12-2011

Use of statistical analysis, data mining, decision analysis and cost effectiveness analysis to analyze medical data : application to comparative effectiveness of lumpectomy and mastectomy for breast cancer.

Beatrice Ugiliweneza
University of Louisville

USE OF STATISTICAL ANALYSIS, DATA MINING, DECISION ANALYSIS AND COST EFFECTIVENESS ANALYSIS TO ANALYZE MEDICAL DATA: APPLICATION TO COMPARATIVE EFFECTIVENESS OF LUMPECTOMY AND MASTECTOMY FOR BREAST CANCER

By

Beatrice Ugiliweneza

B.S. Niamey, NIGER, 2006

A Dissertation

Submitted to the Faculty of the

College of Arts and Sciences of the University of Louisville

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

Department of Mathematics

University of Louisville

Louisville, Kentucky

December 2011

USE OF STATISTICAL ANALYSIS, DATA MINING, DECISION ANALYSIS AND COST EFFECTIVENESS ANALYSIS TO ANALYZE MEDICAL DATA: APPLICATION TO COMPARATIVE EFFECTIVENESS OF LUMPECTOMY AND MASTECTOMY FOR BREAST CANCER

By

Beatrice Ugiliweneza

B.S. Niamey, NIGER, 2006

A dissertation approved on

November 4, 2011

By the following Dissertation Committee:

Dr. Patricia Cerrito (Director)

Dr. Ryan Gill (Co-Director)

Dr. Kiseop Lee

Dr. Jiaxu Li

Dr. Ibrahim Imam

Dr. Mehmed Kantardzic

DEDICATION

This thesis is dedicated to my wonderful family:

My Husband Thacien Ntihinyurwa for his continued support

My daughters Benita Bwiza Ntihinyurwa and Elisa Hora Ntihinyurwa for the inspiration I get from their eyes every morning

My sister Pacifique Munezero, my brothers Desire Isidore Munyeshuli and Romeo Irere Kwihangana for their invaluable encouragement

My beloved mother Annonciata Nyiramahoro for the strength she put in me, for making me who I am today and without whom all this could not have happened

My late father Valere Munyakazi whom I love very much and miss every day for teaching me at a very young age that the most important things in life are to obey God's commandments and respect people

ACKNOWLEDGMENT

I would like to thank my advisor Dr. Patricia Cerrito, for her guidance, help and patience. I would also like to thank my co-advisor Dr. Ryan Gill for invaluable help, suggestions and comments. My thanks go to the other committee members, Dr. Mehmed Kantardzic, Dr. Ibrahim Imam, Dr. Kiseop Lee and Dr. Jiaxu Li as well for their comments. I would also like to express my thanks to my husband, Thacien, who had to be mommy many times, and to my daughters, Benita and Elisa, who had to understand at a very young age that I could not be there every time they needed me. Also, many thanks to my mother, Annonciata, my sister, Pacifique, my brothers, Desire and Romeo, who gave me the encouragement I needed to stay on track. Finally, I would like to thank all other member of my family and friends in Louisville, Kentucky, in San Diego, California, in Dayton, Ohio, in Kenya, in Rwanda, in France, in Belgium, in Niger, and all over the world.

ABSTRACT

USE OF STATISTICAL ANALYSIS, DATA MINING, DECISION ANALYSIS AND COST EFFECTIVENESS ANALYSIS TO ANALYZE MEDICAL DATA: APPLICATION TO COMPARATIVE EFFECTIVENESS OF LUMPECTOMY AND MASTECTOMY FOR BREAST CANCER

Beatrice Ugiliweneza

November 4th, 2011

Statistical models have been the first choice for comparative effectiveness in clinical research. Though effective, these models are limited when the data to be analyzed do not fit the assumed distributions; which is mostly the case when the study is not a clinical trial. In this project, data mining, decision analysis and cost effectiveness analysis methods were used to supplement statistical models in comparing lumpectomy to mastectomy for surgical treatment of breast cancer. Mastectomy has been the gold standard for breast cancer treatment for since the 1800s. In the 20th century, an equivalence of mastectomy and lumpectomy was established in terms of long-term survival and disease free survival. However, short term comparative effectiveness in post-operative outcomes has not been fully explored. Studies using administrative data are lacking and no study has used new technologies of self-expression, particularly the internet discussion board. In this study, data used were from the Nationwide Inpatient

Sample (NIS) 2005, the Thomson Reuter's MarketScan 2000 – 2001, the medical literature on clinical trials and online individuals' posts in discussion boards on breastcancer.org. The NIS was used to compare lumpectomy to mastectomy in terms of hospital length of stay, total charges and in-hospital death at the time of surgery. MarketScan data was used to evaluate the comparative follow-up outcomes in terms of risk of repeat hospitalization, risk of repeat operation, number of outpatient services, number of prescribed medications, length of stay, and total charges per post-operative hospital admission on a period of eight months average. The MarketScan was also used to construct a simple post-operative hospital admission predictive model and to perform short-term cost-effectiveness analysis. The medical literature was used to analyze long term -10 years- mortality and recurrence for both treatments. The web postings were used to evaluate the comparative cost to improve quality of life in terms of patient satisfaction. In NIS and MarketScan data, International Classification of Disease, 9^{th} revision, Clinical Modification (ICD-9-CM) diagnosis codes were used to extract cases of breast cancer ; and ICD-9-CM procedure codes and Current Procedural Terminology, 4^{th} edition procedure codes were used to form groups of treatment.

Data were pre-processed and prepared for analysis using data mining techniques such as clustering, sampling and text mining. To clean the data for statistical models, some continuous variables were normalized using methods such as logarithmic transformation. Statistical models such as linear regression, generalized linear models, logistic and proportional hazard (Cox) regressions were used to compare post-operative outcomes of lumpectomy versus mastectomy. Neural networks, decision tree and logistic regression predictive modeling techniques were compared to create a simple predictive model

predicting 90-day post-operative hospital re-admission. Cost and effectiveness were compared with the Incremental Cost Effectiveness Ratio (ICER). A simple method to process and analyze online postings was created and used for patients' input in the comparison of lumpectomy to mastectomy. All statistical analyses were performed in SAS 9.2. Data Mining was performed in SAS Enterprise Miner (EM) 6.1 and SAS Text Miner. Decision analysis and Cost Effectiveness Analysis were performed in TreeAge Pro 2011.

A simple comparison of the two procedures using the NIS 2005, a discharge-level data, showed that in general, a lumpectomy surgery is associated with a significantly longer stay and more charges on average. From the MarketScan data, a person-level data where a patient can be followed longitudinally, it was found that for the initial hospitalization, patients who underwent mastectomy had a non-significant longer hospital stay and significantly lower charges. The post-operative number of outpatient services, prescribed medications as well as length of stay and charges for post-operative hospital admissions were not statistically significant. Using the MarketScan data, it was also found that the best model to predict 90-day post-operative hospital admission was logistic regression. A logistic regression revealed that the risk of a hospital re-admission within 90 days after surgery was 65% for a patient who underwent lumpectomy and 48% for a patient who underwent mastectomy. A cost effectiveness analysis using Markov models for up to 100 days after surgery showed that having lumpectomy saved hospital related costs every day with a minimum saving of $33 on day 10. In terms of long-term outcomes, the use of decision analysis methods on the literature review data revealed that, 10-years after surgery, 739 recurrences and 84 deaths were prevented among 10,000 women who had

mastectomy instead of lumpectomy. Factoring patients' preferences in the comparison of the two procedures, it was found that patients who undergo lumpectomy are non-significantly more satisfied than their peers who undergo mastectomy. In terms of cost, it was found that lumpectomy saves $517 for each satisfied individual in comparison to mastectomy.

In conclusion, the current project showed how to use data mining, decision analysis and cost effectiveness methods to supplement statistical analysis when using real world non-clinical trial data for a more complete analysis. The application of this combination of methods on the comparative effectiveness of lumpectomy and mastectomy showed that in terms of cost and patients' quality of life measured as satisfaction, lumpectomy was found to be the better choice.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

1.1. Introduction

Statistical models have been the primary choice when comparing lumpectomy to mastectomy for the treatment of breast cancer. This is mainly due to the fact that many studies in this field were clinical trials. Clinical trials can be defined as research on human subjects (medical, biomedical or behavioral) that are designed to answer very specific questions. The inclusion/ exclusion criteria for subject selection for the study ensure that the data to be collected will fit the statistical methods proposed to be used. Clinical trials are the best types of study for medical research. However, they are expensive and not always feasible. Alternatives to clinical trials are observational studies including the use of administrative data. Healthcare administrative data are medical data collected for administrative purposes such as processing health insurance claims. These data represent the real world effects since the subjects are not pre-screened for inclusion. Statistical methods can be used to analyze these data. However, the data have to be cautiously pre-processed to fit the needed statistical assumptions. Sometimes, the assumptions cannot be verified. Other times, the assumptions are verified, but the sample size is too large that all results will be significant. In the current study, methods to supplement statistical models are presented when comparing health outcomes for two groups; in this case, lumpectomy and mastectomy.

Mastectomy was the "gold" standard for breast cancer surgical treatment until the mid-20th century[1]. It was then that clinical trials established the equivalent effectiveness of lumpectomy. Since then, women diagnosed with early stage breast cancer are given the choice between lumpectomy as a minimally invasive surgery or mastectomy, the traditional approach.

Various studies, mainly clinical trials [2-7], in different settings have compared lumpectomy to mastectomy generally in terms of disease-free and overall survival. They all come to the conclusion that lumpectomy is comparable to mastectomy.

Although the long-term benefits of lumpectomy in comparison to mastectomy were assessed, the short-term outcomes following the surgery have not yet been clearly discussed. Moreover, the comparisons of these surgeries using non-clinical trial data are lacking. Clinical outcomes such as in-hospital length of stay, post-operative healthcare resource use (such as the number of hospitalizations, the number of outpatient services, and the number of prescribed medications) and the risk of post-operative hospital admission have yet to be compared.

The purpose of this study was to present methods that can be used to analyze administrative data consisting of hospital data and insurance claims data in comparing lumpectomy to mastectomy in health, clinical outcomes and healthcare utilization. The main goal was to use statistical methods supplemented with data mining techniques to assess the effectiveness of lumpectomy in comparison to mastectomy in short-term, immediate follow-up and long-term outcomes after the surgery. Secondary objectives included comparing these surgical procedures using data extracted from the literature

with decision analysis methods and comments extracted from online discussion board with text mining and cost effectiveness analysis.

1.2. Background Related to Breast Cancer

1.2.1. Breast cancer overview

Breast cancer is a development of malignant cells in the breast. Most commonly, cancer originates from the milk-producing organs (lobules) or the tubules (ducts) that conduct the milk to the nipple [8]. Breast cancers that originate from the ducts are called ductal carcinomas; those originating from the lobules are known as lobular carcinomas. In rare cases, breast cancer can start in other areas of the breast [8].

1.2.2. Breast cancer risk factors

The risk factors for breast cancer can be divided into two categories: factors that can be changes and those that cannot be changed. For breast cancer, risks factors that cannot be changed include [8]:

- Age: breast cancer risk has been found to be increasing with age
- Gender: women are 100 times more at risk of breast cancer than men
- Family history of breast cancer: having a close relative with breast cancer has been found to be associated with a higher risk of breast cancer
- Genes: having genes more prone to developing breast cancer such as *BRAC1* and *BRAC2*
- Early onset of menstruation: having the first period before the age of 12

- Late menopause: having menopause at or after the age of 55

Risk factors that can be changed for breast cancer comprise but are not limited to [8, 9]:

- Child birth: never having a child or having a first child after the age of 30 is a risk of developing breast cancer
- Hormone Replacement Therapy (HRT): Receiving hormone replacement therapy for years for menopausal symptom relief has been associated with a high risk of breast cancer
- Radiation: Being exposed to radiation as a child or as a young adult increased the risk of developing breast cancer
- Obesity, high-fat intake and alcohol use have been linked to a high risk of breast cancer

1.2.3. Symptoms of breast cancer

Very early onset breast cancer may stay unnoticed for a while. Growing breast cancer symptoms comprise [9]:

- A breast lump or a lump in the armpit
- Change in the size, shape or feel of the breast
- Nipple discharge, including blood

Metastatic breast cancer can be noticed by [10]:

- Pain or discomfort in the breast
- Swelling of the arm
- Pain in the bones

- Skin ulcers

1.2.4. Diagnosis

The diagnosis of breast cancer is accomplished with tests looking for any sign or symptom. The usual tests consist of a biopsy of any suspicious lump or a mammogram to identify an observed anomaly [11]. Even when a woman has no signs or symptoms, it is recommended to get regular breast examinations either using a Breast Self-Examination or an exam by a healthcare professional (Clinical Breast Examination). The American Cancer Society recommends an annual mammogram for all women starting at age 40 [10]. All these screening techniques are recommended in an effort to diagnose any malignancies early. Diagnosis of breast cancer at an early stage has been associated with longer survival [11].

For women who present breast cancer symptoms, additional exams (in addition to biopsy and/or mammograms) may be judged necessary. These tests include ultrasounds, computer tomography (CT) and magnetic resonance imaging (MRI). If a diagnosis is confirmed to be breast cancer, then the physicians will perform additional tests to determine if the cancer has spread beyond the breast [8]. This is called breast cancer staging.

1.2.5. Breast cancer staging

The stages of breast cancer describe the extent of the metastasis. The stage depends on whether the cancer is invasive or in-situ, the size of the tumor, how many lymph nodes are involved, and if it has spread beyond the breast [10]. Breast cancer stages range from

zero to four and the higher the stage, the more advanced the cancer [8]. The following table summarizes stage I to IV [11].

Table 1.1: Cancer staging [11]

Stage	Description
1	The tumor size is no more than two centimeters and no cancer cells are found in the lymph nodes
2	The tumor size is more than two centimeters and no more than five centimeters and the cancer has spread to the lymph nodes
3A	The tumor size is more than five centimeters or less than five centimeters but the cancer has spread to the lymph nodes, which have grown into each other
3B	The cancer has spread to tissues near the breast (local invasion), or to lymph nodes inside the chest wall, along the breast bone
4	The cancer has spread to the skin and lymph nodes beyond the axilla or to other organs of the body

1.2.6. Treatment

Treatment depends on many factors, including type and stage of breast cancer, hormonal reception status and whether the cancer overproduces Human Epidermal growth factor Receptor 2 (HER2) [9].

In general, the main treatment is surgery in order to remove all the cancer. In a surgical treatment for breast cancer, the whole infected breast can be removed (mastectomy) or only the tumor along with surrounding healthy tissues (Breast Conserving Surgery) can be taken out [11]. Radiation therapy is offered after surgery to kill locally any microscopic tumors that might have escaped surgery [9]. Drug therapy or chemotherapy can be offered before and/or after surgery. Before therapy (neo-adjuvant), it is intended to shrink the tumor. After surgery (adjuvant therapy), it is intended to kill any remaining cancers. Some women also receive hormonal therapy to block certain hormones.

1.2.7. Prognosis

A prognosis is a prediction of an outcome and the probability of progression-free survival or disease-free survival [9]. The prognosis for breast cancer depends on the type, stage and classification of the cancer [11]. Survival decreases with the stage of the disease. About 80% of patients at stage I are cured and only 70% of patients at stage II survive breast cancer. The five-year survival for patients at stage III is estimated to be 40% and only 20% for stage IV patients. Younger women tend to have a poorer prognosis than post-menopausal women [9].

1.2.8. Epidemiology

The incidence of breast cancer is variable around the world: the highest rates are in more-developed countries while the lowest rates are observed in less-developed countries [8]. In the USA, the annual incidence rates of breast cancer are 128.6 per 100,000 in white women and 112.6 per 100,000 in African American women [11]. These statistics place the USA among the countries with the highest rates in the world [8, 11]. About 45,000 women die of breast cancer each year [11]. Breast cancer is the second-most common cancer (after skin cancer) and the second-most common cancer death (after lung cancer) [11].

1.3. Purpose of the study

1.3.1. Target population

It is well established that breast cancer is much more likely to affect women than it is to affect men [8]. Every woman is at risk of breast cancer and the risk increases with age. At 25 years of age, a woman has a 1 in 19,608 risk of developing breast cancer, but this number increases to 1 in 93 by the age of 45. A woman who lives to age 85 has a 1 in 8 risk of developing breast cancer in her lifetime [11]. This condition disproportionally affects older women, especially in the post-menopausal stage. The majority of all breast cancer cases are found in women over 50 while less than 55 percent of breast cancer cases are discovered in women under the age of 35 [11].

1.3.2. The problem

Breast cancer is a terrible condition that affects a considerable proportion of the US female population. The medical area of breast cancer treatment has gone through considerable improvements, which have resulted in both patient satisfaction and prolonged survival. Breast radical removal (mastectomy), once a gold standard for first step breast cancer treatment is no longer the only choice. Breast Conserving Surgery such as lumpectomy has been proven to be just as effective in terms of overall- survival and disease-free survival. This equivalence is well established and many studies have been undertaken with different population settings and designs and reports have been reported in the medical literature [2-7]. Also, recent studies have suggested that breast conserving treatment improves quality of life.

However, in 2003, research showed that young women who choose breast conserving surgery are at a higher risk of local recurrence [11]. Thus, the controversy in the choice of surgical treatment for breast cancer remains. Moreover, the comparative effectiveness in

terms of health outcomes in the period following the surgery has not yet been fully explored. Patients need to be aware of what to expect after surgical treatment, not only in terms of long term survival but also in terms of short term immediate and follow-up health and clinical outcomes. That is, there is a need to address breast cancer post-operative treatment management by providing statistical comparison of health and clinical outcomes and developing simple predictive models or scores for short term healthcare resources use.

1.3.3. Objectives

The main goal for this study was to use data mining to supplement statistical analysis in comparing breast cancer surgical treatments in terms of health and clinical outcomes healthcare utilization and charges. The comparison of the surgical procedures was performed in two stages: a short-term analysis that compared the outcomes of the hospitalization during the surgery and follow-up analysis that compared the longitudinal data of patients after they undergo the operation. Two groups were compared: patients treated by mastectomy and patients treated by lumpectomy (breast conserving surgery). Patients who underwent both procedures at the same time were excluded from the study.

Health outcomes were in-hospital death during the hospitalization for the surgery. Healthcare utilization was expressed as in-hospital length of stay, number of re-hospitalizations, number of outpatient services and the number of prescribed medications. Healthcare resources charges considered were hospitalization charges, outpatient service charges and medication charges.

Secondary goals included (1) the use of decision analysis to perform a long-term comparative effectiveness of the mastectomy and the breast conservation surgery in terms of deaths averted; (2) the use of cost effectiveness analysis to analyze the incremental cost of lumpectomy in comparison to mastectomy; and (3) the use of text mining to explore patient opinion through online comments on discussion boards.

1.3.4. Data

Data used for the current analysis are the Nationwide Inpatient Sample records of 2005, the MarketScan database records of 2000 and 2001, the medical literature and online posting in discussion board forums. The NIS data was used for the short-term in-hospital comparisons; the MarketScan data were used for the follow-up data analysis as well as in the predictive modeling. The literature was used in the long-term comparative effectiveness and the online comments were used in the exploratory analysis of patient opinion.

The NIS is part of the Healthcare Cost and Utilization Project (HCUP). NIS is a large national database containing hospital discharges from all-payer. It is a 20% sample of all US non-federal hospitals and contains data from approximately 1000 hospitals in 37 states. Hospitals are selected to represent 5 strata of hospital characteristics: ownership-control, bed-size, teaching status, rural-urban location, and geographical region. For a given year, NIS provides information on approximately five million to eight million hospital stays from about 1000 hospitals. Inpatient stay records in NIS include hospital and resource use. Hospital data are provided by the American Hospital Association's

annual survey of hospitals. The large sample size of the NIS provides numerous advantages, including analyses of rare conditions [12].

The Thomson Reuter's MarketScan Database contains person-level information on hospitalization usage, charges and enrollment. The annual datasets include data from about 100 payers and comprises inpatient, outpatient and prescription drugs, from about 45 large employers, health plans, government and public organizations. The collective MarketScan Databases refers to five individual databases: Commercial Claims and Encounters Database, Medicare Supplemental and COB Database, Health and Productivity Management Database, Benefit Design Database and Medicaid Database [13]. For this analysis, the Commercial Claims and Encounters Database were used. Sets were linked using an encrypted enrollee unique identifier named ENROLID [13].

Relevant medical literature was obtained from searching MEDLINE through PUBMED. MEDLINE (Medical Literature Analysis and Retrieval System Online) is a database containing more than 18 million records covering biomedicine and health from 1950 to the present [14]. MEDLINE is freely accessible on the internet via PUBMED.

Topics related to lumpectomy versus mastectomy from discussion board forums in breastcancer.org were obtained and analyzed.

1.3.5. Implication from the study

The current study will provide methods to pre-process and analyze longitudinal administrative data. It will also give insights in how to explore online postings and process them to factor them in our analyses. Finally, it will give an analysis algorithm

that can be used to have a complete study with results from different aspects of the problem, not only from the statistics.

From the healthcare perspective, the current analysis will provide more information to the patients and their physicians which will help them in their choice of surgical procedures for breast cancer, in particular those for whom lumpectomy is an option. The results obtained in the comparison of the clinical outcomes in the index procedure hospitalization and in the follow-up data will help the patients and their families get prepared. The results of the predictive models will provide the patients and their families with an insight into the risk of a subsequent hospital admission. This will guide them in choosing an optimal surgical choice.

1.3.6. Innovation

The ultimate goal of this study was use statistical, data mining, decision analysis and cost effectiveness analysis methods on real-world data in order to provide health outcomes results of the main surgical procedure treatments for breast cancer. The 'real world' data include administrative claims data and online comments. First, the approach used in this study to compare lumpectomy to mastectomy differs from previously published reports in the fact that real data are used in a retrospective fashion as opposed to prospective clinical trials. Second, another innovation is the use of data mining techniques such as cluster analysis and predictive modeling to breast cancer data in addition to classical statistical methods. Third, exploration of online comments in discussion boards provides an algorithm to factor patient opinion into classical analysis such as cost effectiveness analysis.

1.4. Methods

1.4.1. Study methods overview

The current study investigated the health outcomes and healthcare resources utilization for breast cancer patients who undergo a mastectomy or a lumpectomy. First, to analyze the current study data, a descriptive statistical analysis was performed to look at the population involved. Then, an inferential analysis was performed to test for differences in outcomes studies between the two groups. Second, data mining techniques were applied to classify and predict future re-hospitalization in case of surgical treatment for breast cancer. Third, decision analysis was used to analyze data from published reports to assess the effectiveness of a breast conserving surgery in comparison to the traditional mastectomy. Finally, text mining and cost effectiveness methods were used to study patients' postings online in lumpectomy versus mastectomy.

1.4.2. Statistical analysis

Statisticians are faced with more than just advocating a choice that represents the highest probability of a wanted outcome. They aim to discover new relationships in order to make new statements and/or validate old ones.

Statistical analysis mainly aims to describe the data and to make inferences. Most statistical methods assume that data follow pre-defined distributions. A number of other assumptions are also made, including the nature of the sample. These methods are called

"parametric methods". If the assumptions are not satisfied, statisticians rely on the counterpart of "non-parametric methods" when available.

The commonly used parametric models include the Generalized Linear Models (GLM) and the regression analysis models (linear, logistic, Poisson). The most used non-parametric models comprise the Wilcoxon tests and the Kaplan Meier curves. Some models are semi-parametric such as the Proportional Hazard Regression (Cox Regression).

Statistical models have been the benchmark for data analysis for a long time. Even though their usefulness is still effective, their efficiency is limited when it comes to data that do not fit the assumed distributions and also when the size of the data is very large, in which case all tests tend to be statistically significant.

1.4.3. Data mining methods

Nowadays, immense amounts of data are available. New and evolving technologies have given the companies, organization and institutions, including healthcare institutions, the capacity to collect and store very large data sets. These data contain important information, mostly useful for decision making. One way of digging into this information is to use data mining. Data mining can be defined as "the *non-trivial* extraction of *implicit*, previously unknown and potentially useful information from data" [15]. Data mining comprises the classical statistical analysis, artificial intelligence, machine learning and the development of large databases [16].

Data mining has the capabilities of the traditional data analysis methods blended with more sophisticated algorithms to process large volumes of data. With these algorithms,

data mining is able to find new patterns and associations that would have otherwise been unidentified [16]. The most used data mining techniques are cluster analysis and predictive modeling, or classification.

1.4.4. Decision analysis

Decision analysis is a systematic quantitative approach for assessing the value of different alternative choices [17, 18]. The uncertainly associated with each choice is represented through probabilities and probability distributions [18]. If there is a risk involved in the decision making, the attitude to risk is represented through utility functions. Otherwise, the utility functions are replaced by probabilities of achieving the uncertain aspiration levels [18]. As a result, the option that maximizes the probability of achieving the uncertain aspiration level is chosen. In other words, the decision whose consequences yield the maximum expected utility is recommended [18]. The probabilities are estimated using collected data or data from previously published reports. Decision analysis uses decision trees as graphical tools. Initially, decision analysis was developed as a method to help clinicians make decisions on an individual patient's management. It is increasingly used as a policy making tool in the management of groups of patients [17].

1.4.5. Cost effectiveness analysis

Cost Effectiveness Analysis is an economic analysis which compares alternative actions in relative costs and outcomes [19]. It is mostly used in health economics and pharmacoeconomics for evaluation of health programs or interventions. In this context, the cost-effectiveness is the ratio of the cost of the program or intervention to a defined

measure of its effect. It is primarily used for funds allocation but it can also be used for individual decision making [20].

1.5. Organization

In the following chapters, the research agenda will be discussed. Chapter 2 provides an overview of the breast cancer disease, steps from diagnosis to prognosis, including treatment and all the features taken into consideration to make a particular treatment choice. Chapter 3 discusses the literature review and work related to the current research. Chapters 4, 5 and 6 provide the mathematical theory behind the current analysis. Chapters 7 to 11 provide methods used and results obtained comparing lumpectomy to mastectomy. The following methods are used from chapter 7 to chapter 11 respectively: statistical methods, statistical methods and data mining, decision analysis, text mining, data mining and cost effectiveness analysis. Chapter 12 discusses the findings as well as the possible implications. Chapter 12 also presents the limitations and future research and summarizes the study, while providing an overall conclusion to the study.

CHAPTER 2

BREAST CANCER

2.1. Objective

The purpose of this chapter is to provide some basic information about breast cancer. First, its definition and characteristics are presented. Second, the signs and symptoms are evaluated. Third, breast cancer risk and protective factors are reviewed. Fourth, the current methods of diagnosis, screening and treatment are summarized. Finally, the treatment evolution since the early 1900 to the present is presented.

2.2. Breast cancer overview

Breast cancer is a malignant (cancerous) growth that begins in the tissues of the breast. Cancer is a disease in which abnormal cells grow in an uncontrolled way [21]. Figure 2.1 illustrates the mammary gland of the human female breast.

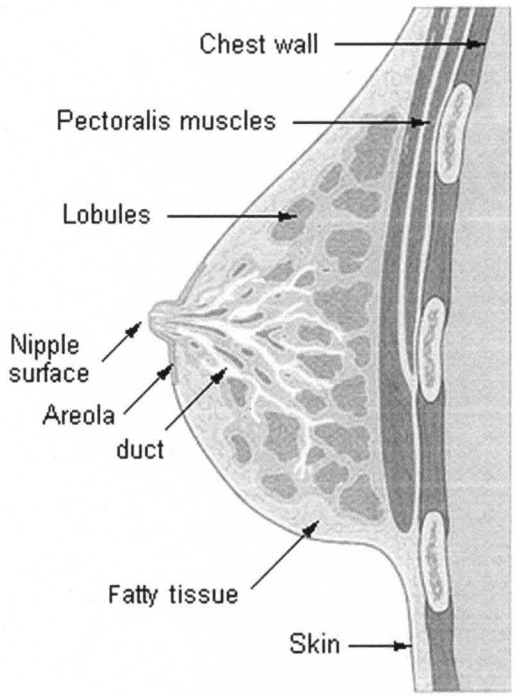

Figure 2.1: Illustration of the mammary gland of the human female breast [22]

Types of breast cancer [23]

The types of breast cancer depend on whether it begins in the lobules (organs that produce the milk) or in the ducts (the tubes that carry the milk from the lobes to the nipple). The cancer that originates from the lobules is called lobular or medullar carcinoma and the one that starts in the ducts are called ductal carcinoma.

The type of breast cancer also refers to the cancer's attitude towards surrounding tissues. If it is contained within its initial tissue, then it is called carcinoma in-situ. If on the other hand, the cancer cells can infiltrate and spread to other tissues, it is called invasive breast cancer.

The combinations of these characteristics define the four main types of breast cancer:

- *Ductal Carcinoma In-Situ (DCIS):* the breast cancer is confined inside the ductal system.
- *Invasive Ductal Carcinoma (IDC):* the breast cancer started in the ductal system and it is infiltrating to surrounding tissues
- *Medullary carcinoma:* the breast cancer started in the lobes and it is in-situ
- *Infiltrating medullary carcinoma:* the breast cancer started in the lobes and is invasive

2.3. Breast cancer classification

The classification of breast cancer is done in an effort to offer a tailored treatment to each patient. It also helps predict treatment response as well as prognosis.

Breast cancer is classified according to its stage, histological appearance, differentiation grade, hormone receptor status and DNA changes.

2.3.1. Breast cancer staging [24, 25]

For breast cancer, most registries use the summary staging, which groups cancer cases into five main categories: in-situ, localized, regional, distant and unknown. Cancer staging is the description of how cancer has spread. The most commonly used cancer staging is the TNM system (T=Tumor, N=the spread to the lymph nodes, M=Metastasis). The different TNM combinations correspond to five main stages which are denoted by the roman numerals I, II, III, IV and the digit 0 (i.e. stage 0, stage I, stage II, stage III and stage IV).

- *T-levels:* T represents different types of tumor evaluation:
  - TX: non evaluable tumor
  - T0: no sign of tumor
  - Tis: early tumor that has not spread
  - T1, T2, T3, T4:size and /or extension of the primary tumor
- *N-levels:* N represents different types of tumors spread to the regional lymph nodes:
  - NX: non evaluable lymph node
  - N0: no cancer in lymph nodes
  - N1, N2, N3: number and/or extent of regional spread
- *M-levels:* M represents different types of distant metastasis:
  - MX: non evaluable metastasis
  - M0: No cancer spread in other parts of the body
  - M1:Cancer has spread to the other parts of the body
- The five main cancer stages:
  - Stage 0: Carcinoma in situ (early cancer, only present in primary cells)
  - Stage I: Cancers are localized into one part of the body
  - Stage II: Cancers are locally advanced
  - Stage III: Cancers are also locally advanced
  - Stage IV: Cancers have spread to other organs

2.3.2. Breast cancer histological grading [26]

A cancer's grade is determined in terms of three factors: the frequency of cell mitosis or division, the tubules formation and the nuclear pleomorphism (change in cell size and uniformity). Pathologists assign a score between one and three to each of these features and add them up. Table 2.1 describes the scoring technique [26].

Table 2.1: Breast Cancer histological grading techniques [26]

Feature	**Levels**	**Score**
Tubule formation (percentage of carcinoma composed of tubular structures)	>75%	1
	10-75%	2
	Less than 10%	3
Nuclear Pleomorphism (change in cell)	Small, uniform cells	1
	Moderate increase in size and variation 2	2
	Marked variation	3
Mitosis count (cell division)	Up to 7	1
	8 to 14	2
	15 or more	3

The final score ranges from 3 to 9. If this sum is 3, 4 or 5, the tumor is considered grade 1. If it is 6 or 7, the assigned grade is 2. A sum of 8 or 9 is a grade 3. Table 2 details the descriptions of each grade.

Table 2.2: Description of cancer grades [26]

Grade	**Scores**	**Description**
1 (lowest)	3, 4, 5	Well-differentiated breast cells; cells are not growing rapidly; cancer cells are arranged in small tubules

Grade	Scores	Description
2	6, 7	Moderately-differentiated breast cells; have characteristics between grade 1 and grade 3 tumors
3 (highest)	8, 9	Poorly-differentiated breast cells; cells do not appear normal and tend to grow and spread aggressively

2.3.3. Breast Cancer Hormone receptor status

In a breast cancer examination, physicians evaluate whether a hormone receptor is present. When the receptor is present, it is noted by the plus sign (+) after the hormone and when it is not present, it is noted by the negative sign (-) after the hormone. Breast cancer may or may not have three important receptors: an estrogen receptor (ER), a progesterone receptor (PR) or the human epidermal growth factor receptor 2 (HER2) [27, 28].

ER+ cancers depend on estrogen for their growth. Thus, blocking this estrogen slows the growth and reproduction of cancerous cells [27]. Not much is known about PR receptors yet. However, it has been noticed that most ER+ breast cancers are also PR+ and that if an ER+ breast cancer is PR-, the patient has a bad prognosis [27]. HER2+ represents an over-expression of HER2, a protein responsible for regulating cell growth. HER2+ responds to drugs [28].

2.3.4. Breast cancer DNA cytometry

Breast Cancer DNA cytometry consists of counting and measuring a breast tumor's DNA in order to determine its *ploidy* (amount of DNA). Ploidy can be defined as a marker that can help predict how quickly a cancer is likely to spread. If a breast cancer cell has the

same amount of DNA as the normal cells, then the cancer is called diploid. If the amount is different, the cancer is called aneuploidy [9, 29].

2.4. Signs and Symptoms [30]

There are signs that may be indicative of presence of breast cancer. These are:

- A new lump in the breast or armpit
- Thickening or swelling of part of the breast
- Irritation or dimpling of the breast skin
- Redness or flaky skin in the nipple area or breast
- Pulling in of the nipple or pain in the nipple area
- Nipple discharge other than breast milk, including blood
- Any change in the size or shape of the breast
- Pain in any area of the breast

It is important to note that most of these symptoms may not represent an underlying breast cancer. Nevertheless, it is recommended to take each and every one of them seriously.

2.5. Risk factors [31]

A risk factor is anything that increases the chance of developing cancer. The most significantly higher risk is a personal history of breast cancer. There are several moderately higher risks:

- Getting older: breast cancer risk increases with age
- Direct family history: having a mother, sister or daughter with breast cancer increases the risk of breast cancer
- Genetics: being a carrier of either of two familial breast cancer genes *BRAC1* or *BRAC2* puts a woman at higher risk of breast cancer
- Breast lesions: a previous breast biopsy result of atypical hyperplasia increases breast cancer risk

Other risks include but are not limited to:

- Distant family history: having a distant relative with breast cancer increases breast cancer risk
- Age at childbirth: having a first child after the age of 35 or never having children is a risk for breast cancer
- Early menstruation: having periods before the age of 12
- Late menopause: beginning menopause after the age of 55
- Weight: being overweight with excess caloric and fat in-take
- Excessive radiation: being exposed to a large amount of radiation before the age of 30
- Family history of other cancer: family history of cancers of the ovaries, cervix, uterus or colon has been associated with higher risk of breast cancer

- Hormone replacement therapy (HRT): long-term use of combined estrogen and progesterone

2.6. Breast cancer diagnosis and screening

In the presence of a breast cancer sign, a physician performs additional exams to determine whether the underlying condition leads to a diagnosis of breast cancer. Since breast cancer is a serious disease, healthy individuals with no signs or symptoms are tested or screened in an effort to achieve an earlier diagnosis.

There are several screening techniques available [10]. Starting at age 20, it is recommended that every woman performs a breast self-examination (BSE) and reports any changes as soon as possible. For those who do not know how to perform a DBS or choose not to, there is an option to have it done by a health care professional, in which case, it is called a clinical breast examination (CBE) [10].

Even though these exams are very important, it has been found that they play a small role compared to other diagnostic screening tests. More sophisticated techniques are recommended for women with a high risk of breast cancer, and for every woman after the age of 40. These methods include mammogram and magnetic resonance imaging (MRI). A mammogram is an x-ray of the breast. The doctor looks for calcifications (tiny mineral deposits) and a mass within a breast. An MRI is usually offered in addition to the mammogram for certain high risk women. The MRI scans use magnet and radio waves as opposed to x-ray and produce high quality images [10].

2.7. Breast cancer treatment

Usually, the main treatment for breast cancer is surgery. Then, there is an association with either medication (such as chemotherapy) or radiation, or both.

2.7.1. Surgery [32]

Surgery is performed in an effort to remove all the breast cancer cells. It constitutes the main line attack in the episode of treatment. There are two types of surgeries: Mastectomy and Breast Conservation Surgery. Mastectomy is the removal of all of the breast tissue. Breast Conservation surgery, mostly referred to as Lumpectomy, is the removal of only the tumor along with an amount of surrounding healthy tissues.

2.7.2. Medication [9]

Medications are usually offered after surgery (adjuvant therapy), but they can also be offered prior to surgery (neo-adjuvant therapy) in order to shrink the tumor. In neo-adjuvant therapy, the medications offered are chemotherapy. In adjuvant therapy, there are three main groups of medications used:

- Hormone Blocking Therapy
- Chemotherapy
- Monoclonal Antibodies

Hormone Blocking Therapy medications are used in women with ER+ and PR+. Chemotherapy consists of a combination of medications for a period of 3 to 6 months. Monodermal Antibodies are used to block the over-expression of HER2+.

2.7.3. Radiation [9]

Radiation therapy (or radiotherapy) is offered to women to target and destroy very small tumors that may have not been seen. It is usually administered after surgery (external beam radiotherapy) for several weeks, but it can also be administered at the time of operation (branchy-therapy or internal radiotherapy).

2.8. Breast cancer evolution [33]

The area of breast cancer has seen tremendous evolution from the early 1900 to the present. In the 1970s, the breast cancer incidence in the USA was about 105 per 100,000 women. At that time in history, mastectomy was the only surgical treatment adopted and the relative survival rate was about 76%. In medical research, there was only one completed clinical trial, the randomized trial of mammography. The clinical investigation of combination chemotherapy and of the drug, tamoxifen, as well as adjuvant therapy had just started.

In 2007, the incidence rate of female breast cancer was estimated to be 125 per 100,000 with a mortality rate of 23 per 100,000. The five-year relative rate was then 91% among white women and 78% among African American women. Mastectomy was no longer the first option as a surgical treatment; it had been replaced by Breast Conservation Surgery

supplemented by radiation. In addition, women had an option of neo-adjuvant (pre-operation) therapy to help shrink the tumors. The use of tamoxifen and other selective estrogen receptor modulators had become a regular treatment for women with type estrogen receptor positive (ER+). The knowledge in genomics had advanced and breast cancer could be classified in terms of gene susceptibility (*BRAC1, BRAC2, TP53* and *PTEN/NMAC1*).

Nowadays, the knowledge in genomics has increased considerably and is in rapid and extraordinary evolution. Studies on more advanced and less toxic treatments are being done. With the exploration of gene expression knowledge, treatments tailored to the tumor characteristics will be developed [33].

The future holds promising advances in breast cancer. Development in breast cancer prevention will be made possible through the increasing knowledge of the immune system. Vaccines of breast cancer are under clinical evaluation. Screening techniques providing detection of breast cancer at the earliest stages will be developed with the constant increase in technology. The development in biology, medicine, and pharmaceutical science will provide treatments that will reduce breast cancer mortality rates and improve survival.

2.9. Summary

Breast cancer is an uncontrolled growth of the breast tissues that can start either in the lobules or in the ducts. There are various classifications that are used, the tumor size, the spread, the cell division and the DNA composition. Breast cancer has many signs and

symptoms that alert the patient. Screening techniques are available and the can help diagnose breast cancer early. When breast cancer is confirmed, many treatment sequences options are available. However, surgery remains the core action. Breast cancer treatment has gone through remarkable improvement through the years thanks to the evolution of technology and scientific knowledge. Because of this, the future in breast cancer treatment research holds promising discoveries.

CHAPTER 3

LITERATURE REVIEW AND RELATED WORK

3.1. Introduction

In this chapter, the literature is reviewed to analyze the types of studies that were done in the comparison of lumpectomy to mastectomy. The methods used for analysis are evaluated along with the types of data. Also, the objectives of the study as well as the results are presented.

Several studies have been conducted to evaluate the long-term effect of breast conservation surgery in comparison to mastectomy. These studies used different designs and analyzed different types of data covering different populations. Most of them achieved the same conclusion that breast conservation surgery has the same disease free survival and overall survival as mastectomy.

