/centre-for-the-north/research.

[85] University of Northern British Columbia, "Dr Robert Olson," [Online]. Available: http://www.unbc.ca/robert-olson.

[86] C. Nyce, "Predictive Analytics White Paper," *American Institute for CPCU. Insurance Institute of America,* pp. 9-10, 2007.

[87] S. Finlay, "Predictive Analytics," in *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*, Springer, p. 237.

[88] IBM, "IBM SPSS Modeler," [Online]. Available: http://www-01.ibm.com/support/docview.wss?uid=swg27050406.

[89] R. Mikut and M. Reischl, "Data mining tools," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 1, no. 5, pp. 431-443, 2011.

[90] IBM, "IBM Cognos Analytics," IBM, [Online]. Available: http://www.ibm.com/analytics/us/en/technology/cognos-software/#what-is-cognos.

[91] Tableau, "Stories," Tableau, [Online]. Available: https://onlinehelp.tableau.com/current/pro/desktop/en-us/stories.html.

[92] J. Brandt, J. P. Garne, I. Tengrup and J. Manjer, "Age at diagnosis in relation to survival following breast cancer: a cohort study," *World Journal of Surgical Oncology,* vol. 13, no. 1, p. 33, 2015.

[93] SEER, "SEER Registeries," Surveillance, Epidemiology, and End Results (SEER) Program, [Online]. Available: https://seer.cancer.gov/registries/.

[94] BreastCancer.Org, "Lymph Node Involvement," BreastCancer.Org, 2017. [Online]. Available: http://www.breastcancer.org/treatment/surgery/lymph_node_removal/axillary_dissection.

[95] Interworks, "Tableau Essesntials: Chart Types - Box-and-Whisker Plot," [Online]. Available: https://interworks.com/blog/ccapitula/2014/12/09/tableau-essentials-chart-types-box-and-whisker-plot/.

[96] Amazon Machine Learning, "Retraining Models on New Data," AWS, [Online]. Available: https://docs.aws.amazon.com/machine-learning/latest/dg/retraining-models-on-new-data.html.

[97] M. L. Lousdal, I. S. Kristiansen, B. Moller and H. Stovring, "Trends in breast cancer stage distribution before, during and after introduction of a screening programme in Norway," *The European Journal of Public Health,* vol. 24, no. 6, pp. 1017-1022, 2014.

[98] A. Aljarrah and W. Miller, "Trends in the distribution of breast cancer over time in the southeast of Scotland and review of the literature," *Ecancermedicalscience,* vol. 8, p. 427, 2014.

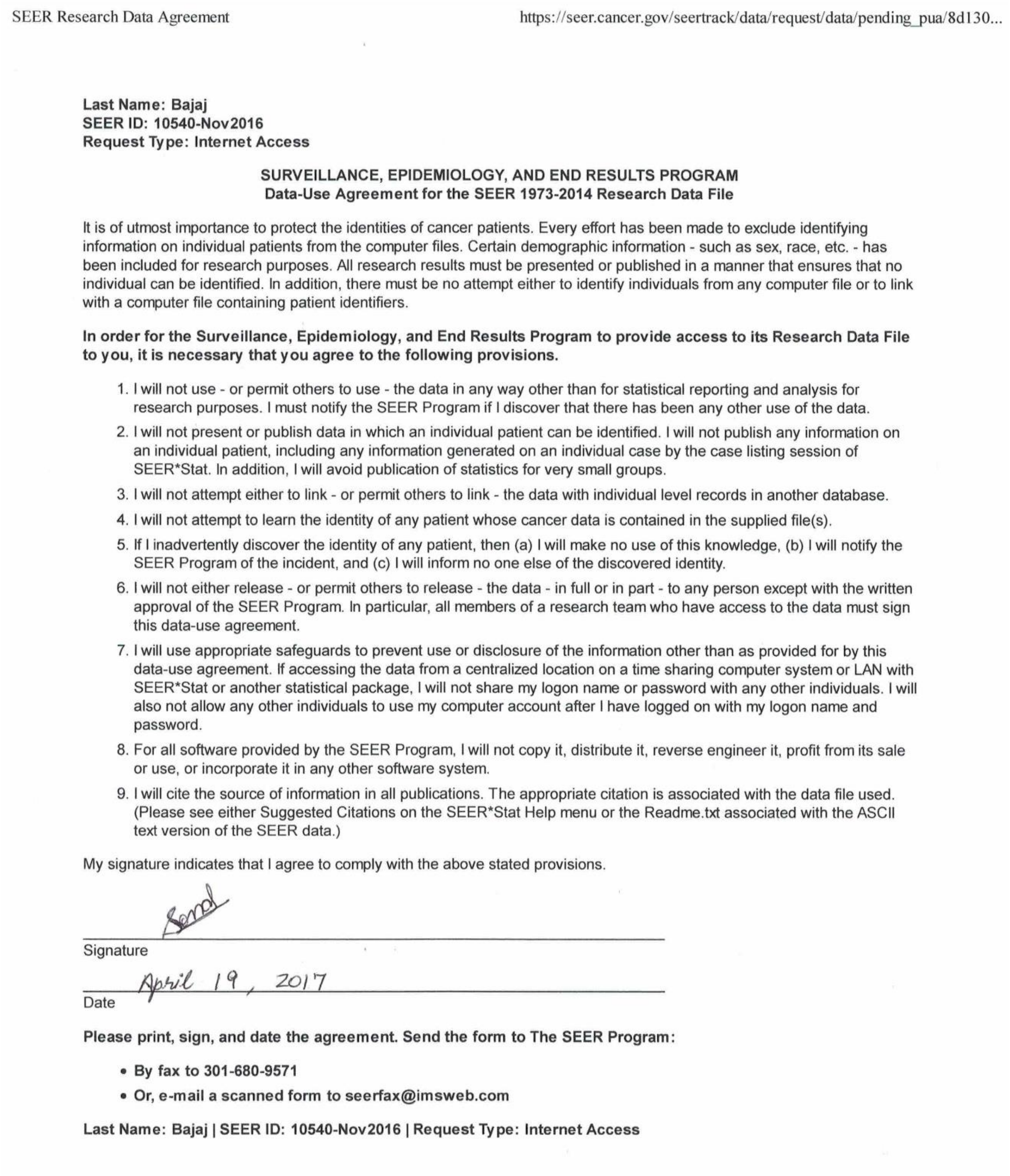
[99] "11 Reasons to start a database instead of Excel and Word," The IT Training Surgery, [Online]. Available: https://theittrainingsurgery.com/11-reasons-start-using-database-instead-excel-word/.

