PERSONALIZED MEDICINE: DEVELOPMENT OF A PREDICTIVE COMPUTATIONAL MODEL FOR PERSONALIZED THERAPEUTIC INTERVENTIONS

by

Nelofar Kureshi

Submitted in partial fulfilment of the requirements
for the degree of Master of Health Informatics

at

Dalhousie University
Halifax, Nova Scotia
August 2013

DEDICATION PAGE

For my mother, Fara and my godmother, Tahira.

Thank you for supporting my decisions and encouraging me to explore all the opportunities that have come my way. Your love, kindness and guidance have profoundly and positively affected my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Lung cancer is the leading cause of cancer-related deaths among men and women. Non-Small Cell Lung Cancer (NSCLC) constitutes the most common type of lung cancer and is frequently diagnosed at advanced stages. In the past decade, discovery of Epidermal Growth Factor Receptor (EGFR) mutations have heralded a new paradigm of personalized treatment for NSCLC. Clinical studies have shown that molecular targeted therapies, such as EGFR tyrosine kinase inhibitors (TKIs), increase survival, lower toxicity and improve quality of life in patients. Despite these advances, the realization of personalized therapies for NSCLC still faces a number of challenges including effective integration of clinical and genetic data and a lack of clinical decision support tools to assist physicians with patient selection. This thesis demonstrates the development of a predictive computational model for personalized therapeutic interventions in advanced NSCLC. The findings of this research suggest that the combination of patient clinical and genetic data significantly improves the model's predictive performance for tumor response than clinical data alone. The decision model is driven by real-world patient data and is a promising step in fostering personalized medical decision-making for patients with advanced NSCLC.

LIST OF ABBREVIATIONS USED

ASCO	American Society of Clinical Oncology
AUC	Area under the curve
CART	Classification and regression tree
CBI	Classifier-based nominal imputation
CDS	Clinical decision support
CEBM	Center for Evidence-based medicine
CENTRAL	Cochrane Central Register of Controlled Trials
COSMIC	Catalogue of somatic mutations in cancer
CR	Complete response
DARE	Database of Abstracts of Reviews of Effect
DDS	Data-driven support
DIRECT	DNA-mutation Inventory to Refine and Enhance Cancer Treatment
EBM	Evidence-based medicine
EED	Economic Evaluation Database
EGFR	Epidermal Growth Factor Receptor
EGFR-TKI	Epidermal Growth Factor Receptor -Tyrosine Kinase Inhibitor
EML4/ALK	Echinoderm Microtubule-associated Protein-Like 4 and Anaplastic Lymphoma Kinase
IDEAL	The Iressa Dose Evaluation in Advanced Lung cancer
IPASS	IRESSA Pan Asia Study
KRAS	Kristin sarcoma virus
MeSH	Medical Subject Headings
NCCCN	National Comprehensive Cancer Network
NFL	No Free Lunch Theorem

NSCLC	Non-small cell lung cancer
ORR	Overall response rate
OS	Overall survival
PD	Progressive disease
PFS	Progression-free survival
PICO	Patient/Population, Intervention, Comparison and Outcome
PR	Partial response
RCT	Randomized controlled trials
RECIST	Response Evaluation Criteria in Solid Tumors
ROC	Receiver operator curve
RR	Response rate
SCC	Small cell lung cancer
SD	Stable disease
SEER	Surveillance, Epidemiology, and End Results Program
SM-EGFR-DB	EGFR Mutations database
TK	Tyrsoine kinase
TKI	Tyrosine Kinase Inhibitor
UICC	International Union against Cancer

ACKNOWLEDGEMENTS

This work would not have been possible without the encouragement and support of many people. First, I extend my utmost gratitude to my supervisors- Dr. Raza Abidi and Dr. Christian Blouin. Thank you for the patient guidance, advice, and mentorship you have provided in the course of this project. I have been blessed to have two extremely enthusiastic researchers watch over me as I journeyed through the last year of the MHI program.

In addition, I wish to thank two bright young students, Fatima Imran and Syed Mohammad, who contributed to the data collection and collation in the initial stages of this research. I am also grateful to Sam Stewart for his helpful advice during the data pre-processing stages and his ongoing assistance for statistical issues. I also want to thank Chris Jones, who was extremely supportive in the later stages of the research analysis and worked through a number of solutions with me.

I appreciate the constructive comments and feedback offered by Mete Erdogan that have helped refine the writing of this thesis. I am grateful to my friends and family for their continuing support and encouragement. Thank you for participating in the joys, frustrations, and eureka moments of my research rollercoaster.

CHAPTER 1: INTRODUCTION

Variability is the law of life, and as no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions we know as disease.

Sir William Osler (1849-1919)

1.1 The Era of Personalized Medicine

Scientific advances since the completion of the Human Genome Project have confirmed that the genetic composition of individual humans has a significant role to play in predisposition to common diseases and therapeutic interventions. The translation of genetic and genomic data into the knowledge of patient care for prevention, diagnosis, prognosis and treatment has introduced a new paradigm for healthcare: personalized medicine. The traditional medicine model has relied on best practices emerging from large population studies and dictates a *one-size-fits-all* approach [1]. Although synthesized evidence is essential to demonstrate the overall safety and efficacy of medical approaches, it falls short in explaining the individual variations that exist among patients. Recent advances in genome-wide association studies have revolutionized the practice of medicine, causing a shift to a *patient-centered* model [2] and offering tailored diagnostic and therapeutic strategies.

The overarching goal of personalized medicine is for physicians to prescribe appropriate medication to the right target of the disease at the right dose for individual patients to achieve maximal therapeutic benefit with minimal, tolerable adverse effects [3].Personalized medicine, frequently called genome-based medicine, offers many distinct advantages over traditional clinical approaches. These benefits include early detection of disease, selection of optimal therapy, reduction in adverse drug reactions and the improvement in selection of targets for drug discovery [1]. Recognition of genetically-determined individual differences to drug response form the cornerstone of personalized medicine [3]. In this context, we introduce and define the terms

pharmacogenetics and pharmacogenomics, which tend to be used interchangeably in the literature. Historically, pharmacogenetics has been defined as the study of germline (or inherited) differences in variation that lead to differences in drug metabolism response [4]. The evolution of pharmacogenetics into pharmacogenomics was instigated by advances in human genome sequencing.