3.2. Methods used to compare lumpectomy to mastectomy in clinical trial data

In 1995, Jacobson et al [5] in an effort to confirm the clinical equivalence of the conservative therapy and mastectomy reported the results of a clinical trial. The main objective was to compare lumpectomy plus radiation with modified radical mastectomy. The recruitment time was from July 1979 to December 1987 and the patients were

followed until November 1993 for a median follow up time of 121 months. A total of 237 eligible participants with clinical stage I and II breast cancer were randomly assigned to either mastectomy (116) or lumpectomy and radiation (121). The investigators were interested in the 10-year overall survival, 10-year disease free survival and the 10-year local-regional recurrence. As results, it was found that the overall survival was 75% in the mastectomy arm and 77% in the lumpectomy arm (p-value=0.89). The 10-year disease free survival rate was 69% in the mastectomy in comparison to 72% in the lumpectomy plus radiation arm (p-value=0.93). Finally, 10% of women treated with mastectomy had a local-regional recurrence compared to only 5% in the lumpectomy arm (p-value=0.17). Jacobson and colleagues concluded that in the management of stage I and II breast cancers, breast conservation with lumpectomy and radiation offer results at 10-years that are equivalent to those with mastectomy. In this research, data were analyzed with Kaplan-Meier estimates to calculate the probability of survival and disease free survival and the Mantel-Haenszel test to determine the significance of the difference between pairs of actuarial curves.

Van Dongen et al [4] published a report in 2000 about a randomized multicenter clinical trial with two arms, mastectomy and breast conservation therapy. They realized that there were no trials with women with large tumors and decided to compare breast conservation surgery with mastectomy in women with stage II breast cancers with tumors up to five centimeters. The study population consisted of 868 patients with a diagnosis of stage I or II invasive carcinoma of the breast, among which, 420 were randomized to undergo a mastectomy and 448 breast conservation therapy. Patients were recruited from May 1980 to May 1986. Follow up time was from the randomization date until May, 1990. The

overall survival rate was 66% in the mastectomy group versus 65% in the breast conservation therapy group (p-value=0.11). The distant metastasis-free rate was 66% in the mastectomy group versus 61% in the breast conservation therapy group (p-value=0.24) and the locoregional recurrence rate was 12% in the mastectomy arm versus 20% in the breast conservation surgery group (p-value=0.01). The investigators concluded that breast conservation therapy and mastectomy demonstrated similar survival rates in a trial in which the majority of participants had stage II breast cancer. Statistical methods used to analyze the data were the Kaplan-Meier curves to estimate duration of survival, time to distant metastasis and time to locoregional recurrence, and the log rank test to compare the results of mastectomy and breast conservation therapy. The Cox proportional hazard models were also used.

In 1969, Veronesi and colleagues [3] conducted a randomized trial for which reports were published in 1977 [34] and 1981 [35]. In 2002, an update of the 20-year follow up was published [3]. It had as an objective to compare the efficacy of radical (Halsted) mastectomy with that of breast conservation surgery. The trial had two arms: the radical mastectomy and breast conservation surgery. About 701 women were enrolled, 349 in the mastectomy arm and 352 in the breast conservation group during the recruitment time (from 1973 to May 1980). The investigators compared the two groups in terms of recurrence of the tumor in the same breast and mortality. Other events included contralateral breast carcinomas, distant metastases and secondary primary cancers. After 20 years of follow up, 30 individuals in the breast conservation arm had had recurrent tumors versus only eight in the mastectomy arm (p-value<0.001). No statistically significant difference in the occurrence of other events or in mortality was observed. The

investigators concluded that the long-term survival of women with breast cancer who were treated with breast conservation surgery was virtually identical to the rate among women who had radical mastectomy. The overall survival and breast cancer-specific survival rates were also similar in the two groups. Data analysis was performed using the Gray test to compare the crude cumulative incidence of recurrent tumor and the local recurrence of tumors and the Kaplan-Meier curves to estimate the survival curves.

Fisher et al [2, 36, 37] initiated a study in 1973 and published reports of the eight year results [37], the 12 year results [36] and the 20 year results [2]. The latter was published in 2002 and aimed to determine whether lumpectomy with or without radiation was as effective as total mastectomy for the treatment of invasive breast cancer. The design was a prospective randomized clinical trial with three arms: total mastectomy, lumpectomy, and lumpectomy with additional breast irradiation. Women were recruited between August 8th 1976 to January 27th 1984 and they were followed up for a mean of 20.8 years (total mastectomy group), 20.6 years (lumpectomy group) and 20.7 years (lumpectomy with irradiation group). The 20-year follow up data were available for 1851 participants, 589 in the total mastectomy arm, 634 in the lumpectomy arm and 628 in the lumpectomy plus radiation arm. The investigators looked at the disease-free survival, distance-disease free survival, and overall survival. It was found that the Hazard Ratio for death was 1.05 (95% confidence interval: 0.90-1.25) for the lumpectomy group and 0.97 (95% confidence interval: 0.83-1.14) for the lumpectomy plus radiation group in comparison to the mastectomy group. As a conclusion, no significant differences were observed among the three groups of women with respect to disease-free survival, distance disease free survival or overall survival. Data were analyzed using the Kaplan-Meier method to

estimate disease-free survival, distant disease-free survival and overall survival rates and the log-rank statistics to determine differences among the treatment groups with respect to death from causes other than breast cancer.

In 2003, Poggi [6] and colleagues conducted a randomized prospective clinical trial in which 237 women with breast cancer stage I and stage II were assigned to either the mastectomy therapy (116 patients) arm or the breast conservation therapy (121 patients). The median age was 50 years old. The participants were followed from 1979 for a median time of 18.5 years. The investigators were interested in the overall survival as well as disease free survival. They were motivated by the fact that there was no long-term randomized data in the literature and they aimed to confirm the fact that there were no detectable differences between the mastectomy therapy and the breast conservation surgery in overall survival and disease-free survival. Their study found that the estimated overall 20-year overall survival rate was 58% for the mastectomy therapy group and 54% for the breast conservation therapy group (p-value=0.67). The overall disease-free survival rate was 67% for the mastectomy group and 63% for the breast conservation surgery group (p-value=0.64). The occurrence of secondary events (isolated chest wall events, regional events, distant events and non-breast cancer events) were also evaluated. It was found that their occurrence was not significantly different between the two groups. Peggi et al concluded that there were no statistically significant differences in overall survival and disease-free survival. In this study, the investigators used the Kaplan-Meier methods to compare the probabilities of overall survival and disease-free survival and Mantel-Haenszel tests to evaluate the differences between pairs of actuarial curves.

3.3. Methods used to compare lumpectomy to mastectomy in administrative data

In the year of 2003, Kroman et al [7], realizing that the published studies comparing breast conservation therapy to radical mastectomy had used data of middle-aged or older women, decided to analyze the effect of age on breast carcinoma survival according to the type of surgical treatment used. They performed a registered data-based research study using the Danish Breast Cancer Cooperative Group database. The study analyzed data from 9285 premenopausal women of 50 years old and under with primary breast carcinoma who had undergone radical mastectomy (7165 women, 77.2%) or breast conservation surgery (2120 women, 22.8%). Women considered for analysis were those diagnosed between January 1st 1982 and December 31st 1998. The variable of interest was the 10-year overall death rates. The study found that in comparison to women who had received the radical mastectomy, women of age less than 45 did not have a different risk of death. The report concluded that for younger women, long-term survival was similar for those who were treated by breast conservation therapy and those who were treated by radical mastectomy. The statistical methods used were the Poisson Regression performed as likelihood ratio tests to analyze the relative risk of death due to breast carcinoma and the Chi-Square tests to analyze the associations between baseline characteristics.

3.4. Use of advanced data mining techniques to compare Mastectomy to Lumpectomy

Martin et al. [38] published a study in 2006 in which a comparison of Mastectomy and Breast Conserving Surgery was made using a classification tree approach. They used data

from the Western Australian data on a population of women diagnosed between 1990 and 2000. The main objective was to identify the most important factors to determine the type of surgery. The total number of women was 2713 among which 39% underwent Mastectomy and 61% underwent Breast Conserving Surgery. The outcome of analysis was type of surgery and the factors of interest included tumor size, age, area of residence, tumor histology, lymph node (nodal) status, country of birth, payment class and marital status. Two models were compared, the decision tree and the logistic regression. They found that tumor size was the primary determinant of patient choice. Patients with smaller tumors (less than 20 mm in diameters) preferred Breast Conserving Surgery. For patients with larger tumors (greater than 20 mm), important factors for choice were age, nodal status and tumor histology. Methods used for analysis were the decision tree and the logistic regression. In terms of model, they found that classification trees performed as well as logistic regression for predicting type of surgery.

3.5. Comparison of Mastectomy and Lumpectomy in terms of patient choice and psychological outcomes

Kirby et al. [39] analyzed the effect of patient choice on the rates of Mastectomy. This was a cross-sectional study in which 203 breast cancer patients who had mastectomy were invited to fill out a questionnaire to assess whether an option of Breast Conserving Surgery was provided and the reason of the choice made. This study found that patients chose Mastectomy because they felt safer (n = 119), wanted to decrease the risk of further surgery (n = 87) and/or wished to avoid radiotherapy (n = 34). It was found that despite being advised that there is no difference between survival rates of Mastectomy and Breast Conserving Surgery, many patients still felt safer with Mastectomy.

Fallowfield et al. [40] evaluated the psychological outcome of different treatment policies in women with early breast cancer who underwent either mastectomy or breast conservation surgery depending on the surgeon's opinion or the patient's choice outside a clinical trial. This was a prospective, multicenter study. The study population comprised 269 women under 75 with a probable early stage cancer who were referred to 22 different surgeons. The main outcomes of interest were anxiety and depression as assessed by standard methods two weeks, three weeks, and 12 months after surgery. It was found that the incidences of anxiety, depression, and sexual dysfunction were high in all treatment groups. There were no significant differences in the incidences of anxiety and depression between women who underwent Mastectomy and those who underwent Lumpectomy. This study was inconclusive as to the question whether women with early breast cancer who undergo breast conservation surgery have less psychiatric morbidity after treatment than those who undergo mastectomy.

3.6. Economic comparison of Mastectomy and Lumpectomy

Barlow et al. [41] compared the cost of Mastectomy and Breast Conserving Therapy for Early Stage Breast Cancer in a total of 1675 women 35 years old or older. Their objective was to evaluate the cost of medical care up to five years after diagnosis in early breast cancer where both treatment had been shown to be equally officious. This study was a retrospective observational longitudinal study using a regional nonprofit health maintenance organization in the period 1990 through 1997. Comparative treatment groups were Mastectomy only (n = 183), Mastectomy with adjuvant hormonal therapy or chemotherapy (n = 417), Breast Conserving Therapy with radiation therapy (n = 405) and Breast Conserving Therapy with radiation therapy and adjuvant hormonal therapy or

chemotherapy (n = 670). In this study, it was found that six months after diagnosis, the differences among the mean total medical care costs for the four treatment groups were statistically significantly (p-value < 0.001), with Breast Conserving Therapy being more expensive than Mastectomy. The adjusted mean costs were $12987, $14309, $14963 and $15779 respectively for Mastectomy alone, Mastectomy with adjuvant therapy, and Breast Conserving Therapy plus radiation therapy, Breast Conserving Therapy plus radiation therapy with adjuvant therapy. One year after diagnosis, the difference in cost was still statistically significant (p-value < 0.001), but costs were influenced more by the use of adjuvant therapy than by the type of surgery. The one-year adjusted mean costs were $16704, $18856, $17344, $19081, respectively for Mastectomy alone, Mastectomy with adjuvant therapy, Breast Conserving Therapy plus radiation therapy, Breast Conserving Therapy plus radiation therapy with adjuvant therapy. By five years, Breast Conserving Surgery was less expensive than Mastectomy (p-value < 0.001). The five-year adjusted mean costs of $41930, $45670, $35787 and $39926 respectively for Mastectomy alone, Mastectomy with adjuvant therapy, Breast Conserving Therapy plus radiation therapy, Breast Conserving Therapy plus radiation therapy with adjuvant therapy. They concluded that had higher short-term costs but lower long-term costs in comparison Mastectomy.

Polsky et al. [42] conducted an economic evaluation comparing Breast Conservation and radiation with Mastectomy. The purpose of their study was to compare the incremental cost effectiveness of Breast Conservation and radiation versus Mastectomy with the restriction of choice to a single therapy versus providing a choice of either therapy. This was a random retrospective cohort study which included a total of 2517 Medicare

beneficiaries with early breast cancer treated in the years 1992 to 1994. The outcome variables of interest were quality-adjusted life-years and 5-year medical costs. In terms of cost, they found that Breast Conservation and radiation had significantly higher costs than Mastectomy in the first year after surgery. Five years after surgery, the adjusted costs were $14,054 (95% confidence interval, $9791 to $18312) higher for Breast Conservation and radiation than for Mastectomy. The incremental cost effectiveness ratio comparing Breast Conservation and radiation to Mastectomy was $219594 per quality adjusted life year for comparison of both treatment strategies. They also found that if possibility of patient choice from maintaining the availability of multiple treatments versus restricting choice to mastectomy alone provides a quality-of-life gain of 0.031 quality adjusted life years, then the cost-effectiveness ratio of this choice option was $80440 per quality adjusted life years. They concluded that the system of providing a choice between Mastectomy and Breast Conservation surgery was economically attractive when the economic analysis includes the benefit of patient choice of treatment.

3.7. Summary

Many breast cancer studies published, comparing breast conservation surgery to mastectomy, were clinical trials and record-based observational analyses. The main variables of interest across many of them were health outcomes expressed in terms of mortality and survival. Data analysis of such variables has been performed with the use of statistical models such as Kaplan-Meier estimates, Log-Rank tests, ANOVA, Logistic Regression, Poisson Regression, etc.

In contrast to these generated data analysis models, the focus has also been into developing mathematical models to assess the risk of breast cancer [43, 44]. These models include 'the risk for familial breast cancer model' [45, 46], 'the individualized probability of developing breast cancer model [47], the log-incidence mathematical model of breast cancer incidence' [48], and the breast cancer prediction model which incorporates familial and personal risk factors [49].

To this investigator's knowledge, there are no studies in which lumpectomy was compared to mastectomy in terms of breast cancer health outcomes and hospitalization usage variables using data mining techniques such as cluster analysis and predictive modeling. Nor is there any study that reported a prediction model for a risk of re-hospitalization in the case of surgical treatment for breast cancer. The current study will provide a new approach of large longitudinal record-data analysis.

CHAPTER 4

STATISTICAL ANALYSIS THEORY

4.1. Objective

The objective of the current chapter is to discuss the theory of statistical analysis behind this examination of breast cancer. Statistical analysis was used to evaluate the contrast of lumpectomy to mastectomy in terms of short-term in-hospital resources use and short-term post-operative follow-up healthcare resources use (hospital, outpatient service and prescribed medication use). First, an overview of the statistical analysis is presented. Then, statistical analysis practical notes are reviewed before summarizing the statistical theory. Since the current study is a comparative, the emphasis is put on the statistical methods and models used in comparing outcomes among groups.

4.2. Statistical analysis overview

Statistics is the branch of the scientific method where limited sample data are used to make inferences about random phenomena [50, 51]. It can simply be defined as the science of the organization of data collection and data interpretation [52]. Statistical methods include every process from the planning of data collection to producing reports of data analysis. The field of statistics has two main areas: mathematical statistics and applied statistics. Mathematical statistics is about the development of new methods

requiring in-depth knowledge of abstract mathematics. Applied statistics is about applying the methods of mathematical statistics to specific subject fields, such as business, human sciences, public health and medicine [51].

In general, the aim of the field statistics is to characterize a population based on the information observed in a sample taken from the same population. The sample information is expressed by functions of the observed data, which are called *statistics.* The field of statistics seeks to determine which functions are the most relevant in the characterization of various populations [53].

4.3. Statistical Analysis Practical Notes

Applied statistics can be viewed as a set of methodologies used to help carry out scientific experiments. In keeping with the scientific method, applied statistics consists of developing a hypothesis, determining the best experiment to test the hypothesis, conducting the experiment, observing the results, and making conclusions. The statistician's responsibilities include: study design, data collection, statistical analysis, and making appropriate inferences from data. In doing so, the statistician attempts to limit bias, maximize objectivity and obtain scientifically valid results [52].

Often, the common goal for a statistical research project is to examine causality: the effect of independent variables (or predictor variables) on dependent variables (or responses). However, no statistical study can make a conclusion as causality; they can only investigate whether or not parameters are related. There are two types of statistical research studies: experimental studies and observational studies. The difference between

the two resides in the conduct of the study and the randomness of the data. In an experimental study, the investigator takes measurements of the system under study, manipulates the system, evaluating how the manipulation modifies the measurement. In contrast, in an observational study, the investigator has no control over the measurements. In this case, data are gathered and the investigator examines the relationship/association between predictor and response. There is no randomness and the attempt of the investigator in an observational study is to approximate the concept of randomness as much as possible.

4.3.1. Statistical methods

Statistical methods used to analyze the sample data in an objective to characterize populations can be classified as descriptive, inferential or exploratory analyses. *Descriptive statistics* are methods used to describe the distribution of the measurements. They consist of estimates of the central tendency, measures of variability, counts, percentages and graphical tools [51, 53]. *Inferential statistics* are methods that use probability to express the level of certainty about estimates and to test specific hypotheses. There are two main inferential statistics methods: confidence interval estimation and hypothesis testing [51, 53]. *Exploratory analyses* methods use both descriptive and inferential techniques to explore potential relationships in data. Given a large data set, it is very likely that at least one statistically significant result can be found by using exploratory analysis. The results found are not used to draw conclusions but they are considered 'hypothesis-generating' because they are not pre-planned. Usually, these results inspire the design of new prospective studies to test these new hypotheses [53].

4.3.2. Probability distributions

Probability distributions play a major role in statistics. In inferential statistics, probability distributions are used to test hypotheses. There are two types of probability distributions: discrete and continuous.

Discrete distributions describe variables that can only take discrete values (discrete random variables). Commonly, the discrete distributions include the binomial, the negative binomial, the poisson and the hypergeometric distributions [53].

Continuous distributions describe variables that can take any value within an interval (continuous random variables). The common continuous distributions used are the normal distribution, the exponential distribution, the chi-square distribution, the F-distribution and the Student's t-distribution [53].

4.4. Statistical Analysis theoretical notes

Mathematical statistics is the study of statistics using mathematical theories, such as probability theory, statistical theory and also linear algebra and analysis [52]. Methods used in applied statistics are developed in mathematical statistics. There are mainly two types of statistical analysis: descriptive and inferential.

4.4.1. Descriptive statistics

Descriptive statistics describe the probability distribution of the population. This is achieved by computing measures of central tendency and dispersion, counts and frequencies and by viewing the shape of the distributions using density estimations and

histograms. Descriptive statistics is the first step in a statistical study; it gives the overall picture of what the data look like [53].

Table 4.1: The common descriptive statistics for central tendency and dispersion [53]

Type of measure	Measure	Formula
Measures of central tendency	Arithmetic mean	$\bar{x}=\frac{\sum x_i}{n}=\frac{x_1+x_2+\cdots+x_n}{n}$
	Median	- The middle value if n, the sample size, is odd - The average of the two middle values if n is even
	Mode	The most frequently occurring value
	Geometric mean	$(\prod x_i)^{1/n} = (x_1 * x_2 * x_3 * \ldots * x_n)^{1/n}$
	Harmonic mean	$\frac{n}{\sum(x_i)^{-1}} = n\{(1/x_1) + (1/x_2) + \ldots + (1/x_n)\}^{-1}$
	Weighted mean	$\bar{x}_w = \frac{\sum w_i x_i}{W}$, where $W = \sum w_i$
	Trimmed mean	Arithmetic mean omitting the largest and the smallest observations
	Winsorized mean	Arithmetic mean after replacing outliers with the closest non-outliers values
Measures of dispersion	Variance	$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
	Standard deviation	$s = \sqrt{s^2}$
	Standard Error (of the mean)	$(s^2 / n)^{1/2}$ = standard deviation of $\bar{x}$
	Range	Largest value – smallest value
	Mean absolute deviation	$\frac{\sum \lvert x_i - \bar{x} \rvert}{n}$
	Inter-Quatile range	75^{th} percentile - 25^{th} percentile
	Coefficient of Variation	$\frac{s}{\bar{x}}$

4.4.2. Inferential statistics: Confidence interval estimation

The two primary statistical methods for making inferences are confidence interval estimation and hypothesis testing. Population parameters can be estimated by point

estimates such as the mean and median using descriptive statistics. This estimate represents the 'best guess' at the value of the true parameter. However, a point estimate does not give any idea about how much information it is based on or how likely it is close to the true parameter. A way to remedy this incompleteness of a point estimate is to obtain an estimate of an interval that is likely to contain the true value of the parameter, also known as a confidence interval. A confidence interval is constructed around the point estimate and it contains the parameter with a specific high probability or confidence level. In general, the confidence interval is finite of the form $[\theta_l, \theta_u]$, where θ_l represents the lower limit and θ_u is the upper limit of the interval [53, 54]. The probability that a confidence interval will contain θ is called the confidence coefficient.

Suppose that $\hat{\theta}_l$ and $\hat{\theta}_u$ are the (random) lower and upper confidence limits, respectively for the parameter θ. Then if

$$P(\hat{\theta}_l \leq \theta \leq \hat{\theta}_u) = 1 - \alpha,$$

the probability $(1 - \alpha)$ is the confidence coefficient. The resulting random interval defined by $[\hat{\theta}_l, \hat{\theta}_u]$ is called a two-sided confidence interval. One widely used method for finding confidence intervals is called the pivotal method. This method depends upon finding a pivotal quantity that possesses two main characteristics: (1) It is a function of the sample measurements and the unknown parameter θ, where θ is the only unknown quantity and (2) its probability distribution does not depend upon the parameter θ [46]. The logic behind this method is that for a random variable X, the probability P ($a \leq X \leq b$) is unaffected by a change of scale or translation on X. Therefore, if the probability

distribution of a pivotal quantity is known, the operations such as scaling and translation may be used to create a confidence interval [55].

Suppose that a pivotal point quantity Q (X, θ) has been determined for a random sample $x_1, x_2, \ldots, x_n$ from a population X with probability density function f (x; θ), where θ is the unknown. For a specific value of α, two numbers a and b that do not depend on θ, can be found to satisfy

$$P(a \leq \theta \leq b) \geq 1 - \alpha$$

With algebraic manipulation of the inequality, θ can be isolated in the middle yielding

$$P(\hat{\theta}_l \leq \theta \leq \hat{\theta}_u) = 1 - \alpha,$$

where $\hat{\theta}_l = \hat{\theta}_l (x_1, x_2, \ldots, x_n)$ and $\hat{\theta}_u = \hat{\theta}_u (x_1, x_2, \ldots, x_n)$ [56].

4.4.3. Inferential statistics: Hypothesis testing

Hypothesis testing is a means of formalizing the inferential process for the purpose of decision-making. It is a way to use logical arguments to test hypothesized statements about population parameters in a statistical approach [53]. In many ways the formal procedure for hypothesis testing is similar to the scientific method [50].

The scientist observes nature, formulates a theory, and then tests this theory against observation. In the statistical context, the scientist poses a hypothesis concerning one or more population parameters. Then, he/she samples the population and compares her observations with the hypothesis. If the observations disagree with the hypothesis, the scientist rejects it. If not, the scientist concludes either that the hypothesis is true or that

the sample did not detect the difference between the real and hypothesized values of the population parameters.

Any statistical test of hypotheses works in exactly the same way and is composed of five essential elements: (1) the null hypothesis (H_0), (2) the alternative hypothesis (H_a), (3) the test statistics and (4) the rejection region or decision rule and (5) the conclusion [46, 49].

In mathematical terms, the hypothesis test is an ordered sequence (X_1, X_2, ..., X_n; H_0, H_a; C) where X_1, X_2, ..., X_n is a random sample from a population X with the probability density function f(x; θ), H_0 and H_a are hypotheses concerning the parameter θ in f(x; θ), and C is a Borel set in $\mathbb{R}^n$ [56].

Borel sets are defined using the notion of σ-algebra. A collection of subsets $\mathcal{A}$ of a set S is called a σ-algebra if (i) $S \in \mathcal{A}$, (ii) $A^c \in \mathcal{A}$, whenever $A \in \mathcal{A}$, and (iii) $\bigcup_{k=1}^{\infty} A_k \in \mathcal{A}$, whenever A_1, A_2, ..., A_n, ... $\in$ A. The Borel sets are the members of the smallest σ-algebra containing all open sets of $\mathbb{R}^n$. Two examples of Borel sets in $\mathbb{R}^n$ are the sets that arise by a countable union of closed intervals in $\mathbb{R}^n$, and a countable intersection of open sets in $\mathbb{R}^n$ [56].

The set C is called the critical region in the hypothesis test. The critical region is obtained using a test statistic $W(X_1, X_2, ..., X_n)$. If the outcome of $(X_1, X_2, ..., X_n)$ turns out to be an element of C, then we decide to accept H_a; otherwise we accept H_0 [56].

4.4.3.1. The null hypothesis and the alternative hypothesis

A statistical hypothesis H is a conjecture about the distribution, f(x; θ), of a population X. This conjecture is usually the parameter [56].

A hypothesis H can be a simple hypothesis if it completely specifies the density $f(x; \theta)$ of the population; otherwise, it is called a composite hypothesis.

The hypothesis to be tested is called the null hypothesis (denoted by H_0) and the negation of the null hypothesis is called the alternative hypothesis (denoted by H_a).

If θ denotes a population parameter, then the general format of the null hypothesis and alternative hypothesis is

$$H_0 : \theta \in \Omega_0 \quad \text{and} \quad H_a : \theta \in \Omega_a \quad [56]$$

where Ω_0 and Ω_a are subsets of the parameter space Ω with

$$\Omega_0 \cap \Omega_a = \emptyset \text{ and } \Omega_0 \cup \Omega_a \subseteq \Omega \quad [56].$$

Most often, $\Omega_0 \cup \Omega_a = \Omega$; thus, the expressions of the null and the alternative hypotheses become

$$H_0 : \theta \in \Omega_0 \quad \text{and} \quad H_a : \theta \notin \Omega_0 \quad [56]$$

If Ω_0 is a singleton set, then H_0 reduces to a simple hypothesis.

4.4.3.2. The test statistic

Broadly speaking, a hypothesis test is a rule that tells us for which sample values we should decide to accept H_0 as true and for which sample values we should decide to reject H_0 and accept H_a as true.

Typically, a hypothesis test is specified in terms of a test statistic, W. There are several methods to find test procedures including: (1) Likelihood Ratio Tests, (2) Invariant Tests,

(3) Bayesian Tests, and (4) Union-Intersection and Intersection-Union Tests. The most commonly used method is the likelihood ratio test. The likelihood ratio test statistic for testing the simple null hypothesis H_0: $\theta \in \Omega_0$ against the composite alternative hypothesis H_a: $\theta \notin \Omega_0$ based on a set of random sample data $x_1, x_2, \ldots, x_n$ is defined as

$$W(x_1, x_2, \ldots, x_n) = \frac{max_{\theta \in \Omega_0} L(\theta_0, x_1, x_2, \ldots, x_n)}{max_{\theta \in \Omega} L(\theta_a, x_1, x_2, \ldots, x_n)} \text{ [56],}$$

where Ω denotes the parameter space, and $L(\theta, x_1, x_2, \ldots, x_n)$ denotes the likelihood function of the random sample, that is

$$L(\theta, x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i, \theta) .$$

A likelihood ratio test (LRT) is any test that has a critical region C (rejection region) of the form

$$C = \{(x_1, x_2, \ldots, x_n)) \mid W(x_1, x_2, \ldots, x_n)) \leq k\} \text{ [56],}$$

where k is a number in the unit interval [0, 1].

If $H_0 : \theta = \theta_0$ and $Ha : \theta = \theta_a$ are both simple hypotheses, then the likelihood ratio test statistic is defined as

$$W(x_1, x_2, \ldots, x_n) = \frac{L(\theta_0, x_1, x_2, \ldots, x_n)}{L(\theta_a, x_1, x_2, \ldots, x_n)} . \text{ [56]}$$

4.4.4. Common inferential statistical techniques for continuous variables

4.4.4.1. One sample t-test [51, 53]

The one-sample t-test is used to infer whether an unknown population mean is different from a constant value. The t-test assumes that the data are normally distributed with a constant variance.

Suppose that a set of n data points, $y_1, y_2, \ldots, y_n$, represents a random sample selected from a normally distributed population with unknown mean, μ. The test statistic t is a function of the deviation between $\bar{y}$ (the sample mean) and μ_0. It is standardized by the standard error of the sample mean, $/\sqrt{n}$. When H_0 is true, t has the Student's t probability distribution with (n-1) degrees of freedom.

The one-sample two-sided t-test is summarized below [53]:

Null hypothesis:	$H_0: \mu = \mu_0$
Alternative hypothesis:	$H_a: \mu \neq \mu_0$
Test statistic:	$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$
Decision rule:	reject H_0 if $\lvert t \rvert > t_{\alpha/2, n-1}$

The value $t_{\alpha/2, n-1}$ represents the 'critical t-value' of the Student's t-distribution at a two-tailed significance level, α, and $(n - 1)$ degrees of freedom.

4.4.4.2. Two sample t-test [51, 53]

The two-sample t-test is used to compare the means, μ_1 and μ_2, of two independent populations, denoted μ_1 and μ_2. The two populations are assumed to be normally distributed with the same variance σ^2.

Let (y_{11}, y_{12}, …, y_{1n_1}) and (y_{21}, y_{22}, …, y_{1n_2}) be two random samples selected, respectively, from Population 1 and Population 2. Let μ_1 and μ_2 the unknown means of the two populations. Then μ_1 and μ_2 are estimated by $\bar{y}_1$ and $\bar{y}_2$, respectively.

The test statistic, t, is a function of the difference between $\bar{y}_1$ and $\bar{y}_2$ standardized by its standard error, $s(1/n_1 + 1/n_2)^{1/2}$. When H_0 is true, t has the Student's t distribution with N -2 degrees of freedom, where $N = n_1 + n_2$. The unknown common variance σ^2 is estimated by the 'pooled' variance ($s_p{}^2$):

$$s_p{}^2 = \frac{(n_1-1)*s_1{}^2 + (n_2-1)*s_p{}^2}{(n_1+n_2-2)}$$

The two-sample t-test is summarized below [53].

Null hypothesis:	H_0: $\mu_1 = \mu_2$
Alternative hypothesis:	H_a: $\mu_1 \neq \mu_2$
Test statistic:	$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_p{}^2(\frac{1}{n_1}+\frac{1}{n_2})}}$
Decision rule:	reject H_0 if $\lvert t \rvert > t_{\alpha/2,\, N-1}$

4.4.4.3. Wilcoxon rank-sum (Mann Whitney U) test [51, 53]

The Wilcoxon rank-sum test is a non-parametric alternative to the two-sample t-test. It does not assume that the data is normally distributed and can be used to make inferences about the mean as well as the median. It does assume a symmetric population distribution.

Let $y_{i1}, y_{i2}, \ldots, y_{in_i}$ (for i = 1, 2) be two random samples of sizes n_1 and n_2, selected from two independent populations. The test is based on the ranking of the $n_1 + n_2$ combined sample and Wilcoxon rank sum is the sum of the ranks of one of the samples. Without loss of generality, assume that it is the sum of the ranks of the first group.

Let r_{1j} = rank of y_{1j} (j = 1, 2, … , n_1) and r_{2j} = rank of y_{2j} (j = 1, 2, … , n_2), and compute

$$R_1 = \sum_{j=1}^{n_1} r_{1j} \text{ and } R_2 = \sum_{j=1}^{n_2} r_{2j}$$

The closer the average ranks R_1/n_1 and R_2/n_2 are, the more likely the hypothesis of equal means will be supported. For large samples, the rank sum test is performed through a normal approximation. With $N = n_1 + n_2$, the sum of the ranks (1 + 2 + … + N) can be expressed as N (N + 1) / 2. When H_0 is true, it is expected that the proportion of the sum of ranks from sample 1 is about n_1/N and the proportion from sample 2 is about n_2/N. Thus, the expected value of R_1 under H_0 is

$$\mu\mu_{R_1} = \left(\frac{n_1}{N}\right)\left(\frac{N\,(N+1)}{2}\right) = \frac{n_1*(N+1)}{2}$$

The variance of R_1 can be expressed as

$$\sigma^2{}_{R_1} = \frac{n_1 * n_2}{12}(N+1)$$

The test statistic based on an approximate normal distribution, using a 0.5 continuity correction, is summarized below [53]:

Null hypothesis: $H_0: \theta_1 = \theta_2$

Alternative hypothesis: $H_a: \theta_1 \neq \theta_2$

Test statistic: $Z = \frac{|R_1 - \mu_{R_1}| - 0.5}{\sigma_{R_1}}$

Decision rule: reject H_0 if $|Z| > Z_{\alpha/2}$

where θ_1 and θ_2 represent the location parameter (mean, median, …) for the two populations on which the inference is being made.

4.4.4.4. One-way ANOVA [51, 53, 54]

One-way ANOVA (analysis of variance) is used to infer the equality of two or more means of independent groups based on selected random samples. One-way ANOVA assumes that the populations (groups) are normally distributed and have all the same variance, σ^2.

Let $y_{i1},\ldots, y_{in_i}$ be the random sample selected from group (population) i(i=1, 2, …,k). The data are displayed in Table 4.2.

Table 4.2: Data display for the group comparisons [53]

GROUP			
Group1	Group2	…	Group k
y_{11}	y_{21}	…	y_{k1}
y_{12}	y_{22}	…	y_{k2}
…	…	…	…
$y_{1n_1}y_{1n1}$	y_{2n_2}	…	y_{kn_k}

The null hypothesis (H_0) stipulates that there is no difference in mean responses of the different groups or, in other words, that there is "no Group effect" (i.e., no difference in mean responses among groups). The alternative hypothesis is that at least two group means are different or that "the Group effect is important". When H_0 is true, the variation among groups and the variation within groups are independent estimates of the same measurement, σ^2, and their ratio should be close to 1. The test statistic F uses this ratio, variation among groups (MSG)/variation within groups (MSE). Let $N = n_1 + n_2 + \ldots + n_k$. Then the test statistic F has the F-distribution with k-1 upper and N-k lower degrees of freedom.

The test summary is given as [53]

Null hypothesis:	H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
Alternative hypothesis:	H_a: not H_0
Test statistic:	$F = \frac{MSG}{MSE}$
Decision rule:	reject H_0 if $F > F_{N-k}^{k-1}(\alpha)$

MSG is an estimate of the variability among groups, and MSE is an estimate of the variability within groups.

4.4.4.4.1. Mean Square Error (MSE)

One of the assumptions of the ANOVA is that the within-group variance is constant across groups. This assumption is also called the 'variance homogeneity'. For the k groups, this can be expressed as

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$$

where σ_i^2 denotes the unknown variance of the i[th] population. The common variance σ^2 is estimated by s^2, computed as the weighted average of the k sample variances:

$$s^2 = \frac{(n_1-1)*s_1^2+(n_2-1)*s_2^2+\cdots+(n_k-1)*s_k^2}{(n_1+n_2+\cdots+n_k)-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{N-k}$$

where:

$$s_i^2 = \sum_{j=1}^{n_i}\frac{(y_{ij}-\bar{y}_i)^2}{(n_i-1)}$$

is the estimate of the variance within group i and $N = n_1 + n_2 + \ldots + n_k$.

$s^2$is called the mean square error (MSE), and its numerator is the sum of squares for error (SSE). The term 'error' is for the deviation of each observation from its group mean.

4.4.4.4.2. Mean Square for the Group factor (MSG)

The 'among groups' variability is a function of the deviations of the group means ($\bar{y}_i$) from the overall average $\bar{y}$. The overall variance is expressed as:

$$\text{MSG} = \frac{\sum_{i=1}^{k}(\bar{y}_i-\bar{y})^2}{k-1} = \frac{SSG}{k-1}$$

where

SSG = $\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2$ is the sum of squares for groups.

4.4.4.5. Kruskal-Wallis Test [51, 53]

The Kruskal-Wallis test is a non-parametric substitute of the one-way ANOVA when the response variable is not normally distributed. Like in the one-way ANOVA case, samples used to perform the test are assumed to be random and independent as well as symmetric. The Kruskal-Wallis test is an extension of the Wilcoxon rank-sum for more than two groups, just as a one-way ANOVA is an extension of the two-sample t-test.

Let $y_{i1},\ldots, y_{in_i}$ be the random sample selected from group (population) i (i=1, 2, …,k) following the display in Table 4.2. The null hypothesis (H_0) is that of equal mean responses among groups. Suppose that the data are ranked, from lowest to highest, over the combined samples; the test statistic is a function of the ranks and sample sizes.

For i = 1, 2, …, k and j = 1, 2, …, n_i, let r_{ij} = rank of y_{ij} over the k combined samples. For each group (i=1, 2, …, k), compute

$$R_i = \sum_{j=1}^{n_i} r_{ij}$$

The average rank of all $N = n_1 + n_2 + \ldots + n_k$ observations is $\bar{R} = (N+1)/2$. When the null hypothesis is true, the average rank for each group, $\bar{R}_i = (R_i/ n_i)$, should be close to this value and the sum-of-squared deviations

$$\sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R})^2$$

should be small.

The Kruskal-Wallis test statistic is a function of this sum of squares, which simplifies algebraically to the quantity

$$h^* = \frac{12}{N\,(N+1)} \left(\sum_{i=1}^{k} \frac{R_i^2}{n_i}\right) - 3\,(N+1)$$

When H_0 is true, h* has an approximate *chi-square* distribution with k – 1 degrees of freedom.

Let θ represent the population location parameter; the test is summarized as shown below.

Null hypothesis:	H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$
Alternative hypothesis:	H_a: $\theta_i \neq \theta_j$ for at least one pair (i, j)
Test statistic:	$h = \frac{h^*}{\left(1 - \frac{C}{N\,(N^2-1)}\right)}$
Decision rule:	reject H_0 if $h > \chi^2_{k-1}(\alpha)$

$\chi^2_{k-1}(\alpha)$ represents the critical value from a chi-square distribution based on k -1 degrees of freedom and a significance level of α.

4.4.4.6. Two-way ANOVA [51, 53, 54]

The two-way ANOVA is used when analyzing two factors that affect a response simultaneously. As in the one-way ANOVA, a two-way ANOVA factors in the analysis, a group effect. In addition, the two-way ANOVA includes another identifiable source of variation called a blocking factor. Because of this blocking factor, the two-way ANOVA layout is sometimes referred to as a 'randomized block design'.

In general, the randomized block design (two-way ANOVA) has g ($g \geq 2$) levels of a 'group' factor and b ($b \geq 2$) levels of a 'block' factor. Independent samples are

measurements taken from each of the g × b cells formed by the group-block combinations. Let n_{ij} represent the number of measurements taken in Group i and Block j (cell i – j), and let N represent the total number of all measurements over all g × b cells. Let y_{ijk} denote the k^{th} response in Cell i – j (k = 1, 2, …, n_{ij}). The general layout of the randomized block design is shown in Table 4.3.

Table 4.3: Layout of the Randomized Block Design [53]

	Group 1	Group 2	…	Group g
Block 1	$y_{111}, y_{112}, \ldots, y_{11n_{11}}$	$y_{211}, y_{212}, \ldots, y_{21n11}$		$y_{g11}, y_{g12}, \ldots, y_{g1n11}$
Block 2	$y_{121}, y_{122}, \ldots, y_{12n11}$	$y_{221}, y_{222}, \ldots, y_{22n11}$		$Y_{g21}, y_{g22}, \ldots, y_{g2n11}$
…	…	…	…	…
Block b	$y_{1b1}, y_{1b2}, \ldots, y_{1bn11}$	$y_{2b1}, y_{2b2}, \ldots, y_{2bn11}$		$Y_{gb1}, y_{gb2}, \ldots, y_{gbn11}$

The general entries in a two-way ANOVA summary table are represented as shown in the following table.

Table 4.4: ANOVA Summary Table for the Two-Way ANOVA [53]

Source	**df**	**SS**	**MS**	**F**
Group (G)	g - 1	SSG	MSG	F_G = MSG/MSE
Block (B)	b – 1	SSB	MSB	F_B = MSB/MSE
G x B (Interaction)	(g-1)(b-1)	SSGB	MSGB	F_{GB} = MSGB/MSE
Error	N - gb	SSE	MSE	
Total	N - 1	TOT(SS)		

The SS represents the sum of squared deviations associated with the factor listed under Source. Sum of squares are computed similarly as in one-way ANOVA.

The mean square (MS) is ratio of the SS by the corresponding degrees of freedom. The MS represents a measure of variability associated with the factor listed under source. When there is no effect due to the specified factor (under source), this variability is the measurement of error variability, σ^2, which is estimated by MSE.