[100] "SAS Enterprise Miner," SAS, [Online]. Available: https://www.sas.com/en_ca/software/enterprise-miner.html.

[101] SAP, "Power Users' Guide," SAP Documentation, [Online]. Available: https://help.sap.com/saphelp_nw70ehp2/helpdata/en/1e/013f420e09b26be10000000a155106/frameset.htm.

Appendix A

Forms



SEER Research Data Agreement https://seer.cancer.gov/seertrack/data/request/data/pending_pua/8d130...

Last Name: Bajaj
SEER ID: 10540-Nov2016
Request Type: Internet Access

SURVEILLANCE, EPIDEMIOLOGY, AND END RESULTS PROGRAM
Data-Use Agreement for the SEER 1973-2014 Research Data File

It is of utmost importance to protect the identities of cancer patients. Every effort has been made to exclude identifying information on individual patients from the computer files. Certain demographic information - such as sex, race, etc. - has been included for research purposes. All research results must be presented or published in a manner that ensures that no individual can be identified. In addition, there must be no attempt either to identify individuals from any computer file or to link with a computer file containing patient identifiers.

In order for the Surveillance, Epidemiology, and End Results Program to provide access to its Research Data File to you, it is necessary that you agree to the following provisions.

1. I will not use - or permit others to use - the data in any way other than for statistical reporting and analysis for research purposes. I must notify the SEER Program if I discover that there has been any other use of the data.
2. I will not present or publish data in which an individual patient can be identified. I will not publish any information on an individual patient, including any information generated on an individual case by the case listing session of SEER*Stat. In addition, I will avoid publication of statistics for very small groups.
3. I will not attempt either to link - or permit others to link - the data with individual level records in another database.
4. I will not attempt to learn the identity of any patient whose cancer data is contained in the supplied file(s).
5. If I inadvertently discover the identity of any patient, then (a) I will make no use of this knowledge, (b) I will notify the SEER Program of the incident, and (c) I will inform no one else of the discovered identity.
6. I will not either release - or permit others to release - the data - in full or in part - to any person except with the written approval of the SEER Program. In particular, all members of a research team who have access to the data must sign this data-use agreement.
7. I will use appropriate safeguards to prevent use or disclosure of the information other than as provided for by this data-use agreement. If accessing the data from a centralized location on a time sharing computer system or LAN with SEER*Stat or another statistical package, I will not share my logon name or password with any other individuals. I will also not allow any other individuals to use my computer account after I have logged on with my logon name and password.
8. For all software provided by the SEER Program, I will not copy it, distribute it, reverse engineer it, profit from its sale or use, or incorporate it in any other software system.
9. I will cite the source of information in all publications. The appropriate citation is associated with the data file used. (Please see either Suggested Citations on the SEER*Stat Help menu or the Readme.txt associated with the ASCII text version of the SEER data.)