Pharmacogenomics is the study of genetic changes such as the somatic changes in cancer tissue or the use of gene expression profiles to predict the likelihood of response to medication, its efficacy, or the probability of adverse drug effects. Pharmacogenetics and pharmacogenomics seek to improve therapeutic processes through analyzing individual differences in genetic variation. The assumption of pharmacogenomics is that although commonly occurring diseases such as cancer, diabetes, atherosclerosis may have a common clinical phenotype, there are distinct genetic differences that constitute variation in drug response [5].

Complex diseases such as cancer are multifactorial, arising from the intricate interaction of genetic and environmental factors. Processes common to the development of malignant neoplasms are uncontrolled cell growth and proliferation, angiogenesis, invasion of cells into local vasculature, and the spread of cancer cells to distant sites (metastasis). Concurrently, the heterogeneity of cancer is acknowledged by the immense variations with regards to etiology, pathogenesis, treatment response, prognosis and survival. The diversity in malignant cancers is also demonstrated through diverse genetic and epigenetic mutations arising in somatic cells and distinct gene-gene and gene-environment interactions which add to the complexity of cancer risk and tumorigenesis. The application of personalized medicine in oncology includes improved diagnosis, stratification into molecular subtypes, pharmacogenetics, tailored therapy and predictive models [6].

In the field of oncology, cancer biomarkers characterize the emerging standard of care in personalized medicine [7]. Biomarkers are molecular changes which occur in association with disease and dysfunction of biological networks. *Prognostic biomarkers* are

associated with the course of clinical outcomes, such as progression-free survival (PFS) and overall survival (OS), independent of the type of treatment [8]. Prognostic biomarkers are often used in untreated patients with early stage cancer to help determine the use of adjuvant therapy. In contrast, *predictive biomarkers* help identify subpopulations of patients who would most likely benefit from individualized therapy, thereby providing information on the efficacy and response from tailored treatment [8]. Predictive biomarkers can inform on drug resistance as well as drug responsiveness. Successful examples of predictive biomarkers include screening for BRAF V600E mutation in advanced melanomas and KRAS in colorectal cancer [9].To further explain the role of predictive biomarkers and their success in personalized therapy, we provide the example of lung cancer.

Non-Small Cell Lung Cancer (NSCLC) comprises 80-90% of all diagnosed lung cancers [10]. Approximately two-thirds of patients are not diagnosed until the late stages of the disease, limiting the role of surgical resection as a treatment option. The discovery of Epidermal Growth Factor Receptor (EGFR), a key molecule in the growth factor signalling pathway advanced our understanding of the molecular basis of lung cancer. EGFR is a receptor tyrosine kinase, and thus requires phosphorylation of its tyrosine residues in order to activate downstream intracellular signaling pathways [11]. In an attempt to target this molecule, agents which compete with ATP binding to the tyrosine kinase domain of EGFR were tested and developed [12]. Currently, gefitinib (Iressa®, AstraZeneca) and erlotinib (Tarceva®, Roche) are the two EGFR Tyrosine Kinase Inhibitors (EGFR-TKI) that have been approved for use in clinical practice [11]. Furthermore, two landmark studies demonstrated that patients with somatic mutations in exons 18-21 of the tyrosine kinase domain of EGFR show marked response to gefitinib and erlotinib [13],[14]. These findings encouraged further investigations of EGFR-TKI mutations and their role in predicting drug sensitivity. At present, the FDA approves erlotinib for the first-line treatment of patients with metastatic NSCLC whose tumors have EGFR exon 19 deletions or exon 21 (L858R) mutations [15]. Erlotinib is also approved for maintenance treatment of patients with locally advanced or metastatic NSCLC whose disease has not progressed after four cycles of platinum-based

chemotherapy. Similarly, the European Commission authorizes the use of gefitinib for the treatment of adults with locally advanced or metastatic NSCLC with activating mutations of EGFR-TK across all lines of therapy [16].

Clinical response to targeted therapy is a key indicator of the effectiveness of a given anticancer treatment. The value and the interpretation of the clinical response must be kept in perspective, taking into consideration the context within which it is being measured and used. A number of groups have developed strict predefined criteria for tumor response evaluation. Of these, the Response Evaluation Criteria in Solid Tumors (RECIST) criteria is the most widely used in population studies and clinical trials [17], [18].

The measurement of tumor response to molecular targeted therapy in NSCLC carries with it both a therapeutic cost and a financial cost. The therapeutic benefits for patients selected on the basis of their EGFR mutation status are favorable toxicity profiles and superior survival outcomes when compared with patients receiving standard platinum-based National Cancer Institute of Canada chemotherapy [19]. In terms of the financial cost of targeted therapy, the average one year cost borne by a US health insurer for treating advanced stage NSCLC with erlotinib is USD $382,418. This figure was evaluated by comparing formularies with and without erlotinib and demonstrated a relatively small impact on the annual healthcare budget [20]. Compared to its US counterpart, Canada performs a cost benefit analysis of new therapies in order to assess the effect on the public healthcare system. According to the (NCIC) Clinical Trials Group Working Group on Economic Analysis, the cost of providing erlotinib to patients previously treated for advanced NSCLC is approximately CAD $95,000 per year of life gained [21]. The drug may appear to be marginally cost-effective in an unselected population, however, when the analysis is restricted to the EGFR mutation positive sub-group, the incremental cost effectiveness ratio is $138,168 versus $87,994. Cost-effectiveness takes into account magnitude of survival and cost of targeted agent, however, it must be remembered that for the healthcare provider, the cost of drug is a negligible factor when compared to the clinical benefit, decreased toxicity and improved

quality of life for patients. Pharmaceutical studies [19] and economic analyses demonstrate that clinical outcomes and cost-effectiveness of EGFR-TKIs are markedly improved in population subgroups and this underlies the importance of patient selection for targeted therapy in advanced NSCLC.

There are a number of challenges facing the implementation and widespread adoption of personalized medicine in healthcare. These obstacles include the lack of quality assurance for genetic sequencing, limited clinical evidence of genomic assays, and restricted genetic and genomic knowledge among healthcare providers [22]. Of note, the authors in [23], bring to light the following barriers:

1. Although structured medical data allow standardization, the majority of medical data exist in an unstructured form [24], including free-text medical records, clinical notes, research papers, and clinical trial documentation. As clinical studies are progressively accruing molecular profiling and routine clinical patient data, the success of personalized medicine is faced with the challenge of integrating structured and unstructured data from genomic and clinical sources [25].