The F-values are ratios of the effect mean squares (MS_) to the mean square error (MSE). When the null hypothesis is true, i.e. there is no effect, the F-ratio should be close to 1. These F-values are test statistics used for testing the null hypothesis of no mean differences among the levels of the factor.

The F-test for group (F_G) tests the primary hypothesis of no group effect. Denoting the mean for the i^{th} group by μ_i, the test is summarized below [53].

Null hypothesis:	H_0: $\mu_1 = \mu_2 = \ldots = \mu_g$
Alternative hypothesis:	H_a: not H_0
Test statistic:	$F = \frac{MSG}{MSE}$
Decision rule:	reject H_0 if $F > F_{N-g}^{g-1}(\alpha)$

4.4.4.7. Repeated Measure ANOVA [51, 53, 54]

Repeated measure ANOVA is used to evaluate a group effect in the case where multiple measurements are taken on the same subject. These type of data (repeated measure measurements) also called 'longitudinal data' have the particularity to be non-independent. Most of the times, the repeated response are measurements taken over time. Comparisons are made through a single F-test form a repeated measure analysis.

In general, there are g independent groups of experimental units who are subjected to repeated measurements of the same response variable, y, at t time periods. Let n_i be the number of subjects in Group i (i=1, 2, …, g), table 4.5 shows the layout for the Repeated Measure ANOVA design.

Table 4.5: Layout for a Repeated Measures Design [53]

Group	Subject	Time			
		1	2	…	T
1	1	y_{111}	y_{112}	⋯	y_{11t}
	2	y_{121}	y_{122}	⋯	y_{12t}
	…	…	…	⋯	…
	n_1	y_{1n_11}	y_{1n_12}	⋯	y_{1n_1t}
2	1	y_{211}	y_{212}	⋯	y_{21t}
	2	y_{221}	y_{222}	⋯	y_{22t}
	…	…	…	⋯	…
	n_2	y_{2n_21}	y_{2n_21}	⋯	y_{12t}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
g	1	y_{g11}	y_{g12}	⋯	y_{g1t}
	2	y_{g21}	y_{g22}	⋯	y_{g2t}
	…	…	…	⋯	…
	n_g	y_{gn_g1}	y_{gn_g2}	⋯	y_{gn_gt}

Repeated measure ANOVA measures can be handled using several analytic approaches. The 'univariate' approach using ANOVA concepts is presented here.

In the 'univariate' method, the Group, Patient, and Time effects shown in the Table 4.5 (above) are considered three factors in an ANOVA. Thus, one can examine the variability within and among these factors, keeping in mind that the Time effect constitutes correlated measurements. The response might vary among groups, among subjects within groups, and among the different measurement times. Therefore, the model includes a

Group effect, a Subject (within-Group) effect, a Time effect as sources of variation in the ANOVA and in addition, a Group-by-Time interaction. Table 4.6 contains the repeated measure ANOVA summary with $N = n_1 + n_2 + \ldots + n_g$.

Table 4.6: ANOVA Summary for Repeated-Measures Design [53]

SOURCE	**df**	**SS**	**MS**	**F**
GROUP	g – 1	SSG	MSG	F_G = MSG/MSS(G)
SUBJECT (within GROUP)	N – g	SSS(G)	MSS(G)	....
TIME	t – 1	SST	MST	F_T = MST/MSE
GROUP-by-TIME	(g-1)(t-1)	SSGT	MSGT	F_{GT} = MSGT/MSE
Error	(N-g)(t-1)	SSE	MSE	...
Total	Nt - 1	TOT(SS)		

For the balanced layout ($n_1 = n_2 = \ldots = n_g$), the sums of squares can be computed in a way similar to that used for the two-way ANOVA. Then, the mean squares (MS) are calculated, by dividing the sums of squares (SS) by the corresponding degrees of freedom.

Variation from subject-to-subject is one type of random error, as estimated by the mean square for Subject (within Group). If there is no difference among groups, the between-group variation merely reflects subject-to-subject variation. Therefore, when the null hypothesis (of no Group effect) is true, MSG and MSS (G) are independent estimates of the among-group variability. Thus, F_G has the F-distribution with g – 1 upper and N – g lower degrees of freedom.

4.4.4.8. Linear Regression [51, 53, 54]

Regression analysis is used in analyzing the relationship response variables, and quantitative factors. When there is only one response and one explanatory variable, the model is a simple linear regression. When there are one or more response variables and/or more than one explanatory variables, the model is a multiple linear regression.

4.4.4.8.1. Simple linear regression

Simple linear regression is used to find the best line that fits through a set of data points, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ drawn as a scatter plot. The objective of the analysis is to determine the significance and the strength of the linear relationship, to estimate mean responses for given predictor values, and to predict future responses. The slope of the line is representative of the relationship between the response and the predictor variable. Thus, inferences are made regarding this slope.

A linear relationship between a response, y, and an independent explanatory variable, x, can be expressed as

$$y = \alpha + \beta x + \varepsilon$$

where α is the intercept and β is the slope. The response y is subject to a random measurement error, the random error ε accounts for this random nature of the response y. This random error ε is assumed to be normally distributed with mean 0 and variance σ^2. This assumption implies that the response y is also normally distributed with mean $\alpha + \beta x$ and variance σ^2. The parameters α, β and σ^2 are unknown and simple linear regression methodologies attempt to estimate them from the observed data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

4.4.4.8.1.1. Parameter estimation

Let $((x_i, y_i)$, for i = 1, 2, …, n) be a set of n pairs where the y_i's are assumed to be independent, normally distributed with the same variance σ^2, for all x values. From these data, the model parameters α and β are estimated by $\hat{\alpha}$ and $\hat{\beta}$, respectively, in such a way that the resulting 'prediction equation'

$$\hat{y} = \hat{\alpha} + \hat{\beta}\ x$$

is the 'best-fitting' line through the measured pairs. When this is achieved, the prediction equation represents the best estimate of the unknown linear regression model. The search for the estimates $\hat{\alpha}$ and $\hat{\beta}$ is done in a way that minimizes the error (y - $\hat{y}$), which is the difference between the actual observed response and the predicted response. The most common method used to satisfy this requirement is the Least Square method.

The Least Squares methods seeks $\hat{\alpha}$ and $\hat{\beta}$ by minimizing the sum of squared differences between y and $\hat{y}$,

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

The derivation of the parameter estimates based on the Least Squares criterion is presented below:

- One of the assumption of the linear regression modeling is that the error has mean 0

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{y}_i\} = \frac{1}{n}\sum_{i=1}^{n}\{y_i - (\hat{\alpha} + \hat{\beta}x_i)\} = 0$$

- By solving the equation above, the estimate for α is obtained

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

- Substituting this value in the sum of square error expression, the SSE is

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\{y_i - (\hat{\alpha} + \hat{\beta}x_i)\}^2 = \sum_{i=1}^{n}\{y_i - (\hat{\beta}\bar{x} + \hat{\beta}x_i)\}^2$$

$$= \sum_{i=1}^{n}\{(y_i - \bar{y})^2 + 2\hat{\beta}(y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}^2(x_i - \bar{x})^2\}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\hat{\beta}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Define the quantities S_{yy}, S_{xx} and S_{xy} as follows:

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,\ S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ and } S_{xy} = \sum_{i=1}^{2}(x_i - \bar{x})(y_i - \bar{y})$$

- The expression for the SSE becomes:

$$\text{SSE} = S_{yy}^2 - 2\hat{\beta}S_{xy}^2 + \hat{\beta}^2 S_{yy}^2$$

- Differentiate this equation with respect to $\hat{\beta}$ and equate the result to 0

$$\frac{d(SSE)}{d\hat{\beta}} = -2S_{xy}^2 + 2\hat{\beta}S_{xx}^2 = 0$$

- By solving this last equation, the estimate of β is obtained:

$$\hat{\beta} = \frac{S_{xy}^2}{S_{xx}^2}$$

The 'best-fitting' line based on the Least Squares criterion is given by

$$\hat{y} = \hat{\alpha} + \hat{\beta}\,x$$

where

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

The best estimate for the variance, σ^2, is

$$s^2 = \frac{SSE}{n-2}$$

4.4.4.8.1.2. Inference on the slope β

The main question in simple linear regression analysis concerns the significance of the slope parameter. Given a number of observed values of normally distributed response variable, $(y_1, y_2, \ldots, y_n)$, the mean, $\bar{y}$, represents the best estimate of a future response. If the slope $\beta = 0$, then the value of x will not improve the prediction of y over the ordinary predicted, $\bar{y}$. A significant slope β indicates that a linear relationship exists between y and x and that the knowledge of the x-values will significantly improve the prediction ability.

Under the assumption that $\varepsilon \sim N(0, \sigma)$, the estimate $\hat{\beta} \sim N\left(\beta, \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right) = N\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right)$.

The statistical test is based on a function of the slope estimate which has the t-distribution with n-2 degrees of freedom, when the null hypothesis of 'slope=0' is true. The test summary is:

Null hypothesis:	$H_0: \beta = 0$
Alternative hypothesis:	$H_a: \beta \neq 0$
Test statistic:	$t = \frac{\hat{\beta}}{s/\sqrt{S_{xx}}}$
Decision rule:	reject H_0 if $\lvert t \rvert > t_{\alpha/2,\, n-1}$

4.4.4.8.2. Multiple linear regression

Multiple linear regression is used when there are one or more response variables and/or more than one predictor variables.

Suppose that there are n response variables $y_1, y_2, \ldots, y_n$ and p explanatory variables $x_1, x_2, \ldots, x_p$. If the relationship between the response variables and the predictor variables is assumed to be linear then the model can be expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\mathbf{X} = (I|\, x_1| \ldots |x_p) = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$ with $\mathbf{I} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$,

$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ and $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$.

A number of assumptions are made in the multiple linear regression: (1) the entries of the error vector $\boldsymbol{\varepsilon}$ are independent from each other and normally distributed with mean 0 and variance σ^2. Also the covariance of any two distinct entries is 0. (2) Consequently, the response vector contains elements that are mutually independent and each with variance σ^2.

The parameters α, $\boldsymbol{\beta}$ and σ^2 are known and they are estimated from the observed data. Just like in the simple linear regression case, inferences are made on the entries of the $\boldsymbol{\beta}$ vector which symbolize the effect of the corresponding predictor variables on the responses. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$; the regression equation is given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}.$$

4.4.4.8.2.1. Parameter estimation

To estimate **β**, the Least Squares approach, analogous to the simple linear regression case, is used. Here, the function to minimize is:

$$\text{SSE} = \sum_i \left(y_i - \alpha - \left(\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)\right)^2$$

Algebraic methodology can be applied to find **β**:

- From the regression equation,

$$\mathbf{Y = X\beta + \varepsilon}$$

- In matrix form, the sum of squares can be expressed as

$$\text{SSE} = \|\mathbf{Y - X\,\beta}\|^2 = (\mathbf{Y - X\,\beta})'(\mathbf{Y - X\beta})$$

where the matrix M' represents the transpose of the matrix M

- Perform matrix operations to obtain

$$\text{SSE} = \mathbf{Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta}$$

$$= \mathbf{Y'Y} - 2\,\mathbf{\beta'X'Y} + \mathbf{\beta'(X'X)\beta}$$

- Differentiate this expression with respect to **β** to obtain

$$\frac{\partial(\text{SSE})}{\partial\beta} = -2\,\mathbf{X'Y} + 2\,\mathbf{X'X\beta}$$

- Equate the last expression to $O_2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and solve the resulting equation to obtain $\hat{\beta}$

$$-2\,X'\hat{Y} + 2\,X'X\hat{\boldsymbol{\beta}} = \mathbf{O_2}$$

$$\mathbf{X'X\hat{\beta} = X'\hat{Y}}$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{Y}}$$

This estimate obtained minimizes the sum of squared errors.

The estimate of σ^2 is given by:

$$s = \sqrt{\frac{SSE}{n-(p+1)}}$$

4.4.4.8.2.2. Inference on the slope vector β entries

Just like in the case of simple linear regression, under the assumption that every error in the error matrix ε is normally distributed with mean 0 and variance σ^2,

$$\hat{\beta}_i \sim \mathrm{N}\left(\hat{\beta}_i, \sigma\sqrt{\frac{1}{S_{x_i x_i}\left(1-R^2_{x_i|x_1,\ldots x_{i-1}x_{i+1}\ldots x_k}\right)}}\right)$$

where $S_{x_i x_i} = \sum_j x_{ij}^2 - n\bar{x}_i^2$ and $R^2_{x_i|x_1,\ldots x_{i-1}x_{i+1}\ldots x_k}$ is the coefficient of determination for the multiple regression of x_i on all other explanatory variables.

The standard error for $\hat{\beta}_i$ is given by

$$s(\hat{\beta}_i) = s\sqrt{\frac{1}{S_{x_i x_i}\left(1-R^2_{x_i|x_1,\ldots x_{i-1}x_{i+1}\ldots x_k}\right)}}$$

With all this information, the test summary for each factor is summarized below [49]

Null hypothesis: H_0: $\beta_i = 0$

Alternative hypothesis: H_a: $\beta_i \neq 0$

Test statistic: $t = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$

Decision rule: reject H_0 if $|t| > t_{\alpha/2,\, n-(p+1)}$

4.4.5. Common inferential statistical techniques for categorical variables

4.4.5.1. Chi-square test [51, 53, 57]

The chi-square test is used to test the equality of two independent binomial proportions, p_1 and p_2. The chi-square test is an approximate test, which assumes that the normal approximation to the binomial distribution is applicable. If this assumption is violated, the alternative is the Fisher's Exact test, which is a test based on exact probabilities.

Assume that there are two independent groups (Group 1 and Group 2) of, respectively, n_1 and n_2 subjects. Suppose that there are X_1 responders in Group 1 and X_2 responders in Group 2 (Table 4.7).

Table 4.7: Layout for the Chi-Square Test [53]

	Number of Responders	**Number of Non-Responders**	**Total**
Group 1	X_1	$n_1 - X_1$	n_1
Group 2	X_2	$n_2 - X_2$	n_2
Combined	$X_1 + X_2$	$N - (X_1 + X_2)$	$N = n_1 + n_2$

The goal of the Chi-square test is to compare population 'response' rates (p_1 vs. p_2) based on these sample data. Compute

$$\text{NUMERATOR} = \frac{(X_1 * n_2 - X_2 * n_1)}{N} \text{ and DENOMINATOR} = \frac{n_1 * n_2 * (X_1 + X_2) * (N - X_1 - X_2)}{N^3}$$

Assuming that the normal approximation to the binomial distribution is applicable, the chi-square test summary is:

Null hypothesis:	H_0: $p_1 = p_2$
Alternative hypothesis:	H_a: $p_1 \neq p_2$
Test statistic:	$\chi^2 = \frac{NUMERATOR^2}{DENOMINATOR}$
Decision rule:	reject H_0 if $\chi^2 > \chi_1^2(\alpha)$

The rejection region is found by obtaining the critical chi-square value based on 1 degree of freedom, denoted as $\chi_1^2(\alpha)$, from chi-square tables.

4.4.5.2. Fisher's exact test [51, 53, 57]

Fisher's exact test is an analogue to the chi-square test for comparing two independent binomial proportions, p_1 and p_2 when the normal approximation to the binomial distribution is not applicable. This usually is the case when the cases sizes are small or in case of extreme proportions.

When the null hypothesis is true, Fisher's exact test method is based on computing exact probabilities of observing a given result or a more extreme result. The same notation as in the Chi-square test and same layout as in Table 4.7 are used here. Given equal proportions, $p_1 = p_2$, the probability of observing the configuration shown in Table 4.7, when the marginal totals are fixed, is found by the 'hypergeometric probability distribution' as

$$\text{probability} = \frac{\binom{n_1}{X_1} * \binom{n_2}{X_2}}{\binom{N}{X_1+X_2}}$$

where $\binom{a}{b} = \frac{a!}{b! * (a-b)!}$ is the combinatorial symbol that represents "the number of ways 'b' items can be simultaneously selected from a set of 'a' items" without replacement.

The p-value for the test, Fisher's exact probability, is the probability of the observed configuration (Table 4.7) plus the sum of the probabilities of all other configurations with a more extreme result for fixed row and column totals.

4.4.5.3. Logistic Regression [51, 53, 57]

The logistic regression method is used to analyze the effect of one or more factors (predictor variables) on a dichotomous response.

4.4.5.3.1. The Logit Model: One covariate

Consider a response variable, Y, taking two possible values, which can be coded as 0 and 1. Let a response of Y=1 indicate that the event of interest occurs (event), and a response of Y=0 indicate that the event does not occur (non-event). Suppose that a variable x is included in the model.

Logistic regression models apply a transformation of the response variable using the 'logit function' as follows:

$$Y^* = \text{Ln}\left(\frac{P}{1-P}\right)$$

where P is the expected value of Y for a specified set of X-values and 'Ln' represents the natural-logarithm function.

Since Y takes only two values 0 and 1, the mean of Y is the Probability that Y=1. Denote this probability by P_x. As a probability P_x, $0 \leq P_x \leq 1$. In logistic regression, it is assumed that the relationship of P_x with X is sigmoidal:

$$P_x = \frac{1}{1+e^{-(\alpha+\beta X)}}$$

With algebraic computation, this can be re-expressed as

$$\text{Ln}\left(\frac{P_x}{1-P_x}\right) = \alpha + \beta X$$

the left of which is the logistic transformation or 'logit' function (Y*). The expression $P_x/(1 - P_x)$ represents the 'odds' that Y=1, i.e., the odds that the event of interest occurs. The logit is sometimes referred to as the 'log-odds'. In logistic regression methodologies, the log-odds becomes a linear function of the covariate, X, assuming a sigmoidal relationship between X and P_x.

4.4.5.3.2. The Odds Ratio

Applying the log function on both sides of the logistic regression model, the odds for a specific covariate value of X = x is

$$\text{Odds}_x = e^{\alpha+\beta x}.$$

Replacing x by x+1, the odds for a covariate value of X = x + 1 is

$$\text{Odds}_{x+1} = e^{\alpha+\beta(x+1)}.$$

Hence, the ratio of the odds based on a 1-unit increment in X is

$$\frac{Odds_{x+1}}{Odds_x} = e^{\beta}.$$

This value is called the 'odds ratio' (OR) for the covariate X. 100(1- OR) represent the percent change in the odds of event occurrence when the covariate (X) increases by 1 unit. When X is a dichotomous variable with values 0 and 1, the OR represents the factor by which the odds of event increases level X=1 relative to X=0.

4.4.5.3.3. Model Estimation

For a fixed value of the covariate, X=x, P_x can be estimated by $\hat{p}_x = \frac{y_x}{n_x}$, where n_x is the number of observations at X = x and y_x is the number of events out of the n_x observations. Since y_x is a binomial random variable, $\hat{p}_x$ is a better estimate when n_x is large.

Most often, an estimate of P_x is obtained by using the 'Maximum likelihood' method. This method uses the data to estimate the model parameters, in a way that maximizes the likelihood of observing the data collected. Estimates obtained are called 'maximum likelihood estimates' (MLE). A model obtained with the maximum likelihood estimation can be used for prediction purposes.

The mathematical derivations are based on the maximum likelihood method and they yield a set of simultaneous equation that can be solved for the estimates of α and β. These equations, which do not have a closed solution, are of the form

$$\sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} (1 + e^{-(\hat{\alpha}+\hat{\beta}x_i)})^{-1} \text{ and } \sum_{i=1}^{N} X_i \times Y_i = \sum_{i=1}^{N} X_i \times (1 + e^{-(\hat{\alpha}+\hat{\beta}x_i)})^{-1}.$$

There are numerical techniques, such as iteratively weighted least-squares and Newton-Raphson algorithms, which are used by computer software to solve these equations.

4.4.5.3.4. The Logit Model: Multiple Covariates

In general, the logistic regression layout has N responses and k covariates, X_1, X_2, …, X_k, and a typical data set can have the following layout:

Table 4.8: Layout for Logistic Regression [53]

Response	Covariates			
Y	X_1	X_2	…	X_k
y_1	x_{11}	x_{21}	…	x_{k1}
y_2	x_{12}	x_{22}	…	x_{k2}
…	…	…	…	…
y_N	x_{1N}	x_{2N}	…	x_{kN}

The X_i's can be continuous, ordinal categorical or nominal categorical.

The model for the probability of 'event', P, becomes

$$P_x = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}$$

Thus, the logit becomes the linear function

$$\text{Ln}\left(\frac{P_x}{1-P_x}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Exponentiating both sides, the odds are obtained and can be expressed as

$$\left(\frac{P}{1-P}\right) = e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}$$

For a particular factor X_i , the odds ratio is

$$OR_{x_i} = e^{\beta_i}$$

with the interpretation that 100 (e^{β_i} – 1) represents the percent increase in the odds of 'event' occurrence when X_i increases by 1 unit and all other X's are held constant.

4.4.5.3.5 Inference on the β_i coeficients

The parameter coefficient β_i measures the effect of the covariate (X_i) on the event probability. Estimates (b_i) of these coefficients can be obtained as discussed above by the maximum likelihood methods. If the sample size is large, the estimates have approximately a normal distribution with mean β_i. If s_b represents the standard error of the estimate b_i, then b_i/s_b has an approximate standard normal distribution when the null hypothesis β_i=0 is true. Its square has the chi-square distribution with 1 degree of freedom. The test summary for each model parameter, β_i, is based on this Wald chi-square, and is summarized as follows [53]:

Null hypothesis:	H_0: $\beta_i = 0$
Alternative hypothesis:	H_a: $\beta_i \neq 0$
Test statistic:	$\chi_w^2 = (\frac{b_i}{s_b})^2$
Decision rule:	reject H_0 if $\chi_w^2 > \chi_1^2(\alpha)$

4.4.6. Common inferential statistical techniques for time-to-an event data

4.4.6.1. Log Rank test [51, 53, 58]

The log-rank test is used to compare distributions of 'time until the occurrence of an event of interest' among different independent populations (or groups). In medical research, the event is often death, but it can be any outcome, such as cure, response, relapse, failure, etc. The elapsed time from initial observation time until the occurrence of the event is called 'survival time' even when the event of interest is not 'death'.

The log-rank test does not make any assumption about the distributions of the event times. Thus, it is non-parametric method. One important factor in the log-rank test is that it adjusts for censoring. An individual is said to be censored when the event of interest does not occur during the observation period. If all the individuals experienced the event during the observation time, then data could be modeled and analyzed with the Wilcoxon-Rank Sum test. However, some subjects may drop out or experience the event before or after the study time. Event times for these individuals are estimates of the unknown event times.

The null hypothesis tested by the log-rank test is that of equal event time distributions among groups. Equality of the distributions of event times implies similar risk-adjusted event rates among groups. Rejection of the null hypothesis indicates that the event rates differ among groups at one or more time points during the study.

Without loss of generality, two independent groups are examined here. The method can easily be extended to more than two groups. Let Y be the time from initial observation to the event occurrence and let 'c' indicate a censored value. Table 4.9 represents the layout for a log-rank test with two groups

Table 4.9: Layout for a log-rank test for two groups [53]

GROUP1		**GROUP2**	
Subject Number	**Event Time**	**Subject Number**	**Event Time**
101	Y_{11}	102	Y_{21} 'c'
103	Y_{12} 'c'	105	Y_{22}
104	Y_{13}	106	Y_{23}
.	.	.	.
.	.	.	.
.	.	.	.
N_1	Y_{1N1} 'c'	N_2	Y_{2N2} 'c'

'c' indicates censored time

Suppose that the study is divided into k distinct time periods, $t_1, t_2, \ldots, t_k$, where t_j (j = 1, 2, …, k) represents the j^{th} time point when one or more patients in the combined samples experiences the event. Let d_{ij} represent the number of subjects in Group i (i = 1, 2) who first experience the event at time period t_j, and let n_{ij} represent the number of subjects in Group i who are at risk at the beginning of time period t_j. At risk represents the subjects who have not yet experienced the event and are still in the study. Let $d_j = d_{1j} + d_{2j}$ and let $n_j = n_{1j} + n_{2j}$. For j = 1, 2, …, k, compute

$$e_{ij} = \frac{n_{ij}\, d_j}{n_j} \quad \text{and} \quad v_j = \frac{n_{1j}\, n_{2j}\, d_j\, (n_j - d_j)}{n_j^2\, (n_j - 1)}$$

and compute,

$$O_1 = \sum_{j=1}^{k} d_{1j}\,, \qquad E_1 = \sum_{j=1}^{k} e_{1j} \quad \text{and} \quad V = \sum_{j=1}^{k} v_j.$$

Denote by Y_i a random variable that represents the event time for Group i (i = 1, 2), and let $S_i(t) = \text{Prob}(Y_i \geq t)$. The test summary for the log-rank test is as follows:

Null hypothesis:	H_0: $S_1(t) = S_2(t)$ (for all times, t)
Alternative hypothesis:	H_a: $S_1(t) \neq S_2(t)$ (for at least one time, t)
Test statistic:	$\chi^2 = \frac{(O_1 - E_1)^2}{V}$
Decision rule:	reject H_0 if $\chi^2 > \chi_1^2(\alpha)$

4.4.6.2. Cox Proportional Hazard [51, 53, 58]

The Cox proportional hazards model, like the log-rank test is another method used to analyze event or 'survival' times with no assumption on their distribution. The inverse of the time to event occurrence is called the 'hazard'. The hazard of some events, such as death, might be likely to increase with the passage of time. Thus, the Cox proportional hazards model adopts the a reasonable assumption that the event hazard rate changes over time, but an assumption that the ratio of event hazards between two individuals is constant, is made. This is known as the 'proportional hazards' assumption, and it stipulates that the ratio of hazards between any two values of a covariate does not vary with time.

Let $X_1, X_2, \ldots, X_k$ be k covariates in a Cox proportional hazard model on the event times. The X_i's can be continuous covariates, numerically coded ordinal responses or dummy variables. The model for the hazard function of the Cox proportional hazards method has the form

$$h(t) = \lambda(t)\, e^{(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}$$

where h(t) is the hazard function, the β_i's represent the parameter coefficients of the X_i's, and $\lambda(t)$ represents an unspecified initial hazard function.

Just like in the regression analysis, the magnitudes of the β_i's show the importance of the covariate's effect on survival times. Thus, inferences are made about these parameters. β_i's can be estimated by b_i based on a 'maximum partial likelihood', a modification of the maximum likelihood method. For large samples, these estimates (b_i) have an approximate normal distribution.

Suppose that s_b is the standard error of the estimate b_i, then b_i/s_b has an approximate standard normal distribution under the null hypothesis that $\beta_i = 0$, and its square has the chi-square distribution with 1 degree of freedom. The test summary for each model parameter, β_i, can be summarized as follows [53]:

Null hypothesis:	H_0: $\beta_i = 0$
Alternative hypothesis:	H_a: $\beta_i \neq 0$
Test statistic:	$\chi_w^2 = (\frac{b_i}{s_b})^2$
Decision rule:	reject H_0 if $\chi_w^2 > \chi_1^2(\alpha)$

The magnitude of the effect of a covariate is often expressed as a hazard ratio similar to the odds ratio. The ratio of hazards for a 1-unit increase in X_i (all other covariates held constant) is e^{β_i}.

4.5. Summary

Statistical methods and models are widely used for inference. Their efficacy and effectiveness has been proven by their extensive use in research projects. When the data to be analyzed does not follow the assumed distribution, statistical methods are limited because not all models have equivalent non-parametric options. Data mining offers data exploration methods that do not rely on data distributions.

CHAPTER 5

DATA MINING THEORY

5.1. Objective

The aim of this chapter is to provide an overview of data mining theory. Data mining was used for data grouping into clusters and for predictive modeling purposes in an objective to contrast lumpectomy to mastectomy in terms of short-term post-operative follow-up health outcomes. First, data mining is reviewed in general. Then, the data mining practical notes are presented. Finally, data mining is summarized from a theoretical standpoint, focusing on predictive modeling and cluster analysis.

5.2. Data mining overview

Data mining is '*the non-trivial* extraction of *implicit*, previously unknown, and potentially useful information from data' [15]. Data mining can also be defined as the process by which patterns are extracted and/or discovered from large amount of data [16].

Data mining is a composite science that uses principles of algorithms used in statistics, artificial intelligence, pattern recognition, and machine learning. Processes such as sampling, estimation and hypothesis testing derive from statistics while processes such as

search algorithms, modeling methods and learning theories derive from artificial intelligence, pattern recognition and machine learning [16].

Data mining tasks are mostly either predictive or descriptive in nature. Predictive tasks seek to predict the value of a particular attribute based (called target, dependent or response variable) on the values of other attributes (called predictors, independent or explanatory variables). Descriptive tasks, usually exploratory in nature, are used to derive patterns (i.e. correlations, clusters, trends) summarizing the unknown underlying relationships in the data.

5.3. Data mining practical notes

There are two types of data mining procedures: supervised learning and unsupervised learning. In supervised learning, there are variables measured that are assumed to have an influence on one or more other variables [59]. In the unsupervised learning, there is no specific output variable [59].

5.3.1. Supervised learning

In supervised learning, there is an outcome measure-target variable (quantitative or qualitative) and the goal is to predict it based on feature measurements-predictor variables. There is training set of data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ where both the outcome (x_i) and feature (y_i) measurements are observed and this set is used to build a prediction model that enables to predict new outcome.

Supervised learning presents an analogy with 'learning with a teacher'. Under this metaphor, the 'student' presents an answer $\hat{y}_i$ for each x_i in the training sample, and the

supervisor or 'teacher' provides either the correct answer and/or an error associated with the student's answer. This is usually associated with some loss function L (y, $\hat{y}$) [59].

Supervised methods include regression, nearest neighbor methods, discriminant analysis, logistic regression, kernel methods, decision trees, neural networks, etc. [60]

5.3.2. Unsupervised learning

As opposed to supervised learning, in unsupervised learning, there is no outcome measure. This is analogous to 'learning without a teacher'. In this case, there is a set of observations $(x_1, x_2, \ldots, x_N)$ and the goal is to infer the properties of the underlying distribution without the help of a supervisor or a 'teacher' providing correct answers or degrees of error for each observation.

The absence of 'supervision' is not without consequences. With supervised learning there is a clear measure of success that can be used to judge the performance of a model and to compare the effectiveness of different methods. In the case of unsupervised learning, there is no such direct measure of success. It is difficult to determine the validity of conclusions drawn from the output of most unsupervised learning methods [59].

The most common unsupervised learning techniques include association rules, cluster analysis, principal components, etc. [59]

5.3.3. Overview of commonly used data mining techniques

5.3.3.1. Predictive Modeling

Predictive modeling is one of the most commonly used techniques in data mining. It is the process of using the patterns found in the training data set to predict future outcomes.

Predictive modeling algorithms build a model for the dependent variable as a function of the independent variables. There are two types of predictive modeling tasks: classification and regression. Classification tasks are used when the target or dependent variable is qualitative. Regression tasks are used when the target variable is quantitative.

Predictive modeling problems are comprised of four things: a dependent variable, independent variables, a learning/training set, and a test data set. The learning/training data set contains values for both the dependent and independent variables, and is used to build the model. This model is then applied to the test set for evaluation. The performance of the model is based on the counts of the test records that are correctly or incorrectly predicted or classified.

There are several predictive modeling techniques. The most commonly used methods are decision trees, regression and neural networks.

Decision trees [16]: A decision tree is a predictive model in which the results are structured as a tree. A decision tree consists of a collection of decision nodes, which are connected by branches, descending from the root node until coming to an end at the leaf nodes. Each branch of the tree is a classification question and the leaves are the partitions or segments of the datasets with the classification. These decision trees are different from the decision trees used in Decision Analysis. While growing the tree, a question is asked at each branch or split point in the tree. The tree stops growing when there is either only one record in the segment, each of the records in the segment are the same, or there is not any significant gain in making a split. In order to apply a decision tree algorithm, the target variable must be discrete. There are several algorithms that are used to produce

decision trees. These include: ID3, C4.5. Classification and Regression Trees (CART), and Chi-Square Automatic Interaction Detector (CHAID).

Neural Networks [16]: Neural networks try to mimic the capabilities of the human brain. The brain can recognize patterns, make predictions, and learn. Neural networks are data mining methods with pattern recognition and machine learning algorithms to build predictive models. There are two main structures in a neural network: nodes and links. Nodes are artificial neurons and links are the connections between them. To make a prediction, the neural network accepts values for the independent variables or predictors at the input nodes. The values of these nodes are then multiplied by values stored in the links. These values are added together at the output node, after which some threshold function is used and the resulting number is the prediction. Most neural network usually have a hidden layer of nodes between the input and output nodes. They are deemed 'hidden' because their contents are not made known to the end user. It is also possible to have more than one hidden layer, thus making the network very complex.

Regression [61]: Regression analysis is a popular method used in many data mining projects for building predictive models. *Linear regression* models are used to predict a continuous response and *logistic regression* models are used to predict binary responses. *Linear regression* models define the linear relationship between a series of independent variables and a single response variable. If there is one independent variable, the linear regression is referred to as a *simple linear regression.* In this case, the model can be visualized as a straight line overlaid on a scatterplot. *Multiple linear regression* analysis involves understanding the relationship between more than one independent variable and a single response variable. *Logistic regression* models are built from one or more

independent variables that can be continuous, discrete, or a mixture of both. In addition to classifying observations into these categories, logistic regression models also calculate a probability that reflects the likelihood of a positive outcome.

5.3.3.2. Cluster analysis

Clustering [16, 59, 61] can be defined as a division of data into groups of similar objects. Instances within these groups or clusters are more similar to each other than instances belonging to other clusters. Clustering differs from predictive modeling in the fact that there is no target or dependent variable in clustering. Clustering algorithms try to segment the whole data set into homogenous clusters. The more similarity within a cluster and the bigger the difference between clusters the better the clustering. There are two main types of clustering algorithms: hierarchical and partitional.

Hierarchical Clustering: Hierarchical clustering creates a tree of clusters known as a dendrogram. In the dendrogram, the smallest clusters in the tree join together to create the next level of clusters. The top or root of this tree (dendrogram) is the cluster that contains all the records. There are two types of hierarchical clustering: agglomerative and divisive. *Agglomerative clustering* algorithms begin with each record consisting of a cluster. At this level, there are as many clusters as there are records. Then, based on some distance criteria, the clusters that are closest to one another are joined together to create the next largest cluster. This process in continued until the hierarchy is built with a single cluster, which contains all records. *Divisive clustering* algorithms work in an opposite way. These algorithms start with all of the records in one cluster and then, based on some suitable

distance choice, split it into two clusters. This process continues until some stopping criteria are met.

Depending on how similar or dissimilar items are in each cluster, the merging or splitting of the clusters occurs. The distance used to measure similarity between individual records is generalized to a robust between-cluster measure which is then used to evaluate the need for merging or splitting. This between-cluster measure is called a linkage metric. The major linkage metrics include: Single Linkage, Complete Linkage, and Average Linkage. *Single linkage* (nearest neighbor) is based on the minimum distance between any record in one cluster and any record in another cluster. Cluster similarity is based on the similarity of the most similar members from each cluster. *Complete linkage* (farthest neighbor) is based on the maximum distance of any record in one cluster and any record in another cluster. Cluster similarity is based on the similarity of the most dissimilar members from each cluster. *Average linkage* was designed to decrease the dependence of the cluster linkage criteria on extreme values. The criteria here is the average distance of all the records in one cluster from all the records in another cluster.

Partitional Clustering: Partitional clustering is dividing the data set into clusters that are exhaustive and mutually exclusive. In contrast to the hierarchical clustering, a single partition of the data is produced. Partitional clustering algorithms begin with a randomly picked or user defined number of clusters. The algorithms then optimize each cluster based on some validity measure. There are several partioning clustering approaches. The most common are K-Means and Expectation Maximization (EM).

K-Means is one of the oldest and most widely used clustering techniques. The name K-Means stands for the fact that each of the K clusters is represented by the mean point of that cluster (the centroid). Basically, the K-mean algorithm is as follows:

1. Select K points as the initial centroids
2. **Repeat**
    a. Form K clusters by assigning all points to the closest centroid
    b. Recompute the centroid of each cluster
3. **Until** the centroids do not change

First step: K initial centroids are randomly selected. Second step: each point is assigned to the closest centroid. The resulting groupings constitute clusters. Third step: for each cluster, the centroids are recomputed. Fourth step: data points are re-assigned based on the new centroids. The second, third and fourth steps are repeated until the centroids do not change. The assignments of points to centroids are made based on a proximity measure that quantifies the closeness of points. There are several types of proximity measures such as Euclidian distance and Cosine similarity.

Expectation Maximization (EM): The EM technique also starts with a random guess of the k clusters. In this case, the k clusters are represented by a set of probability distributions. The EM algorithm is a repetitive two-step process: Expectation and Maximization. The Expectation part consists of calculating cluster expected class values. The Maximization part consists of finding distribution parameter estimates that maximizes the expectation given the data. These steps are repeated until the log-likelihood converges.

5.3.3.3. Text mining

Text mining [62, 63] is a variation on the field of data mining that tries to find interesting patterns from large databases in character format. The patterns in effect provide information that can be extracted to derive summaries of the words contained in the documents, or to compute summaries for the documents based on the words contained in them. One of the main themes supporting text mining is the transformation of text into numerical data.

Usually, data sets used in data mining consists of attributes or columns and records (rows) chosen before data are collected. In the case of text mining, records (rows) are text documents and features (columns) are elements extracted from these documents. These elements can be single words or combination of words. Most commonly, these elements consists of simple words. The method used to transform text into instances counting the frequency of simple words is called the *bag of words* representation [62]. Without loss of generality, the bag of words representation will be discussed. With this representation, each document is a set of words, some occurring more than once [62]. The values stored in each feature (or column) are the number of times the element occurs in the corresponding document.

Transformation of documents into a data set [62]: In the presence of many documents, usually, a word dictionary is constructed. This dictionary contains all the words that occur at least once in every document. One way to construct the data set is to consider a column for each word. The problem with this method is that for example a dictionary may have 10,000 words and a particular document only 200 words. The representation

(or row) of this document will have 9,800 columns with 0's for the unused words. Ways to resolve this issue include the *'stop words'* and the *'stemming'* techniques. These methods help reduce the size of the word dictionary which is representative of the feature (column) space size.

Stop words: The stop words methods is an approach that removes words more likely to be useless from the word dictionary. There is no universal list of stop words; they vary by language. In English, stop words choices includes words such as: 'a', 'an', 'the', 'is', 'I', 'you', 'of', etc…

Stemming: Stemming is another method used to reduce the number of words in the word dictionary by putting together all the words that have the same linguistic root. For example, the words 'computing', 'computer', 'computation', 'computes', 'computational', 'computable 'and 'computability' can be reduced to the root (stem) 'comput'.

5.3.3.4. Association rules

Association rules analysis or discovery is a useful technique of finding important relationships in large data sets. Association rules are based on frequencies of the occurrence of items (or attributes) alone or in combination with other items (or attributes) [16, 64]. The relationships are expressed in the form: (X $\Rightarrow$ Y: support (X, Y), confidence (X, Y)) where X and Y are disjoint item sets. X (the left hand side) is called the antecedent and Y (the right hand side) is called the consequent. The confidence and support are elements that measure the strength of the association or the performance of the rule discovery [16, 64]. The association (X $\Rightarrow$ Y: support (X, Y), confidence (X, Y))

means 'If item X is part of an event, then item Y is also part of an event x percent of the times'.

The support measures how often the items X and Y occur together. The confidence assesses how many times Y appear in instances that contain X. There are two more measures of the goodness of the association rules: the expected confidence and the lift. The expected confidence of an association rule (X $\Longrightarrow$ Y: support (X, Y), confidence (X, Y)) quantifies the number of records that contain Y. The lift is the ratio of the association rule's confidence to the association rule's expected confidence. A good rule has a large confidence, a large support and a lift greater than 1.

Mathematical expressions

$$\text{Support} = \frac{records\ that\ contain\ both\ X\ and\ Y}{all\ records}$$

$$\text{Confidence} = \frac{records\ that\ contain\ both\ X\ and\ Y}{records\ that\ contain\ X}$$

$$\text{Expected confidence} = \frac{records\ that\ contain\ Y}{all\ records}$$

$$\text{Lift} = \frac{confidence}{expected\ confidence}$$

Association rules analysis is useful for finding interesting relationships that are hidden in large data sets. The goal of this type of analysis is to uncover rules (or associations) for quantifying the relationship between two or more attributes. These relationships are displayed in the form of an association rule. An association rule is an implication

expression of the form: $X \Rightarrow Y$: support (X, Y), confidence (X, Y) where X and Y are disjoint item sets.

5.4. Data mining theoretical notes

Predictive modeling and cluster analysis are discussed in mathematical terms below.