My signature indicates that I agree to comply with the above stated provisions.

Sonal
Signature

April 19, 2017
Date

Please print, sign, and date the agreement. Send the form to The SEER Program:

- **By fax to 301-680-9571**
- **Or, e-mail a scanned form to seerfax@imsweb.com**

Last Name: Bajaj | SEER ID: 10540-Nov2016 | Request Type: Internet Access

Appendix B

The list of all the variables selected with their definition[9]and coding are listed below:

1) CS mets at DX-bone (2010+)

“Identifies the presence of distant metastatic involvement of bone at time of diagnosis. The presence of metastatic bone disease at diagnosis is an independent prognostic indicator, and it is used by Collaborative Staging to derive TNM-M codes and SEER Summary Stage codes for some sites. This field should be coded for all solid tumors, Kaposi sarcoma, Unknown Primary Site, and Other and Ill-Defined Sites. Only available for 2010+ diagnosis. This includes only the bone, not the marrow.”

Codes

0: No
1: Yes
8: N/A
9: Unknown
14: Blank(s)

2) CS mets at DX-lung (2010+)

“Identifies the presence of distant metastatic involvement of the lung at the time of diagnosis. The presence of metastatic lung disease at diagnosis is an independent prognostic indicator, and it is used by collaborative Staging to derive TNM-M codes and SEER Summary Stage codes for some sites. Only available for 2010+ diagnosis. *Note:* This includes only the lung, not pleura or pleural fluid.”

Codes

0: No
1: Yes
8: N/A
9: Unknown
14: Blank(s)

[9] All the definitions are from SEER’s documentation for text files [49].

3) CS mets at DX-liver (2010+)

"Identifies the presence of distant metastatic involvement of the liver at time of diagnosis. The presence of metastatic liver disease at diagnosis is an independent prognostic indicator, and it is used by Collaborative Staging derive TNM-M codes and SEER Summary Stage codes for some sites. Only available for 2010+ diagnosis."

Codes

0: No
1: Yes
8: N/A
9: Unknown
14: Blank(s)

4) CS mets at DX-brain (2010+)

"Identifies the presence of distant metastatic involvement of the brain at the time of diagnosis. The presence of metastatic brain disease at diagnosis is an independent prognostic indicator, and it is used by Collaborative Staging to derive TNM-M codes and SEER Summary Stage codes for some sites. This field should be coded for all solid tumors, Kaposi sarcoma, Unknown Primary Site, and Other and Ill-Defined Primary Sites. Only available for 2010+ diagnosis."

Codes

0: No
1: Yes
8: N/A
9: Unknown
14: Blank(s)

5) Breast Subtype (2010+)

"Created with combined information from ER Status Recode Breast Cancer (1990+), PR Status Recode Breast Cancer (1990+), and Derived HER2 Recode (2010+)."

Codes

1: Her2+/HR+
2: Her2+/HR-
3: Her2-/HR+
4: Triple Negative
5: Unknown
9: Not 2010+ Breast

6) Vital status recode (study cutoff used)

“Any patient that dies after the follow-up cut-off date is recoded to alive as of the cut-off date.”

Codes

1: Alive
4: Dead

7) Age recode with < 1 year old

“The age recode variable is based on Age at Diagnosis (single-year ages). The groupings used in the age recode variable are determined by the age groupings in the population data. This recode has 19 age groups in the age recode variable (< 1 year, 1-4 years, 5-9 years, 85+ years).”

Codes

00: Age 00
01: Age 01-04
02: Age 05-09
03: Age 10-14
04: Age 15-19
05: Age 20-24
06: Age 25-29
07: Age 30-34
08: Age 35-39
09: Age 40-44
10: Age 45-49
11: Age 50-54
12: Age 55-59
13: Age 59-60
14: Age 65-69
15: Age 70-74
16: Age 75-79
17: Age 80-84
18: Age 85+
99: Unknown Age

8) Radiation

“The method of radiation therapy performed as part of the first course of treatment.”

Codes

0: None; diagnosed at autopsy
1: Beam radiation

2: Radioactive implants
3: Radioisotopes
4: Combination of 1 with 2 or 3
5: Radiation, NOS – method or source not specified
6: Other radiation (1973-1987 cases only)
7: Patient or patient's guardian refused radiation therapy
8: Radiation recommended, unknown if administered
9: Unknown if radiation administered

9) Radiation sequence with surgery

"The order in which surgery and radiation therapies were administered for those patients who had both surgery and radiation"

Codes

0: No radiation and/or surgery as defined above
2: Radiation before surgery
3: Radiation after surgery
4: Radiation both before and after surgery
5: Intraoperative radiation therapy
6: Intraoperative radiation therapy with other radiation given before or after surgery
9: Sequence unknown, but both surgery and radiation were given

10) N value-based on AJCC 3rd (1998-2003)

"Derived by algorithm from extent of disease (EOD). N value denotes the degree of nearby lymph nodes involved."