2. New scientific discoveries infiltrate regular clinical practice in approximately 17 years, where the success rate remains less than 15% [26] In addition, the evolution of advances in the field of molecular oncology outpace the rate at which clinicians are able to keep up with new findings. It is necessary to ensure that proper mechanisms are in place to successfully translate research into practice and also keep physicians up to date with the latest advances in genomic medicine [27]. The need for supportive clinical decision making tools that combine genomic research with clinical data has been identified as an essential requirement for realizing the promise of personalized medicine [28],[29].

1.2 Research Hypotheses

Despite the overwhelming amounts of clinical and genomic data being captured and collected, by and large, these data are not being analyzed in a manner that allows for the production of actionable information [1]. This represents lost opportunities for making data-driven improvements to personalized healthcare. Although clinical practice guidelines for NSCLC recommend EGFR testing prior to treatment with EGFR-TKIs [30], these evidence-based guidelines are formulated using population-based statistics and do not speak to the individual variability of tumor response that is seen among NSCLC patients. In the absence of EGFR testing, purely clinical factors are used to determine patient selection for TKIs. While, the addition of molecular testing for patient selection has been shown to improve the response rates to targeted therapy, the ideal combination of factors that most accurately predict tumor response in patients with NSCLC has yet to be identified. To address these issues, we formulated the following two research hypotheses:

1. Data-driven decision support models can be used to accurately predict tumor response to EGFR-TKIs in patients with advanced NSCLC.

2. The combination of clinical and genetic factors can better predict tumor response to EGFR-TKIs in patients with advanced NSCLC, than clinical or genetic factors alone.

1.3 Research Objectives

1. Creating a data-driven decision support model for personalized treatment selection in advanced NSCLC

The primary objective of this thesis is to develop a proof of concept data-driven decision model for personalized treatment patient selection in advanced NSCLC. The objective is to demonstrate that the predictive power of an integrated clinicogenomic model has the

potential to serve as a data-driven decision support that can be embedded at the point-of-care.

2. Identifying frequent patterns of patient characteristics and tumor response in advanced NSCLC

The secondary objective of this research is to determine the relationships that exist among predictor attributes (age, gender, smoking status, mutation status) and the relationships existing between these predictor attributes and the target attribute of tumor response. The aim of this objective is to explore the strength of associations among clinical and molecular factors in patients with advanced NSCLC.

The relationship of clinical predictor variables such as gender, ethnicity, histology and environmental risk factors such as smoking history, to the clinical response has been explored by many studies. High response rates are observed in female non-smoking patients of East Asian ethnicity with lung adenocarcinoma histology [31-33]. Moreover, several biomarkers such as EGFR, KRAS, MET and ALK have been examined for their clinical application in molecular targeted therapy for NSCLC. Of these, somatic mutations in the tyrosine kinase domain of EGFR are associated with the greatest sensitivity to EGFR-TKI response, demonstrated by an 87% response and disease control rate [34].

1.4 Contribution

The success of personalized medicine will depend on the accurate identification of patients who can benefit from targeted therapies [35]. This work leverages the research on successful predictive modeling in NSCLC that has been previously established by several researchers [36-38]. In this thesis, we demonstrate the development of data-driven decision support for patient selection in NSCLC using real-world patient data. This type of data-driven decision support has the potential to rapidly implement research findings into clinical practice and help clinicians accurately plan and deliver individualized treatment. Furthermore, accurately identifying patients who will respond to targeted

EGFR-TKI therapy will lead to improved patient outcomes, i.e., increased survival, decreased drug toxicity, and better quality of life.

1.5 Thesis Organization

The remainder of the thesis is organized as follows: Chapter 2 presents the background, research motivation, and core concepts related to the research question. Chapter 3 outlines the data collection and preparation methodology with a description of the dataset. Chapters 4 and 5 describe the application and results from the techniques applied on the NSCLC dataset. Finally, Chapter 6 includes a discussion of the major findings and limitations of this research, potential directions for future research, and concluding remarks.

CHAPTER 2: BACKGROUND

2.1 Non-Small Cell Carcinoma

Lung cancer is the most commonly diagnosed cancer worldwide and also the leading cause of death among cancers [39]. According to Canadian cancer statistics, in 2012, an estimated 25,600 Canadians will be diagnosed with lung cancer and 20,100 will die of it [40]. Lung cancer is broadly divided into small cell cancer and Non-Small Cell Cancer (NSCLC); 85% to 90% of lung cancers are non-small cell [10] [41].Microscopically, NSCLC can be sub classified into adenocarcinoma, large cell carcinoma and squamous cell carcinoma, each differing with respect to histological features. Less frequent subtypes of NSLC include adenosquamous, pleomorphic, bronchogenic and sarcomatoid carcinomas [42].

Following NSCLC diagnosis, lung cancer staging is used to assess tumor growth. The assessment of tumor size and spread in lung cancer staging helps to determine treatment options and disease prognosis. Lung cancer staging for NSCLC is based on the Tumor Node Metastasis (TNM) classification of primary tumor size, lymph nodal status and metastases and falls between Stages 0-4. Stage 3 cancer is a heterogeneous group which is further broken down into Stages 3A and 3B. Stage 3A is considered locally advanced cancer where there is spread to lymph nodes on the same side of the chest as the primary tumor. Stage 3B cancers have nodal involvement on the opposite side of the primary tumor site, whereas in stage 4, there is distant metastasis. Stage 3B and Stage 4 together are often referred to as “advanced NSCLC”. Approximately two-thirds of NSCLC presents in the advanced stage, where surgery is not an option and chemotherapy remains the mainstay of treatment [43]. Overall, the prognosis and survival for advanced NSCLC cancer remains poor. 5- year survival rates for Stages 3 and 4 are between 1-14% [44].

2.1.1 NSCLC Risk Factors

2.1.1.1 Environmental risk factors

Cigarette smoking is the strongest environmental risk factor for the subsequent development of lung cancer. Similarly, smoking light cigarettes, pipes, and cigars have a similar effect of the risk of developing lung cancer [45], [46]. Studies show that the risk of developing lung cancer increases proportionally to the number of cigarettes smoked per day and "pack years". Prolonged exposure to second hand smoke also increases the risk of lung cancer in non-smokers. In NSCLC, smoking history is more common in squamous cell carcinoma and is less frequently found in adenocarcinomas.

In addition to smoking, other environmental risk factors include exposure to radiation, radon, arsenic, chromium, nickel, tar, and air pollution. Furthermore, exposure to a combination of these factors may have a synergistic carcinogenic effect, increasing the risk of developing lung cancer.