5.4.1. Predictive modeling

The definitions, objectives and main use of data mining predictive modeling have been presented in section 5.3.3.1. Below, the mathematical background of predictive modeling is presented. In general, predictive modeling tasks can all be viewed as a function approximation task [59]. First, the difference between predictive modeling and linear models is discussed.

5.4.1.1. Difference between predictive modeling and linear models

As discussed in the sections above, one of the core theories of data mining is statistical analysis. Predictive modeling methods used in data mining are very identifiable to the ones used in statistical inference [63]. The main difference of predictive modeling and statistical inference techniques resides in the fact that it is possible, in data mining to fit many different models and compare their performance on a testing set; in statistical inference, usually, a single model is fit and its performance is judged through p-values [63].

5.4.1.2. Quantitative outputs

Let $X \in \mathbb{R}^p$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable. Let P (X, Y) be their joint distribution. A predictive modeling task is to

search a function f(X) which best predicts Y given values of the input X. The errors in prediction are measured through the loss function L(Y, f(X)). The most common loss function used is the squared error loss:

$$L(Y, f(X)) = (Y - f(X))^2 \text{ [60]}$$

For a given f under this squared error loss, the expected prediction error is:

$$EPE(f) = E[L(Y, f(X))^2] = \int (y - f(x))^2 P(dx, dy) = EPE(f) = E_X E_{Y|X}([Y - f(X)]^2)$$

[60]

Predictive modeling attempts to minimize this function. It is minimized by

$$f(x) = \text{argmin}_c E_{Y|X}([Y - c]^2 | X = x) = E(Y|X = x) \text{ [60]}$$

which is a conditional expectation, also called the regression function. This solution shows that the best prediction of Y at any point X = x is the conditional mean.

- The best solution is measured by average squared error:

$$\hat{f}(x) = \text{Average}\ (y_i | x_i = x) \text{ [56]}.$$

- Using the Nearest Neighbor for classification, the solution is given by

$$\hat{f}(x) = \text{Average}\ (y_i | x_i \in N_k(x)) \text{ [60]}.$$

where $N_k(x)$ is the neighborhood containing the k points closest to x.

- Using linear regression, the regression function f(x) is approximately linear to its arguments

$$f(x) \approx x^T \beta \text{ [60]}$$

where β is estimated using least square methods.

5.4.1.3. Qualitative outputs

Typically, qualitative variables are transformed into numerical codes using coding such as dummy variables [59]. In general, when the output is qualitative, a different type of loss function is used to measure the errors in prediction. Let G denote a qualitative output and $\mathcal{G}$, the set of all possible classes. Let $\hat{G}$ be the predicted value of G. The loss function L (G, $\hat{G}$(X)) can be represented by a K × K matrix L, where K = card ($\mathcal{G}$). L contains zeros on the diagonal and nonnegative values elsewhere. A matrix input L (k, l) is the error for classifying an observation belonging to class $\mathcal{G}_k$ as $\mathcal{G}_l$. Most often, the zero-one loss function is used, where all misclassifications are charged a single unit. For a given G, the expected prediction error is

$$\text{EPE} = \text{E}\,[\text{L}\,(\text{G}, \hat{G}(\text{X}))]\ [59].$$

where the expectation is taken with respect to the joint distribution P (G, X). If it is conditioned, the EPE can be written as

$$\text{EPE} = \text{E}_\text{X} \sum_{k=1}^{K} L[\mathcal{G}_k, \hat{G}(\text{X})]\text{P}(\mathcal{G}_k|\text{X})\ [59].$$

Pointwise, the EPE is minimized by

$$\hat{G}(\text{x}) = \text{argmin}_{g\in\mathcal{G}} \sum_{k=1}^{K} L(\mathcal{G}_k, \text{g})\ \text{P}(\mathcal{G}_k|\text{X} = \text{x})\ [59].$$

Using the 0 -1 loss function, this solution simplifies to

$$\hat{G}(\text{x}) = \text{argmin}_{g\in\mathcal{G}}[1 - \text{P}(\text{g}|\text{X} = \text{x})]\ [59]$$

or simply

$$\hat{G}(X) = \mathcal{C}_k \text{ if } P(\mathcal{C}_k|X = x) = \max_g P\ (g|X = x) \text{ [59].}$$

This solution is called the Bayes classifier. In this case also, different approaches attempt to provide an optimal solution.

5.4.2. Cluster analysis

As in the case of predictive modeling in the section above, definitions, objectives and main uses of cluster analysis have been discussed in section 5.3.3.2. The center common notion of the clustering objectives is the degree of closeness (or similarity) or difference (or dissimilarity) between individual objects being clustered [59]. Thus, the mathematical theory behind the measure of similarity and dissimilarity is presented below. Also, the mathematical backgrounds of combinatorial algorithms, agglomerative and divisive algorithms are presented.

5.4.2.1. Dissimilarity of individual measurements in the same attribute

Consider measurements x_{ij} for i = 1, 2, …, N, on variables X_j, j = 1, 2, …, p (also called attributes). In most common case, dissimilarity is defined as $d_j\ (x_{ij}, x_{i'j})$ between values of the j^{th} attribute. Define

$$D\ (x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij},\ x_{i'j}) \text{ [59]}$$

as the dissimilarity between objects i and i'. There are several choices for $d_j(x_{ij},\ x_{i'j})$ but the most common choices is the squared distance

$$d_j\,(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2 \text{ [59]}$$

However, the choice depends on attribute type:

- *Quantitative variables.* Quantitative variables have measurements which are continuous real-valued numbers. In this case, the dissimilarity is measured as a distance between measurements as a monotone-increasing function of their absolute difference

$$d_j\,(x_{ij}, x_{i'j}) = l\,(|x_i - x_{i'}|). \text{ [59]}$$

  Beside the squared distance $(x_i - x_{i'})^2$, a common choice is the identity.

- *Ordinal variables.* Ordinal variables are qualitative variables which constitute an ordered set. Distance measures for ordinal variables are generally defined by replacing their M original values with

$$\frac{i-1/2}{M}, i = 1, 2, \ldots, M \text{ [59]}$$

  in the prescribed order of their original values. They are then treated as quantitative variables on this scale.

- *Categorical variables.* With unordered categorical variables, the degree-of-difference between pairs of values must be delineated explicitly. If the variable assumes M distinct values, these can be arranged in a symmetric M×M matrix

with elements $L_{rr'} = L_{r'r}, L_{rr} = 0, L_{rr'} \geq 0$. The most common choice is $L_{rr'} = 1$.

5.4.2.2. Dissimilarity of different attributes in an object

Most often, a single overall measure of dissimilarity D $(x_i, x_{i'})$ is done by computing a weighted average of the p-individual attribute dissimilarities d_j $(x_{ij}, x_{i'j})$, j= 1, 2, …, p:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} w_j \times d_j(x_{ij}, x_{i'j}); \quad \sum_{j=1}^{p} w_j = 1 \text{ [59]}$$

Here w_j is a weight assigned to the j^{th} attribute.

5.4.2.3. Clustering algorithms

5.4.2.3.1. Combinatorial algorithms

Combinatorial algorithms work directly on the observed data with no direct reference to an underlying probability model. Each observation is uniquely labeled by an integer i ∈ {1, …, N}. A prespecified number of clusters K < N is postulated, and each one is labeled by an integer k ∈ {1,…, K}. Each observation is assigned to one and only one cluster. These assignments can be characterized by a many-to-one mapping, or encoder k = C* (i), that assigns the i^{th} observation to the k^{th} cluster. One seeks the particular encoder C* (i) that achieves the required goal, based on the dissimilarities d $(x_i, x_{i'})$ between every pair of observations. These are specified by the user. Generally, the encoder C (i) is explicitly delineated by giving its value (cluster assignment) for each observation i. The "parameters" of the procedure are the individual cluster assignments for each of the N observations. These are adjusted so

as to minimize a "loss" function that characterizes the degree to which the clustering goal is not met.

One approach is to directly specify a mathematical loss function and attempt to minimize it through some combinatorial optimization algorithm. Since the goal is to assign close point to the same cluster, a natural loss function would be

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'}) \text{ [59]}$$

This criterion characterizes the extent to which observations assigned to the same cluster tend to be close to one another. It is sometimes referred to as the "within cluster" point scatter since

$$T = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{ii'} = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \left(\sum_{C(i')=k} d_{ii'} + \sum_{C(i')\neq k} d_{ii'} \right) \text{ [59]}$$

or

$$T = W(C) + B(C) \text{ [59]},$$

where $d_{ii'} = d(x_i, x_{i'})$. Here T is the total point scatter, which is constant given the data, independent of cluster assignment. The quantity

$$B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d_{ii'} \text{ [59]}$$

is the between cluster point scatter. This will tend to be large when observations assigned to different clusters are far apart. Thus one has

$$W(C) = T - B(C) \text{ [59]}$$

and minimizing W(C) is equivalent to maximizing B(C).

Cluster analysis by combinatorial optimization is simple and straightforward. One simply minimizes W (which is equivalent to maximizing B) over all possible assignments of the N data points to K clusters. Since most clustering problems involve very large data sets, such optimization by complete enumeration is practically difficult. More practical combinatorial clustering algorithm are able to examine only a very small fraction of all possible encoders k = C (i). The goal is to identify a small subset that is likely to contain the optimal one, or at least a good suboptimal partition.

Such feasible strategies are based on iterative greedy descent. An initial partition is specified. At each iterative step, the cluster assignments are changed in such a way that the value of the criterion is improved from its previous value. Clustering algorithms of this type differ in their prescriptions for modifying the cluster assignments at each iteration. When the prescription is unable to provide an improvement, the algorithm terminates with the current assignment as its solution. Since the assignment of observations to clusters at any iteration is a perturbation of that of the previous iteration, only a very small fraction of all possible assignments are examined.

K-means: The K-means algorithm is one of the most popular iterative descent cluster method. It is used in cases where all variables are quantitative, and the dissimilarity measure is the squared Euclidian distance

$$d(x_i, x_{i'}) = \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2 \text{ [59].}$$

The within-point scatter can be written as

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i)\neq k}||x_i - x_{i'}||^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k}||x_i - \bar{x}_k||^2 \text{ [59]},$$

where $\bar{x}_k = (\bar{x}_{1k}, \ldots, \bar{x}_{pk})$ is the mean vector associated with the k^{th} cluster, and $N_k = \sum_{k=1}^{K} I(C(i) = k)$. Thus, the criterion is minimized by assigning the N observations to the K clusters in such a way that within each cluster the average dissimilarity of the observations from the cluster mean, as defined by the points in that cluster, is minimized.

An iterative descent algorithm for solving

$$C^* = \min_C \sum_{k=1}^{K} N_k \sum_{C(i)=k}\|x_i - \bar{x}_k\|^2 \text{ [59]}$$

can be obtained by noting that for any set of observations S

$$\bar{x}_S = \operatorname{argmin}_m \sum_{C(i)=k}\|x_i - m\|^2 \text{ [59]}.$$

Hence we obtain C* by solving the enlarged optimization problem

$$\min_{C,\{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k}\|x_i - m_k\|^2 \text{ [59]}.$$

This can be minimized by an alternating optimization procedure below.

Step1: For a given cluster assignment C, the total cluster variance is minimized with respect to $\{m_1, \ldots, m_K\}$ yielding the means of the current assigned clusters.

Step 2: Given a current set of means$\{m_1, \ldots, m_K\}$, the enlarged optimization problem above is minimized by assigning each observation to the closest (current) cluster mean. That is

$$C(i) = \operatorname{argmin}_m \sum_{C(i)=k}\|x_i - m\|^2.$$

Step 3: Steps 1 and 2 are iterated until the assignments do not change.

5.4.2.3.2. Hierarchical algorithms

Hierarchical clustering [55] produce hierarchical representations in which clusters at each level of the hierarchy are created by merging clusters at the next lower level. At the lowest level, each cluster contains a single observation. At the highest level there is only one cluster containing all the data points.

Agglomerative clustering (bottom-up) [55]: Agglomerative clustering algorithms begin with every observation representing a cluster with one element. At each of the N-1 steps the closest two clusters are merged into a single cluster, producing one less cluster at the next higher level. Let G and H represent two such groups. The dissimilarity d (G, H) between G and H is computed form the set of pairwise observation dissimilarities $d_{ii'}$ where one member of the pair i is in G and other i' is in H. *Single linkage (SL)* agglomerative clustering (also called the *nearest neighbor* technique) takes the intergroup dissimilarity to be that of the closest (least dissimilar pair)

$$d_{SL}(G,H) = \min_{i\in G, i'\in H} d_{ii'}$$

Complete linkage (CL) agglomerative clustering (*furthest-neighbor* technique) takes the intergroup dissimilarity to be that of the furthest (most dissimilar) pair

$$d_{CL}(G,H) = \min_{i\in G, i'\in H} d_{ii'}$$

Group average (GA) clustering uses the average dissimilarity between the groups

$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

Where N_G and N_H are the respective number of observations in each group.

Divisive clustering (top-down) [55]: Divisive clustering algorithms begin with the entire data set as a single cluster, and recursively divide one of the existing clusters into two daughter clusters at each iteration in a top-down fashion.

The divisive paradigm can be employed by recursively applying any of the combinatorial methods to perform the splits at each iteration. One divisive algorithm was proposed by Macnaughton Smith et al. in 1965. It begins by placing all observations in a single cluster G. it then chooses that observation whose average dissimilarity from all the other observations is largest. This observation forms the first member of a second cluster H. At each successive step that observation in G whose average distance from those in H, minus that for the remaining observations in G is largest, is transferred to H. This continues until there are no longer any observations in G that are, on average, closer to those in H. the result is a split of the original cluster into two daughter clusters. The observations transferred to H, and those remaining in G. these two clusters represent the second level of the hierarchy. Each successive level is produced by applying this split procedure to one of the clusters at the previous level. Kaufman and Rousseeuw (1990) suggest choosing the cluster at each level with the largest diameter

$$D_G = \max_{i \in G, i' \in G} d_{ii'} \text{ [59]}$$

for splitting. An alternative would be to choose the one with the largest average dissimilarity among its members

$$\bar{d}_G = \frac{1}{N_G}\sum_{i\in G}\sum_{i'\in G} d_{ii'} \text{ [59]}$$

This recursive splitting continues until all clusters either become singletons or all members of each one have zero dissimilarity from one another.

5.5. Summary

One of the bases of data mining is statistics. However, data mining methods for data analysis go further than statistical methods because they mostly do not assume any underlying distribution. Also, although some techniques are exploratory, they have the power to generate and validate hypotheses simultaneously. This quality makes them very promising. Data mining methods have been used and validated in business. Their use in healthcare research is still evolving and is yet to be validated in the literature.

CHAPTER 6

DECISION ANALYIS AND COST EFFECTIVENESS ANALYSIS THEORY

6.1. Objective

The aim of this chapter is to present the theory of decision analysis and cost-effectiveness analysis. Decision analysis was used to compare lumpectomy to mastectomy in terms of long-term (10-years) comparative effectiveness. Cost-effectiveness analysis was used to analyze the incremental cost per satisfied patient after lumpectomy versus mastectomy. First, the decision analysis theory is reviewed. Then the cost-effectiveness theory is presented.

6.2. Decision analysis

6.2.1. Decision analysis overview

Decision analysis can be defined as an organized quantitative method for measuring the relative value of different decision options. Decision analysis results are aimed to provide information on which strategy has the best 'outcome' of interest. Historically, it was used as an approach to help physicians make decisions on how to treat individual patients. Nowadays, it is increasingly used to help policy makers in decisions about the management of groups [17].

6.2.2. Decision analysis practical notes

Decision Analysis (DA) is performed in five steps: (DA 1) Identifying and bounding the problem, (DA 2) Structuring of the problem and construction of the decision tree, (DA 3) Gathering the information to fill out the decision tree, (DA 4) Analysis of the decision tree using probability and estimation methods and (DA 5) Sensitivity analysis [17].

6.2.2.1. Identifying and bounding the Problem

In this step, the first component is to state the alternative strategies to be compared. Then, the events that follow the different alternative methods are identified. The last component is to define the outcome [17].

6.2.2.2. Structuring the Problem

Structuring the problem mainly involves the construction of the decision tree. The decision tree is a graphical representation of the different alternatives, their subsequent consequences and the resulting outcomes [17]. In general, decision trees are made of nodes (decision nodes and chance nodes), branches and outcomes. Decision nodes identify points where there are alternative actions that are under the control of the decision maker. In the simplest problem, the decision node describes the problem. Chance nodes identify points where events that are not in the control of the decision maker may occur [17]. There are conventions that guide the construction of the decision tree. First, decision trees are built from left to right. Second, when time is involved, earlier events or choices are represented first (put on the left) left and later ones are presented next (on the right) [17]. Third, decision nodes are drawn as squares, chance nodes as circles and outcomes as large triangles. Fourth, branches are drawn at right

angles to nodes; they connect nodes with nodes and nodes with outcome [17]. Fifth, chance nodes for the same events should line up horizontally. Probabilities are associated with the events that are mutually exclusive and jointly exhaustive [17]. Figure 6.1 illustrates a hypothetical decision tree.

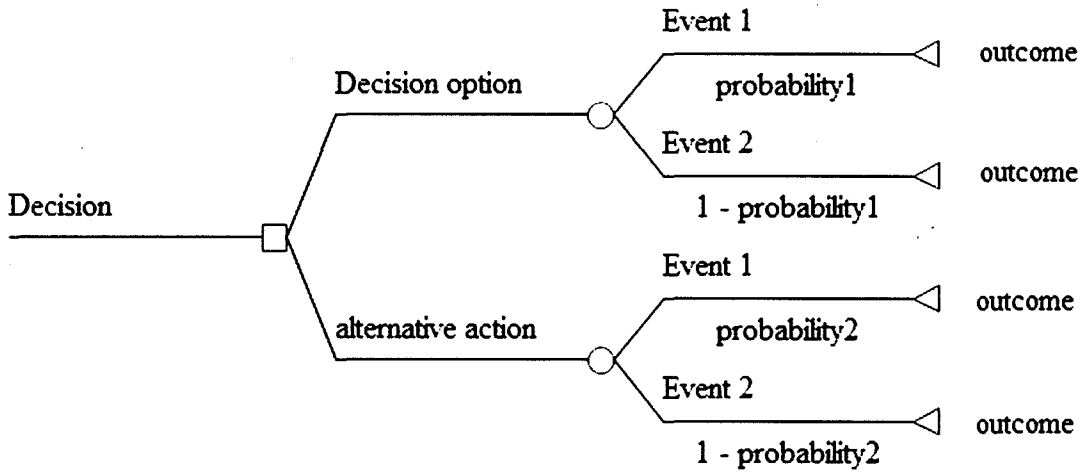

Figure 6.1: A hypothetical simple decision tree [17]

In the decision tree, outcomes are the consequences of the final events depicted in the tree. An outcome can be life or death; disability or health; or any variety of other risks or benefits of the strategy [17]. An outcome may also be the extension in life or the quality of life. In most current decision analysis studies, the outcome measures are life expectancy or quality adjusted life expectancy. Estimation of quality adjusted life expectancy involves the measurement of utilities. A utility is a measure of preference for the outcome (condition) to society or to an individual [17].

6.2.2.3. Gathering information to Fill in the Decision Tree

Information to fill the decision tree can be gathered from a literature review (including a meta-analysis), primary data collection, consultation with experts or all of the above [17].

6.2.2.4. Analyzing the Decision Tree

The decision tree is analyzed by a process called folding back and averaging. This method results in an estimate of the probability of the expected outcome of each alternative. There are a number of computer software available to perform decision analysis (such as TreeAge, etc.). However, the computations necessary to analyze simple decision trees are simple arithmetic operations [17].

6.2.2.5. Sensitivity analysis

A sensitivity analysis is conducted in order to measure the stability of the conclusion with respect to the assumptions made. In a sensitivity analysis, probabilities on which assumptions were made are varied. [17]

6.2.3. Decision Analysis Theoretical Notes

The task of the analyst is to present the systems along with their estimated effectiveness to the decision-maker [65]. From a mathematical point of view, effectiveness can be viewed as either a random variable or not.

6.2.3.1. Decision analysis when effectiveness in a non-random variable

If effectiveness is a non-random variable, its values for the different alternatives can be known with certainty in advance of acquiring the strategy. In this case, the comparison is straightforward.

6.2.3.2. Decision analysis when effectiveness is a random variable

If effectiveness is a random variable, its values for each strategy are not known but the probability distribution of different levels of effectiveness can be estimated.

Assume two methods A and B are compared. Let a (e) and b (e) be the probability that A and B respectively, if acquired would have a level of effectiveness e. Let the probability that the methods' level of effectiveness is ultimately e* or greater be A (e*) and B (e*). Then,

A (e*) = P (e $\geq$ e*) = $\int_{e*}^{\infty} a\ (e)\ de$ and B (e*) = P (e $\geq$ e*) = $\int_{e*}^{\infty} b\ (e)\ de$ for every possible e* [65].

Thus, the decision-maker can select B in preference over A if B (e*) > A (e*) or if:

$\int_{e*}^{\infty} b\ (e)\ de > \int_{e*}^{\infty} a\ (e)\ de$. [65]

6.3. Cost Effectiveness Analysis

6.3.1. Cost Effectiveness Analysis overview

Cost Effectiveness Analysis is a full economic evaluation that compares decision options in term of their monetary cost per unit of effectiveness [17, 66]. In this type of analysis, both the cost and the consequences of the alternatives are examined [66]. In health economics, cost effectiveness analysis is mostly used for allocating limited funds but it can also be used in decision making by groups or individuals [20].

6.3.2. Cost Effectiveness Analysis Practical notes

A Cost Effectiveness Analysis (CEA) study consists of seven steps [18]: (CEA 1) stating the problem, (CEA 2) describing the conceptual model, (CEA 3) defining the perspective, (CEA 4) identifying contributors to cost and gathering data to value costs, (CEA 5) identifying outcomes and gathering data to value outcomes, (CEA 6) estimating cost effectiveness, and finally (CEA 7) performing a sensitivity analysis.

Cost effectiveness analysis is closely related to decision analysis. Decision analysis is concerned with comparing outcomes in terms of their effectiveness. Cost effectiveness analysis goes beyond just effectiveness to include the cost. Decision analysis is somewhat factored into cost effectiveness analysis and some steps of analysis are similar. In fact, (CEA 1) is identical to (DA1), (CEA 2) to (DA 2), (CEA 5) to (DA 3), and (CEA 7) to (DA 5). Thus, only steps (CEA 3), (CEA4) and (CEA 6) are presented below.

6.3.2.1. Defining the perspective

Costs and outcomes included in a cost effectiveness analysis may differ considering which angle or point of view is taken for analysis. For example, the cost of a hospitalization is different for a insurance provider and for a hospital. For the insurance, it is the amount of money that the insurance pays. For a hospital, it is the sum of money paid to the caregivers, money to run the hospital and other direct and indirect charges [17]. These points of views are called perspectives. In a cost effectiveness analysis, the perspective must be stated explicitly. Popular perspectives used are the societal and the program. The societal perspective includes all costs, and the program's perspective includes only the cost to implement the program.

6.3.2.2. Identifying contributors to cost and gathering data to value costs

The next step is to identify contributors to the cost. These contributors depend on the perspective. Contributors to costs include direct and indirect costs [17]. Direct costs are costs disbursed such as cost of treatment, cost paid to administer the treatment, etc. Indirect costs comprise cost of travel for patients, costs of lost wages for patients, etc. After the contributors to cost have been identified, data on these costs are gathered. Cost data can be obtained from primary data collection (micro-cost) or from the medical literature or in an electronic data set (gross cost) [17, 67].

6.3.2.3. Estimating cost effectiveness

The cost effectiveness of an action relative to its alternative is the ratio of the net cost to the net effectiveness [17]. The net cost is the difference between costs of different alternatives. The net benefit is the difference between effectiveness of different alternatives. Consider two alternatives 1 and 2, the incremental cost effectiveness ratio (ICER) is of the form

$$\text{ICER} = \frac{cost\ (alternative\ 2) - cost\ (alternative\ 1)}{effectiveness\ (alternative\ 2) - effectiveness\ (alternative\ 1)}$$

6.3.3. Cost Effectiveness Analysis Theoretical notes

The need of a cost effectiveness analysis usually arises when there is no fixed cost constraint and no fixed effectiveness requirement. In fact, if the cost is fixed, then all the alternatives are eliminated but the one that yields the greatest effectiveness [65]. If the effectiveness is fixed, the question would be a cost-minimization one. Most often, there is an acceptable range of cost and an acceptable range of effectiveness [65]. In mathematical terms, cost and effectiveness can be random or non-random variables.

6.3.3.1. Cost effectiveness analysis in the case where cost and effectiveness are both non-random variables

If cost and effectiveness are non-random variables, then their values are known with certainty for the different alternatives. Suppose that an alternative B is compared to an alternative A; three cases are possible: (1) B costs more and is less effective than A. In this case B is said to be dominated and it is rejected. (2) B costs less and is more effective. In this case A is dominated and B is adopted. (3) B costs more and is more effective. In this case, the ICER is computed and, based on the value obtained, the decision maker has to determine whether the additional effectiveness is worth the additional cost.

6.3.3.2. Cost effectiveness analysis in the case where cost is a non-random variable and effectiveness is a random variable

Assume that the cost is known with certainty and that effectiveness is a random variable. This case is similar to the case of decision analysis when effectiveness is a random variable.

6.3.3.3. Cost effectiveness analysis in the case where cost is a random variable and effectiveness is a non-random variable

Assume that effectiveness is known with certainty and cost is a random variable for which the probability distribution can be estimated. The solution here is analogous to the case where effectiveness is a random variable and cost is known with certainty. Let a (c) and b (c) be the probability density function of cost for the alternatives A and B respectively. B is chosen if, for every c*,

$$\int_0^{c*} b\,(c)\,dc \;<\; \int_0^{c*} a\,(c)dc.\ [65]$$

6.3.3.4. Cost effectiveness analysis in the case where both cost and effectiveness are random variables

In the case where both cost and effectiveness are random variables, a joint probability distribution associated with both cost and effectiveness can be derived for each alternative. Let a (e, c), b (e, c) represent the joint distribution of effectiveness and cost for alternatives A and B, respectively. Then B is chosen if, for every e* and c*,

$$\int_{e*}^{\infty}\int_0^{c*} a\,(e,c)dc\,de \;\leq\; \int_{e*}^{\infty}\int_0^{c*} b\,(e,c)dc\,de.\ [65]$$

6.4. Deterministic model versus Markov model

The model presented so far is called a deterministic model. This model is represented by simple decision trees such as the one in Figure 6.1. In this type of model, time is not a component of the alternatives.

When time is a factor in the alternative, then it is more appropriate to use a Markov model. In a Markov model, there is a recursive component that repeats over time. In this case, individuals can shift from one state of the recursive component to the other with a certain probability. A Markov model is also called a state transition model [67]. In this type of model, the system can be represented by two types of matrices: the state matrix and the transition probability matrix. The transition probability matrix, noted M, is a square matrix of order n where n is the number of different states [68, 69]. An entry m_{ij} of this matrix represents the probability of moving from state i to state j in one step. The

state matrix, noted A is a 1×n matrix containing probabilities of being in each of the n states.

6.5. Summary

Decision analysis is used in presenting the relative effectiveness of one alternative in comparison to another. Cost effectiveness presents the comparison in terms of monetary value. Although decision analysis can be performed alone, it is usually a first step to cost-effectiveness analysis.

CHAPTER 7

USE OF STATISTICAL METHODS TO COMPARE SHORT TERM IN-HOSPITAL OUTCOMES FOR LUMPECTOMY AND MASTECTOMY

7.1. Objective

The objective of this chapter was to use the classical statistical methods to compare lumpectomy to mastectomy. Data used are administrative; thus, the first task is to use a good algorithm of patient selection and the second task is to clean and manipulate the data to fit the assumptions of the statistical models. The difference of the real world data, such as the one used in this section, and clinical trial data, is that the clinical trial data are already nicely selected and cleaned from the inclusion-exclusion criteria. Here, SAS is used to select cases of mastectomy and lumpectomy then data are prepared and finally statistical methods are applied to the data. Lumpectomy is compared to Mastectomy in terms of clinical outcomes such as length of stay and hospital charges and health outcomes in terms of in-hospital death.

7.2. Data –Nationwide Inpatient Sample (NIS) Database

This study used the Nationwide Inpatient Sample (NIS) for the year 2005 [12]. The NIS was described in section 1.3.4.

7.3. Data pre-processing

7.3.1. Case selection

The NIS database contains up to 15 diagnoses (DX1 – DX15) and 15 procedures (PR1 – PR15) for each discharge record. The coding uses the International Classification of Disease, 9th edition, Clinical Modification (ICD-9-CM) diagnosis and procedure codes. ICD-9-CM translation is publically available at www.icd9cm.chrisendres.com.

A cohort of hospital stays with any malignant neoplasm of the breast was generated with the extraction of observation with any of the following ICD-9-CM codes: 174.0, 174.1, 174.2, 174.3, 174.3, 174.4, 174.5, 175.6 and 174.8 (Table 7.1). Among patients with a diagnosis of breast cancer, only those with a surgical treatment procedure were retained for the analysis. Mastectomy was recognized by the presence of codes 85.41, 85.42, 85.43, 85.44, 85.45, 85.46, 85.47, and 85.48 in at least one of the procedures and, for Lumpectomy, procedure code 85.21 was used. Cases with missing age or race were excluded.

Table 7.1: ICD-9-CM codes for breast cancer

Code type	ICD-9-CM codes	Description
Diagnosis	174 174.0 174.1 174.2 174.3 174.4 174.5 174.6 174.8 174.9	Malignant neoplasm of the female breast Malignant neoplasm of the female breast nipple and areola Malignant neoplasm of the female breast central portion Malignant neoplasm of the female breast upper-inner quadrant Malignant neoplasm of the female breast lower-inner quadrant Malignant neoplasm of the female breast upper-outer quadrant Malignant neoplasm of the female breast lower-outer

Code type	ICD-9-CM codes	Description
		quadrant Malignant neoplasm of the female breast lower-outer quadrant Malignant neoplasm of the female breast axillary tail Malignant neoplasm of the female breast, other specified sites
Procedure		
Mastectomy	85.4	Mastectomy
	85.41	Unilateral simple mastectomy
	85.42	Bilateral simple mastectomy
	85.43	Unilateral extended simple mastectomy
	85.44	Bilateral extended simple mastectomy
	85.45	Unilateral radical mastectomy
	85.46	Bilateral radical mastectomy
	85.47	Unilateral extended mastectomy
	85.48	Bilateral extended mastectomy
Lumpectomy	85.21	Local excision of lesion of breast

Cases of mastectomy and lumpectomy were selected using the following SAS code:

Code 7.1: selection of mastectomy and lumpectomy cases

```
/*extract breast cancer patients*/
DATA INPATBC;
      SET NIS2005.NIS_2005_CORE;
      DX1=TRIM(DX1); DX2=TRIM(DX2); DX3=TRIM(DX3);
            DX4=TRIM(DX4); DX5=TRIM(DX5); DX6=TRIM(DX6);
            DX7=TRIM(DX7); DX8=TRIM(DX8); DX9=TRIM(DX9);
            DX10=TRIM(DX10); DX11=TRIM(DX11); DX12=TRIM(DX12);
            DX13=TRIM(DX13); DX14=TRIM(DX14); DX15=TRIM(DX15);
      ARRAY DXX {15} $ DX1-DX15;
      DO I=1 TO 15;
            IF DXX[I]='1740' OR DXX[I]='1741' OR DXX[I]='1742' OR
DXX[I]='1743' OR DXX[I]='1744'
                  OR DXX[I]='1745' OR DXX[I]='1746' OR
DXX[I]='1748' THEN BC=1;
      END;
      OUTPUT;
RUN;
DATA INPATBC;
      SET INPATBC;
      DXCLUSTER=CATX(' ', OF DX1-DX15);
      IF BC=1;
```

```
RUN;

/*get the groups*/
DATA INPATBC2OLD;
      SET INPATBC;
      PR1=TRIM(PR1); PR2=TRIM(PR2); PR3=TRIM(PR3); PR4=TRIM(PR4);
           PR5=TRIM(PR5); PR6=TRIM(PR6); PR7=TRIM(PR7);
PR8=TRIM(PR8);
           PR9=TRIM(PR9); PR10=TRIM(PR10); PR11=TRIM(PR11);
PR12=TRIM(PR12);
           PR13=TRIM(PR13); PR14=TRIM(PR14); PR15=TRIM(PR15);
      ARRAY PRR {15} $ PR1-PR15;
      MAST=0; LUMP=0;
      DO I=1 TO 15;
           IF (PRR[I]='8541' OR PRR[I]='8542' OR PRR[I]='8543'
                     OR PRR[I]='8544' OR PRR[I]='8545' OR
PRR[I]='8546')
                                THEN MAST=1;
           IF (PRR[I] = '8521') THEN LUMP=1;
      END;
      OUTPUT;
RUN;

DATA INPATBC2OLD;
      SET INPATBC2OLD;
      IF MAST=1 AND LUMP=0 THEN GROUP=1;
      IF MAST=0 AND LUMP=1 THEN GROUP=2;
      IF MAST=1 AND LUMP=1 THEN GROUP=3;
      IF MAST=0 AND LUMP=0 THEN GROUP=0;
      DXCLUSTER=CATX(' ', OF DX1-DX15);
RUN;

DATA INPATBC2;
      SET INPATBC2OLD;
      IF GROUP NOT IN ('0' '3');
RUN;
```

From this code, 8333 cases of mastectomy and 892 cases of lumpectomy were obtained. Some had missing values on some variables. It was decided to exclude cases with missing age or race using the following code:

Code 7.2: Elimination of cases with missing age or race

```
DATA SURGERY;
      SET INPATBC2;
      IF RACE GE 3 THEN RACEGP=3; ELSE RACEGP=RACE;
      IF AGE LT 40 THEN AGEGP=1;
           ELSE IF 40 LT AGE LE 60 THEN AGEGP=2;
```

```
            ELSE AGEGP=3;
      IF AGE NE . AND RACE NE .;
RUN;
```

With the use of this code, the size of the mastectomy group was reduced to 673 cases. The sizes were very different and to address this issue, a random sample of the mastectomy group of the size of the lumpectomy group was chosen. The code below shows the process.

Code 7.3: Selection of a random sample from the mastectomy group with the size of the lumpectomy group.

```
/*get the equal sizes*/
DATA LUMPECTOMY;
      SET SURGERY;
      IF GROUP=2;
RUN;

DATA MASTECTOMY;
      SET SURGERY;
      IF GROUP=1;
RUN;

PROC SURVEYSELECT DATA=MASTECTOMY N=673
      METHOD=SRS SEED=1234 OUT=SAMPLEMAST;
RUN;

DATA ANALYSISDATA;
      SET SAMPLEMAST LUMPECTOMY;
RUN;
```

The resulting set was used for analysis. It contained 673 cases of mastectomy and 673 cases of lumpectomy. Next, the variable preparation is discussed.

7.3.2. Outcome variables

The main outcome variables were in-hospital length of stay (LOS) at the time of the procedure and hospital total charges (TOTCHG). In-hospital death (DIED) was also

analyzed. All these variables are recorded in the data. LOS and TOTCHG are continuous and DIED is categorical.

7.3.3. Input variables

Demographics: Demographics factored in the analysis were the patient's age, gender and race which are present in the data.

Charlson index: the Charlson index is a measure of the burden of comorbidities [70]. Deyo's modification of the Charlson index for administrative data was used [71]. Deyo developed an algorithm to compute the charlson index in data where diagnoses are recorded with ICD-9-codes. Table 7.2 provides the translation of the charlson comorbidity index components [70] into ICD-9-CM codes used by Deyo et al. [71]

Table 7.2: Translation of charlson comorbidity index component into ICD-9-CM codes from Deyo's paper [71]

Diagnostic category	**ICD-9-CM codes**	**Assigned weight**
Myocardial infarction	410-410.9, 412	1
Congestive heart failure	428-428.9	1
Peripheral vascular disease	430-438	1
Cerebrovascular disease	430-438	1
Dementia	290-290.9	1
Chronic pulmonary disease	490-496, 500-505, 506.4	1
Rheumatologic disease	710.0, 710.1, 710.4, 714.0-714.2, 714.81, 725	1
Peptic ulcer disease	531-534.9, 531.4-531.7, 532.4-532.7, 533.4-533.7, 534.4-534.7	1

Diagnostic category	ICD-9-CM codes	Assigned weight
Mild liver disease	571.2, 571.5, 571.6, 571.4-571.49	1
Diabetes	250-250.3, 250.7	1
Diabetes with chronic complications	250.4-250.6	1
Hemiplegia or paraplegia	344.1, 342-342.9	2
Renal disease	582-582.9, 583-583.9, 585, 586, 588-588.9	2
Any malignancy, including leukemia and lymphoma	140-172.9, 174-195.8, 200-208.9	2
Moderate or severe liver disease	572.2-572.8, 456.0-456.21	3
Metastatic solid tumor	196-199.1	6
AIDS	042-044.9	6

The following code was used to compute the Charlson index using Deyo's adaptation to the ICD -9-CM code:

Code 7.4: Code to compute the Charlson index using Deyo's adaptation

```
DATA ANALYSISDATA;
      LENGTH ICD9CODE $5.;
      LENGTH INDEX 3;
      IF _N_=1 THEN DO;
            DECLARE HASH H(DATASET:"BEA.DEYO_CHARLSON_INDEX",
ORDERED:"NO");
            H.DEFINEKEY ("ICD9CODE");
            H.DEFINEDATA ("INDEX", "ICD9CODE");
            H.DEFINEDONE();
                  CALL MISSING(ICD9CODE, INDEX);
      END;

      SET ANALYSISDATA;
      ARRAY DXX {14} $ DX2-DX15;
      CHARLSON=0;
      DO I=1 TO 14;
            IF H.FIND(KEY: DXX[I])=0 THEN CHARLSON=CHARLSON+INDEX;
```

```
      END;
      OUTPUT;
RUN;

DATA ANALYSISDATA;
      SET ANALYSISDATA;
      ARRAY PRR {15} $ PR1-PR15;
      DO I=1 TO 15;
            IF PRR[I]='3848' THEN CHARLSON=CHARLSON+1;
      END;
      OUTPUT;
RUN;

DATA ANALYSISDATA;
      SET ANALYSISDATA;
      IF CHARLSON GE 3 THEN CHARLSON1=3; ELSE CHARLSON1=CHARLSON;
RUN;
```

7.4. Statistical methods

Categorical variables were tabulated using the frequencies and continuous variables were visualized using kernel density estimation. Comparison of the two surgical procedures with respect to in-hospital length of stay and hospital charges in the short term analysis was studied with ANOVA models on log-transformed variables. These variables were log-transformed to approach the normal distribution assumed by the ANOVA models. The risk of in-hospital death was analyzed with univariate logistic regression. The SAS codes used are presented below.

Code 7.5: Frequency

```
PROC FREQ DATA=ANALYSISDATA;
      TABLES CHARLSON1*GROUP AGEGP*GROUP RACEGP*GROUP DIED*GROUP;
RUN;
```

Code 7.6: Kernel Density Estimation

```
/*overall distributions*/
ODS GRAPHICS ON;
```

```
PROC KDE DATA=ANALYSISDATA;
     UNIVAR LOS/GRIDL=0 GRIDU=20 OUT=ALLLOS;
     UNIVAR TOTCHG/GRIDL=0 GRIDU=100000 OUT=ALLTOTCHG;
RUN;
ODS GRAPHICS OFF;

/*distributions per group*/
PROC SORT DATA=ANALYSISDATA;
     BY GROUP;
ODS GRAPHICS ON;
PROC KDE DATA=ANALYSISDATA;
     UNIVAR LOS/GRIDL=0 GRIDU=20 OUT=LOS;
     UNIVAR TOTCHG/GRIDL=0 GRIDU=100000 OUT=TOTCHG;
     BY GROUP;
RUN;
ODS GRAPHICS OFF;
```

Code 7.7: ANOVA model

```
/*log transformation of the variables*/
DATA ANALYSISDATA;
     SET ANALYSISDATA;
     LOS1=LOG(LOS);
     TOTCHG1=LOG(TOTCHG);
RUN;

/*ANOVA*/
PROC GLM DATA=ANALYSISDATA;
     CLASS GROUP AGEGP RACEGP;
     MODEL LOS1=GROUP AGEGP RACEGP CHARLSON1;
     MEANS GROUP;
     CONTRAST 'GP2 vs. GP1' GROUP 1 -1 0;
     CONTRAST 'GP3 vs. GP1' GROUP 1 0 -1;
     CONTRAST 'GP3 vs. GP2' GROUP 0 1 -1;
RUN;

PROC GLM DATA=ANALYSISDATA;
     CLASS GROUP AGEGP RACEGP;
     MODEL TOTCHG1=GROUP AGEGP RACEGP CHARLSON1;
     MEANS GROUP;
     CONTRAST 'GP2 vs. GP1' GROUP 1 -1 0;
     CONTRAST 'GP3 vs. GP1' GROUP 1 0 -1;
     CONTRAST 'GP3 vs. GP2' GROUP 0 1 -1;
RUN;
```

Code 7.8: Logistic Regression

```
PROC LOGISTIC DATA=ANALYSISDATA DESCENDING;
     CLASS GROUP /REF=FIRST;
     MODEL DIED=GROUP ;
     ODDSRATIO GROUP;
RUN;
```

7.5. Results

7.5.1. Data description

NIS 2005 contains discharge information on 7,968,569 patients among whom 4,692,644 (58.89%) are women. Of these women, 15,437 were diagnosed with breast cancer, and 9403 (60.91%) among the breast cancer discharges were associated with a surgical procedure for breast cancer treatment. From this group, 178 observations were associated with procedures that involved both mastectomy and lumpectomy; they were excluded from the analysis, yielding an analysis set of 9225 hospital discharges. Only 11 (0.12%) deaths were recorded from all of these procedures.