Codes

00:	N0
10:	N1
11:	N1a
12:	N1b
19:	N1x
20:	N2
21:	N2a
22:	N2b
23:	N2c
30:	N3
70:	NXr
80:	Nxu
90:	NX

11) M value-based on AJCC 3rd (1998-2003)

"Derived by algorithm from extent of disease (EOD). M value denotes the presence of distant metastasis."

Codes

00: M0
10: M1
99: MX
11: M1a
12: M1b

12) T value-based on AJCC 3rd (1998-2003)

"Derived by algorithm from extent of disease (EOD). T value denotes the size of the original (primary) tumor."

Codes

00: Tis
01: Ta
10: T1
11: M1a
12: M1b
13: T1c
16: T1a1
17: T1a2
19: T1x
20: T2
21: T2a
22: T2b
23: T2c
29: T2x
30: T3
31: T3a
32: T3b
33: T3c
39: T3x
40: T4
41: T4a
42: T4b
43: T4c
44: T4d
49: T4x

70: T0
71: T0a
72: T0b
81: Txa
82: TXb
83: TXc
84: TXd
99: TX

13) Regional Nodes Positive

“Records the exact number of regional lymph nodes examined by the pathologist that were found to contain metastases.”

Codes

0: All nodes examined are negative
01-89: Exact number of nodes positive
90: 90 or more nodes are positive
95: Positive aspiration of lymph node(s) was performed
97: Positive nodes are documented, but number is unspecified
98: No nodes were examined
99: Unknown whether nodes are positive; not applicable; not stated in patient record

14) Regional Nodes Examined

“Records the total number of regional lymph nodes that were removed and examined by the pathologist.”

Codes

0: No nodes were examined
01-89: Exact number of nodes examined
90: 90 or more nodes were examined
95: No regional nodes were removed, but aspiration of regional nodes was performed
96: Regional lymph node removal was documented as a sampling, and the number of nodes is unknown/ not stated
97: Regional lymph node removal was documented as a dissection, and the number of nodes is unknown/not stated
98: Regional lymph node were surgically removed, but the number of lymph nodes is unknown/not stated and not documented as a sampling or dissection, nodes were examined, but the number is unknown
99: Unknown whether nodes are positive; not applicable; not stated in patient record

15) CS mets at dx (2004+)

“Information on distant metastasis. Available for 2004+. Earlier cases may be converted and new codes added which weren't available for use prior to the current version of CS.”

Codes

0: No distant metastasis
5: No clinical or radiographic evidence of distant metastasis, but deposits of molecularly or microscopically detected tumor cells in circulating blood, bone marrow or other non-regional nodal tissue that are 0.2 millimeters (mm) or less in a patient without symptoms or signs of metastasis
7: Stated as M0(i+) with no other information on distant metastasis
10: Distant lymph node(s): Cervical; NOS, Contralateral/bilateral axillary and/or internal mammary, Other than above. Distant lymph node(s), NOS
40: Distant metastasis except distant lymph node(s) (code 10) Carcinomatosis
42: Further contiguous extension: Skin over: Axilla, Contralateral (opposite) breast, Sternum, Upper abdomen.
44: Metastasis: Adrenal (suprarenal) gland, Bone, other than adjacent rib, Contralateral (opposite) breast - if stated as metastatic Lung, Ovary, Satellite nodule(s) in skin other than primary breast
50: (40 - 44) + 10
60: Distant metastasis, NOS. Stated as M1 with no other information on distant metastasis
99: Unknown; distant metastasis not stated. Distant metastasis cannot be assessed. Not documented in patient record.
126: Unknown

16) CS tumor Size

“Information on tumor size. Available for 2004+. Earlier cases may be converted and new codes added which weren't available for use prior to the current version of CS.”

Codes

0: Indicates no mass or no tumor found; for example, when a tumor of a stated primary site is not found, but the tumor has metastasized
1-989:1-989 millimeters
990: Microscopic focus or foci only; no size of focus is given
991: Described as less than 1 cm
992: Described as less than 2 cm
993: Described as less than 3 cm
994: Described as less than 4 cm
995: Described as less than 5 cm
996: Site-specific codes where needed
997: Site-specific codes where needed
998: Site-specific codes where needed
999: Unknown; size not stated; not stated in patient record

17) CS Extension

"Information on extension of the tumor. Available for 2004+. Earlier cases may be converted and new codes added which weren't available for use prior to the current version of CS."