2.1.1.2 Clinical risk factors

Approximately 70% of lung cancer patients are over the age of 65 [47] and the median age of diagnosis in NSCLC is 71 years [48]. An analysis of the Surveillance, Epidemiology, and End Results registry demonstrated that overall survival is better in patients <40 years compared to those who are >40 years [48].

In addition to cigarette smoking and other environmental factors, gender differences influence the risk of lung cancer [49], [50]. After controlling for smoking, females have been shown to carry triple the risk of lung cancer compared to men and non-smoking females have higher risk for adenocarcinoma [50]. Evidence supports the biological hypothesis that tobacco carcinogens increase the metabolism of steroidal estrogens which can modify both tumor suppressor and proto-oncogenes in female smokers [51].

2.1.1.3 Genetic risk factors

In addition to environmental and clinical risk factors, a number of genetic risk factors have been shown to contribute to the development of NSCLC. Of these genetic variables, three of the most well studied and documented are the Epidermal Growth Factor Receptor (EGFR), gene fusion of Echinoderm Microtubule-associated Protein-Like 4 and Anaplastic Lymphoma Kinase (EML4/ALK) and Kirsten Rat Sarcoma viral oncogene homolog (KRAS).

Epidermal Growth Factor Receptor

The Epidermal Growth Factor Receptor (EGFR) belongs to the ErbB family of receptor tyrosine kinases. Mutations of the Tyrosine Kinase (TK) domain of EGFR results in activation of signalling pathways leading to increased proliferation, angiogenesis, metastasis, and decreased apoptosis [52]. EGFR TK mutations occur in exons 18-21 and are classified into three main categories. Class I mutations are in-frame deletions in exon 19, where there is usually a loss of 4-6 amino acids (746-752) and these account for nearly 44% of all EGFR TK genetic mutations. Class II mutations are single-nucleotide substitutions which may occur in any of the four exons of the TK domain, the most common of which is the L858R substitution in exon 21.Class III mutations are in-frame duplications and/or insertions in exon 20 and these constitute only 5% of mutations in EGFR TK domain [53], [54]. Deletions in exon 19 and L858R constitute the classical activating mutations in EGFR TK domain [53] which occur in female East Asian never smokers with adenocarcinomas [54].

EML4-ALK

EML4-ALK results from the fusion of the Anaplastic Lymphoma Kinase (ALK) with the Echinoderm Microtubule-Associated Protein-Like 4 (EML4) on chromosome 2p. This fusion oncogene occurs in a unique subset of NSCLC patients; mostly young males [55] who are never/former light smokers [56], [57].

EML4-ALK translocations occur more frequently in NSCLC adenocarcinoma histological subtypes than in squamous cell carcinomas [58]. Patients with EML4-ALK are resistant to EGFR-TKI targeted therapy and their clinical response to platinum-based chemotherapy is similar to EGFR wildtype patients [55]. ALK inhibitors may serve as therapeutic targets for EML4-ALK and ongoing studies will validate their use for this distinct cohort of NSCLC [57].

KRAS

KRAS is a member of the RAS family of oncogenes, a class of guanosine triphosphate (GTP)-binding proteins involved in cellular signalling transduction. RAS activation promotes a number of mutagenic events including cell proliferation, upregulation of autophagy, and suppression of apoptosis to support oncogenic transformation [59]. Mutations in the KRAS protooncogene are found in nearly 30% adenocarcinomas and 5% squamous cell carcinomas of NSCLC and amino acid substitutions at residues G12 and G13 are the most commonly reported [60].

KRAS , EGFR and ALK mutations are almost always mutually exclusive and the presence of KRAS mutations occurs in EGFR wildtype and EML 4-ALK translocation negative individuals. In advanced metastatic NSCLC , KRAS mutations are negative predictors of therapy from EGFR TKIs. Since KRAS mutations occur downstream of EGFR, they remain uncontrolled by inhibitors of EGFR TK.

2.1.2 Histological and Molecular Classification of NSCLC

Traditionally, oncologists have divided lung cancer into small cell and non-small cell. The three main subtypes of NSCLC include adenocarcinoma, squamous cell carcinoma, and large cell carcinoma while the less common subtypes are adenosquamous, bronchioalaveolar, carcinoid, and undifferentiated tumors [61]. This traditional distinction of small cell and non-small cell is no longer sufficient to diagnose lung cancer and there is a need to rethink this traditional paradigm of morphological diagnosis.

Molecular sub-classification testing is becoming increasingly important necessity to determine the response to targeted therapy. Interestingly, histology often guides mutation/biomarker testing since specific mutations occur more commonly in NSCLC subtypes [62].

Nine lung cancer molecular classifications have been proposed, where tumors are grouped according to genetic alterations and molecular pathway aberration. Each division is associated with biomarkers assays, targeted therapies and guidelines for clinical decision making. Subtype 1 of this classification consists of NSCLC tumors with an EGFR genetic defect or specific protein signature. This includes Class I-III mutations described above as well as the T790M mutation in exon 20 [63].

2.1.3 NSCLC Molecular Therapeutics

Platinum-based chemotherapy has long been the standard of care for advanced NSCLC. The advances in tumor biology have led to the discovery of agents which target specific molecular defects involved in carcinogenesis, referred to as targeted therapies. EGFR Tyrosine Kinase Inhibitors (TKIs) such as erlotinib and gefitinib target EGFR mediated signalling and have shown the most promising results for advanced NSCLC in clinical trials [14], [31], [32], [64-67].

EGFR sensitizing mutations such as exon 19 deletions and L858R have shown the greatest benefit in advanced NSCLC and thus molecular testing is increasingly being used to determine targeted therapy [68]. A number of studies have examined and established the clinicopathological and molecular profile of advanced NSCLC patients in association with response to EGFR TKIs. However, there is a lack of predictive and prognostic models to assist in this clinical decision making process and achieve better health outcomes.

Despite these developments, EGFR mutation testing remains underutilized [69] in NSCLC patients. The National Comprehensive Cancer Network (NCCN) conducted a

survey on clinicians' patterns of care and preferences for testing of patients with NSCLC for EGFR mutations at the time of presentation with locoregional, distant recurrence, or Stage 4 metastatic disease. 65% respondents indicated that their patients were "sometimes" or "often" tested for EGFR mutations in advanced or metastatic disease [70].