After elimination of cases with missing age and race, there were 673 observations with lumpectomy and 6327 with mastectomy. To overcome potential problems that may result from the differences in sample sizes, a random sample of 673 discharges with mastectomy was selected to use in the comparison.

The average age was 63 (see Table 7.3) and all the records were of patients 40 years and older. The white population made up 77.19% and the black population represented 11.00% of all the records. More discharges were of patients with a Charlson index of 0 (42.35%); however, a considerable proportion had a Charlson index of 3 or more (41.31%).

Table 7.3: NIS 2005 Data description for the short term analysis

Variable	all sample (n=1346)	Mastectomy (n=673)	Lumpectomy (n=673)
Age [mean(std)]	63 (14)	62 (14)	64 (14)
Age [n (%)]			
<40	0 (0.00)	0 (0.00)	0 (0.00)
40 – 60	518 (38.48)	269 (39.97)	249 (37.00)
>60	828 (61.52)	404 (60.03)	424 (63.00)
Race [n (%)]			
White	1039 (77.19)	522 (77.56)	517 (76.82)
Black	148 (11.00)	68 (10.10)	80 (11.89)
Other	159 (11.81)	83 (12.33)	76 (11.29)
Charlson [n(%)]			
0	570 (42.35)	312 (46.36)	258 (38.34)
1	124 (9.21)	63 (9.36)	61 (9.06)
2	96 (41.31)	28 (4.16)	68 (10.10)
≥3	556 (41.31)	270 (40.12)	286 (42.50)
Died	6 (0.45)	2 (0.30)	4 (0.59)

7.5.2. Outcomes variable description

The main outcome variables for the short-term analysis were in-hospital death, length of stay and total charges.

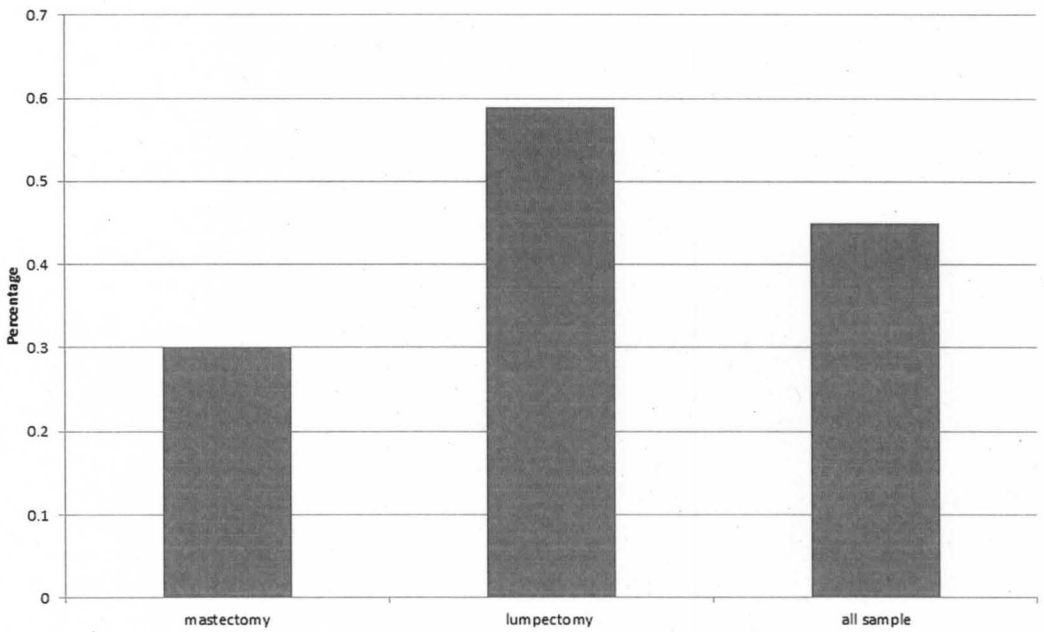

Figure 7.1: In-hospital death distribution

The in-hospital death following a surgical procedure for breast cancer treatment was in general very low. Percentages were higher in the lumpectomy group.

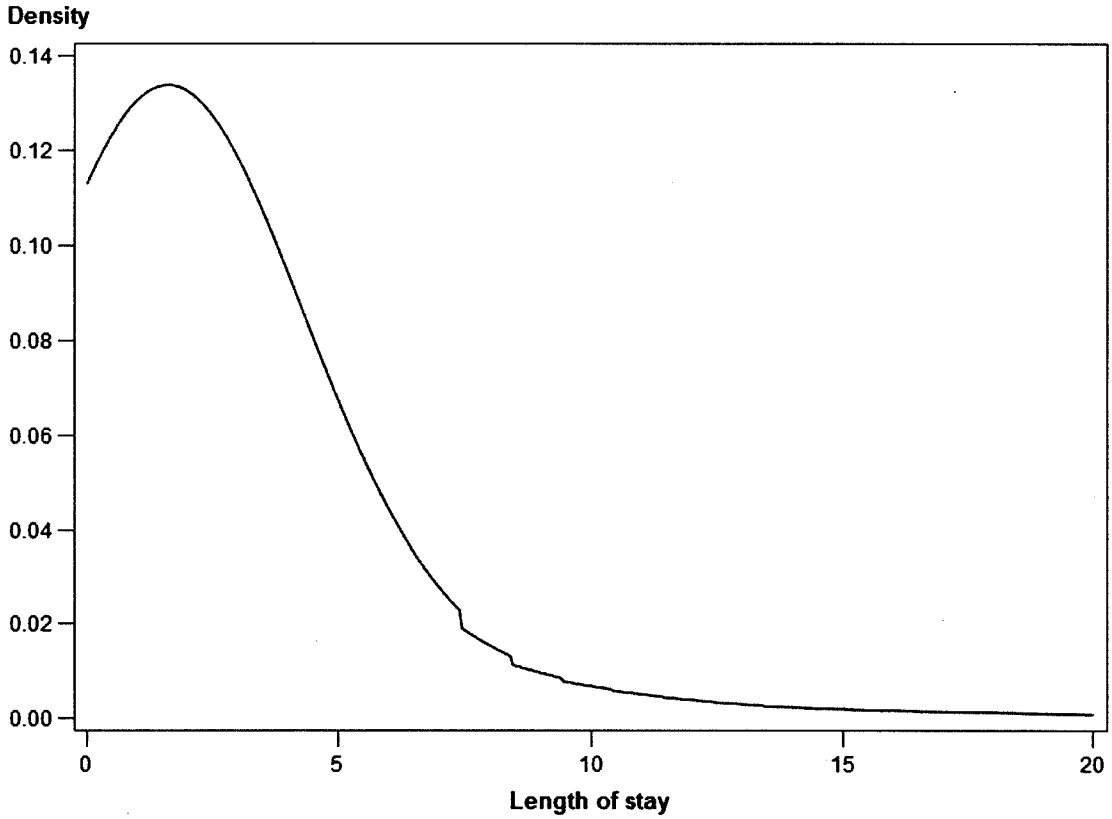

Figure 7.2: Kernel density estimation for the in-hospital length of stay for all observations

The average length of stay during hospitalization was less than 5 days for most of the observations.

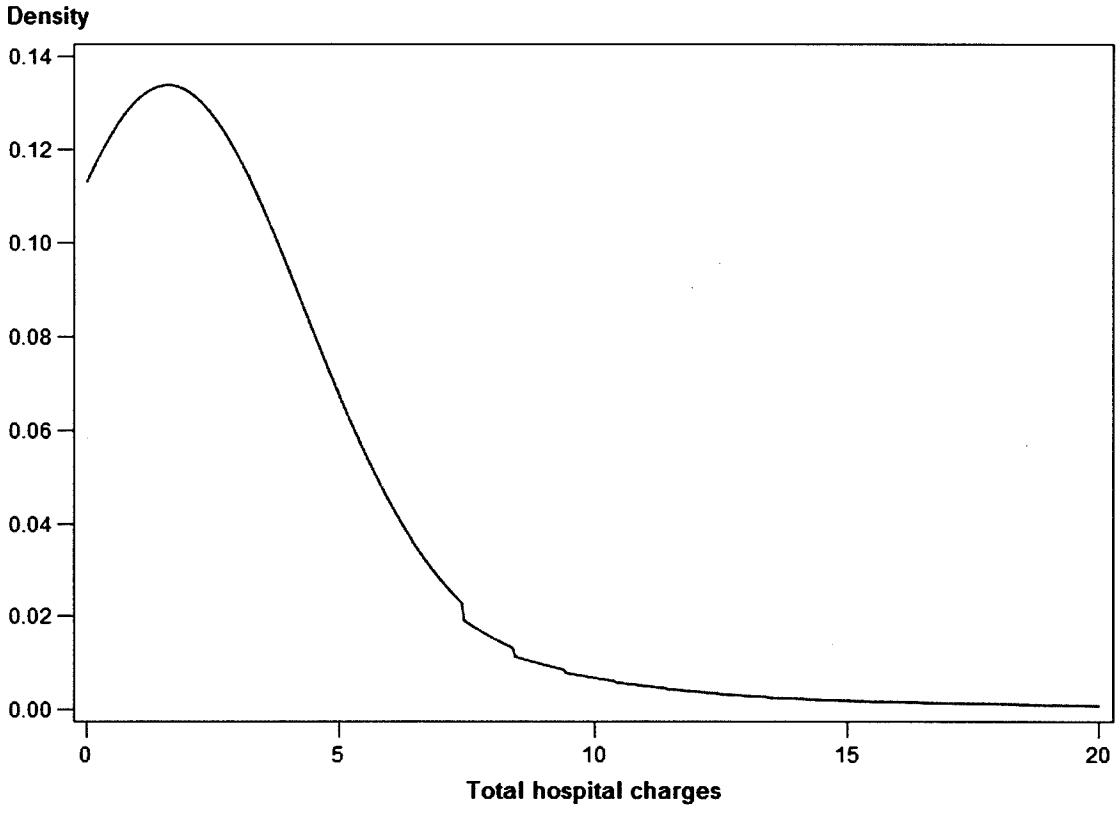

Figure 7.3: Kernel density estimation for the hospital total charges for all the observations

Most of the discharges were associated with hospital charges that are less than $20,000.

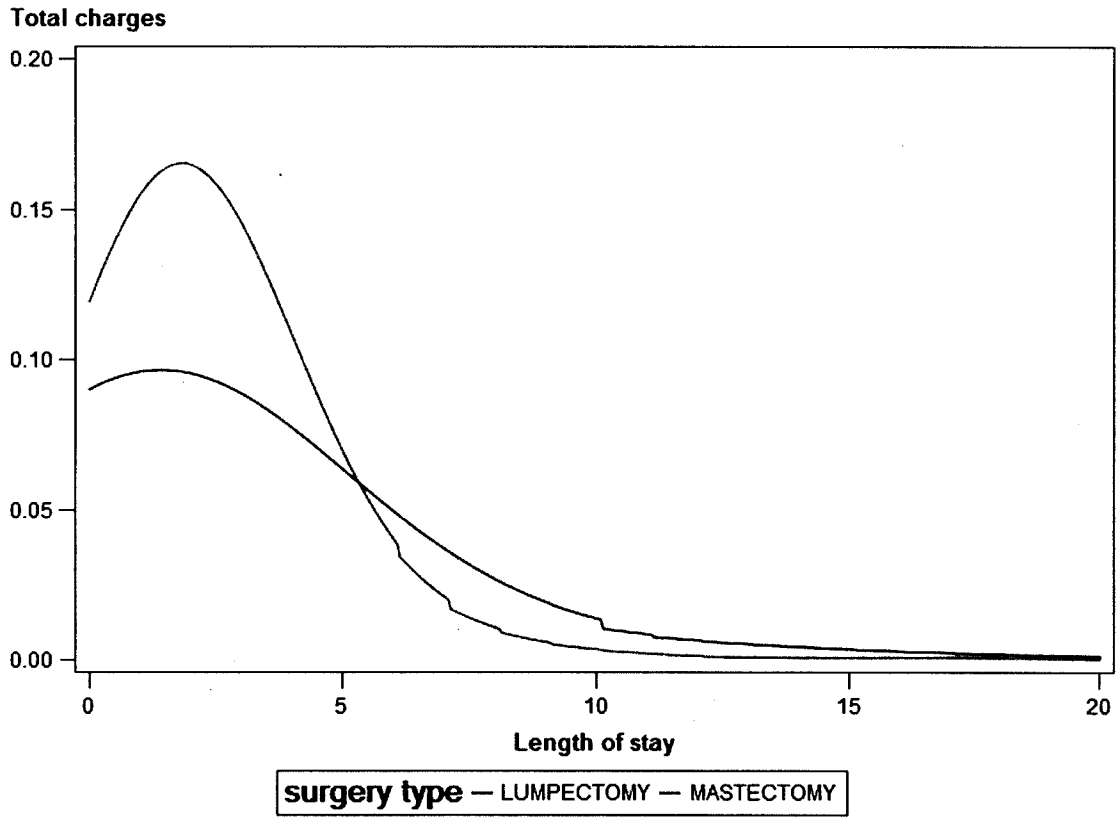

Figure 7.4: Comparative in-hospital length of stay distributions in the two groups

More discharges in the mastectomy group were associated with a shorter length of stay in comparison to the lumpectomy group. A high probability of a longer stay was observed in the lumpectomy group.

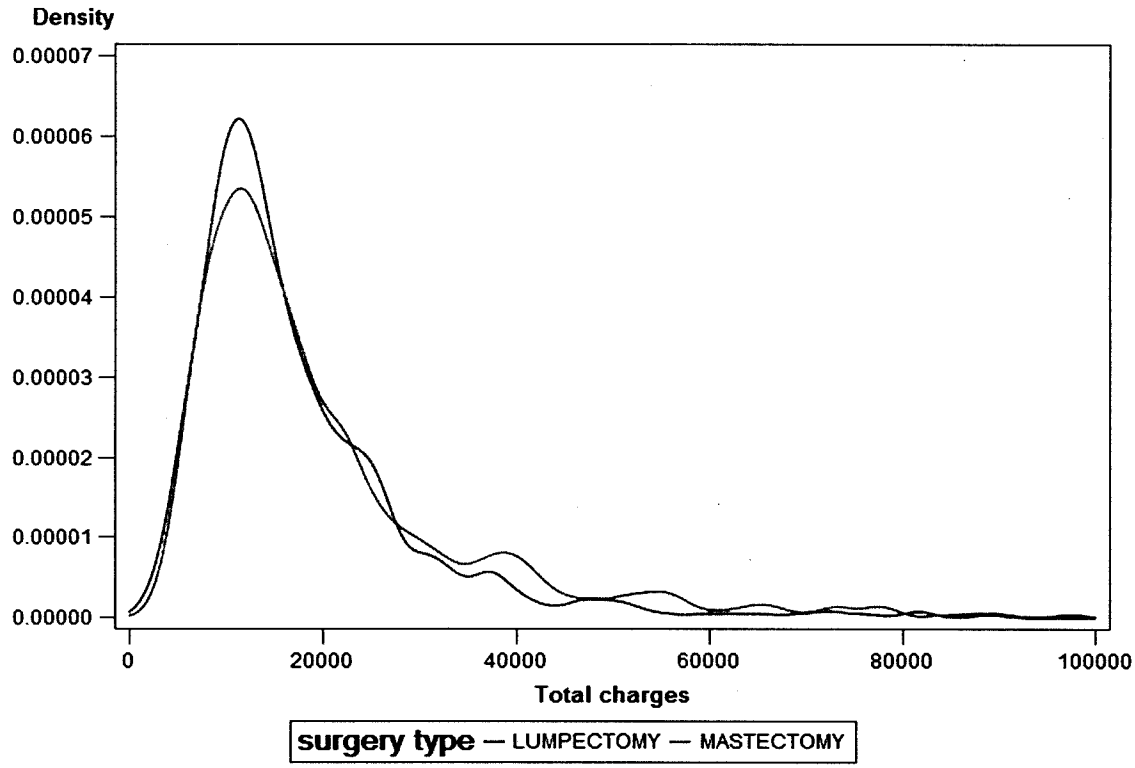

Figure 7.5: Comparative hospital total charges distributions in the two groups

The total charges were found to be comparable in the two groups. However, lumpectomy had a slightly higher probability of lower charge.

7.5.3. Inferential statistics: Group effect comparisons

In this part of the study, differences observed in the descriptive statistics above, were tested for their statistical significance using ANOVA models and Chi-square tests. Tests were performed at a 0.05 significance level. The Table 7.4 below contains a summary of the statistical results.

Table 7.4: Short-term two group comparison

Outcome variable	Mastectomy (n=673)	Lumpectomy (n=673)	p-value

Outcome variable	Mastectomy (n=673)	Lumpectomy (n=673)	p-value
Length of stay [mean (std)]	2.42 (3.28)	2.74 (5.87)	<0.0001
Total charges [mean (std)]	20796 (16950)	21159 (33364)	0.02
In-hospital death [OR (95% CI)]	Ref	0.5 (0.09, 2.73)	0.42

OR: odds ratio, CI: confidence interval

Comparison of the in-hospital length of stay: patients who had a lumpectomy had a longer stay in comparison to those who had a mastectomy. This difference was found to be statistically significant with a p-value less than 0.0001 (see Table 7.4). *Comparison of the hospital total charges:* the average total charges were above $20,000 in both groups. The lowest charges were observed in the mastectomy group. It was found that the type of procedure had a significant effect on the hospital charges (see Table 7.4); lumpectomy and mastectomy had significantly different total hospital charges (p-value=0.02).

Comparison of the in-hospital death proportions: The probability of in-hospital death was very low in all three groups. The risk of in-hospital death was about 50% lower in the lumpectomy group than in the mastectomy group, but this reduced risk was not statistically significant (OR = 0.5, CI: 0.09 to 2.73, p-value: 0.42, see Table 7.4).

7.6. Summary

In this chapter, statistical methods were used to analyze both health and clinical outcomes after surgery. Two procedure groups were compared: lumpectomy and mastectomy. In terms of outcomes it was found that patients in the lumpectomy group had a significantly longer stay and more hospital charges.

CHAPTER 8

USE OF DATA MINING, STATISTICAL METHODS AND COST EFFECTIVENESS TECHNIQUES TO COMPARE SHORT TERM POST-OPERATIVE FOLLOW-UP OUTCOMES FOR LUMPECTOMY AND MASTECTOMY

8.1. Objective

In this chapter, classical statistical models are used to compare lumpectomy to mastectomy in terms of clinical outcomes (healthcare resources use and cost). In addition, data mining methods were used to enhance statistical methods in the analysis of the longitudinal data. First, clusters of diagnoses are used as a factor in statistical comparison models. Then, data mining predictive models are contrasted to find and build a simple 90-day post-operative hospital re-admission. Finally cost effectiveness techniques are used to evaluate the short-term costs.

8.2. Data –MarketScan Database [13]

Additional data used are from the Reuter's MarketScan database records of 2000 and 2001. The MarketScan data was described in section 1.3.4.

8.3. Data pre-processing

8.3.1. Patient selection

First, cases of breast cancer were extracted from the MarketScan records of 2000 using ICD-9-CM diagnosis codes described in Table 7.1. Next, observations with a procedure code of mastectomy or lumpectomy were retained. Procedures were queried using ICD-9-CM procedure codes (see table 7.1) as well as the Current Procedural Terminology, 4th edition (CPT-4) codes [72] (Table 8.1).

Table 8.1: CPT-4 codes for mastectomy and lumpectomy

Procedure	CPT-4 codes	Description
Mastectomy	19303	Simple, complete mastectomy
	19304	Subcutaneous mastectomy
	19305	Radical mastectomy
	19307	Modified radical mastectomy
Lumpectomy	19301	Partial mastectomy (lumpectomy)

For this purpose of patient selection, the following SAS codes were used.

Code 8.1: Selection of cases of breast cancer and of mastectomy and lumpectomy

```
/*isolate breast cancer cases*/
DATA DATA;
      SET BEA.CCAEI001 BEA.CCAEI012;
      DX1=TRIM(DX1); DX2=TRIM(DX2); DX3=TRIM(DX3); DX4=TRIM(DX4);
      DX5=TRIM(DX5); DX6=TRIM(DX6); DX7=TRIM(DX7); DX8=TRIM(DX8);
      DX9=TRIM(DX9); DX10=TRIM(DX10); DX11=TRIM(DX11);
      DX12=TRIM(DX12); DX13=TRIM(DX13); DX14=TRIM(DX14);
      DX15=TRIM(DX15);
      ARRAY DXX {15} $ DX1-DX15;
      BC=0;
      DO I=1 TO 15;
            IF DXX[I] IN ('1740' '1741' '1742' '1743' '1744'
            '1745'
            '1746' '1748' '1749') THEN BC=1;
```

```
      END;
      IF DSTATUS IN ('20' '21' '22' '23' '24' '25' '26' '27' '28'
      '29') THEN DIED=1; ELSE DIED=0;
RUN;

DATA BREASTCANCER;
      SET DATA;
      IF BC=1;
RUN;

/*procedure groups*/
DATA BREASTCANCER;
      SET BREASTCANCER;
      PROC1=TRIM(PROC1); PROC2=TRIM(PROC2); PROC3=TRIM(PROC3);
      PROC4=TRIM(PROC4); PROC5=TRIM(PROC5); PROC6=TRIM(PROC6);
      PROC7=TRIM(PROC7); PROC8=TRIM(PROC8); PROC9=TRIM(PROC9);
      PROC10=TRIM(PROC10); PROC11=TRIM(PROC11);
      PROC12=TRIM(PROC12); PROC13=TRIM(PROC13);
      PROC14=TRIM(PROC14); PROC15=TRIM(PROC15);
      ARRAY PR {15} $ PROC1-PROC15;
      MAST=0;
      LUMP=0;
      DO I=1 TO 15;
      IF PR[I] IN ('8541' '8542' '8543' '8544' '8544' '8546'
      '8547' '8548' '19303' '19304' '19305' '19307') THEN MAST=1;
      IF PR[I] IN ('8521' '19301') THEN LUMP=1;
      END;

      IF MAST=1 AND LUMP=0 THEN GROUP=1;
      IF MAST=0 AND LUMP=1 THEN GROUP=2;
      IF MAST=1 AND LUMP=1 THEN GROUP=3;
      IF MAST=0 AND LUMP=0 THEN GROUP=0;
RUN;

DATA BCSURGERY;
      SET BREASTCANCER;
      IF GROUP NE 0;
RUN;
```

8.3.2. Selection of post-operative data

The selection algorithm described in the section above queried all the hospitalizations in which a lumpectomy or a mastectomy was performed. For each patient retained, the first

occurrence was considered the initial procedure hospitalization and the first observation of the follow up.

Using the ENROLID, all records of 2000 and 2001 were then extracted from the inpatient admission, inpatient service, outpatient service, outpatient pharmaceutical claims and enrollment tables. Below, the codes used for extraction are presented.

Code 8.2: Separation of pre- from post-operative inpatient records

```
DATA INPATIENT;
      SET INPATIENT;
      BY ENROLID;
      RETAIN SURGERY;
      IF FIRST.ENROLID THEN SURGERY=0;
      IF MAST=1 OR LUMP=1 THEN SURGERY+1;
RUN;

/*post-surgery data*/
DATA POSTOP;
      SET INPATIENT;
      IF SURGERY NE 0;
RUN;

/*surgery hospitalization*/
DATA OPDAY;
      SET POSTOP;
      BY ENROLID;
      IF FIRST.ENROLID AND GROUP=0 THEN DELETE;
      IF FIRST.ENROLID;
RUN;
```

Code 8.3: Outpatient and medication sets

```
/*inpatient data*/
DATA OUTPATIENT;
      SET CCAEO00A CCAEO00B CCAEO00C CCAEO00D CCAEO00E CCAEO00F
            CCAEO01A CCAEO01B CCAEO01C CCAEO01D CCAEO01D CCAEO01F
                  CCAEO01G CCAEO01H;
      FORMAT DATE MMDDYY10.;
      DATE=SVCDATE;
RUN;
/*medication data*/
DATA MEDICATION;
      SET CCAED00A CCAED00B CCAED00C CCAED00D
            CCAED01A CCAED01B CCAED01C CCAED01D CCAED01E CCAED01F;
      FORMAT DATE MMDDYY10.;
```

```
      DATE=SVCDATE;
RUN;
```

Code 8.4: Post-surgery outpatient records

```
/*data with date of operation date*/
DATA OPDATE;
      SET OPDAY (KEEP=ENROLID GROUP ADMDATE);
      FORMAT DATE MMDDYY10.;
      DATE=ADMDATE;
RUN;

/*outpatient data*/
PROC SORT DATA=OPDAY;
      BY ENROLID;
PROC SORT DATA=BEA.OUTPATIENT;
      BY ENROLID;
DATA OUTPATBCSURGERY;
      MERGE OPDATE (IN=INA) OUTPATIENT (KEEP=ENROLID DATE SVCDATE
PAY);
      BY ENROLID;
      IF INA;
      DIFF = DATDIF (ADMDATE,DATE,'ACT/ACT');
RUN;

PROC SORT DATA=OUTPATBCSURGERY;
      BY ENROLID DATE;
RUN;

DATA OUTPATBCSURGERY;
      SET OUTPATBCSURGERY;
      BY ENROLID;
      RETAIN OP;
      IF FIRST.ENROLID THEN DO;
            OP=0;
      END;
      IF DIFF GE 0 THEN OP=1;
RUN;

DATA OUTPATPOSTOP;
      SET OUTPATBCSURGERY;
      IF OP NE 0;
RUN;

DATA OUTPATPOSTOP;
      SET OUTPATBCSURGERY;
      BY ENROLID;
      IF FIRST.ENROLID THEN NVIS=0;
      NVIS+1;
RUN;
```

Code 8.5: Post-surgery medication records

```
/*medication data*/
PROC SORT DATA=OPDAY;
     BY ENROLID;
PROC SORT DATA=MEDICATION;
     BY ENROLID;
DATA MEDBCSURGERY;
     MERGE OPDATE (IN=INA) MEDICATION (KEEP=ENROLID DATE SVCDATE
PAY);
     BY ENROLID;
     IF INA;
     DIFF = DATDIF(ADMDATE,DATE,'ACT/ACT');
RUN;

PROC SORT DATA=MEDBCSURGERY;
     BY ENROLID DATE;
RUN;

DATA MEDBCSURGERY;
     SET MEDBCSURGERY;
     BY ENROLID;
     RETAIN OP;
     IF FIRST.ENROLID THEN DO;
          OP=0;
     END;
     IF DIFF GE 0 THEN OP=1;
RUN;

DATA MEDPOSTOP;
     SET MEDBCSURGERY;
     IF OP NE 0;
RUN;

DATA MEDPOSTOP;
     SET MEDPOSTOP;
     BY ENROLID;
     IF FIRST.ENROLID THEN NMED=0;
     NMED+1;
RUN;
```

Code 8.5: Computation of post-operative follow-up time

```
/*with 2000 enrollment records*/
DATA ENROL00;
     SET CCAET00A(KEEP=ENROLID DTEND) CCAET00B (KEEP=ENROLID
DTEND);
     BY ENROLID;
     IF LAST.ENROLID;
```

```
RUN;

DATA F00;
      MERGE OPDAY ENROL00;
      BY ENROLID;
      FY00=YRDIF(ADMDATE,DTEND,'ACT/ACT');
      FD00=DATDIF(ADMDATE,DTEND, 'ACT/ACT');
RUN;

/*with 2006 enrollment records*/
DATA ENROL01;
      SET CCAET01A (KEEP=ENROLID DTEND) CCAET01B(KEEP=ENROLID
DTEND) BEA.CCAET01C(KEEP=ENROLID DTEND);
      BY ENROLID;
      IF LAST.ENROLID;
RUN;

DATA F01;
      MERGE OPDAY ENROL01;
      BY ENROLID;
      FY01 = YRDIF (ADMDATE,DTEND,'ACT/ACT');
      FD01 = DATDIF (ADMDATE,DTEND,'ACT/ACT');

      FORMAT ENDDATE MMDDYY10.;
      ENDDATE = '31DEC2001'D;

      FY01END = YRDIF(ADMDATE,ENDDATE,'ACT/ACT');
      FD01END = DATDIF(ADMDATE,ENDDATE,'ACT/ACT');
RUN;

/*post-surgery enrollment time*/
DATA POSTOPFTIME;
      MERGE OPDAY(KEEP=ENROLID IN=INA) F00(DROP=DTEND)
F01(DROP=DTEND) ;
      BY ENROLID;
      IF INA;
      FYTIME=MAX(FY00,FY01);
      FDTIME=MAX(FD00,FD01);
      IF FYTIME LT 0 THEN FYTIME=FY01END;
      IF FDTIME LT 0 THEN FDTIME=FD01END;
RUN;
```

8.3.3. Outcome variables

Healthcare resources use: The healthcare resources considered were the post-operative length of stay, hospital re-admissions, outpatient services and prescribed medications

(new and refills). These clinical outcomes were compared between patients who underwent the lumpectomy and those who underwent the mastectomy initially.

Healthcare resource cost: The costs per encounter for all healthcare resource use were compared.

Post-operative hospital re-admission: the proportions of patient re-admitted at least once were compared in both surgical groups.

Re-operation: Re-operation rates were evaluated for patients in both groups of the breast cancer surgical treatment.

8.3.4. Input variables

Age: the age at time of initial surgical procedure

Charlson index: the Charlson index of the patient at time of the initial procedure. Here also, Deyo's adaptation of the Charlson index to ICD-9-CM codes was used (refer to Table 7.2 and Code 7.4).

Disease cluster: the diagnoses of the initial surgery for all the patients were grouped into clusters. Each patient was assigned to a disease cluster that symbolized the medical condition burden. Disease clusters were obtained using SAS Enterprise Miner 6.2. Text Mining and Cluster analysis were performed in the data mining interface of SAS, SAS Enterprise Miner (EM) 6.2 on the diagnoses. First, all the diagnoses for each patient were concatenated in one variable and transformed into a string (Code 8.6). Then, this string of diagnoses was set to be considered as a text document in EM. The Text Miner node in EM (Figure 8.1) was used to discover number patterns and cluster them using the

Expectation Maximization cluster algorithm. The maximum number of cluster to be formed was set to five.

Code 8.6: Use of the catx function in SAS to concatenate all the diagnoses into one string

```
DATA SASUSER.CLUSTERS;
     SET OPDAY;
     DX=CATX(' ', OF DX1-DX15);
     KEEP ENROLID DX;
RUN;
```

Figure 8.1: Diagram flow of the Text Miner node in SAS Enterprise Miner

Region: the region in which the initial procedure took place was considered as an input for analysis. In the MarketScan data, five categories are given to the region variable: (1) Northeast, (2) North Central, (3) South, (4) West and (5) Unknown.

8.4. Use of statistical methods and cluster analysis to analyze clinical outcomes

8.4.1. Summary of statistical methods

Descriptive statistics and data visualization were performed using codes similar to Codes 7.5 and 7.6. Comparison of the two surgical procedures with respect to in-hospital length of stay and hospital charges for the initial procedure was studied with ANOVA models

on log-transformed variables. Longitudinal data (post-operative length of stay and hospital charges per hospitalization) were compared with Repeated measure ANOVA on log-transformed variables. The risk of re-operation and hospital re-admissions was analyzed with univariate logistic regression. The time to re-operation and time to re-hospitalization were compared using the log-rank test. The number of re-operation and the number of re-hospitalizations were compared using the Kruskal-Wallis test. All the statistical analyses were performed in SAS [72] using SAS codes [73] in the interface SAS Enterprise Guide (EG) 4.3. To perform ANOVA and univariate logistic regression, codes similar to Codes 7.7 and 7.8 were used. Below, SAS codes for the Kruskal-Wallis test and the Repeated Measure ANOVA codes are presented.

Code 8.7: the Kruskal-Wallis Test

```
PROC NPAR1WAY WILCOXON DATA=ANALYSISDATA1;
      CLASS GROUP;
      VAR NVIS;
RUN;
```

Code 8.8: Repeated Measure ANOVA

```
DATA OUTPATPOSTOP1;
      MERGE ANALYSISDATA1(KEEP=ENROLID GROUP IN=INA)
            OUTPATPOSTOP;
      BY ENROLID;
RUN;
/*log-transformation of the variables*/
DATA OUTPATPOSTOP1;
      SET OUTPATPOSTOP1;
      PAY1=LOG(PAY);
RUN;
/*ANOVA*/
PROC GLM DATA=OUTPATPOSTOP1;
      CLASS GROUP ENROLID NVIS;
      MODEL PAY1=GROUP ENROLID(GROUP) NVIS GROUP*NVIS/SS3;
      RANDOM ENROLID(GROUP);
      TEST H=GROUP E=ENROLID(GROUP);
      MEANS GROUP;
```

```
RUN;

/*medication charges*/
DATA MEDPOSTOP1;
      MERGE ANALYSISDATA1(KEEP=ENROLID GROUP) MEDPOSTOP;
      BY ENROLID;
RUN;
/*log-transformation of the variables*/
DATA MEDPOSTOP1;
      SET MEDPOSTOP1;
      PAY1=LOG(PAY);
RUN;
/*ANOVA*/
PROC GLM DATA=MEDPOSTOP1;
      CLASS GROUP ENROLID NMED;
      MODEL PAY1=GROUP ENROLID(GROUP) NMED GROUP*NMED/SS3;
      RANDOM ENROLID(GROUP);
      TEST H=GROUP E=ENROLID(GROUP);
      MEANS GROUP;
RUN;
```

8.4.2. Results

8.4.2.1. Data description

The inpatient data sets of the years 2000 and 2001 contained 494,106 records for a total of 137,890 patients. The females in this dataset were 236,001 (63.12%) of all patients among which 3919 (1.66% of all females) were diagnosed with breast cancer. Among these women with breast cancer, 1284 (89.29%) were treated with mastectomy, 154 (10.70%) were treated by lumpectomy. The rest of the patients, 2481 (63.31%) did not receive either a mastectomy or a lumpectomy and they were dropped from the analysis. Thus, the surgery set contained 1438 patients among which the majority of 89.29% underwent a mastectomy and 10.70% underwent a lumpectomy (see Table 8.2).

Table 8.2: Summary of the database of analysis

	Size and percentage
Database size	137,890
Females in the database	236,001 (63.12% of the total sample)
Females with breast cancer	3919 (1.66% of all females)
Breast cancer patients treated by surgery	1438(36.69% of all patients with breast cancer)
Mastectomy alone	1284 (89.29% of the breast cancer surgery sample)
Lumpectomy alone	154 (10.70% of the breast cancer surgery sample)

In order to have comparable group sample sizes, a random sample of size 154 was selected from the mastectomy group and used in the study. Thus, the study set had a total of 308 patients. After extracting all the records of 2000 and 2001 for the 308 patients in the analysis, the total number of inpatient records after the initial procedure was 4090. The post-operative outpatient services were 782,259 all together and the post-operative medication claims were a total of 95,181.

8.4.2.2. Cluster analysis

Cluster analysis was performed to obtain a grouping for the diagnosis at the time of the initial procedure. Patients presented many different conditions and in the database, up to 15 diagnoses are recorded per visit. Clustering the diagnoses provided a way to define them in a small number of groups. Four diagnosis clusters were obtained. Cluster 1 contained diagnoses related to breast cancer with metastasis in other sites, cluster 2 grouped patients with breast cancer with affected lymph nodes, cluster 3 contained patients who had breast cancer localized only in the breast and cluster 4 contained patients who were diagnosed with breast cancer and had a personal history of breast cancer (see Table 8.3). The clustering process elaborated in section 8.3.4 was used.

Table 8.3: Description of the diagnosis clusters

Diagnosis cluster	ICD 9 codes and description	Cluster name
1	174.8: breast cancer in the outer specified sites 611.9: unspecified breast disorder 198.89: secondary malignant in other sites 233.0: breast carcinoma in-situ 174.6: breast cancer in Axillary tail	Breast cancer with metastasis in the other sites
2	174.4: breast cancer in the upper-outer quandrant 196.3: secondary and unspecified malignant neoplasm of lumph nodes of axilla and upper limp 198.89: secondary malignant in other sites 174.9: breast cancer, unspecified breast site 174.8: breast cancer in the outer specified sites	Breast cancer with affected lymph nodes
3	174.2: breast cancer in the upper-inner quadrant 239.3: neoplasm of unspecified nature in the breast 174.1: breast cancer in central portion 229.0: benign neoplasm of the lymph nodes 611.72: lump or mass in breast	Breast cancer in the breast only
4	611.8: other specified disorders of breast 174.0: breast cancer in the nipple and areola 401.9: unspecified Essential hypertension 174.3: malignant neoplasm of female breast in the lower inner quadrant V10.3: personal history of breast cancer	Breast cancer with a personal history of breast cancer

8.4.2.3. Descriptive statistics: distribution of the input variables

The sample data comprised a cohort of 308 between the ages of 27 to 67 years old with an average value of 52 (standard deviation 8, see Table 8.4). More than half of the patients had a Charlson index of 2 (40.26%) and less than 1% had a Charlson index of 1. Many patients were in diagnosis cluster 3 (36.21%) and among the rest, the majority were in diagnosis cluster 4 (27.59%). Most operations took place in the north central (39.94%) and the area with the least number of initial breast cancer surgical operations was the northeast with 13.64%. The post-operative continuous enrollment follow-up time was on

average 249 days (standard deviation 178) varying from 0 to 728 with a median of 218 (interquartile range 116 - 337).

Table 8.4: Description of the analysis data

Variable	**All patients (n=308)**	**Mastectomy (n=154)**	**Lumpectomy (n=154)**
Age [mean (standard deviation)]	52 (8)	53 (8)	52 (8)
Age n (%)			
<40	22 (7.14)	11 (7.14)	11 (7.14)
40-60	228 (74.03)	109 (70.78)	119 (77.27)
>60	58 (18.83)	34 (22.08)	24 (15.58)
Charlson index n (%)			
0	71 (23.05)	32 (20.78)	39 (25.32)
1	3 (0.97)	1 (0.65)	42 (1.30)
2	124 (40.26)	65 (42.21)	59 (38.31)
>=3	110 (35.71)	56 (36.36)	54 (35.06)
Post-operative follow-up days [mean (standard deviation)]	249 (178)	253 (173)	245 (183)
Disease cluster n (%)			
Breast cancer with metastasis	12 (20.69)	5 (16.67)	7 (25.00)
Breast cancer with affect lymph nodes	9 (15.52)	4 (13.33)	5 (17.86)
Breast cancer in the breast only	21 (36.21)	8 (26.67)	13 (46.43)
Breast cancer with personal history of breast cancer	16 (27.59)	13 (43.33)	3 (10.71)
Region of initial procedure n (%)			
Northeast	42 (13.64)	21 (13.64)	21 (13.64)
North central	123 (39.94)	60 (38.96)	63 (40.91)
South	78 (25.32)	53 (34.42)	25 (16.23)
West	64 (20.78)	20 (12.99)	44 (28.57)

8.4.2.4. Descriptive statistics: distribution of the outcome variable of interest

Breast cancer treatments, especially surgery, trigger hospitalizations and frequent outpatient services. Even though each treatment sequence has its own characteristics and each patient has a different and specific reaction, similar treatments will have similar patterns in terms of healthcare resource usage. In the current study, statistical analysis

tests and models were used to compare lumpectomy to mastectomy in length of stay per post-operative hospitalization, charges per post-operative hospitalization, charges per post-operative outpatient service and charges per post-operative medication. The analysis was performed on a longitudinal data set where each patient could have more than one record. During this period of post-operative follow-up time, only 12 individuals were re-operated, four in the mastectomy group and eight in the lumpectomy group.