Codes

0: In situ: noninfiltrating; intraepithelial. Intraductal WITHOUT infiltration. Lobular neoplasia.

50: Paget disease of nipple WITHOUT underlying tumor

70: Paget Disease disease of nipple WITHOUT underlying invasive carcinoma pathologically

100: Confined to breast tissue and fat including nipple and/or areola Localized, NOS

110: Stated as T1mi with no other information on extension

120: Stated as T1a with no other information on extension

130: Stated as T1b with no other information on extension

140: Stated as T1c with no other information on extension

170: Stated as T1 (NOS) with no other information on extension or size

180: Stated as T2 with no other information on extension or size

190: Stated as T3 with no other information on extension or size

200: Invasion of subcutaneous tissue. Local infiltration of dermal lymphatics adjacent to primary tumor involving skin by direct extension. Skin infiltration of primary breast including skin of nipple and/or areola.

300: Attachment or fixation to pectoral muscle(s) or underlying tissue. Deep fixation. Invasion of (or fixation to) pectoral fascia or muscle

380: OBSOLETE DATA CONVERTED V0203. See code 790. Stated as T4 (NOS) with no other information on extension

390: OBSOLETE DATA CONVERTED V0203. See code 410. Stated as T4a with no other information on extension.

400: Invasion of (or fixation to): Chest wall, Intercostal or serratus anterior muscle(s), Rib(s). See codes 610 (obsolete), 612-615, and 620 (obsolete) for combinations with this code.

410: Stated as T4a with no other information on extension

510: OBSOLETE DATA RETAINED V0200. Extensive skin involvement, including: Satellite nodule(s) in skin of primary breast, Ulceration of skin of breast. Any of the following conditions described as involving not more than 50% of the breast, or amount or percent

512: Extensive skin involvement, including: Satellite nodule(s) in skin of primary breast, Ulceration of skin of breast.

514: Any of the following conditions described as involving less than one-third (33%) of the breast WITHOUT a stated diagnosis of inflammatory carcinoma. WITH or WITHOUT dermal lymphatic infiltration: Edema of skin, En cuirasse, Erythema, Inflammation of skin,

516: 514 + 512

518: Any of the following conditions described as involving one third (33%) or more but less than or equal to half (50%) of the breast WITHOUT a stated diagnosis of inflammatory carcinoma. WITH or WITHOUT dermal lymphatic infiltration: Edema of skin, En cuira

519: 518 + 512

520: Any of the following conditions described as involving more than 50% of the breast WITHOUT a stated diagnosis of inflammatory carcinoma. WITH or WITHOUT dermal lymphatic infiltration: Edema of skin, En cuirasse, Erythema, Inflammation of skin, Peau d'ora
575: 520 + 512
580: Any of the following conditions with amount or percent of breast involvement not stated and WITHOUT a stated diagnosis of inflammatory carcinoma. WITH or WITHOUT dermal lymphatic infiltration: Edema of skin, En cuirasse, Erythema, Inflammation of skin, P
585: 580 + 512
590: OBSOLETE DATA CONVERTED V0203. See code 605. Stated as T4b with no other information on extension.
600: Diagnosis of inflammatory carcinoma WITH a clinical description of inflammation, erythema, edema, peau d'orange, etc., involving less than one-third (33%) of the skin of the breast, WITH or WITHOUT dermal lymphatic infiltration.
605: Stated as T4b with no other information on extension
610: OBSOLETE DATA RETAINED V0200. (400) + (510)
612: Any of (512-516) + 400
613: Any of (518-519) + 400
615: Any of (520-585) + 400
620: OBSOLETE DATA RETAINED V0200. (400) + (520)
680: Stated as T4c with no other information on extension
710: OBSOLETE DATA RETAINED V0200. Diagnosis of inflammatory carcinoma WITH a clinical description of inflammation, erythema, edema, peau d'orange, etc., involving not more than 50% of the skin of the breast, WITH or WITHOUT dermal lymphatic infiltration. Infl
715: OBSOLETE DATA RETAINED V0202. Diagnosis of inflammatory carcinoma WITH a clinical description of inflammation, erythema, edema, peau d'orange, etc., involving not more than one-third (33%) of the skin of the breast, WITH or WITHOUT dermal lymphatic infilt
720: OBSOLETE DATA CONVERTED V0102. Description: Diagnosis of inflammatory carcinoma WITH a clinical diagnosis of inflammation, erythema, edema, peau d'orange, etc., of not more than 50% of the breast, WITH or WITHOUT dermal lymphatic infiltration. Inflammator
725: Diagnosis of inflammatory carcinoma WITH a clinical description of inflammation, erythema, edema, peau d'orange, etc., involving one-third (33%) or more but less than or equal to one-half (50%) of the skin of the breast, WITH or WITHOUT dermal lymphatic i
730: Diagnosis of inflammatory carcinoma WITH a clinical description of inflammation, erythema, edema, peau d'orange, etc., involving more than one-half (50%) of the skin of the breast, WITH or WITHOUT dermal lymphatic infiltration.
750: Diagnosis of inflammatory carcinoma WITH a clinical description of inflammation, erythema, edema, peau d'orange, etc., but percent of involvement not stated, WITH or WITHOUT dermal lymphatic infiltration. Note: If percentage is known, code to 600, 725
780: Stated as T4d with no other information on extension
790: Stated as T4 (NOS) with no other information on extension
950: No evidence of primary tumor
999: Unknown; extension not stated. Primary tumor cannot be assessed. Not documented in patient record.
1022: Unknown