2.2 Data-Driven Decision Models

Data-driven predictive models in medicine have become popular through the use of medical informatics. Adequately modeling of clinical domain problems allows predictive models to assist clinicians with medical prevention, diagnosis, treatment and management. A number of predictive models have been designed for diagnosis and prognosis of lung cancer; the following sub-sections describe the evolution of predictive modeling in NSCLC using clinical factors and their various combinations with pathological and molecular variables.

2.2.1 Clinical Predictive Models

The most common predictive models in medicine employ patient demographic and clinical variables such as age, gender, ethnicity and smoking status. Lee et al used the Cox proportional hazard regression model to analyze multiple clinical factors such as age, gender, stage, tumor size, neoadjuvant therapy and adjuvant therapy to predict recurrence of NSCLC after surgical resection. The resulting model had reliable predictability in distinguishing high and low risk recurrence, however the exclusion of biologic markers weakened its robustness [71]. Using univariate and multivariate analyses, Huang et al identified survival prognostic clinical factors in chemonaive advanced NSCLC and constructed a normogram to predict survival [72]. The Cox proportional hazards model has been used for both univariate and multivariate analyses of clinical prognostic variables in advanced NSCLC. Jeremic et al demonstrated that age, gender performance status, pretreatment weight loss, and number of metastatic sites were prognostic indicators of survival in advanced NSCLC [73]. In addition, Mandrekar

et al showed that hemoglobin levels and white blood cell count are also prognosticators of OS and TTP [74] and Tsao et al found that never-smokers had lower rates of progressive disease and improved overall l survival than smokers who received chemotherapy or chemoradiation [75].

2.2.2 Clinicopathological Predictive Models

Pathological variables such as TNM staging and histological sub-typing are critical for cancer diagnostic assessment and clinical decision making. Highlighting the importance of these variables, a number of researchers have incorporated them into their predictive models. For example, a predictive model combining clinicopathological factors involved in peri-operative mortality of NSCLC was developed by Strand et al [76] using multivariate analyses with multiple logistic regression models. In another study, five clinicopathological factors (vascular invasion, lymphatic permeation, histological subtype, papillary carcinoma component, and smoking status) involved in the recurrence of small adenocarcinomas of male patients were identified in a logistic predictive model by Sakuma et al [77]. Other studies have also reported on the role of combining demographic, clinical and pathological variables to produce predictive models [78-81].

2.2.3 Gene Expression Prediction Models

Several gene signatures derived from microarray expression profiling have been identified to predict clinical outcomes in NSCLC [82-87]. Gene signatures for the histological classification and post-surgery survival of NSCLC patients were developed and validated by Hou et al using Cox proportional hazards regression analysis [88]. A 4-gene Cox model was developed by Mitra et al to determine the prognostic outcome of Stages 1-3 non-small cell carcinoma using microarray profiling of American and Korean patient tumor samples. These markers were associated with recurrence in both demographics and remained independent of patient clinical characteristics [89]. Baty et al demonstrated that a 13-gene metagene obtained using expression profiling from

bronchoscopic and surgical biopsies was correlated with both histological classification and prediction of survival in NSCLC. These 13 genes were independent predictors of survival when compared with International Union against Cancer, 6th edition (UICC) stages, especially in patients with less than 1-year survival [90]. Although gene expression profiling holds remarkable potential for personalized cancer prediction models, its application is currently limited by the reliability and reproducibility of results for the diagnosis, classification and prognosis of NSCLC.

2.2.4 EGFR Mutation Prediction Models

Advanced NSCLC patients with somatic mutations of EGFR show dramatic response to treatment with EGFR-TKIs. At present, only two EGFR-TKIs are approved specifically for treatment of NSCLC: erlotinib (Tarceva®) and gefitinib (Iressa®). The National Comprehensive Cancer Network (NCCN) clinical practice guideline recommends EGFR mutation testing for advanced NSCLC patients who are candidates for TKI therapy such as erlotinib and gefitinib [30]. Following this update, the American Society of Clinical Oncology (ASCO) also issued a provisional clinical opinion recommending EGFR testing for patients considering TKI chemotherapy with a view to improving response and progression-free survival in this cohort. EGFR mutations have been used for genotype-oriented risk prediction and therapeutic response to TKIs in NSCLC [91-96].

2.2.5 Integrated Predictive Models

In the era of genomic and personalized medicine, the development of predictive models that integrate molecular and clinical data can provide guidance and recommendations to clinicians on individualized risk classification and disease management to help improve health outcomes. The analysis and impact of complex factors involved in disease development and progression carries with it both predictive and prognostic value.

Predictive models that combine both clinical and molecular patient information have been successfully developed in many areas of oncology including breast cancer

recurrence [97], large B-cell lymphoma survival [98], and prostate cancer recurrence after radical prostactectomy [99]. In lung cancer, mixed models that fuse the multifactorial features have been shown to provide superior prognostic benefit [100-103]. A few of these composite predictive models that combine anatomical, clinical, and molecular factors have been developed specifically for NSCLC. Lopez et al developed a prognostic survival model for early stage NSCLC using a supervised learning classification algorithm and clearly demonstrated that the prognostic discrimination of integrated models surpasses that of individual risk factors [36]. Spira et al constructed a gene expression biomarker model to predict lung cancer in smokers and then tested it in combination with clinical information, suggesting that an integrated model provided superior specificity for diagnosis[37], [38].Furthermore, a number of studies have explored the interaction of clinical features, EGFR mutations and TKI treatment response using logistic regression [104]. Many clinical trials have investigated factors associated with TKI sensitivity, suggesting that sensitizing EGFR mutations are associated with response to TKIs and an improvement in overall survival in NSCLC patients [92], [105], [106]. Multivariate logistic regression analysis is the most popular technique to test the impact of clinicopathologic variables and genetic mutations on response as assessed by disease control rate or objective response.

2.3 Predictive Analytics and Clinical Intelligence in NSCLC

In an effort to support clinicians and oncologists in medical decision-making, commercial and freely available products are being developed to promote the utility of pertinent and actionable information by clinicians at the point of care. The Vanderbilt-Ingram Cancer Center’s approach to address the challenge of translating genomic discoveries to the bedside is the integration of a clinical decision support system into their electronic health record which uses My Cancer Genome as a knowledge base. This knowledge base is continuously updated with NSCLC driver mutations, their clinical significance, genome directed therapies and relevant clinical trials at both at Vanderbilt and world-wide centers [107]. IBM has signed a recent collaboration with Memorial Sloan Kettering Cancer Center, to build an intelligence engine using natural language processing capabilities to convert the medical center's free text consult notes into usable data. Their first project

focuses on NSCLC, in which the technology will use 14-20 data elements including patient tumor size, metastasis, and genetic mutations to return a list of possible diagnostic tests and chemotherapy protocols to choose from [108].