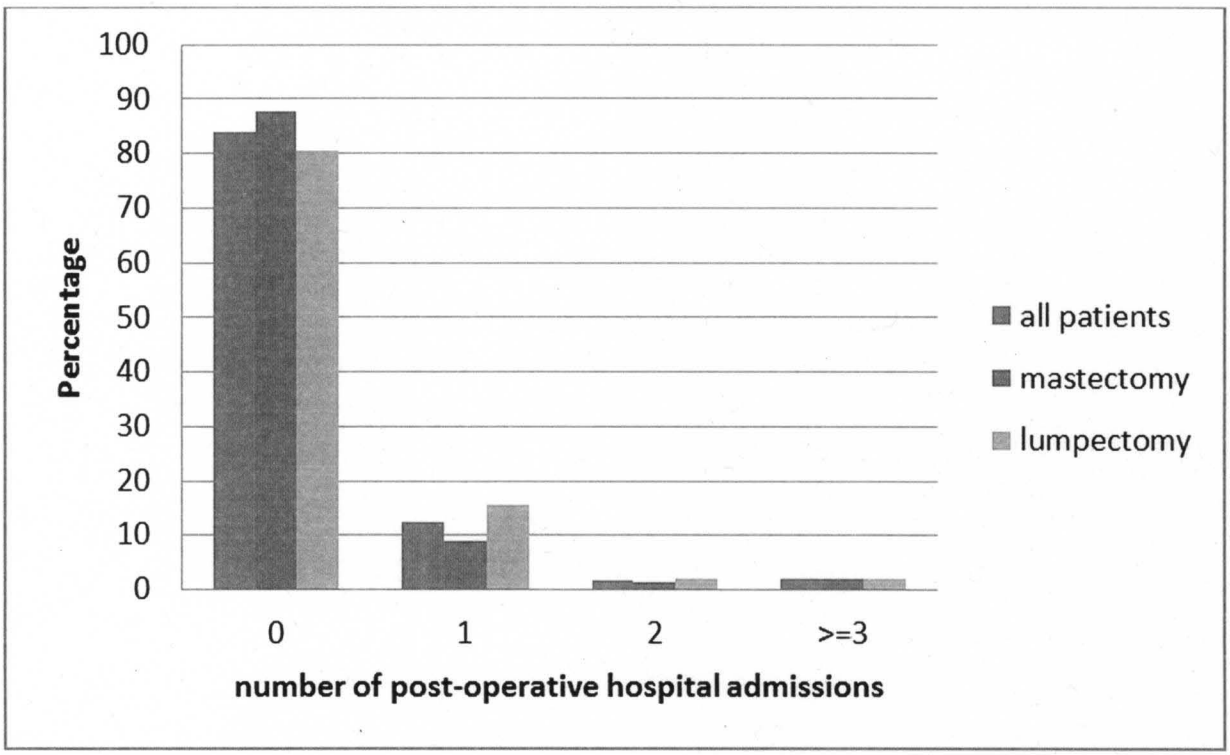

Figure 8.2: Number of post-operative hospital admissions

Throughout the two years 2000 and 2001, a majority of the patients (84.09%) were not re-hospitalized. About 12.34% were re-hospitalized only once; among the rest who were re-hospitalized, only six patients had more than three hospital-stays (see Figure 8.2). A higher percentage in the lumpectomy group was re-hospitalized at least once in comparison to the mastectomy group.

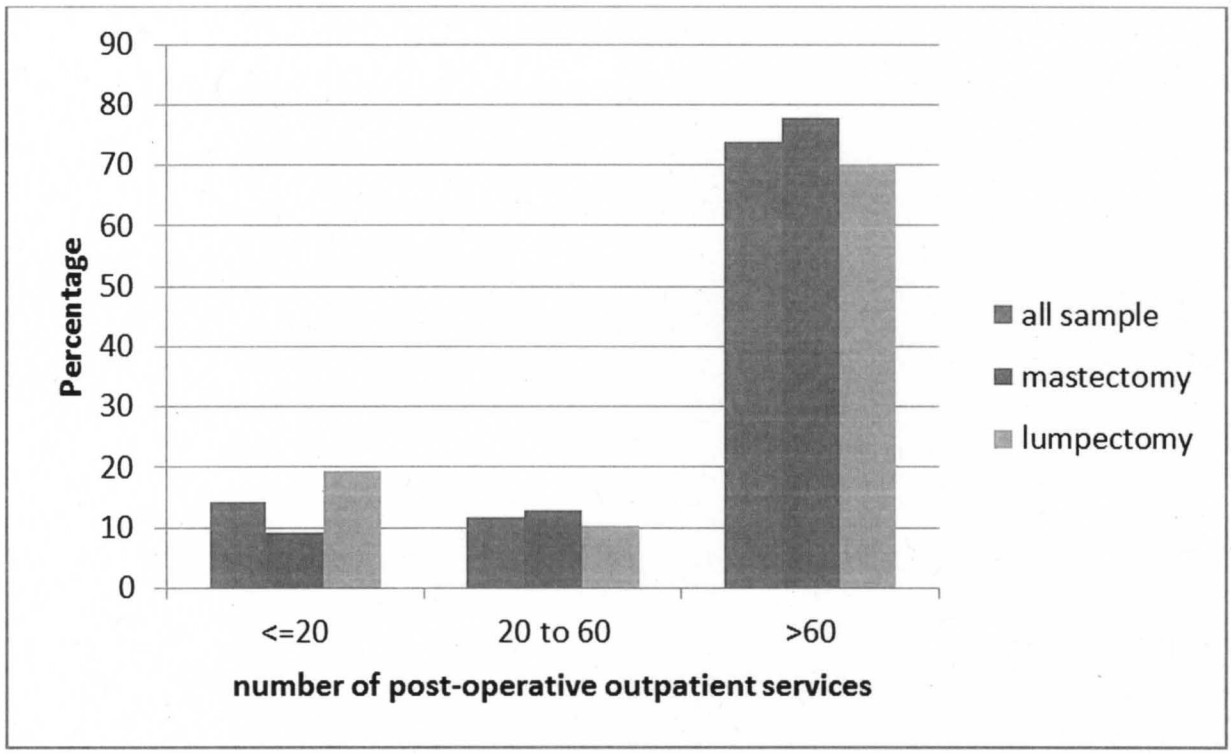

Figure 8.3: Number of post-operative outpatient services

During the 249 mean follow-up days (standard deviation 178, Table 8.4), most of the patients in the data analysis had over 60 outpatient services. In this category, the percentages were higher in the mastectomy group.

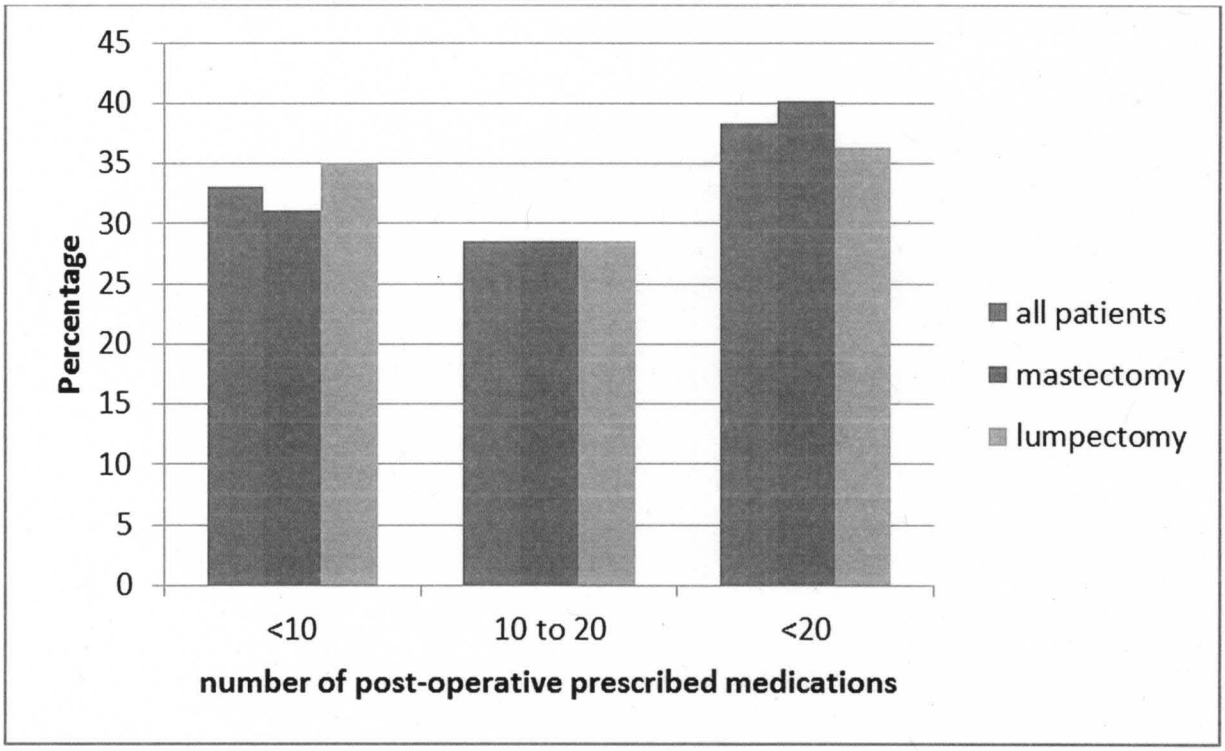

Figure 8.4: Number of post-operative prescribed medications

In the follow-up time (249 days on average), more patients had over 20 prescribed medications. The percentages were higher in the mastectomy group.

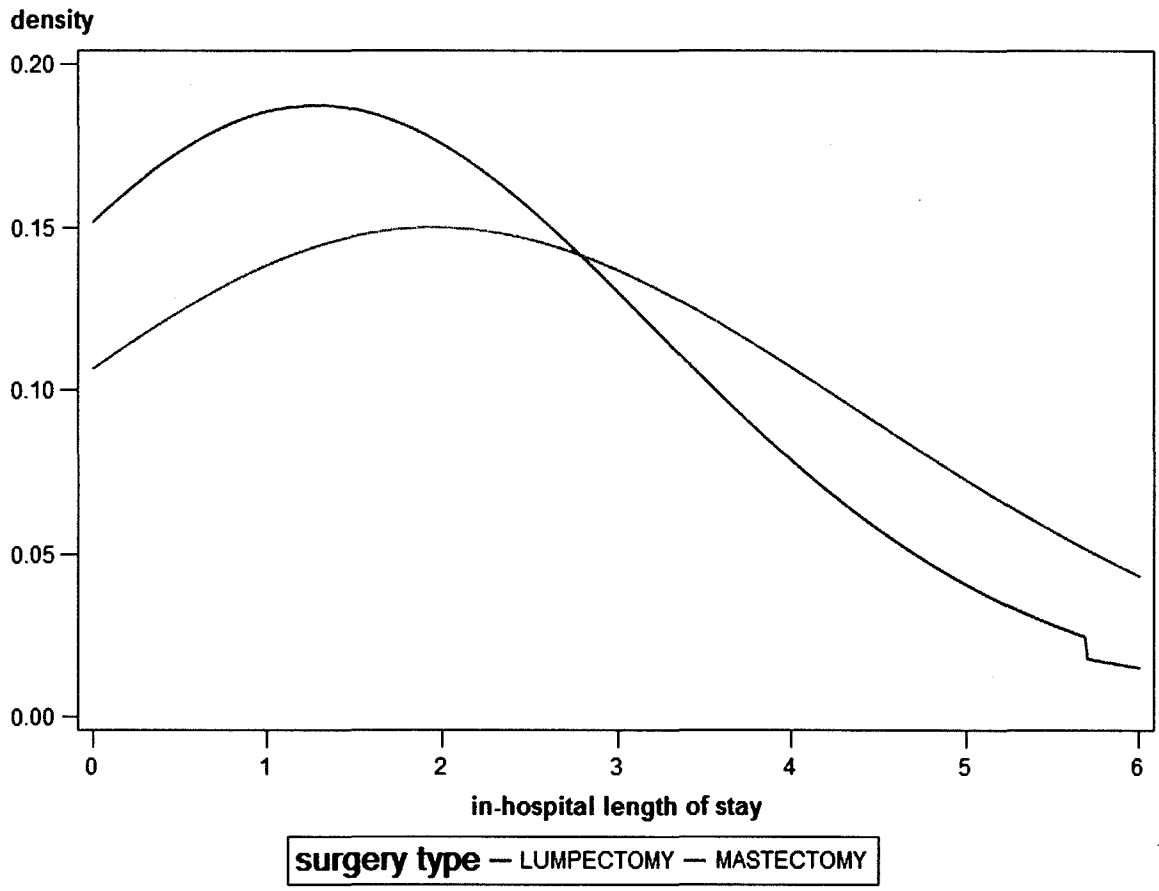

Figure 8.5: Length of stay per post-operative hospitalization

This graph shows that the probability of a shorter hospital stay was higher in the mastectomy group while the probability of longer stay was higher in the mastectomy group. The distributions cross around three days.

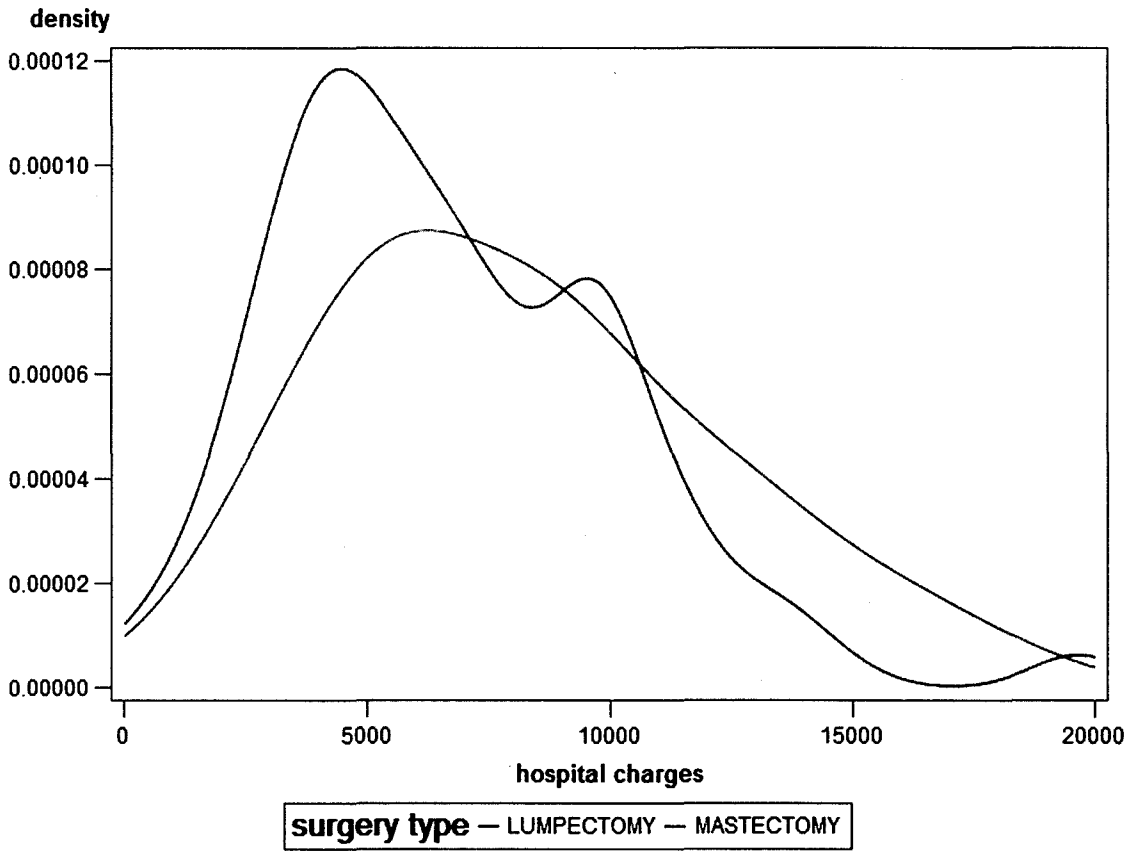

Figure 8.6: Charges per post-operative hospital stay

The distributions of the total hospital charges for each post-operative hospital admission were skewed to the right for both procedures. The probability of lower charges was higher for the lumpectomy group and the probability of higher charges was higher for the mastectomy group.

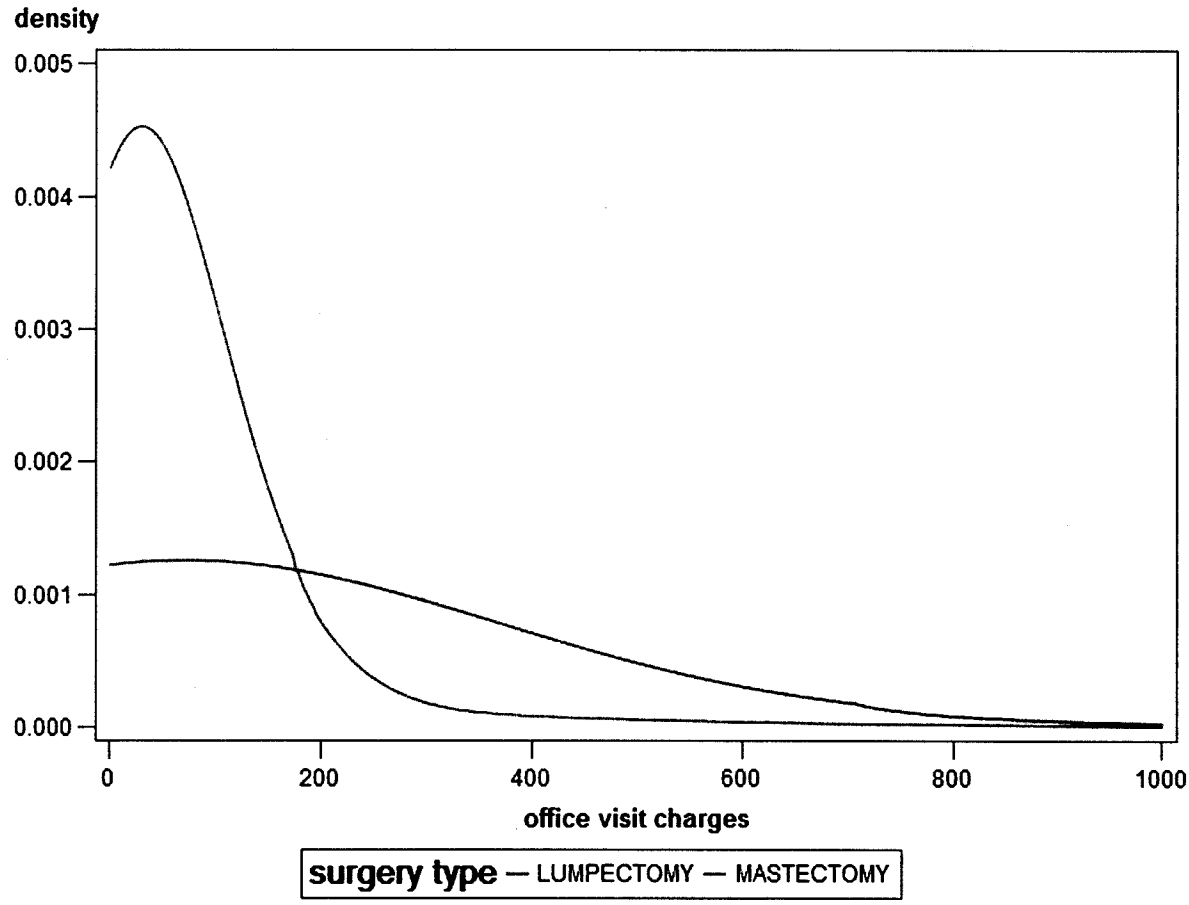

Figure 8.7: Charges per post-operative outpatient service

The distributions of outpatient services crossed around $200 for mastectomy and lumpectomy. Lumpectomy had a lower probability of charges under $200 and a higher probability of charges over $200.

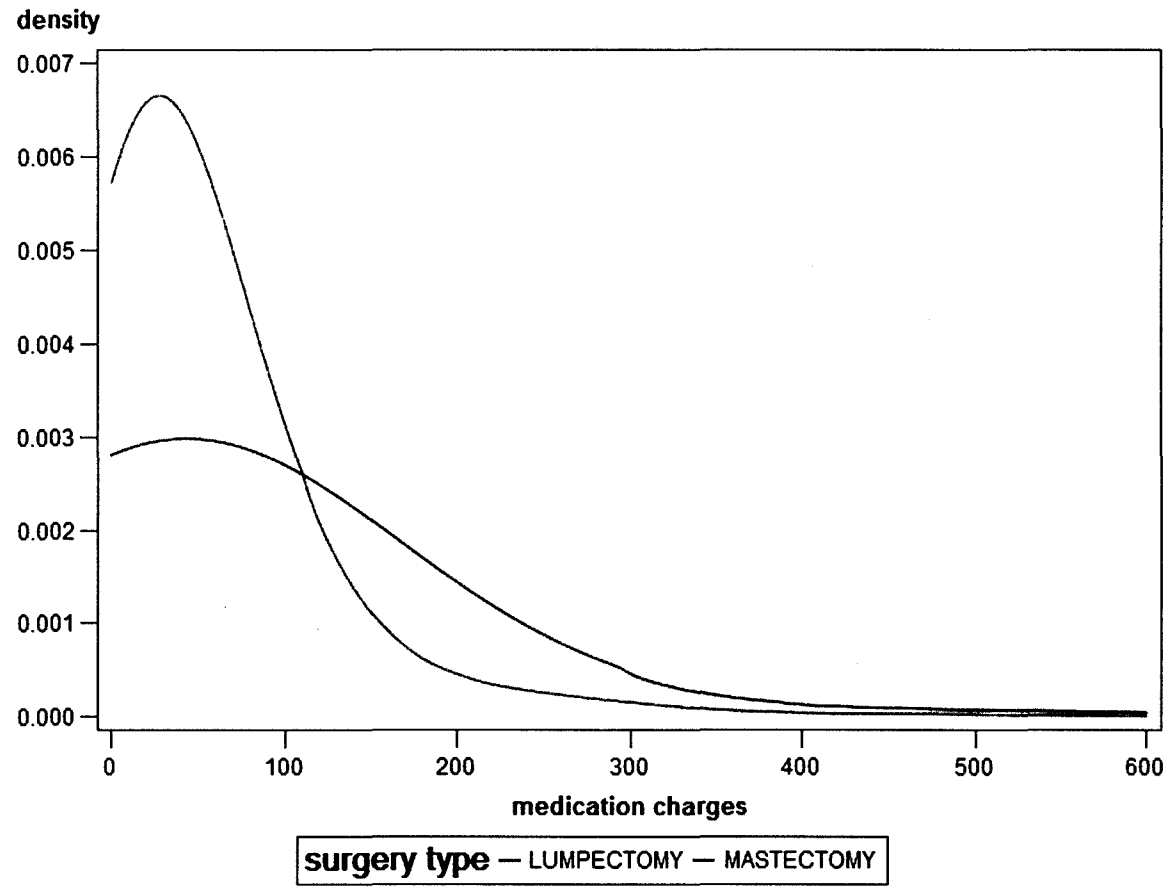

Figure 8.8: Charges per post-operative prescribed medication

The group of patients who underwent mastectomy had higher probability for higher medication costs over $100. The probability of costs below $100 was higher in the lumpectomy group.

8.4.2.5. Inferential statistics results: Comparison of the effect of the surgical procedure groups

The hypothesis was that in terms of clinical resource usage, lumpectomy will be comparable to mastectomy and in terms of clinical charges; mastectomy would have the least charges. In terms of risk of re-hospitalization, it was expected that patients who had

mastectomy will have a higher risk of post-operative hospital admission. The descriptive statistics above suggested differences in outcome variables of interest; however, it was necessary to test these differences for statistical significance.

To compare the in-hospital stay and charges per post-operative hospitalization, the charges per post-operative outpatient service, the cost per post-operative prescribed medication, the *Repeated Measure ANOVA models* discussed in section 8.4 were used. To evaluate the significance of the difference between the number of hospitalizations, the number of outpatient services and the number of prescribed medications, *Mann Whitney tests (Wilcoxon Rank-Sum tests)* were used, and to compare the difference of the rates of re-hospitalization, *Logistic Regression models* were used. All of the statistical tests were two-sided and the significance level was set to 0.05. For continuous variables, means and standard deviations were calculated to describe the central tendency. For categorical variables, counts and percentages were presented. Since the continuous variables were highly skewed to the right, logarithmic transformations were performed and the transformed variables in the form of new-variable = log (old-variable) were used for analysis. The transformation was used in an attempt to normalize these variables and comply to the assumption of the ANOVA models.

8.4.2.5.1. Comparison of lumpectomy and mastectomy with respect to healthcare resources use

Clinical resource use was considered to be the hospital resource utilization, the outpatient service utilization and the prescribed drug use. The analysis used the log-transformed variables.

Comparison of in-hospital stay during the index hospitalization (Table 8.5): During the initial procedure hospitalization, patients who underwent the mastectomy stayed an average two days (standard deviation two days) while patients in the lumpectomy group stay an average one day (standard deviation one day). This difference of about one day was found to be statistically significant (p-value < 0.0001).

Comparison of in-hospital stay per post-operative admission (Table 8.5): The average length of stay per all-type hospitalization was found to be three days (standard deviation: three days) for a patient in the mastectomy group and two days (standard deviation: two days) for a patient in the lumpectomy group. For breast cancer related hospitalizations, the length of stay was on average about two days for both groups (standard deviation: three days for mastectomy and two days for lumpectomy. The comparison of the two groups showed that the group did not have a significant effect on the length stay for all-type post-operative hospitalization (p-value: 0.33) and for post-operative breast cancer related hospitalization (p-value: 0.57).

Comparison of the number of post-operative hospitalizations (Table 8.5): Patients in both procedure groups had on average less than one post-operative hospitalization (all-type as well as breast cancer related). Statistical analysis of the effect of procedure type on post-operative hospitalizations resulted in non-significant results (p-value: 0.09 for all type admissions and 0.23 for breast cancer related admissions).

Comparison of the number of post-operative outpatient services (Table 8.5): The mastectomy group came out with the highest number of all-type outpatient services with an average of 166 services per patient (standard deviation: 132) in comparison to 140

services per patient (standard deviation: 141) in the lumpectomy group. The number of breast cancer related outpatient services was also larger for the mastectomy group (average: 92, standard deviation: 109 versus average: 78, standard deviation 103). However, these differences were not statistically significant (p-value: 0.05 for all-type and 0.15 for breast cancer related).

Comparison of the number of post-operative prescribed medications (Table 8.5): Patients in the mastectomy group had more prescribed medications than in the lumpectomy group (25, standard deviation 28 versus 23, standard deviation 25), but this difference was not statistically significant (p-value: 0.43).

8.4.2.5.2. Comparison lumpectomy and mastectomy with respect to healthcare resources use charges

Comparison of the hospital charges for the initial procedure hospitalization (Table 8.5): On average, the hospitalization of the initial procedure cost $9191 (standard deviation $5410) for a mastectomy and $6911 (standard deviation $3856) for a lumpectomy. Statistical analyses revealed that the type of procedure had an effect on the hospital charge for the index hospitalization (p-value <0.0001).

Comparison of the hospital charges per post-operative admission (Table 8.5): The average hospital charges per all-type hospitalization per patient were highest in the Mastectomy group ($9187, standard deviation $5872 in comparison to $7404, standard deviation $5765). For breast cancer related post-operative hospitalization, this difference was also observed. For the Mastectomy group, the average was $9481 (standard deviation $7044) and for Lumpectomy $7194 (standard deviation: $5553). The effect of the

surgical procedure group on the hospitalization charges was found to be not significant (p-value: 0.06 for all-type re-admissions and 0.67 for breast cancer related re-admissions).

Comparison of the outpatient charges per post-operative service (Table 8.5): Patients in the lumpectomy group had the highest charges per all-type post-operative outpatient service with an average of $196 (standard deviation $914) as opposed to $89 (standard deviation $458). However, this difference was not statistically significant (p-value: 0.19). For breast-cancer related post-operative outpatient services, patients in the lumpectomy group still had the highest charges with an average of $230 (standard deviation: 974) as opposed to $170 (standard deviation: 589). This difference was found to be statistically significant (p-value: 0.0006).

Comparison of the medication charges per post-operative prescription (Table 8.5): The effect of the surgery group on the average medication charges was not significant (p-value: 0.14) even though the charges were on average higher in the lumpectomy group ($94, standard deviation $259 versus $62, standard deviation $183).

8.4.2.5.3. Comparison lumpectomy and mastectomy with respect to hospital re-admission

During the follow-up time, 19 patients (12.34%) who had Mastectomy and 30 patients who had Lumpectomy (19.48%) were re-hospitalized at least once for any reason. This re-hospitalization rate difference of about 7% was not statistically significant (p-value: 0.09, Table 8.5). In general, patients who had the mastectomy were re-hospitalized much sooner (208 days, standard error 6 days) than those who underwent the lumpectomy (282

days, standard error 7 days) but this time difference of more than 80 days was not statistically significant (p-value: 0.07, Table 8.5).

During this post-operative follow-up time, four patients (2.60%) in the Mastectomy group and eight patients (5.19%) in the Lumpectomy group were re-hospitalized with breast cancer as the primary cause (diagnosis). This difference was not statistically significant (p-value: 0.24, Table 8.5).

8.4.2.5.4. Comparison lumpectomy and mastectomy with respect to re-operation

In the study data, four patients from the Mastectomy group (2.60%) and eight patients from the Lumpectomy group (5.19) were re-operated. The difference in re-operation rate was found to be not significant (p-value: 0.24, Table 8.5). In general, patients in the mastectomy group were re-operated non-significantly sooner than patients in the lumpectomy group (60 days versus 240 days on average, p-value: 0.12).

Table 8.5: Comparison of healthcare resources use and charges

Outcome variable	**Mastectomy (n=154)**	**Lumpectomy (n=154)**	**p-value**
Healthcare resources use [mean (SD)]			
All-type			
Initial procedure length of stay	2 (2)	1 (1)	<0.0001
Length of stay per post-operative stay (days)	3 (3)	2 (2)	0.33
Number of hospitalizations	0.19 (0.62)	0.26 (0.62)	0.09
Number of outpatient services	166 (132)	140 (141)	0.05
Number of prescribed medications	25 (28)	23 (25)	0.43
Breast cancer related			
Length of stay per post-operative stay (days)	2 (3)	2 (2)	0.57
Number of hospitalizations	0.02 (0.16)	0.07 (0.34)	0.23
Number of outpatient services	92 (109)	78 (103)	0.15
Healthcare resources charges ($)[mean (SD)]			

Outcome variable	Mastectomy (n=154)	Lumpectomy (n=154)	p-value
All-type			
Initial procedure charges	9191 (5410)	6911 (3856)	<0.0001
Hospital charges per post-operative stay	9187 (5872)	7404 (5765)	0.06
Outpatient charges	89 (458)	196 (914)	0.19
Medication charges	62 (183)	94 (259)	0.14
Breast cancer related			
Hospital charges per post-operative stay	9481 (7044)	7194 (5553)	0.67
Outpatient charges	170 (589)	230 (974)	0.0006
Re-hospitalization			
N (%)	19 (12.34)	30 (19.48)	0.09
OR (95% CI)	REF	1.72(0.92,3.21)	0.09
Days to re-admission [mean (SE)]	208 (6)	282 (7)	0.07
Re-operation			
All-type			
N (%)	4 (2.60)	8 (5.19)	0.24
OR (95% CI)	REF	2.05(0.6,6.97)	0.06
Days to re-operation [mean (SE)]	60 (0.6)	240 (2)	0.12
Breast cancer related			
N (%)	4 (2.60)	8 (5.19)	0.24
OR (95% CI)	REF	2.05 (0.6, 6.97)	0.25

Abbreviations: SD: standard deviation, OR: odds ratio, CI: confidence interval, SE: standard error

8.5. Use of predictive modeling to analyze hospital re-admission

8.5.1. Objective

A secondary aim to the current analysis was to predict the risk of 90-day hospitalization after undergoing a mastectomy or a lumpectomy. A hospitalization shortly after surgery can be a signal of surgery complications. Patients were followed until they were no longer enrolled with their insurance or until the end of the study period, which is December 31st, 2001. The objective was, more specifically, to provide a simple predictive model.

8.5.2. Summary of methods

EM was used to compare different predictive modeling techniques on the 90-day post-operative hospital admission. The input variables were the age divided into three groups (less than 40 years old, 40 to 60 years old, over 60 years old), disease cluster, type of procedure (lumpectomy or mastectomy) and the Charlson index at the time of the initial procedure in three groups (0 or 1, 2, and at least 3). Age was categorized following the recommendation of the American Cancer Society that all women should have an annual mammogram starting at age 40. The cut-point of age 60 was included to analyze if there exist differences in outcomes for women who are past the age range of menopause. The Charlson index was categorized to evaluate the outcomes of patients with comorbidities (Charlson index ≥ 1) to those with no burden of comorbidities (Charlson index = 0). EM was used to fit and compare three predictive model methods, the logistic regression, the neural network and the decision tree. Goodness of fit was evaluated using the misclassification rate, the average squared error and the area under the receiver operating characteristic curve (the ROC index). The misclassification rate quantifies in term of percentage the number of elements that were misclassified by the predictive model. A small misclassification rate is desired. The average squared error assesses the performance of the model by mean squared difference between the actual and the predicted value [59]. The smaller the average squared error, the better the fit. The ROC curve plots the sensitivity (true prediction rate) against 1-specificity (false prediction rate) for consecutive cut-offs points for the probability of the target [74]. The area under the ROC curve or c-statistic is a common measure of model discrimination. A c-statistic of 0.5 is similar to the probability of tossing a coin and is considered to be a poor fit with

equal group sizes. Any c-statistic greater than 0.5 indicates better fit; however, it is desirable to have a value of at least 0.7 [74]. A value of 1 signifies a perfect fit. A flow chart of the comparative performance of predictive models is presented in Figure 8.2.

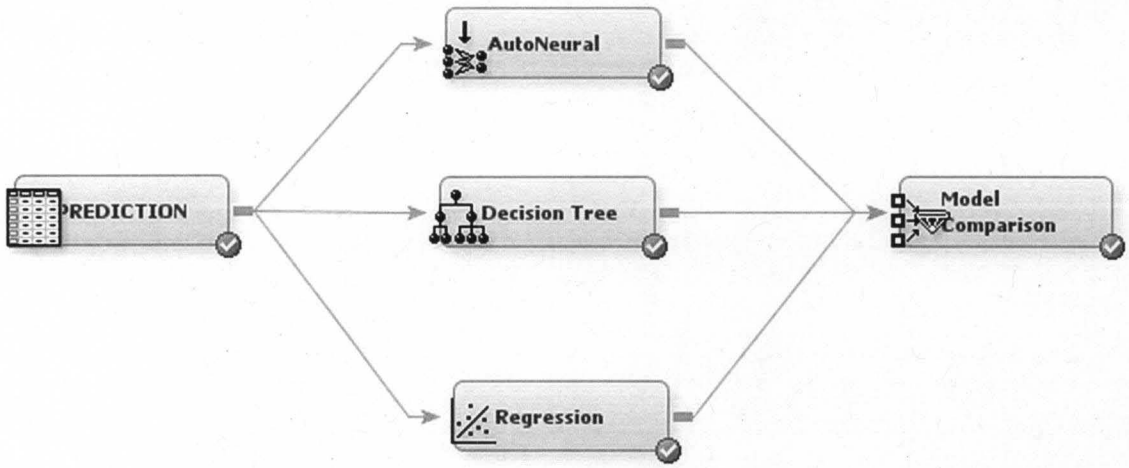

Figure 8.9: Flow chart of predictive model comparison

8.5.3. Results

8.5.3.1. Data description

Due to the small size of the data available, the data were not divided into a training set, a validation set and a testing set; it was used as a whole for model construction and model evaluation. The input variables used were the surgical procedure, the disease cluster, the Charlson index and the age group.

8.5.3.2. Model selection

Among the three models, Logistic Regression discriminated the best with an ROC index of 0.58. The Neural Network and the Decision Tree did not have a good discrimination

(c-statistic: 0.5). The misclassification rate was 12% for the logistic regression and 13% for both the Neural Network and the Decision Tree. The average squared error values were close in the three models (0.107 for logistic regression and 0.113 for both Neural Networks and Decision Tree, Table 8.6). The best predictive model was found to be the logistic regression.

Table 8.6: predictive model comparison

Model	**Misclassification rate**	**Average Squared Error**	**Roc index**	**Choice**
Logistic Regression	0.12	0.107	0.58	*
Neural Networks	0.13	0.113	0.5	
Decision Tree	0.13	0.113	0.49	

8.5.3.3. Predictive model with the logistic regression

First, a bivariate analysis was performed to evaluate which of the input variables had an effect on the 90-day post-operative hospital re-admission. Then, the effects of the variables with a significant association were analyzed in a multivariate logistic regression.

Table 8.7: Association with post-operative re-hospitalization

Input variable	**p-value**
Surgical procedure	0.009
Charlson index	0.7
Age group	0.41
Disease cluster	0.9

Only the surgical procedure had a significant association with 90-day post-operative hospital admission (Table 8.7). The risk of 90-day post-operative hospital admission was about three times higher for a patient who had a lumpectomy in comparison to a patient who had a mastectomy (unadjusted odds ratio: 2.031, 95 % confidence interval: 1.016-4.06, adjusted odds ratio: 124.669, 95% confidence interval: 1.059-124.669, Table 8.8). Unadjusted models were constructed using only the procedure group as input. Adjusted results were obtained from the analysis of models which included the Charlson index, age group and disease cluster in addition to procedure type.

Table 8.8: Odds ratio

Risk measure	**Surgical procedure**		**p-value**
	Mastectomy	**lumpectomy**	
Unadjusted OR (95% CI)	REF	2.031 (1.016, 4.06)	0.0449
Adjusted* OR (95% CI)	REF	11.49 (1.059, 124.669)	0.0447

*Adjusted for disease cluster, Charlson index and age group

8.5.3.4. Construction of a simple predictive model

In the analysis data set, 128 patients who had the mastectomy (47.76%) and 26 patients who had the lumpectomy (65%) were re-hospitalized within 90-days after surgery. The unadjusted risk of 90-day post-operative re-hospitalization was about twice higher for a patient in the lumpectomy group compared to a patient in the mastectomy group. The risk of re-hospitalization as a function of the procedure only can be expressed as

RE-HOSPITALIZATION RISK = 48% MASTECTOMY + 65% LUMPECTOMY

where mastectomy and lumpectomy are dichotomous variables, taking the value 1 if a patient undergoes the procedure.

Using this simple equation, the following risks were obtained

Table 8.9: Risk of 90-day hospital re-admission as a function of the procedure type

Procedure	90-day post-operative hospital admission risk
Mastectomy	48%
Lumpectomy	65%

8.6. Post-operative short-term Cost Effectiveness Analysis

8.6.1. Objective

A cost effectiveness analysis using the claims data was another secondary objective of the current study. The aim was to evaluate the comparative cost impact when lumpectomy is chosen instead of mastectomy.