18) CS Lymph Nodes

"Information on involvement of lymph nodes. Available for 2004+. Earlier cases may be converted and new codes added which weren't available for use prior to the current version of CS."

Codes

0: No regional lymph node involvement OR isolated tumor cells (ITCs) detected by immunohistochemistry/immunohistochemical (IHC) methods or molecular methods ONLY.

50: Evaluated pathologically: None; no regional lymph node involvement BUT ITCs detected on routine hematoxylin and eosin (H and E) stains.

130: Evaluated pathologically: Axillary lymph node(s), ipsilateral, micrometastasis ONLY detected by IHC ONLY (At least one micrometastasis greater than 0.2 mm or more than 200 cells AND all micrometastases less than or equal to 2 mm)

150: Evaluated pathologically: Axillary lymph node(s), ipsilateral, micrometastasis ONLY detected or verified on H&E (At least one micrometastasis greater than 0.2 mm or more than 200 cells AND all micrometastases less than or equal to 2 mm) Micrometastasis, NOS

155: Evaluated pathologically: Stated as N1mi with no other information on regional lymph nodes

250: Evaluated pathologically: Movable axillary lymph node(s), ipsilateral, positive with more than micrometastasis (At least one metastasis greater than 2 mm)

255: Evaluated pathologically: Movable axillary lymph node(s), ipsilateral, positive with more than micrometastasis (At least one metastasis greater than 2 mm)

257: Evaluated clinically: Clinically stated only as N1 (Clinical assessment because of neoadjuvant therapy or no pathology)

258: Evaluated pathologically: Pathologically stated only as N1 [NOS], no information on which nodes were involved

260: Stated as N1 [NOS] with no other information on regional lymph nodes

280: OBSOLETE DATA RETAINED V0104. Stated as N2, NOS

290: OBSOLETE DATA CONVERTED V0203. See code 610. Clinically stated only as N2, NOS (clinical assessment because of neoadjuvant therapy or no pathology)

300: OBSOLETE DATA CONVERTED V0203. See code 620. Pathologically stated only as N2 NOS; no information on which nodes were involved

500: OBSOLETE DATA RETAINED V0104. Fixed/matted ipsilateral axillary nodes, positive with more than micrometastasis (i.e., at least one metastasis greater than 2 mm). Fixed/matted ipsilateral axillary nodes, NOS

510: Evaluated clinically: Fixed/matted ipsilateral axillary nodes clinically (Clinical assessment because of neoadjuvant therapy or no pathology). Stated clinically as N2a (Clinical assessment because of neoadjuvant therapy or no pathology)

520: Evaluated pathologically: Fixed/matted ipsilateral axillary nodes clinically with pathologic involvement of lymph nodes WITH at least one metastasis greater than 2 mm

600: Axillary/regional lymph node(s), NOS Lymph nodes, NOS

610: Evaluated clinically: Clinically stated only as N2 [NOS] (Clinical assessment because of neoadjuvant therapy or no pathology)

620: Evaluated pathologically: Pathologically stated only as N2 [NOS]; no information on which nodes were involved