2.4 Chapter Summary

Currently, the rate of production and collection of biomedical data surpasses current resources to determine its actionability in clinical practice**.** There is a demand to develop clinical decision support tools as genetic variations and their associations to disease subtypes and treatment response are discovered [109]. Using statistical and computational approaches, predictive tools can assimilate rich sources of data from patient records and biomedical databases to analyze therapeutic options and their effectiveness. A review of the literature reveals that Cox proportional hazard models have been used to relate several risk factors considered simultaneously, to survival time in advanced NSCLC patients. Additionally, multiple regression analysis is extensively employed to demonstrate the impact of clinical, pathological and genetic risk factors on objective response to EGFR-TKI therapy. Despite these developments, limited studies have explored the role of predictive modeling using EGFR mutations and clinicopathological risk factors in advanced NSCLC. The wealth of NSCLC patient data can be transformed into novel, potentially useful and understandable information using knowledge discovery methods. The subsequent interpretation, visualization and consolidation of this discovered knowledge will support scientific and ultimate healthcare goals.

This work illustrates the integration of various sources to create a small but representative sample of advanced NSCLC, complete with demographics, pathological diagnosis, and EGFR mutation data. It then presents the application of knowledge discovery techniques on this integrated dataset to detect associations and patterns which may be indicative of biomedical relations. This process is valuable not only to validate many of the recognized associations, but to reveal new and unexpected connections between variables affecting EGFR-TKI response. The research demonstrates and confirms the additional predictive

power of classifiers using integrated data versus clinical parameters alone to predict responsiveness to EGFR-TKIs in advanced NSCLC. Specifically, the significance and explanatory power of decision trees in designing a learning model for clinical management is explored.

CHAPTER 3: DATA COLLECTION AND PREPARATION

This chapter will present the strategy that was devised to collect, collate, and prepare data based on the practical approach to successful secondary data analysis [110]. The data collection strategy involved the use of secondary sources of data to generate research data sets suitable for answering the research questions.

Data collection begins by first defining the attributes of interest. In order to do this, steps from evidence-based medicine (EBM) and practice were borrowed, as depicted in Figure 3.1. The first step is the identification of the clinical problem and the development of specific research question(s). Following this step, the research question is decomposed into its constituent concepts in order to facilitate a focussed literature review which is conducted using the highest levels of evidence. Next, systematic reviews and meta-analyses guided the determination of the most relevant attributes to further explore pursuing the research question. After identification of attributes, multiple secondary sources of data were ascertained. The data extraction processes from each of these sources is described, followed by their integration to formulate a single dataset. To prepare the dataset for use, a series of data pre-processing steps were taken including dimensionality reduction, discretization and imputation of missing values.

3.1 Determination of Attributes

Before commencing data collection, steps were taken to determine the attributes which were most relevant to the research problem. The steps followed in this process were to identify the clinical problem, generate a specific research question, identify components of the question, assess the quality of the evidence, retrieve the highest quality of evidence, and select the most relevant attributes for further exploration. These steps are discussed in greater detail below.

1. Identify the clinical problem

The development of molecular targeted therapies has revolutionized treatment options and clinical outcome for patients with advanced stage NSCLC. Multiple factors such as

gender, ethnicity, smoking status and EGFR mutations are associated with greater sensitivity to these drugs. Despite tremendous progress on the research front, there is little headway being made for clinical decision support tools for patient selection in the personalized treatment for NSCLC.

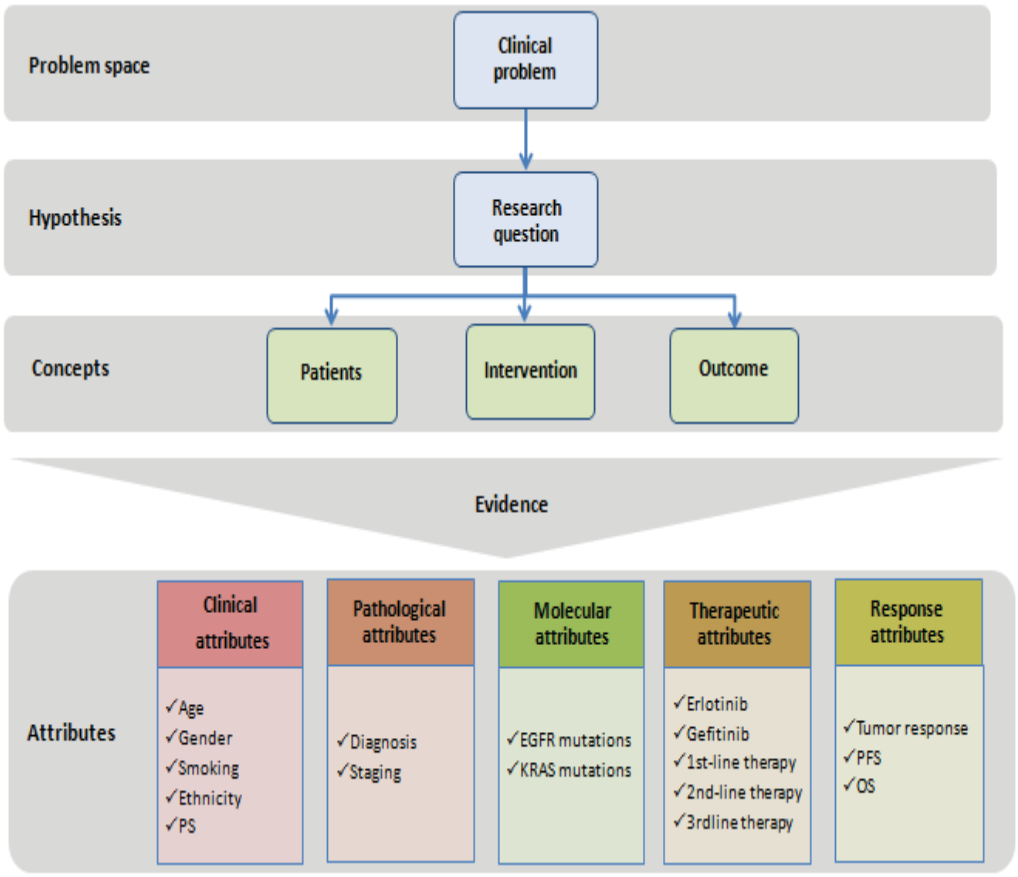

Figure 3.1 Overview of steps used in the determination of attributes

2. Generate specific research question(s)

After determining the clinical problem, a specific research question was generated:

Can data-driven decision support predict the response to EGFR-TKI therapy (such as erlotinib and gefitinib) in patients with advanced NSCLC?