8.6.2. Summary of methods

A special case of stochastic model, Markov Chain was used here. The definition of a stochastic process and a Markov Chain were provided in section 6.4. In previous sections of the current chapter, it was found that the 90-day post-operative hospital re-admission rate was 48% for mastectomy and 65% for lumpectomy. From these values, the probability of a post-operative hospital re-admission per day can be estimated to be 0.48/90 = 0.005 for the mastectomy procedure and 0.65/90 = 0.007 for the lumpectomy procedure. Using the data of the procedure hospitalization, it was found that the maximum length of stay was 20 days for mastectomy and 6 days for lumpectomy. Thus, the probability of being discharged per day when in the hospital can be estimated by 1/20 = 0.05 for mastectomy and 1/6 = 0.17 per day for lumpectomy. In Table 8.5, the average

charges for the initial procedure of $9191 (average $5410) for Mastectomy and $6911 (standard deviation $3856) for lumpectomy were found. Using these estimates computed from the data, the state transition matrices M (M) and M (L) can be written as

M (M) =

$$\begin{bmatrix} P\,(not\ hospitalized|not\ hospitalized) & P\,(not\ hospitalized|hospitalized) \\ P\,(hospitalized|not\ hospitalized) & P\,(hospitalized|hospitalized) \end{bmatrix} =$$

$$\begin{bmatrix} 0.995 & 0.005 \\ 0.05 & 0.95 \end{bmatrix} \text{ for Mastectomy}$$

M (L) =

$$\begin{bmatrix} P\,(not\ hospitalized|not\ hospitalized) & P\,(not\ hospitalized|hospitalized) \\ P\,(hospitalized|not\ hospitalized) & P\,(hospitalized|hospitalized) \end{bmatrix}$$

$$= \begin{bmatrix} 0.993 & 0.007 \\ 0.17 & 0.83 \end{bmatrix} \text{ for Lumpectomy}$$

Here, entries represent probabilities for moving from one state to the next in any given post-operative day. At day 0 (the index hospitalization), all patients are hospitalized. Thus, the initial state matrix is $A_0 = [P\,(not\ hospitalization) \quad P\,(hospitalized)] = [0 \quad 1]$ for both procedures. Day n after surgery, the state matrix can be obtained by

$$A_n\,(M) = [0 \quad 1]\begin{bmatrix} 0.995 & 0.005 \\ 0.05 & 0.95 \end{bmatrix}^n \text{ and } A_n\,(L) = [0 \quad 1]\begin{bmatrix} 0.993 & 0.007 \\ 0.17 & 0.83 \end{bmatrix}^n$$

Measure of Effectiveness can be estimated as follows:

E (M) = P_M (not hospitalized) * 154 and E (L) = P_L (not hospitalized) * 154 and the incremental effectiveness of lumpectomy compared to mastectomy is

$$E\,(L) - E\,(M) = 154^*\,(P_L\,(\text{not hospitalized}) - P_M\,(\text{not hospitalized}))$$

where E (M) is the effectiveness of mastectomy and E (L) is the effectiveness of lumpectomy.

Measures of cost can be estimated as by the average charges for the procedure hospitalization. Hence, the incremental cost effectiveness ratio at day n ($ICER_n$) can be computed as follows:

$$ICER_n = \frac{C(L) - C(M)}{E(L) - C(M)} = \frac{6911 - 9191}{154 * (P_L \text{ (not hospitalized)} - P_M \text{ (not hospitalized)})}$$

where C (M) represents the cost associated with mastectomy and C (L) represents the cost associated with lumpectomy.

8.6.3. Results

In Table 8.10, different values of the ICER are presented for different values of n. To illustrate how computations are performed, $ICER_{30}$ is presented here. At day 30, the state matrices are

$$A_{30}(M) = A_0 * [M(M)]^{30} = [0 \quad 1] \begin{bmatrix} 0.995 & 0.005 \\ 0.05 & 0.95 \end{bmatrix}^{30} = [0.742 \quad 0.257] \text{ for Mastectomy}$$

and

$$A_{30}(L) = A_0 * [M(L)]^{30} = [0 \quad 1] \begin{bmatrix} 0.993 & 0.007 \\ 0.17 & 0.83 \end{bmatrix}^{30} = [0.958 \quad 0.042] \text{ for Lumpectomy.}$$

With these estimates, the corresponding incremental cost effectiveness ratio of lumpectomy in comparison to mastectomy is

$$ICER_{30} = \frac{6911-9191}{154*(0.958-0.742)} = -67.30 \simeq -67,$$

which means that lumpectomy saved about $67 per patient who had it instead of mastectomy each day after hospitalization at day 30.

Table 8.10: Incremental cost effectiveness ratio values for different days after the initial surgery

n	A_n (M) = $[P_M(\bar{h})$ $P_M(h)]$	A_n (L) = $[P_M(\bar{h})$ $P_M(h)]$	$ICER_n$	Short interpretation
1	[0.005 0.95]	[0.17 0.83]	-123	Savings of $123
2	[0.09 0.903]	[0.31 0.69]	-67	Savings of $67
3	[0.142 0.858]	[0.425 0.575]	-50	Savings of $50
4	[0.184 0.816]	[0.52 0.48]	-43	Savings of $43
5	[0.224 0.776]	[0.598 0.402]	-38	Savings of $38
10	[0.393 0.607]	[0.823 0.177]	-33	Savings of $33
15	[0.52 0.48]	[0.909 0.091]	-37	Savings of $37
20	[0.616 0.384]	[0.9445 0.055]	-43	Savings of $43
30	[0.742 0.257]	[0.958 0.042]	-67	Savings of $67
40	[0.814 0.185]	[0.96 0.04]	-99	Savings of $99
50	[0.855 0.145]	[0.96 0.04]	-137	Savings of $137
60	[0.879 0.121]	[0.96 0.04]	-178	Savings of $178
70	[0.892 0.108]	[0.96 0.04]	-212	Savings of $212
80	[0.899 0.101]	[0.96 0.04]	-237	Savings of $237
90	[0.903 0.096]	[0.96 0.04]	-253	Savings of $253
100	[0.906 0.094]	[0.96 0.04]	-267	Savings of $267

Abbreviations: h = hospitalized, $\bar{h}$ = not hospitalized

Table 8.11 shows that lumpectomy is associated with cost savings in comparison to mastectomy after surgery in terms of hospital use.

8.7. Summary

In this chapter, with the use of statistical analysis, it was found that patients in the mastectomy group had a longer stay and more charges during the initial procedure in comparison to patients in the lumpectomy group. However, in terms of post-operative

healthcare resource use, the outcomes were statistically comparable. The use of cluster analysis revealed that the patients could be classified into four clusters that represent the extent of the breast cancer condition. Predictive modeling was used to construct a simple model of post-operative hospital admission. It was found that the risk of 90-day post-operative hospital admission was 65% with lumpectomy while it was 48% with mastectomy. The use of cost effectiveness methods showed that lumpectomy was associated with cost savings each day after surgery in terms of hospital usage. At day 1 after surgery, it was found that the savings amount was $127. At day 100 after surgery, the savings were $267. The lowest savings were observed around day 10 for an amount of about $33.

CHAPTER 9

USE OF DATA MINING AND COST EFFECTIVENESS ANALYSIS TO COMPARE LUMPECTOMY TO MASTECTOMY USING ONLINE COMMENTS OF SATISFACTION AS A MEASURE OF QUALITY OF LIFE

9.1. Objective

In this chapter, the aim was to perform a cost effectiveness analysis of lumpectomy in comparison to mastectomy. The outcome was patient satisfaction as a measure of quality of life. An assumption was made that satisfaction with surgery will improve satisfaction with personal conditions and thus improving the quality of life. Costs were estimated from the Nationwide Inpatient Sample (NIS) data. Thus, the hospital perspective was used. Only direct (gross) costs of the hospitalization in which the procedure is performed were considered. Since time was not a factor in the two alternatives, a deterministic model was used. The objective was to use SAS functions and SAS Text Miner tools to measure an estimate of the probabilities of satisfaction with mastectomy or lumpectomy from comments that patients post on online discussion boards. These probabilities were then used in a cost effectiveness model to evaluate the cost of satisfying a patient with the surgery performed.

9.2. Decision tree

The two alternatives were mastectomy and lumpectomy. In the decision tree, the discharge status was included as an event since that may influence patient satisfaction. The outcome was the fact that either the patient was satisfied or not. Figure 9.1 contains the decision tree with the two alternatives, the events and the outcomes.

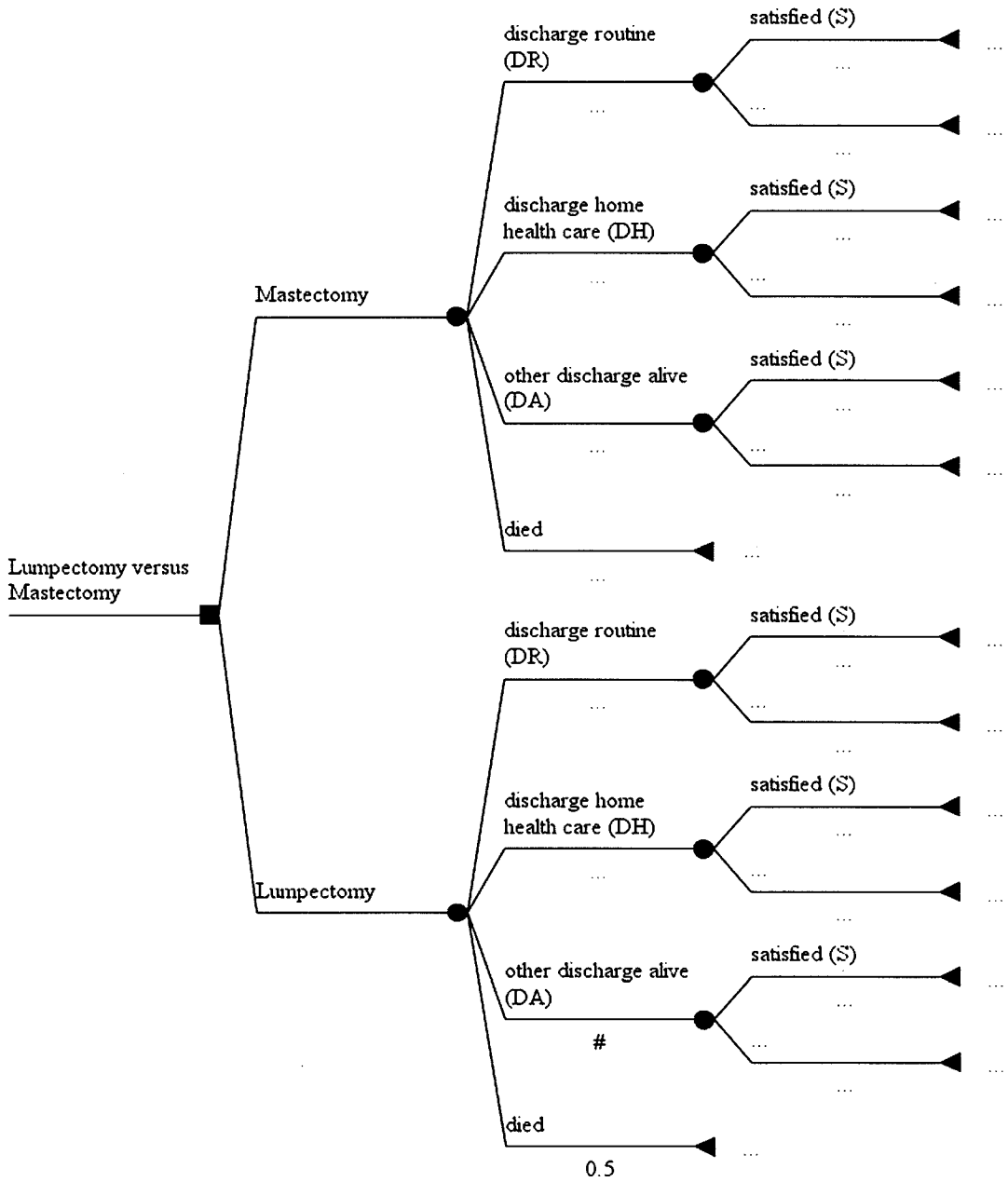

Figure 9.1: Decision tree for the cost-effectiveness model

9.3. Data –Online comments

In the current analysis, web posts from www.breastcancer.org were studied. The website www.breastcancer.org contains general information on breast cancer: symptoms, diagnoses, treatments, post-treatment advice, etc. One of the things the website offers to its viewers is a set of discussion boards. The discussion boards contain many subjects related to the breast cancer journey. Each subject contains many forums and in each forum, there are many topics. Here, the forum of 'surgery-before, during and after' in the subject of 'test, treatments and side effects' was chosen. Within this forum, the interest was on topics that explored the comparison or the choice of lumpectomy or mastectomy. A topic can be created by an individual creating an account and posting a concern, a question, or a quest of advice. By responding to this posting, participants or viewers post their comments. The comments analyzed were about what patients responding with information of what type of choice they made, why they made this choice, whether they are satisfied and what kind of advice they can give to the person asking the question.

9.4. Pre-processing

9.4.1. Transformation of comments into a table

Here, each posting constituted a document. Because of this nature of the data, the macro %TMFILTER was not used since it stores each web page as a document and in this case, there were many documents (comments) in one web page. Instead, the topics were screened to evaluate which ones talked about the comparison of lumpectomy to mastectomy. The following topics were found:

- Lumpectomy or mastectomy
- Future follow-up –lumpectomy vs. MX*
- Lumpectomy vs. mastectomy (7 times)
- What were the reasons that you had a mx versus a lumpectomy
- Lumpectomy vs. skin-sparing mastectomy for cosmetic purposes
- How to decide between lumpectomy and mastectomy
- Mastectomy vs. lumpectomy
- Need to decide: lumpectomy vs. mastectomy
- Lumpectomy or mastectomy
- Lumpectomy: re-excision or mastectomy
- Mastectomy vs. lumpectomy decision
- Mastectomy or lumpectomy
- For those who "could" have lumpectomy… why choose Mast*
- Mastectomy or Lumpectomy
- Lumpectomy or Mastectomy
- Mastectomy vs. Lumpectomy MRI

*mx, mast = mastectomy

The content of the pages of these topics was copied, pasted and formatted in Microsoft Word and then Excel to a table. The data contained a total of 337 web posts discussing the choice between mastectomy and lumpectomy.

9.4.2. Use of SAS to format the table in appropriate analysis format

The final table obtained in Excel contained multiple rows per document (comment). The next step was to use SAS functions to form a table in which a row represented the whole comment of an individual. Next, the SAS codes used are presented

Code 9.1: Creation of the common identification number (ID) for all the rows of a same comment

```
DATA COMMENTS1;
     SET COMMENTS;
     IF NAME NE ' ' THEN ID+1;
RUN;
```

Code 9.2: Transpose the data per person to obtain one row and many columns

```
PROC TRANSPOSE DATA=COMMENTS1 OUT=COMMENTS2;
     VAR POST;
     BY ID;
RUN;
```

Code 9.3: Concatenate all the columns into one to obtain the whole comment in one cell per person

```
DATA COMMENTS3;
     SET COMMENTS2;
     LENGTH POSTS $ 32767;
     POSTS=CATX(' ', OF COL1-COL12);
RUN;
```

Table 9.1: Random sample of the pre-processed data

	ID	POSTS
1	12	Thanks everyone! Has anyone had a lumpectomy and then plastic surgery (particulary size reduction) on the non-effected breast...
2	13	blondelawyer - When I was working my way through this decision I went to the suppoort group at the hospital a couple times. I w...
3	29	Hi everyone. I was dx with bc in Feb08 @ 42 yrs old. I had a lumpectomy, chemo and currently in radiation. I'm doing great! But...
4	31	Tiff, you and I were dx'd at about the same time. I had a left-side mastectomy/SNB in Feb '08 (without recon), and I'm dealing wit...
5	35	Tiff, It's true that many women do choose to have a bilateral and are happy with the decision, but this is a complicated, difficult de...
6	46	It was suggested by a fellow member that I post this as a seperate topic for those who are trying to decide what to do. Before rea...
7	78	I started out with Lumpectomy but after they went in again for clear margins they found a second cancer so mastectomy is was s...
8	84	It was suggested by a fellow member that I post this as a seperate topic for those who are trying to decide what to do. Before rea...
9	124	Janet, did you find that quote on the NCI website? I've looked there and couldn't find it, although I did find the quote on a number...
10	147	thanmks for the input Alyad and Cat. I have been looking for photos of lumpectomies that also removed the nipple that still look g...
11	152	Everything turned out fine with my left mx and te....I'm glad I did it as it turned out that my tumour was much larger that what the...
12	304	Hi everyone. My husband and I are trying to decide between having a lumpectomy v. mastectomy and would appreciate any exp...
13	325	As you I was dx with stage 0 DCIS in late April this year. Because of previous dx of ADH in both breasts, I opted for bilateral mas...
14	326	Hello, I am 28 years old and was diagnosed with breast cancer on 9/14/07. I am scheduled for a lumpectomy on 10/19/07. I am...
15	335	With 'widespread DCIS' it will likely be impossible to get clear margins. I had that problem in Dec. 06 and ended up going with a...

9.4.3. Exploring the resulting data to obtain analysis variable values

The comments analyzed in chapter 10 were used for the purpose of this chapter as well. After the data were preprocessed and in the final data each row presented a comment, the data were entered into Enterprise Miner. The sample node (Figure 9.2) was used to select a random sample of 20 comments (5.93%) that were read completely to evaluate how patients expressed their satisfaction and how they announced what procedure they underwent.

Figure 9.2: Flow diagram for comment sampling

Two variables needed to be created from the comments: (1) the type of procedure the person underwent and (2) whether this person was satisfied. One way to do this would be to read all the comments and to note the values of these variables. However, this method is not efficient, as it will require a lot of time. For a few comments, it is can be done but

as the number of comments increase, it becomes difficult. Here, only 10 comments were completely read and evaluated. It was noted that patients used different expressions but some similarities were present: (1) the verbal part of the expression was similar to 'had', 'chose', 'opted for', 'ended up going with', 'decided on'; (2) the procedure choice was announced with or without the article 'a' as 'mastectomy', 'mx', 'bilateral mastectomy', 'bmx', 'blmx', 'lumpectomy', 'lump'; (3) satisfaction with the surgery was recognized by the presence of words such as 'happy', 'satisfied', 'OK', 'fine'. All these words and expressions were used in a SAS index function to create the procedure type and the satisfaction in the code below.

Code 9.1: SAS code to create procedure type and satisfaction variables

```
DATA DATA;
     SET SASUSER.DATA;
     IF (INDEX(POSTS,'had mastectomy')>0
          OR INDEX(POSTS,'chose mastectomy')>0
          OR INDEX(POSTS,'prefered mastectomy')>0
          OR INDEX(POSTS,'opted for mastectomy')>0
          OR INDEX(POSTS,'ended up going with mastectomy')>0
          OR INDEX(POSTS,'decided on mastectomy')>0) THEN
GROUP=1;
          *GROUP=1 represents the mastectomy procedure;
          *repeat the code above replacing 'mastectomy' by 'mx',
          'bilateral mastectomy', 'bmx', 'blms';
          *repeat the whole resulting code adding the article
'a'
               in front of the procedure;

     IF (INDEX(POSTS,'had lumpectomy')>0
          OR INDEX(POSTS,'chose lumpectomy')>0
          OR INDEX(POSTS,'prefered lumpectomy')>0
          OR INDEX(POSTS,'opted for lumpectomy')>0
          OR INDEX(POSTS,'ended up going with lumpectomy')>0
          OR INDEX(POSTS,'decided on lumpectomy')>0) THEN
GROUP=2;
          *GROUP=1 represents the mastectomy procedure;
          *repeat the code above replacing 'lumpectomy' by
'lump';
          *repeat the whole resulting code adding the article
'a'
               in front of the procedure;
```

```
      IF (INDEX(POSTS,'happy')>0 OR INDEX(POSTS,'satisfied')>0
            OR INDEX(POSTS,'OK')>0 OR INDEX(POSTS,'ok')>0
            OR INDEX(POSTS,'fine')>0)
            THEN SATISFIED=1; ELSE SATISFIED=0;
RUN;
```

Code 9.1 covered a total of 84 comments (25%) which is more than four times the comments read completely. The results are tabulated below.

Table 9.2: Satisfaction probability estimates by procedure

Satisfied	Mastectomy [N (%)]	Lumpectomy [N (%)]
No	19 (63.33)	26 (48.15)
Yes	11 (36.67)	28 (51.85)

Clearly the percentage of satisfied individuals is higher in the lumpectomy group than in the mastectomy group. A chi-square test revealed however, that this difference is not statistically significant (p-value = 0.18).

9.5. Data - NIS

The NIS, which is an in-hospital discharge database was used to compute an estimate of the probability of the discharge status. The discharge status in NIS is recorded in a variable called DISPUNIFORM, which has the following categories [13]: (1) routine; (2) transfer to short term hospital; (5) other transfers, including skilled nursing facility, intermediate care, and another type of facility, (6) home health care, (7) against medical advice, (20) died in hospital, (99) discharge alive, destination unknown. The following code was used to create the categories corresponding to the events in the decision tree:

Code 9.2: SAS code to create the categories of the event 'discharge status' in the decision tree.

```
DATA ANALYSISDATA;
      SET SASUSER.ANALYSISDATA;
      IF DISPUNIFORM=1 THEN DISCHARGE=1;
            ELSE IF DISPUNIFORM=6 THEN DISCHARGE=2;
            ELSE IF DISPUNIFORM=20 THEN DISCHARGE=4;
            ELSE DISCHARGE=3;
RUN;
```

The results are tabulated below.

Table 9.3: Probability estimates of the event 'discharge status'

Discharge status	Mastectomy [n (%)]	Lumpectomy [n (%)]
Discharge routine	505 (75.04)	531 (78.90)
Discharge home health care	136 (20.21)	103 (15.30)
Other discharge alive	30 (4.46)	35 (5.20)
Died	2 (0.30)	4 (0.59)

The NIS was also used to compute the cost of each alternative as the total hospital charges. To compute the charges, the SAS code PROC MEANS procedure was used.

Code 9.3: code used to compute the total costs per procedure.

```
PROC MEANS DATA=ANALYSISDATA SUM;
      VAR TOTCHG;
      BY GROUP;
RUN;
```

9.6. Using probability values to fill out the decision tree

The probability estimate obtained from the analysis of comments and from the NIS data were used to fill out the decision tree (Figure 9.2).

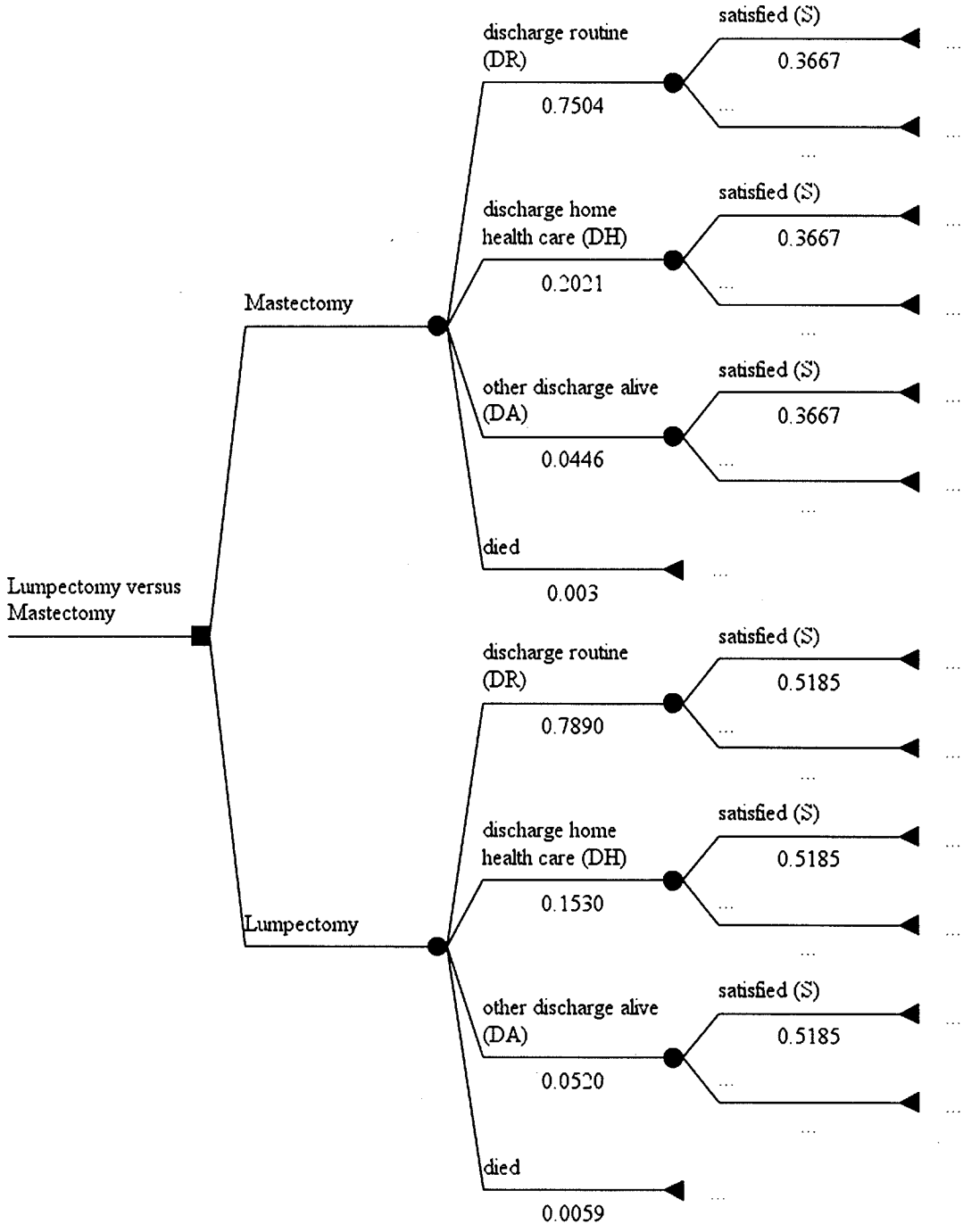

Figure 9.3: The decision tree filled with obtained probability estimates

9.7. Measure of Effectiveness

Effectiveness was measured as the number of satisfied individuals in each procedure type. To compute effectiveness, the probability estimates obtained in the tree (Figure 9.3) were used.

9.7.1. Effectiveness in the Mastectomy group

Let P represent the probability measure, S represent satisfaction, M the mastectomy group, DR the discharge routine, DH the discharge to home health care, DA the other discharge alive to define the probability of satisfaction given that the mastectomy procedure is computed below.

$$P(S|M) = P(S|DR|M) * P(DR|M) + P(S|DH|M) * P(DH|M) + P(S|DA|M) * P(DA|M)$$

$$= (0.3667)*(0.7504) + (0.3667)*(0.2021) + (0.3667)*(0.046) = 0.36615$$

The mastectomy group from the NIS data contained 673 individuals; thus, the expected number of satisfied patients in this group was

$$(673) * (0.36615) = 246$$

9.7.2. Effectiveness in the Lumpectomy group

Using the same definition as in the paragraph above and letting L represent Lumpectomy, the probability of satisfaction given that the lumpectomy procedure is computed below.

$$P(S|L) = P(S|DR|L) * P(DR|L) + P(S|DH|L) * P(DH|L) + P(S|DA|L) * P(DA|L)$$

$$= (0.5185)*(0.7890) + (0.5185)*(0.1530) + (0.5185)*(0.0520) = 0.515389$$

The lumpectomy group from the NIS data contained 673 individuals as well. In fact, a random sample of the mastectomy group was selected to match the lumpectomy group sample size of 673 (refer to chapter 7 for details). Thus the expected number of satisfied patients in this group was

(673) * (0.515389) = 347

The incremental effectiveness of lumpectomy with respect to mastectomy was found to be 347 – 246 = 101 satisfied individuals.

9.8. Measure of cost

The cost was measured computed from the NIS data and it was shown that the total charges for all the patients in the mastectomy group was $13,704,584 and in the lumpectomy group $13,647,285. Thus, the incremental cost of lumpectomy with respect to mastectomy was $13,647,285 - $13,704,584 = -$52,299

9.9. Incremental Cost Effectiveness Ratio (ICER)

The incremental cost effectiveness ration is the relative cost by the relative effectiveness. Here, it was -$52,299 / 101 = -$517.812 . This value represents the incremental cost per satisfied patient if the lumpectomy is used instead of a mastectomy. Hence, in this case, lumpectomy saves $517 for each satisfied individual in comparison to mastectomy.

9.10. Summary

Data mining sample node did a good job of sampling a small set that was representative of the population (comments). With a sample of 20 comments, it was possible to screen 84 comments for what type of procedure they underwent and whether they were satisfied. Although, 84 comments are still a small subset of the 337 comments, it was four times the size of the comments read. There is room for improvement on this method; it can be explored on how to use text mining to screen the sampled comments. This would help increase the size of the sample and thus covering a larger number of comments. Patient satisfaction is very important in measuring treatment success. This chapter provides a simple method that permits the use of open source discussion board comments to factor patients' input into analysis and especially in cost effectiveness analysis.

CHAPTER 10

USE OF TEXT MINING TO ANALYZE PATIENT OPINION WITH LUMPECTOMY OR MASTECTOMY IN THE FORM OF ONLINE COMMENT POSTING

10.1. Objective

The objective of the current section was to use online discussion boards' comments to analyze patients' satisfaction after mastectomy or lumpectomy. In today's society, the analysis of patient input goes beyond a simple participation in a survey study or filling a post-treatment survey. A survey study, though effective and still a valuable means of analysis of the patient's opinion, requires all the hustle that comes with a research study-adequate preparation, and inconvenience of a small sample size due to lack of volunteers. Very few individuals actually fill out the post-treatment satisfaction forms that are either given or sent to them. With the expansion of the internet to the actual extent, most people turn to it to express their feeling about many things including healthcare. Their comments can be about experience, satisfaction or frustration, advice, etc. Today, there are forums of about everything. Here, the forums of lumpectomy versus mastectomy are analyzed. Patient input is very important in analyzing treatment effectiveness. The use of web posts in discussion boards' forums offers a quick and easy way to factor patient opinion into effectiveness analysis.

10.2. SAS Text Miner in SAS Enterprise Miner

SAS Text Miner is an option offered by SAS Enterprise Miner to analyze text documents. In addition to being able to transform documents submitted to it into knowledge, it has the capability to crawl the web with the use of %TMFILTER macro. This macro finds the documents related to an initial website and subsequent links [59] and saves them in a directory on the local computer. These documents can then be analyzed by SAS Text Miner techniques and tools.

10.3. Data – Online comments

The comments pre-processed and analyzed in chapter 9 were used for the purpose of this chapter as well. Here, the data obtained after the preprocessing of section 9.4.2 were used.

10.4. Analysis with SAS Text Miner

The pre-processed data were entered into SAS Enterprise Miner and were analyzed with the Text Miner node (see Figure 10.1 for the process flow).

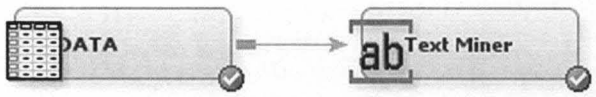

Figure 10.1: Process flow for Text Mining analysis

Most settings of the Text Miner node were left at default (Figure 10.2). The parsing was done using a synonym list provided with the SAS Enterprise Miner software package and

Text Miner was set to create clusters automatically with a total number of no more than four clusters and no more than five descriptive terms per cluster.

Parse	
Parse Variable	POSTS
Language	ENGLISH
Stop List	SASHELP.STOPLST
Start List	
Stem Terms	No
Terms in Single Document	No
Punctuation	No
Numbers	No
Different Parts of Speech	No
Ignore Parts of Speech	
Noun Groups	Yes
Synonyms	Sashelp.engsynms
Find Entities	No
Types of Entities	
Transform	
Compute SVD	Yes
SVD Resolution	Low
Max SVD Dimensions	100
Scale SVD Dimensions	No
Frequency weighting	Log
Term Weight	Entropy
Roll up Terms	No
No. of Rolled-up Terms	100
Drop Other Terms	No
Cluster	
Automatically Cluster	Yes
Exact or Maximum Number	Maximum
Number of Clusters	4
Cluster Algorithm	EXPECTATION-MAXIMIZATION
Ignore Outliers	Yes
Hierarchy Levels	.
Descriptive Terms	5
What to Cluster	SVD Dimensions

Figure 10.2: Enterprise Miner Text Miner node settings for the analysis

10.5. Results – Clusters and concept links

A total of three clusters were created out of 280 documents (83.09% of the total sample size). Most comments (36% of the 280 documents) were about things to take into considerations before making the choice (second opinion, breast cancer, local recurrence

possibility). The second big cluster (29%) contained comments about the fact that the best decision is personal. The smallest cluster (23%) contained web posts about responders who decided to have a breast reconstruction after the surgery.

Table 10.1: Clusters of the comments (Total number of documents = 280)

Cluster number	**Descriptive terms**	**N (%)**	**Cluster name**
1	Second opinion, breast cancer, local recurrence, good luck	115 (36)	Give advice on what to consider to make a choice
2	Personal decision, survival rate, invisible threads, best choice	91 (29)	Made a personal decision
3	Whole reconstruction, straight away making	74 (23)	Had breast reconstruction right after surgery

The clusters obtained reflect the main thought while making the decision. Text Miner has a tool, concept links, that helps to evaluate the strength of association between terms. Next the concept links of the terms related to mastectomy, advice and decision are analyzed.

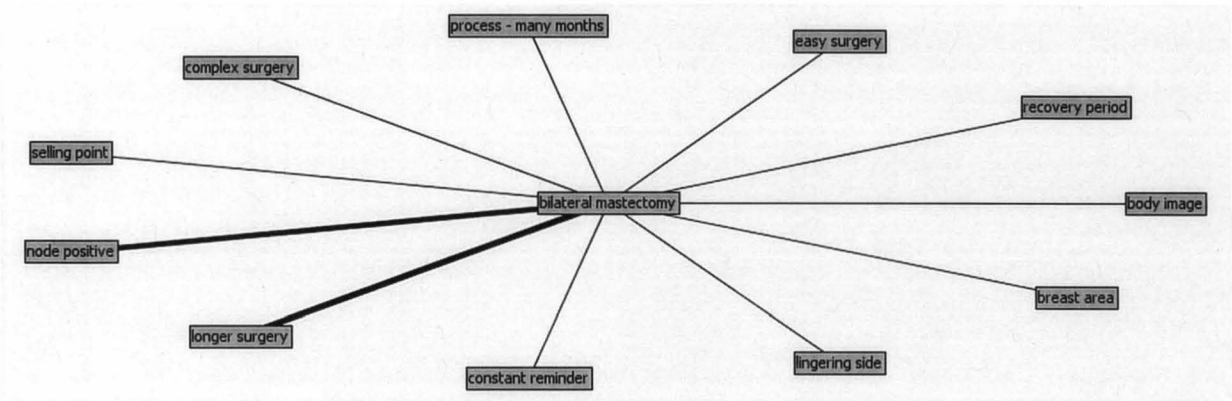

Figure 10.3: Concept links of the term 'bilateral mastectomy'

The term 'bilateral mastectomy' had a strong association with the terms 'node' and 'longer surgery'.

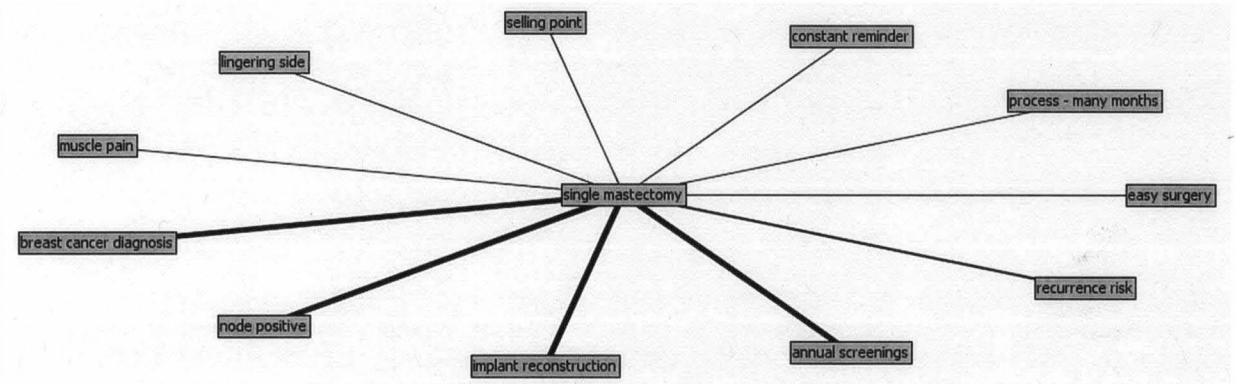

Figure 10.4: Concept links of the term 'single mastectomy'

The term 'single mastectomy' was strongly associated with 'breast cancer diagnosis', 'node positive', 'implant reconstruction' and 'annual screenings'.

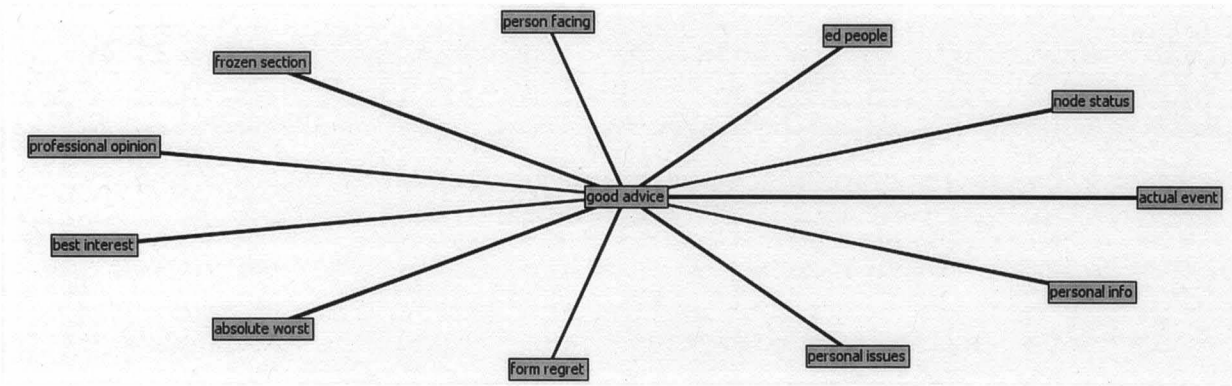

Figure 10.5: Concept links of the term 'good advice'

The term 'good advice' was strongly associated with all the terms it was linked to. However, the strongest association was with the term 'actual event'.

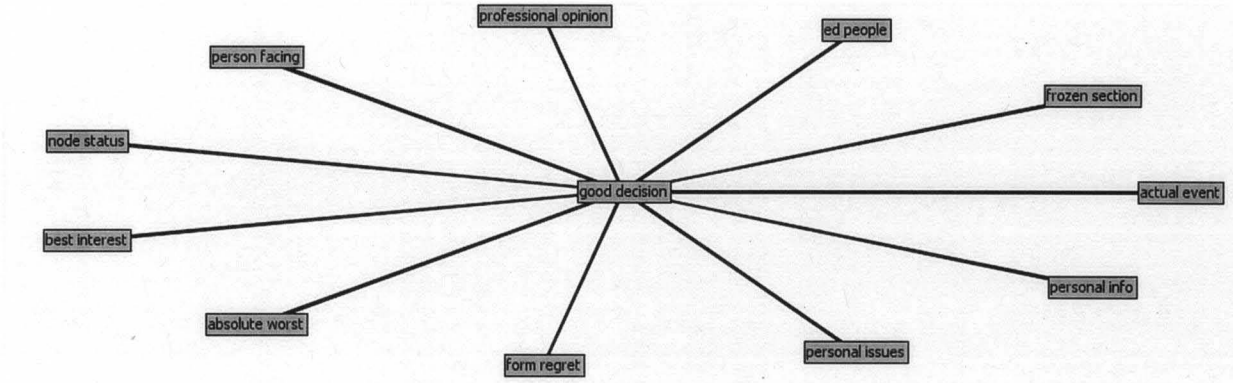

Figure 10.6: Concept links of the term 'good decision'

The term 'good decision' was also strongly associated with all the terms it was linked to. Just like the term 'good advice', the strongest association was with the term 'actual event'.

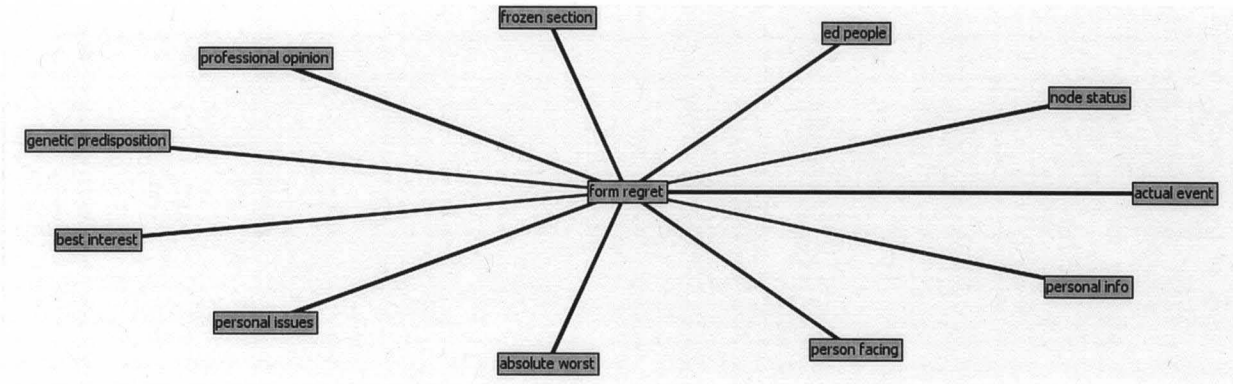

Figure 10.7: Concept links of the term 'form regret'

The term 'form regret' was strongly associated with all the terms it was linked to. These terms include: 'personal issues', 'absolute worst', and 'genetic predisposition'.