630: Stated as N2 [NOS] with no other information on regional lymph nodes
710: Evaluated pathologically: Internal mammary node(s), ipsilateral, positive on sentinel nodes but not clinically apparent (No positive imaging or clinical exam)
WITHOUT axillary lymph node(s), ipsilateral
720: Evaluated pathologically: Internal mammary node(s), ipsilateral, positive on sentinel nodes but not clinically apparent (No positive imaging or clinical exam) WITH axillary lymph node(s), ipsilateral
730: Evaluated pathologically: Internal mammary node(s), ipsilateral, positive on sentinel nodes but not clinically apparent (No positive imaging or clinical exam) UNKNOWN if positive axillary lymph node(s), ipsilateral
735: Evaluated clinically: Internal mammary node(s), ipsilateral, positive on sentinel nodes but primary not resected WITHOUT axillary lymph node(s), ipsilateral OR UNKNOWN if positive axillary lymph node(s)
740: Internal mammary node(s), ipsilateral, clinically apparent (On imaging or clinical exam) WITHOUT axillary lymph node(s), ipsilateral
745: Internal mammary node(s), ipsilateral, clinically apparent (On imaging or clinical exam) UNKNOWN if positive axillary lymph node(s), ipsilateral
748: Stated as N2b with no other information on regional lymph nodes
750: Infraclavicular lymph node(s) (subclavicular) (level III axillary nodes) (apical), ipsilateral WITH or WITHOUT axillary nodes(s)
WITHOUT internal mammary node(s)
755: Stated as N3a with no other information on regional lymph nodes
760: OBSOLETE DATA RETAINED AND REVIEWED V0203. See codes 763 and765. Internal mammary node(s), ipsilateral, clinically apparent (on imaging or clinical exam) WITH axillary lymph node(s), ipsilateral, codes 150 to 600 WITH or WITHOUT infraclavicular (level III axillary nodes) (apical) lymph nodes.
763: Internal mammary node(s), ipsilateral, clinically apparent (On imaging or clinical exam) WITH axillary lymph node(s), ipsilateral, codes 150 to 600 WITHOUT infraclavicular (level III axillary nodes) (apical) lymph nodes or unknown if infraclavicular (level III axillary nodes) (apical) lymph nodes involved
764: Internal mammary node(s), ipsilateral, clinically apparent (On imaging or clinical exam) WITHOUT axillary lymph node(s), ipsilateral WITH infraclavicular (level III axillary nodes) (apical) lymph nodes involved
765: Internal mammary node(s), ipsilateral, clinically apparent (On imaging or clinical exam) WITH axillary lymph node(s), ipsilateral. WITH infraclavicular (level III axillary nodes) (apical) lymph nodes involved
768: Stated as N3b with no other information on regional lymph nodes
770: OBSOLETE DATA RETAINED V0200. Internal mammary node(s), ipsilateral, clinically apparent (on imaging or clinical exam). UNKNOWN if positive axillary lymph node(s), ipsilateral
780: OBSOLETE DATA RETAINED V0200. (750) + (770)
790: OBSOLETE DATA CONVERTED V0203. See code 820. Stated as N3, NOS
800: Supraclavicular node(s), ipsilateral
805: Stated as N3c with no other information on regional lymph nodes
810: Evaluated clinically: Clinically stated only as N3 [NOS] (Clinical assessment because of neoadjuvant therapy or no pathology)

815: Evaluated pathologically: Pathologically stated only as N3 [NOS]; no information on which nodes were involved
820: Stated as N3, NOS with no other information on regional lymph nodes
999: Unknown; regional lymph nodes not stated. Regional lymph node(s) cannot be assessed. Not documented in patient record
1022: Unknown

19) Marital Status at diagnosis

“This variable identifies the patient’s marital status at the time of diagnosis for the reportable tumor.”

Codes

1: Single (never married)
2: Married (including common law)
3: Separated
4: Divorced
5: Widowed
6: Unmarried or domestic partner (same sex or opposite sex or unregistered)
9: Unknown

20) ER Status Recode Breast Cancer (1990+)

“Created by combining information from Tumor marker 1 (1990-2003), with information from CS site-specific factor 1 (2004+).”

Codes

1: Positive
2: Negative
3: Borderline
4: Unknown
9: Not 1990+ Breast

21) PR Status Recode Breast Cancer (1990+)

“Created by combining information from Tumor marker 2 (1990-2003), with information from CS site-specific factor 2 (2004+). This field is blank for non-breast cases and cases diagnosed before 1990.”

Codes

1: Positive
2: Negative
3: Borderline

4: Unknown
9: Not 1990+ Breast

22) SEER cause-specific death classification

"This variable designates that the person died of their cancer for cause-specific survival."

Codes

0: Alive or dead of other cause *(Filtered out from the extracted dataset used in this research)*
1: Dead
9: N/A not first tumor

23) Survival Months

"Created using complete dates, including days, therefore may differ from survival time calculated from year and month only"

24) Laterality

"Laterality describes the side of a paired organ or side of the body on which the reportable tumor originated. Starting with cases diagnosed January 1, 2004 and later, laterality is coded for select invasive, benign, and borderline primary intracranial and CNS tumors."