As part of this investigation, the relationships that exist between the attributes of interest will also be explored.

Note here that the research question has both a clinical and informatics portion. The clinical issue is the prediction of response to EGFR-TKIs in patients with advanced NSCLC - this will be referred to as the clinical research question in the remainder of the chapter. Data-driven decision support is the tool [111] that will be developed to solve the clinical problem.

3. Identify components of question

Evidence-based medicine proposes that clinical problems in research, teaching and practice can be generated using the PICO (Patient/Population, Intervention, Comparison and Outcome) model [112]. This model identifies the critical components used in the construction of a research question, especially in EBM [113], however its use can also be extended to experimental and predictive research. Using the PICO model, the clinical research question was decomposed into its main components as follows:

Patient: the population of interest are patients with diagnosed advanced NSCLC

Intervention: the therapeutic intervention they receive is erlotinib or gefitinib.

Comparison: comparison is the only optional component of the PICO framework and the current research does not consider an alternative to EGFR-TKI therapy. The research scope is limited to patient selection in NSCLC, and does not extend to the suggestion or comparison of alternative treatment options.

Outcome prognosis**;** tumor response to treatment

Clearly defining the building-blocks of the research question made it possible to identify key concepts and terms to explore further by performing a literature search.

4. Assessing Quality of Evidence

Before commencing a literature search, the quality of the evidence gathered was assessed. Evidence that is systematically acquired, analyzed and critically appraised is the cornerstone of EBM [112]. To formulate a response to a clinical research question, EBM

relies on the highest quality of evidence for clinical use in patient care. The Center for Evidence-based medicine (CEBM) promotes the practice of EBM by providing support and resources for the teaching, training and development of to practicing clinicians and researchers. CEBM has developed details for levels and quality of evidence [114]; the hierarchy of evidence is popularly depicted in pyramid form as shown in Figure 3.2.

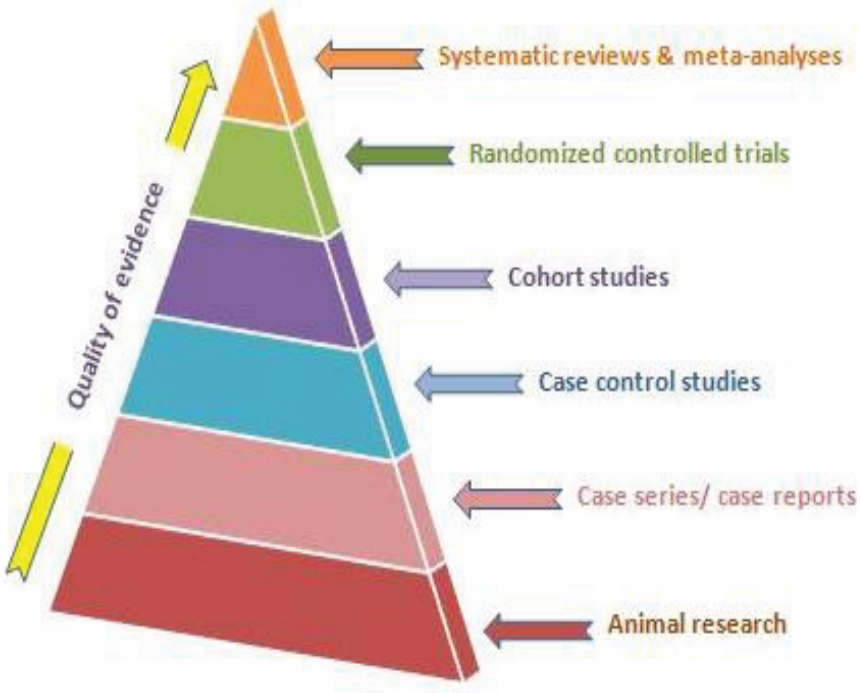

Figure 3.2 Hierarchy of quality of evidence

The literature review presented in Chapter 2 was based mainly on primary sources retrieved from the MEDLINE and EMBASE databases, including experimental studies and original research publications. Review and examination of these primary sources was the initial step in understanding the attributes and associations relevant to the clinical research question. The validity of a primary source is determined solely by the individual researchers who report the personal experience of their discoveries. Often, such reports include novel and unique variables and associations related to the research. Determination of variables from primary sources of evidence would result in a lengthy

list, possibly with invalid or spurious associations. For example, several unique gene signatures that predict clinical outcomes in NSCLC have been developed [83-86], however, they have not been validated as reliable markers in the clinical setting.

To determine the most valid and reliable predictors and outcome attributes, a number of secondary sources of appraised and synthesized evidence were employed, forming the peak of the evidence pyramid as shown in Figure 3.2. These sources were used as the basis in determining and defining the most relevant attributes to study.

5. Retrieval of Best Evidence

A literature search was initiated through the Cochrane Library, a key resource for synthesized evidence in medicine and healthcare [115], [116]. Advanced searching options allow four ways to search and browse the Library databases. The Search manager creates a search strategy which allows multiple entries of key words or thesaurus terms. Cochrane Library's thesaurus is constructed from the MeSH (Medical Subject Headings) thesaurus which groups medical concepts appearing as various terms in literature, under subjects headings to allow a standardized vocabulary search. MeSH is displayed as a tree structure and permits the selection of an exact term or an explosion of terms that appear above or below in the hierarchy.

From the concepts identified using the PICO framework, a MeSH search for key words "non-small cell lung cancer", "tyrosine kinase inhibitors", and "treatment outcome" was conducted. For NSCLC, the MeSH exact term "Carcinoma, Non-Small-Cell Lung" was found and selected. The closest match phrase for EGFR-TKIs, such as erlotinib and gefitinib, was "Protein Kinase Inhibitors". Treatment evaluation and efficacy was matched to "Prognosis". This was the only MeSH term where the search was expanded in the hierarchy to include disease free survival as well as treatment outcome.