10.6. Summary

In this section, Text Miner was found to be an important tool that can be used to analyze online comments posted. It helps reduce many posts into a handful of groups gathered by the similarity of their content. Also, it helps in analyzing the links and the strengths of associations of the terms used in comments. Here, the analysis of web posts on the choice between mastectomy and lumpectomy showed that more commenters wrote about things taken into consideration to make a choice. The second largest group contained patients who advocated that the best decision is personal. The use of concept links to analyze the association of different terms showed that the term 'bilateral mastectomy' was strongly linked to the terms 'node' and 'longer surgery'.

CHAPTER 11

USE OF DECISION ANALYSIS METHODS TO EVALUATE THE LONG-TERM -10 YEARS- COMPARATIVE EFFECTIVENESS FOR LUMPECTOMY AND MASTECTOMY

11.1. Objective

The purpose of this chapter is to complement the statistical analysis with decision science to provide more details on the comparison of outcomes for mastectomy and lumpectomy. Here, the literature review is used to estimate effectiveness for the long term. The studies used are here [2-6] also presented in the literature review in chapter 3. With the use of statistical analysis, these studies concluded that lumpectomy was equivalent to Mastectomy in terms of long term overall survival as well as disease free survival. Decision analysis is based on conditional probability and its theory is reviewed in chapter 6. The effectiveness is expressed as no tumor recurrence and survival in 10-years after surgery.

11.2. The decision trees

Decision analysis was performed in terms of comparative effectiveness of the lumpectomy in contrast to mastectomy. Comparative effectiveness can be defined as a comparison of strategies in terms of their health and clinical outcomes. The effectiveness

for the current study was measured as deaths averted or local-regional recurrence averted by the use of mastectomy in comparison to lumpectomy. The trees in Figures 11.1 and 11.2 were used to evaluate the alternatives.

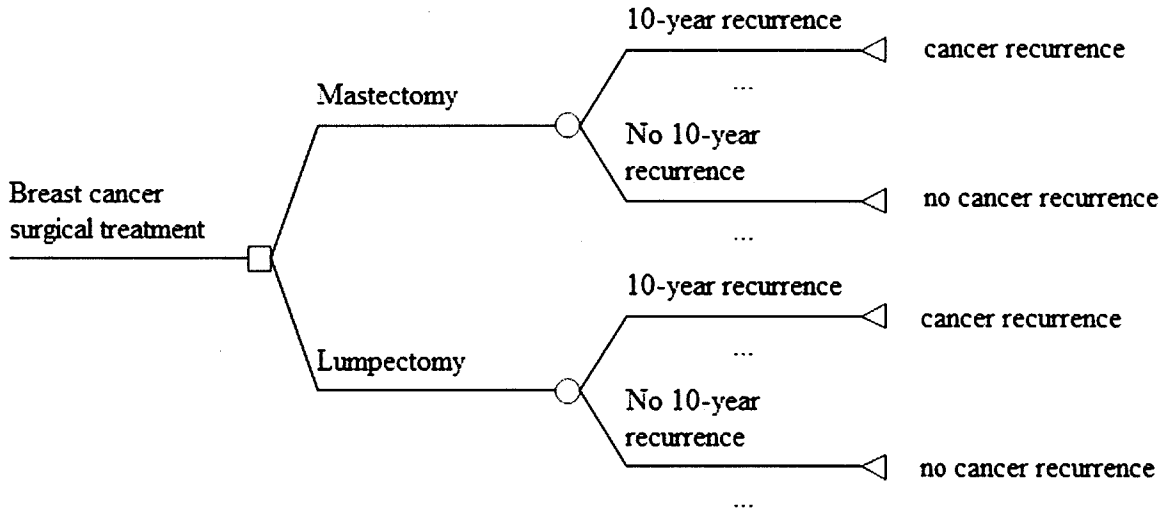

Figure 11.1: Decision tree to evaluate lumpectomy in comparison to mastectomy in terms of tumor recurrence averted.

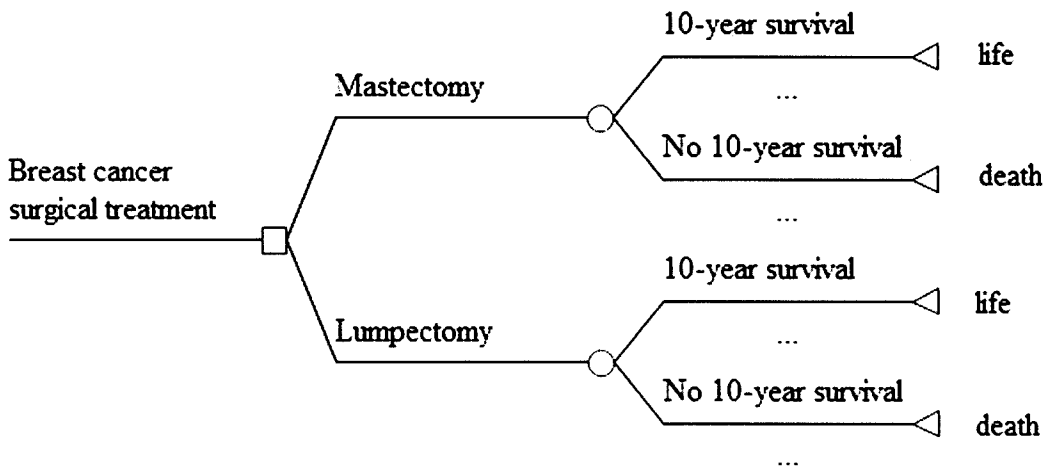

Figure 11.2: Decision tree to evaluate lumpectomy in comparison to mastectomy in terms of death averted.

11.3. Literature search strategy, selection and data extraction

11.3.1. Literature search strategy and selection

The MEDLINE (Medical Literature Analysis and Retrieval System Online) database was searched with the use of PUBMED. MEDLINE is a National library database of life and biomedical science research bibliographies [15]. It covers numerous medical fields including medicine, nursing, dentistry, veterinary medicine, the healthcare system and the pre-clinical sciences. PUBMED is a free database that accesses MEDLINE.

Of interest were studies that focused on the comparison of breast conserving surgery and mastectomy published in English between 2000 and 2010. The searching phrase was 'lumpectomy versus mastectomy in USA'. The process returned a total of 240 results. After reviewing the titles, 56 abstracts were retrieved and analyzed. Next, the abstracts were reviewed and seven articles were fully retrieved and used for the literature review and related work.

The outcome of interest for the current analysis was 10-year mortality and 10-year recurrence. Eligible for the data abstraction were articles that reported the number of individuals who died and the number of recurrences. Other selection criteria included study design, study population, sample size and main outcome of interest. Studies that were not clinical trials and studies that did not report percentages for mortality or recurrence were excluded. Finally, five articles were considered to be relevant for the current analysis.

11.3.2. Data extraction

For these studies, the following data were extracted: the total sample size and the size of each arm, the follow-up time, the recurrence and the mortality percentages. These data were used to estimate effectiveness in the analysis. Table 11.1 provides details on these articles and on the data extracted.

Table 11.1: Studies used in the Comparative Effectiveness Analysis

Study reference and size	**Follow-up time**	**Group**	**Group size**	**Local-regional recurrence (%)**	**Death (%)**
[6] (n=237)	About 20 years	Mastectomy	116	33	42
		BCS	121	37	48
[2] (n=1851)	About 20 years	Total mastectomy	589	37.2	Not available
		Lumpectomy	634	42.4	
		Lumpectomy and irradiation	628	34.1	
[3] 6 (n=701)	About 20 years	Radical mastectomy	348	8	Not available
		BCS	352	30	
[4] (n=868)	10 years	Mastectomy	428	12	34
		BCS	448	20	35
[5] (n=237)	Median 121 months (about 10 years)	Mastectomy	116	10	25
		Lumpectomy and Radiation	121	5	23

*BCS: Breast conservation surgery

In study [2], the local-regional recurrence results were presented in terms of individuals who developed a recurrent tumor. Corresponding percentages were computed and presented here.

Only two groups were considered for the current study, the mastectomy and lumpectomy. Some articles used in the current project, designed their studies with three groups. For these studies, groups that involved the same main procedure were merged and results were estimated by pooled values.

The follow-up time was chosen to be 10 years. Hence, all the 20-year results were adjusted. The 10-year percentages were estimated by half of the 20-year percentages. Table 11.2 contains the values from pooled, adjusted results for studies [6], [2] and [3].

Table 11.2: Adjusted data for studies [6], [2] and [3]

Study reference	**Group**	**Local-regional recurrence (%)**	**Death (%)**
[6]	Mastectomy	16.5	21
	BCS	18.5	24
[2]	Total mastectomy	18.6	Not available
	Lumpectomy	19.13	
[3]	Radical mastectomy	4	Not available
	BCS	15	

11.3.3. Computation of probability estimates

From the data presented in Tables 11.1 and 11.2, a pooled population of 3894 patients in total was obtained, among which, 1598 (41.04%) underwent mastectomy (see Table 11.3).

Table 11.3: Estimated group sizes of the pooled data for comparative effectiveness

	Size	**Percentage**
Total	3894	100
Mastectomy	1598	41.04
Lumpectomy	2300	59.06

Expected mortality and local and regional recurrences associated with each surgery treatment were considered for the current analysis and they were computed using data values summarized below from Tables 11.1 and 11.2.

Table 11.4: Summary of probability estimates from the literature

Study reference	Mastectomy			Lumpectomy		
	Size	Recurrence	Mortality	Size	Recurrence	Mortality
[6]	116	0.165	0.21	121	0.18	0.24
[2]	589	0.186	NA	1262	0.19	NA
[3]	348	0.4	NA	352	0.15	NA
[4]	428	0.12	0.34	448	0.20	0.35
[5]	116	0.10	0.25	121	0.5	0.23

11.4. Results

11.4.1. Effectiveness in terms of 10-year local/regional recurrence: Expected local/regional recurrences

The expected probability of local/regional recurrence of breast cancer was calculated from the data in Table 11.4.

Mastectomy: (116*0.165 + 589*0.186 + 349*0.4 + 428*0.12 + 116*0.10) / (116 + 589 + 349 + 428 + 116) = 0.1287

Lumpectomy: (121*0.18 + 1262*0.19 + 352*0.15 + 448*0.20 + 121*0.5) / (121 + 1262 + 352 + 448 + 121) = 0.2026

The difference in expected probability of local and regional recurrence of breast cancer in 10 years after surgery treatment between lumpectomy and mastectomy is 0.2026 – 0.1287 = 0.0739. That is in 10 years, 739 women out of 10,000 would be expected to not have recurrent breast cancer by undergoing mastectomy instead of lumpectomy.

11.4.2. Effectiveness in-terms of 10-year mortality: Expected deaths

The expected probabilities of death from breast cancer in the mastectomy and lumpectomy groups are computed below. Only data extracted from studies [6], [4] and [5] were used.

Mastectomy: (116*0.21 + 428*0.34 + 116*0.25) / (116 + 428 + 116) = 0.3013

Lumpectomy: (121*0.24 + 448*0.35 +121*0.23) / (121 + 448 + 121) = 0.3097

The difference in expected probability of death from breast cancer in the 10 years after surgical treatment between lumpectomy and mastectomy is 0.3097 – 0.3013 = 0.0084. Thus, an expected number of 84 deaths are prevented in 10 years per 10,000 breast cancer patients who undergo mastectomy.

11.5. Summary

In summary, decision science complements statistical analysis here, in providing the number of deaths and recurrences averted by the alternative procedure. In a population of 10,000 women undergoing surgery for breast cancer, 10 years down the road, mastectomy was found to prevent 739 recurrences and 84 deaths. While statistical analysis provides significance or no significance, the use of decision science was found to give more specific details to the decision maker.

CHAPTER 12

DISCUSSION AND CONCLUSION

12.1. Overview

In the current study, statistical methods complemented with data mining, decision analysis and cost effectiveness analysis methods were used to compare lumpectomy to mastectomy in administrative data, literature review data and online comment postings. The comparison was made to determine whether lumpectomy outcomes statistically differ from mastectomy outcomes, to evaluate if a simple predictive model could be constructed to predict post-operative hospital admission within 90 days after surgery using type of procedure as input and to evaluate patient satisfaction with the procedure. Traditional statistical analyses methods were used to test the differences between groups of procedures. Data mining methods were utilized to create and evaluate the performance of a predictive model. Decision analysis techniques were used to evaluate the effectiveness of lumpectomy in comparison to the effectiveness of mastectomy on a 10 year period. Cost effectiveness methods were used to evaluate the incremental cost per satisfied patient of lumpectomy in comparison to mastectomy. As a result, primary and secondary analyses were performed using the NIS data, the Thomson Reuter's Market Scan data, data from the medical literature and data from online postings in breast cancer forums.

12.2. Description of findings

The NIS data was used to compare health and clinical outcomes during the hospitalization for the surgery, the MarketScan longitudinal data was used to compare clinical and healthcare utilization outcomes after the surgery for a period of 249 days average, the literature was used to assess the differences in health outcomes 10 years after the surgical treatment and the online comments were used to explore patient satisfaction after surgery.

With the use of statistical methods, it was found that at the procedure hospitalization, patients who had lumpectomy had a significantly longer stay (p-value < 0.0001), higher hospital charges (p-value: 0.02) but similar in-hospital death (p-value: 0.42). The analysis of longitudinal data, with statistical methods, showed that post-operative healthcare utilizations were similar in both surgery groups. The use of a predictive model, chosen by data mining to be the best, revealed that patients who undergo lumpectomy are twice more likely to have at least one hospital admission in 90 days after surgery (p-value: 0.0449). Lumpectomy, or in general breast conservation surgery, has been extensively compared to mastectomy in clinical trials in terms of disease free survival and overall survival [2-7, 34-37]. Their conclusions have been concordant; lumpectomy was found to be comparable to mastectomy in terms of these outcomes. Methods used for data analysis were statistical models, in particular survival analysis models. The current analysis did not compare lumpectomy to mastectomy in terms of long term survival. Thus, a contrast of these studies to the current one cannot be performed in these terms. Instead, a long term comparison was made in terms of deaths in recurrence averted using decision analysis methodologies. It was found that in 10 years, an estimate of 84 would be

prevented death and 739 would be prevented to have recurrence by the use of mastectomy on 10,000 breast cancer women in contrast to lumpectomy. No studies are available that compared lumpectomy to mastectomy in terms of immediate and short term follow-up health outcomes after surgery using real world data. Numerous studies compared mastectomy and breast conservation surgery in terms of cost. Barlow et al. [41] found that breast conservation surgery had higher short-term ($\leq$ 1 year) costs but lower long-term ($\geq$ 5 years) after diagnosis. Polsky et al. [42] found that after surgery, providing a choice is economically attractive in comparison to restricting the choice to a single therapy. In the current study, it was found that short after surgery (average 8 months), mastectomy was associated with higher post-operative hospital charges, outpatient services and medications. The difference of these results with earlier published reports may be explained by the small sample size of the current study and a relatively short follow-up time. In addition, the capitation status was unknown for all the patients in the cohort.

The analysis of online comments from breastcancer.org discussion board forums revealed that a high number of posts (36%) were from patients who were advising on what to consider making a choice (second opinion, local recurrence). In terms of patient's choice, Kirby [39] found that many patients, given the choice between breast conserving surgery and mastectomy, choose mastectomy because they feel much safer and/or want to decrease the risk of further surgery. In the current study, it was found that patients express that the choice is personal and that local recurrence is one factor for making a decision.

An analysis of the patient satisfaction reveals that only 36.67% of the patients who indicate that they underwent mastectomy were satisfied while up to 51.85% of the patients who indicate that they underwent lumpectomy were satisfied. A cost effectiveness analysis showed that lumpectomy saved $517 per satisfied patient in comparison to mastectomy.

12.3. Implications

12.3.1. Implications in data analysis

In the medical and public health literature, statistical methods are the ones mostly used for research. They perform well in clinical trial data but are limited when it comes to large administrative databases. Yet, a look in recent published reports shows that the use of administrative data to answer some key research questions is increasing. Methods used in the current study will provide statisticians in the medical and public health field as a way to augment data mining to statistical methods. Also, the algorithm used to evaluate patient satisfaction and find the type of procedure in posted comments will help researchers with limited funds conduct studies on patient opinion.

12.3.2. Implications in breast cancer surgical treatment

Every woman is at risk of breast cancer and the risk of developing breast cancer increases with age [10]. Medical science is evolving and today, more than ever before, women have many options of treatment sequence. Surgery remains, however, the core line of treatment action. Breast conserving surgery and specifically lumpectomy have been

proven to have the same effect as mastectomy in long-term disease free and overall survival. Nevertheless, its comparative effectiveness in health outcomes and healthcare resources use in the period following surgery is still to be discussed. The current study will help patients and their families as well as healthcare providers and those financially responsible in their choice decision making and their preparation to enter the treatment phase.

12.4. Limitation

12.4.1. Limitation related to the methods and the analysis

Although the predictive modeling comparison isolated a best model in terms of performance, the chosen model was still weak with only a c-statistic of 0.58 when a desirable value is at least 0.7. This model needs to be improved. Inclusion of cancer characteristics may help improve the performance.

The method used to extract the type of procedure and whether satisfied covered only 25% of the comments. This low percentage can partially be explained by the fact that some patients did not specifically say what procedure they had. Nevertheless, these methods need to be improved to rely completely on the software and less on the user reading comments.

12.4.2. Limitation related to the data and the variables

The analysis used administrative data in which patients were selected using ICD-9-CM codes and CPT-4 codes. An assumption that these were accurate and correctly entered

was made. For the follow-up analysis, the study population comprised individuals and their dependents with employer sponsored insurance. This portion of the population is more likely to be younger and relatively healthier. Hence, the results should be cautiously interpreted in terms of generalizability to the entire US population. In addition, these data lack information on some demographics variables such as race. Also, these data do not have information on breast cancer characteristics such as staging and class which have been found to influence the choice of surgical procedure and may influence the clinical resource usage as well as charges. Another limitation encountered associated to the data is its nature. The patients are not randomly assigned to procedures; which introduces selection bias to the analysis. Propensity score matching or other matching techniques that attempt to mimic the randomization processes could be used to address this problem if the dataset available is initially large enough.

The analysis mainly considers all-cause hospital usage, all-cause clinical usage and long-term all-cause mortality. A more sophisticated analysis may wish to adjust the cause of the hospital usage, the cause of the clinical usage and the cause of death and to compare the treatments with respect to these causes. It was not known which capitation status the patient's insurance plans were. Hence, the cost analysis may be affected by the fact that some patients have capitated insurance plans while others did not.

12.5. Contribution of the current study to research

12.5.1. Contribution to comparative effectiveness analysis

In the field of comparative effectiveness, the current study innovates the current techniques by giving methodologies to supplement statistical models with data mining, decision analysis and cost effectiveness analysis methods to obtain a complete picture of the comparison containing short and long term outcomes when administrative data are used.

12.5.2. Contribution to breast cancer surgical treatment

The current analysis provides a complement to the results already available in the area of comparison of lumpectomy to mastectomy. It is already established that lumpectomy is equivalent to mastectomy for breast cancer surgical treatment in terms of long term overall survival and disease-free survival. The current study complements these results by providing the following comparison:

(1) Hospital usage during the surgery

(2) Healthcare resource use (hospital admissions, outpatient services, prescribed medications) and charges for up to eight months after surgery

(3) Post-operative hospital re-admission in 90 days

(4) Patient satisfaction

(5) Deaths and recurrences prevented in 10 years

(6) Cost effectiveness each day after surgery in the payer's perspective

12.6. Areas of future research

12.6.1. Further research in methodologies

Further research regarding use of data mining in comparing lumpectomy to mastectomy includes construction of a predictive model for repeat operation. For this purpose, data providing a long follow-up are needed and if cancer characteristics are not available, cluster analysis can be used to group patients in different breast cancer categories. An improvement of the algorithm to analyze online comments should be done. Text mining should be explored and methods developed to obtain detailed information such as type of procedure and/or satisfaction, without the user having to read any comment.

12.6.2. Further research in health outcomes

Further research regarding the comparison of the breast cancer surgical treatments will warrant a deeper analysis of these procedures. To accomplish this, a larger sample and a longer follow-up time should be considered and the procedure groups should be compared adjusting for important breast cancer characteristics, including but not limited to staging. Moreover, propensity score matching should be used to obtain comparative groups in order to achieve a conclusion with limited bias and confounding. The repeat operation should be analyzed to evaluate the group with a greater risk of undergoing a subsequent surgery. Also, a better and more accurate post-operative hospital admission predictive model should be constructed using patient demographics, breast cancer characteristics, hospital characteristics, pre-operative disease cluster and comorbidity condition. A decision analysis and a cost effectiveness analysis should be performed from the analysis data. Further research also includes the analysis of access to breast cancer surgical treatment, looking especially at age disparities since new data show that some patients are denied surgery because they may be too old [75].

12.7. Summary of the current study

12.7.1. Summary of methods

The use of administrative data for medical and public health research is increasing and due to its low cost, it is taking the place once held by clinical trials. These data are now very large and more than ever, it is necessary to develop methods to analyze them adequately. The main objective of this study was to provide methodologies and statistical models, supplemented by data mining, decision analysis and cost effectiveness methods to analyze these data.

Statistical models are still of great importance when the data are processed, cleaned and transformed right. They do come short though, when the assumptions cannot be satisfied, the randomness required cannot be assumed or achieved through some mimic processes and when the data are just too large and statistical tests are more likely to be significant no matter what.

Data mining looks promising in this research field. It can help accomplish much more than statistical analyses since it does not make any assumptions. It has exploratory capabilities; it can generate hypotheses and test them at the same time. The interest of using data mining in medical and public health research is growing.

Decision analysis and Cost Effectiveness methods have been used in medical and healthcare research before. However, this is the first time effectiveness is measured using patient opinion expressed through discussion boards.

12.7.2. Summary of results

Lumpectomy and mastectomy are the two main options available as a first line for breast cancer treatment episode, especially in the case of early stage diagnosis. It is of great importance for the patients that an optimal choice be made to maximize their short term and long term outcomes. The purpose of this study was to compare lumpectomy to mastectomy in peri and post-operative health and clinical outcomes as well as healthcare resources use. The ultimate goal was to complement published reports that have evaluated the long-term outcomes, but do not consider short-term outcomes. Mastectomy is the removal of the whole infected breast while lumpectomy is a minimally invasive procedure during which only the tumor and surrounding healthy tissues are removed from the infected breast. Lumpectomy was introduced and recommended for early stages as an alternative to mastectomy given that it was shown to achieve the same survival effectiveness. Thus, the one objective of the current study was to analyze whether this effectiveness translates into real world cases in short term outcomes; in other words, whether lumpectomy was as beneficial as mastectomy shortly after surgery as well. This goal was realized by evaluating the differences in these procedures and testing these differences for statistical significance. In addition, a long-term comparative effectiveness was performed to evaluate the deaths and tumor recurrences. Overall, this study found that lumpectomy was associated with longer in-hospital stay, higher hospital charges, higher in-hospital death rates and a higher risk of 90-day post-operative hospitalization. It was also found that, in a period of 10 years after surgery, mastectomy would prevent deaths and recurrences if used instead of lumpectomy. In terms of cost, lumpectomy saved money per satisfied patient and more patients who underwent lumpectomy were satisfied compared to those who underwent mastectomy.

12.8. Conclusion

The current study encountered limitation in model development: a weak predictive model, a small sample extracted from the comments online. Other limitations were due to the use of administrative and insurance data, the lack of cancer characteristics, an overall small size, a short follow-up period and lack of information on capitation status of the insurance plans. These limitations imply that the methods should be applied and the results should be interpreted with caution. Despite these limitations, this analysis provides useful methodologies for administrative and web data and important information on the existing debate of the comparison of lumpectomy and mastectomy in health outcomes. A more in depth analysis considering a larger sample size, a longer post-operative follow-up time and adjusted procedure groups would lead to better methods and to a better understanding of the surgery type effect on immediate and follow-up post-operative outcomes.

REFERENCES

1. **A Brief History of Breast Cancer** [http://ezinearticles.com/?A-Brief-History-of-Breast-Cancer&id=609519]

2. Fisher B, Anderson S, Bryant J, Margolese R, Deutsch M, Fisher E, Jeong J, Wolmark N: **Twenty-Year Follow-up of a Randomized Trial Comparing Total Mastectomy, Lumpectomy, and Lumpectomy plus Irradiation for the Treatment of Invasive Breast Cancer**. *N Eng J Med* 2002, **347**(16):1233 - 1241.

3. Veronesi U, Cascinelli N, Mariani L, Greco M, Saccozzi R, Luini A, Aguilar M, Marubini E: **Twenty-Year Follow-up of a Randomized Study Comparing Breast-Conserving Surgery with Radical Mastectomy for Early Breast Cancer**. *New England Journal of Medicine* 2002, **347**(16):1227-1232.

4. van Dongen JA, Voogd AC, Fentiman IS, Legrand C, Sylvester RJ, Tong D, van der Schueren E, Helle PA, van Zijl K, Bartelink H: **Long-Term Results of a Randomized Trial Comparing Breast-Conserving Therapy With Mastectomy: European Organization for Research and Treatment of Cancer 10801 Trial**. *Journal of the National Cancer Institute* 2000, **92**(14):1143-1150.

5. Jacobson JA, Danforth DN, Cowan KH, d'Angelo T, Steinberg SM, Pierce L, Lippman ME, Lichter AS, Glatstein E, Okunieff P: **Ten-Year Results of a Comparison of Conservation with Mastectomy in the Treatment of Stage I and II Breast Cancer**. *New England Journal of Medicine* 1995, **332**(14):907-911.

6. Poggi MM, Danforth DN, Sciuto LC, Smith SL, Steinberg SM, Liewehr DJ, Menard C, Lippman ME, Lichter AS, Altemus RM: **Eighteen-year results in the treatment of early breast carcinoma with mastectomy versus breast conservation therapy**. *Cancer* 2003, **98**(4):697-702.

7. Kroman N, Holtveg H, Wohlfahrt J, Jensen M-B, Mouridsen HT, Blichert-Toft M, Melbye M: **Effect of breast-conserving therapy versus radical mastectomy on prognosis for young women with breast carcinoma**. *Cancer* 2004, **100**(4):688-693.

8. **Breast Cancer epidemiology** [http://www.news-medical.net/health/Breast-cancer-Epidemiology.aspx]

9. **Breast Cancer -From Wikipedia, the free encyclopedia** [http://en.wikipedia.org/wiki/Breast_cancer]

10. **Breast cancer: Early detection, diagnosis, and staging topics** [http://cancer.org/cancer/breastcancer/detailedguide/breast-cancer-staging]

11. **Breast cancer- definition of breast cancer in Medical dictionary** [http://medical-dictionary.thefreedictionary.com/breast+cancer]

12. **Nationwide Inpatient Sample**. HCUP Central Distributor; 2005.

13. **Thomson Reuters MarketScan database**.

14. **MEDLINE** [http://en.wikipedia.org/wiki/MEDLINE]

15. Nisbet R, Elder J, Miner G: *Handbook of Statistical Analysis and Data Mining Applications*. California: Elsevier Inc.; 2009.

16. Williams PK: **A Clustering Rule Based Approach for Classification Problems**. *Dissertation.* Auburn University, Computer Science and Software Engineering; 2010.

17. Petitti DB: *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for quantitative synthesis in Medicine*. Second edition. New York: Oxford University Press; 2000.

18. **Decision Analysis** [http://en.wikipedia.org/wiki/Decision_analysis]

19. **Cost Effectiveness Analysis** [http://en.wikipedia.org/wiki/Cost-effectiveness_analysis]

20. Gold MR, Siegel JE, Russell LB, Weinstein MC: *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press, Inc.; 1996.

21. **Breast Cancer -By category** [http://breastcancer.about.com/od/definition/a/bc_definition.html]

22. **Illustration of the mammary gland**.

23. **Types of Breast Cancer** [http://www.nationalbreastcancer.org/About-Breast-Cancer/Types.aspx]

24. **Cancer staging** [http://en.wikipedia.org/wiki/Cancer_staging]

25. **Cancer** [http://www.answers.com/topic/cancer]

26. **Histologic Grades of Breast Cancer: Helping Determine a Patient's Outcome** [http://www.imaginis.com/breast-health/histologic-grades-of-breast-cancer-helping-determine-a-patient-s-outcome-2]

27. **Hormone Receptor Status and Diagnosis - Estogen and Progesterone** [http://breastcancer.about.com/od/diagnosis/p/hormone_status.htm]

28. **HER2 and breast cancer** [http://www.macmillan.org.uk/Cancerinformation/Cancertypes/Breast/Symptomsdiagnosis/HormoneandHER2receptors/HER2andbreastcancer.aspx]

29. Ross J, Linette G, Stec J, Ross M, Anwar S, Boguniewicz A: **DNA ploidy and cell cycle analysis in breast cancer**. *American journal of clinical pathology* 2003, **120 Suppl**:S72-S84.

30. **How To Detect Early Breast Cancer Signs And Symptoms** [http://www.neomatrix.com/breast-cancer-signs-and-symptoms.aspx]

31. **Risk Factors for Breast Cancer** [http://www.webmd.com/breast-cancer/guide/overview-risks-breast-cancer]

32. **Surgery** [http://www.breastcancer.org/treatment/surgery/]

33. **Cancer Advances in focus** [http://www.cancer.gov/cancertopics/factsheet/cancer-advances-in-focus/breast]

34. veronesi U, Banfi A, Saccozzi R, Salvadori B, Zucali R, Uslenghi C, Greco M, Luini A, Rilke F, Sultan L: **Conservative treatment of breast cancer. A trial in progress at the Cancer Institute of Milan.** *Cancer* 1977, **39**(6):2822-2826.

35. Veronesi U, Saccozzi R, Del Vecchio M, Banfi A, Clemente C, De Lena M, Gallus G, Greco M, Luini A, Marubini E *et al*: **Comparing Radical Mastectomy with Quadrantectomy, Axillary Dissection, and Radiotherapy in Patients with Small Cancers of the Breast**. *New England Journal of Medicine* 1981, **305**(1):6-11.

36. Fisher B, Anderson S, Redmond CK, Wolmark N, Wickerham DL, Cronin WM: **Reanalysis and Results after 12 Years of Follow-up in a Randomized Clinical**

Trial Comparing Total Mastectomy with Lumpectomy with or without Irradiation in the Treatment of Breast Cancer. *New England Journal of Medicine* 1995, **333**(22):1456-1461.

37. Fisher B, Redmond C, Poisson R, Margolese R, Wolmark N, Wickerham L, Fisher E, Deutsch M, Caplan R, Pilch Y *et al*: **Eight-Year Results of a Randomized Clinical Trial Comparing Total Mastectomy and Lumpectomy with or without Irradiation in the Treatment of Breast Cancer**. *New England Journal of Medicine* 1989, **320**(13):822-828.

38. Martin M, Meyricke R, O'Neill T, Roberts S: **Mastectomy or breast conserving surgery? Factors affecting type of surgical treatment for breast cancer - a classification tree approach**. *BMC Cancer* 2006, **6**(1):98.

39. Kirby R, Basit A, Manimaran N: **Patient choice significantly affects mastectomy rates in the treatment of breast cancer**. *International Seminars in Surgical Oncology* 2008, **5**(1):20.

40. Fallowfield LJ, Hall A, Maguire GP, Baum M: **Psychological Outcomes Of Different Treatment Policies In Women With Early Breast Cancer Outside A Clinical Trial**. *BMJ: British Medical Journal* 1990, **301**(6752):575-580.

41. Barlow WE, Taplin SH, Yoshida CK, Buist DS, Seger D, Brown M: **Cost Comparison of Mastectomy Versus Breast-Conserving Therapy for Early-Stage Breast Cancer**. *Journal of the National Cancer Institute* 2001, **93**(6):447-455.

42. Polsky D, Mandelblatt JS, Weeks JC, Venditti L, Hwang Y-T, Glick HA, Hadley J, Schulman KA: **Economic Evaluation of Breast Cancer Treatment:**

Considering the Value of Patient Choice. *Journal of Clinical Oncology* 2003, **21**(6):1139-1146.

43. Freedman AN, Seminara D, Gail MH, Hartge P, Colditz GA, Ballard-Barbash R, Pfeiffer RM: **Cancer Risk Prediction Models: A Workshop on Development, Evaluation, and Application**. *Journal of the National Cancer Institute* 2005, **97**(10):715-723.

44. Gail MH: **Personalized estimates of breast cancer risk in clinical practice and public health**. *Statistics in Medicine* 2011, **30**(10):1090-1104.

45. Ottman R, Pike R, King M, Henderson B: **Practical guide for estimating risk for familial breast cancer**. *Lancet* 1983, **2**(8349):556-558.

46. Anderson D, Badzioch M: **Risk of familial breast cancer**. *Cancer* 1985, **56**:383-387.

47. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ: **Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually**. *Journal of the National Cancer Institute* 1989, **81**(24):1879-1886.

48. Rosner B, Colditz GA: **Nurses' Health Study: Log-Incidence Mathematical Model of Breast Cancer Incidence**. *Journal of the National Cancer Institute* 1996, **88**(6):359-364.

49. Tyrer J, Duffy SW, Cuzick J: **A breast cancer prediction model incorporating familial and personal risk factors**. *Statistics in Medicine* 2004, **23**(7):1111-1130.

50. Wackerly DD, III WM, Scheaffer RL: *Mathematical Statistics with applications*. 6 edition. California: Duxbury advanced series; 2002.

51. Rosner B: *Fundamentals of Biostatistics*. 6 edition. California: Duxbury; 2006.

52. **Mathematical Statistics** [http://en.wikipedia.org/wiki/Mathematical_statistics]

53. Walker GA: *Common Statistical Methods for Clinical Research with SAS examples*. Second edition. Cary, NC: SAS Institute, Inc.; 2002.

54. Hocking RR: *Methods and Applications of Linear Models; Regression and Analysis of Variance*. 2 edition. New Jersey: John Wiley & sons, Inc.; 2003.

55. Petrou CS: **Use of Text Mining to predict patient compliance**. *Dissertation.* University of Louisville, Department of Mathematics; 2008.

56. Sahoo P: **Probability and Mathematical Statistics**. Louisville, KY: University of Louisville; 2008.

57. Agresti A: *Categorical Data Analysis*. Second edition. New Jersey: John Wiley & sons, Inc.; 2002.

58. Klein JP, Moeschberger ML: *Survival Analysis; Techniques for Censored and Truncated Data*. Second edition. New York: Springer; 2003.

59. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*. New York: Springer; 2001.

60. Gill R: **Methods of Classification**. Louisville, KY: University of Louisville; 2008.

61. Myatt GJ, Johnson WP: *Making sense of data II: A practical guide to data mining methods and applications*. Hoboken, New Jersey: John Wiley & sons, Inc.; 2009.

62. Bramer M: *Principles of data mining (Undergraduate topics in computer science)*. London: Springer-Verlag; 2007.

63. Cerrito P: *Introduction to Data Mining using SAS Enterprise Miner*. Cary, NC: SAS Institute, Inc.; 2006.

64. SAS: **SAS Enterprise Miner User's guide**. Cary, NC: SAS Institute, Inc.

65. Fox PD: **A Theory of Cost-Effectiveness for Military Systems Analysis**. Menlo Park, California: Stanford Research Institute; 1964.

66. Drummond MF, O'Brien B, Stoddart GL, Torrance GW: *Methods for the Economic Evaluation of Health Care Programmes*. Second edition. New York: Oxford University Press; 1997.

67. Muenning P: *Cost-Effectiveness Analysis in Health: A practical approach.* Second edition. New Jersey: John Wiley & sons, Inc.; 2008.

68. Nowrouzi F: **Cost Shifting of the Drug-Eluting Stent**. *Dissertation.* University of Louisville, Department of Mathematics; 2007.

69. Abraham M, Ahlam JT, Boudreau AJ, Connelly JL, Evans DD, Glenn RL, Green G, Hayden D, Kotowicz GM, Lumakovska E *et al*: *2011 CPT. Current Procedural Terminology*. Professional Edition edition: 2010 American Medical Association; 2011.

70. Charlson M, Pompei P, Ales K, MacKenzie C: **A new method of classifying prognostic comorbidity in longitudinal studies: development and validation**. *Journal of chronic diseases* 1987, **40**(5):373-383.

71. Deyo R, Cherkin D, Ciol M: **Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases**. *Journal of clinical epidemiology* 1992, **45**(6):613-619.

72. Inc. SI: **SAS**. Cary, NC, USA.

73. Cody R: *Learning SAS by Example: A Programmer's Guide*. Cary, NC: SAS Institute Inc.; 2007.

74. Steyerberg EW: *Clinical Prediction Models*. NY: Springer Science+Business Media, LLC; 2009.

75. **The breast cancer patients TOO OLD to save: Thousands are being denied surgery by 'ageist' doctors** [http://www.dailymail.co.uk/health/article-2004040/Breast-cancer-Thousands-denied-life-saving-surgery-doctors-base-treatment-age.html#]

CURRICULUM VITAE

Beatrice Ugiliweneza

8503 Roseborough Road

Louisville, KY 40228

502-224-6777 (celphone)

ubeatric@yahoo.com, b0ugil01@louisville.edu

Education

Jan. 2007- Dec. 2011 **University of Louisville,** Louisville, KY

PhD/ MSPH (anticipated)

PhD in Applied and Industrial Mathematics/ Master of Public Health in Biostatistics

Applied Mathematics dissertation topic

Use of statistical analysis, data mining, decision analysis and cost effectiveness analysis to analyze medical data: Application to comparative effectiveness of lumpectomy and mastectomy for breast cancer

Public Health-Biostatistics thesis topic

Determining the most efficient way in which to manage Congestive Heart Failure Patients

Overall GPA: 3.949/4.0

Jan. 2003-Mar. 2006 **Université de Niamey,** Niamey, NIGER (WEST AFRICA)

Bachelor of Science in Mathematics

Profile

Excellent understanding of Data analysis, Data Mining and Predictive Modeling

Good perceptiveness of Data Forecasting and Health Care program economic evaluation

Strong research and reporting abilities

Polyglot: English, French, Swahili, Kinyarwanda

Computer skills

Familiar with Microsoft Word, Excel, Power Point, SAS, SAS Enterprise Miner

Basic knowledge of SPSS, R

Awards

2008 SAS Scholarship winner for the MWSUG2008 conference

2009 SAS Scholarship winner for the MWSUG2009 conference

Experience

Dec. 2010 – Present **University of Louisville,** Department of Neurosurgery, Louisville, KY

Program Analyst/Statistician

Aug. 2007-Dec. 2010 **University of Louisville,** Department of Mathematics, Louisville, KY

Graduate Teaching Assistant

Publications

Feb. 2010 Ugiliweneza, B. **Book Chapter**. Analysis of Breast Cancer and Surgery as a treatment option. In: Cerrito, P. (Editor), *Cases on Health Outcomes and Clinical Data Mining: Studies and Frameworks.* Hershey, PA: IGI Publishing. 2010.

Paper presentation

Nov. 2010 **APHA 2010 (anticipation),** Denver, CO

Administrating TDaP during pregnancy increases a Newborn's Protection against Pertussis, Diphtheria and Tetanus

May 2010 **ISPOR2010 OUTCOMES RESEARCH DIGEST,** Atlanta, GA

Analysis of health care outcomes for Congestive Heart Failure (CHF) patients

Mar. 2010 **KPHA 2010,** Louisville, KY

Optimal Management of Congestive Heart Failure patients

Oct. 2009 **MWSUG2009,** Cleveland, OH

Analysis of breast cancer and surgery as treatment option

Oct. 2008 **MWSUG2008,** Indianapolis, IN

Mastectomy versus Lumpectomy in breast cancer treatment

Mar. 2008 **SAS Global Forum 2008,** San Antonio, TX

Analysis of breast cancer cost and treatment using SAS

Nov. 2007 **SESUG2007,** Hilton Head Island, SC

Use of ARIMA Time Series and Regressors to Forecast the sale of electricity

Poster presentation

Oct. 2010 **M2010 (anticipation),** Las Vegas, NV

Best Data Mining model for commercial health insurance companies to detect and profitably retain unsatisfied customers

Oct. 2009 **M2009,** Las Vegas, NV

Analysis of Medications used by Mastectomy-Lumpectomy patients using SAS

Oct. 2008 **M2008,** Las Vegas, NV

Breast Cancer summary statistics from the MarketScan data- A preprocessing analysis

May 2008 **ISPOR2008,** SC

Analysis of Mastectomy in breast cancer treatment

Nov. 2007 **INFORMS2007,** Seattle, WA

Analysis of breast cancer cost and treatment using SAS