Codes

0: Not a paired site
1: Right: origin of primary
2: Left: origin of primary
3: Only one side involved, right or left origin unspecified
4: Bilateral involvement, lateral origin unknown; stated to be single primary- Both ovaries involved simultaneously, single histology, Bilateral retinoblastomas,Bilateral Wilms's tumors
5: Paired site: midline tumor
9: Paired site, but no information concerning laterality; midline tumor

25) Histologic Type ICD-O-3

"Histologic Type describes the microscopic composition of cells and/or tissue for a specific primary. The tumor type or histology is a basis for staging and determination of treatment options. It affects the prognosis and course of the disease. The International Classification of Diseases for Oncology, Third Edition (ICD-O-3) is the standard reference for coding the histology for tumors diagnosed in 2001 and later. All ICD-O-2 histologies for 1973-2000 were converted to ICD-O-3."

Codes

0: Benign (Reportable for intracranial and CNS sites only)
1: Uncertain whether benign or malignant, borderline malignancy, low malignant potential, and uncertain malignant potential (Reportable for intracranial and CNS sites only)
2: Carcinoma in situ; intraepithelial; noninfiltrating; noninvasive
3: Malignant, primary site (invasive)

26) Race/ethnicity

"Recode which gives priority to non-white races for persons of mixed races."

Codes

01: White
02: Black
03: American Indian, Aleutian, Alaskan Native or Eskimo (includes all indigenous populations of the Western hemisphere)
04: Chinese
05: Japanese
06: Filipino
07: Hawaiian
08: Korean (Effective with 1/1/1988 dx)
10: Vietnamese (Effective with 1/1/1988 dx)
11: Laotian (Effective with 1/1/1988 dx)
12: Hmong (Effective with 1/1/1988 dx)
13: Kampuchean (including Khmer and Cambodian) (Effective with 1/1/1988 dx)
14: Thai (Effective with 1/1/1994 dx)
15: Asian Indian or Pakistani, NOS (Effective with 1/1/1988 dx)
16: Asian Indian (Effective with 1/1/2010 dx)
17: Pakistani (Effective with 1/1/2010 dx)
20: Micronesian, NOS (Effective with 1/1/1991)
21: Chamorran (Effective with 1/1/1991 dx)
22: Guamanian, NOS (Effective with 1/1/1991 dx)
25: Polynesian, NOS (Effective with 1/1/1991 dx)
26: Tahitian (Effective with 1/1/1991 dx)
27: Samoan (Effective with 1/1/1991 dx)
28: Tongan (Effective with 1/1/1991 dx)
30: Melanesian, NOS (Effective with 1/1/1991 dx)
31: Fiji Islander (Effective with 1/1/1991 dx)
32: New Guinean (Effective with 1/1/1991 dx)
96: Other Asian, including Asian, NOS and Oriental, NOS (Effective with 1/1/1991 dx)
97: Pacific Islander, NOS (Effective with 1/1/1991 dx)
98: Other
99: Unknown

27) Year of Diagnosis

"The year of diagnosis is the year the tumor was first diagnosed by a recognized medical practitioner, whether clinically or microscopically confirmed."

28) Behavior Code ICD-O3

"SEER requires registries to collect malignancies with in situ /2 and malignant /3 behavior codes as described in ICD-O-3. SEER requires registries to collect benign /0 and borderline /1 intracranial and CNS tumors for cases diagnosed on or after 1/1/2004. Behavior is the fifth digit of the morphology code after the slash (/)."

Codes

0: Benign (Reportable for intracranial and CNS sites only)
1: Uncertain whether benign or malignant, borderline malignancy, low malignant potential, and uncertain malignant potential (Reportable for intracranial and CNS sites only)
2: Carcinoma in situ; intraepithelial; noninfiltrating; noninvasive
3: Malignant, primary site (invasive)

29) Surgery of Primary Site

"Surgery of Primary Site describes a surgical procedure that removes and/or destroys tissue of the primary site performed as part of the initial work-up or first course of therapy."

Codes

00: None; no surgical procedure of primary site; diagnosed at autopsy only
10-19: Site-specific codes. Tumor destruction; no pathologic specimen or unknown whether there is a pathologic specimen
20-80: Site-specific codes. Resection; pathologic specimen
90: Surgery, NOS. A surgical procedure to the primary site was done, but no information on the type of surgical procedure is provided.
98: Special codes for hematopoietic, reticuloendothelial, immunoproliferative, myeloproliferative diseases; illdefined sites; and unknown primaries, except death certificate only
99: Unknown if surgery performed; death certificate only

30) Reason no cancer-directed surgery

"This variable documents the reason that surgery was not performed on the primary site"

Codes

0: Surgery performed
1*: Surgery not recommended
2*: Contraindicated due to other conditions (1973-2002)
5: Patient died before recommended surgery
6: Unknown reason for no surgery
7* Patient or patient's guardian refused
8: Recommended, unknown if done
9: Unknown if surgery performed; Death Certificate Only case; Autopsy only case (2003+)
*Codes not used prior to 1988. Code ‘2’ used only for Autopsy only cases prior to 1988