A total of 43 results from the six databases were retrieved using this search. No Cochrane Systematic Reviews were identified; however, eight reviews from Database of Abstracts of Reviews of Effect (DARE), 31 trials from the Cochrane Central Register of Controlled

Trials (CENTRAL) and four economic evaluations from NHS Economic Evaluation Database (EED) were retrieved. A summary of the relevant provisional and structured abstracts from DARE is provided in Table 3.1.

Type of abstract	Original article title	Year
Structured abstract	EGFR-targeted therapies combined with chemotherapy for treating advanced non-small-cell lung cancer: a meta-analysis	2011
	Epidermal growth factor receptor-tyrosine kinase inhibitor therapy is effective as first-line treatment of advanced non-small-cell lung cancer with mutated EGFR: a meta-analysis from six phase III randomized controlled trials	2010
	Erlotinib and pemetrexed as maintenance therapy for advanced non-small-cell lung cancer: a systematic review and indirect comparison	2012
	Maintenance therapy with continuous or switch strategy in advanced non-small cell lung cancer: a systematic review and meta-analysis	2011
	Efficacy of erlotinib in patients with advanced non-small cell lung cancer: a pooled analysis of randomized trials	2011
Provisional abstract	Somatic EGFR mutation and gene copy gain as predictive biomarkers for response to tyrosine kinase inhibitors in non-small cell lung cancer	2010

Table 3.1 Structured and provisional abstracts from Cochrane library

6. Selection of attributes

The Cochrane Library databases provided a rich source of appraised evidence. Provisional and structured abstracts, studies included within the systematic reviews, and

trials were examined in order to discover the most important attributes relevant to the clinical research question. These high quality sources were used to guide the selection of important attributes, their relationships and associated context. For example, in [117] the authors performed a pooled analysis of randomized controlled trials (RCT), reporting on the efficacy of erlotinib-based regimens in advanced or metastatic NSCLC. A summary of the trial characteristics included the attributes of age, gender, performance status, adenocarcinoma, and smoking history. The primary endpoints of the analysis were PFS and OS and the authors concluded that erlotinib-based regimens increased response rates and improved PFS as first-line maintenance therapy or as a second/third-line therapy compared with placebo. An examination of this systematic review highlighted some of the attributes (age, gender, smoking, performance status, PFS, and OS) that must be considered when seeking to determine responsiveness to EGFR-TKIs.

Table 3.2 provides a summary of the attributes, their definitions, function, and measurement scale. The definitions are in no way meant to provide an absolute explanation of the term, but merely serve to familiarize the reader with an operational definition, as they will be used in the remainder of the thesis. All variables are categorized into groups derived from the work by [118], [119], [36]. Clinical attributes broadly include patient demographics such as age and gender as well as other features recorded in history taking including patient ethnicity, smoking status, and performance status (PS). Pathological attributes consist of the features resulting from surgical pathology procedures such as tissue biopsy for diagnosis, histological examination for cancer classification, and cancer staging. The research question focuses on the outcome from EGFR-TKIs which are molecularly targeted agents. The assumption is that patients receiving such therapies will be screened for genetic mutations. The Molecular attribute group includes the results of EGFR and KRAS mutation testing. Studies report the results of mutation testing in various manners; this is may be as simple as mutation positive/wildtype or details of the exact amino acid sequence change. Intervention attributes provide details of the targeted therapy (erlotinib, gefitinib or a combination of these with platinum agents) as well as details of the drugs used in first, second and third-line treatment. Finally, the response attributes include the various features used to

measure prognosis and the efficacy of the targeted therapy such as objective tumor response, progression-free survival and overall survival.

In the context of the current research, attribute function is defined as either predictor or outcome. The underlying research hypotheses of reviewed randomized clinical trials assumed a causal association between attributes that occurred prior to the outcome of interest. Attributes that occurred prior to the final effect are called predictive attributes and the measured response of the predictors is called the outcome attribute [120].

Measurement scale is generally divided into nominal and numeric, where nominal variables are names or classes with unique distinguishing characteristics and categorical nominal variables have one or more categories or levels with no inherent ranking assigned. Numerical attributes hold numeric values and among these, discrete numerical attributes takes on distinct possible numeric values whereas continuous numerical attributes can take on infinite number of real values.

	Attribute	**Definition**	**Function**	**Measurement scale**
Clinical attributes	Age	Time elapsed since birth [121]	Predictor	Continuous numerical
	Gender	Socially constructed identity of male or female [122]	Predictor	Categorical nominal
	Smoking status	Never smokers= smoked <100 cigarettes over their life-time. Former smokers= smoked ≥100 cigarettes in their lifetime but had stopped smoking for ≥ 1 year before the diagnosis of lung cancer [123]	Predictor	Categorical nominal
	Ethnicity	Group of people with common cultural heritage	Predictor	Categorical nominal
	Performance status (PS)	Standard way of measuring cancer patients ability to perform ordinary tasks [124]	Predictor	Discrete numerical

Pathological attributes	Diagnosis	Process of identifying disease [125]; often by diagnostic surgical pathology of cancer tissue specimen	Predictor	Categorical nominal
	Stage	The extent of a cancer in the body. Staging is usually based on the size of the tumor, whether lymph nodes contain cancer, and whether the cancer has spread from the original site to other parts of the body [126]	Predictor	Discrete numerical
Molecular attributes	EGFR mutation status	Detectable change in EGFR gene that causes a change in genotype	Predictor	Categorical nominal
	KRAS mutation status	Detectable change in KRAS gene that causes a change in genotype	Predictor	Categorical nominal
Intervention attributes	Targeted therapy	Drugs that block the growth and spread of cancer by interfering with specific molecules involved in tumor growth and progression	Predictor	Categorical nominal
	First line therapy	The first treatment given for a disease. When used by itself, first-line therapy is the one accepted as the best treatment [127]	Predictor	Categorical nominal
	Second line treatment	Treatment that is given when initial treatment (first-line therapy) doesn't work, or stops working [128]	Predictor	Categorical nominal
	Third line treatment	Treatment that is given when both initial treatment (first-line therapy) and subsequent treatment (second-line therapy) don't work, or stop working [129]	Predictor	Categorical nominal

Response attributes	Progression-free survival (PFS)	The length of time during and after the treatment of a disease, that a patient lives with the disease but it does not get worse[130]	Outcome	Continuous numerical
	Overall survival (OS)	Patients with a specific type and stage of cancer who are still alive—that is, have not died from any cause—during a certain period of time after diagnosis [131]	Outcome	Continuous numerical
	Tumor response	The observation of therapeutic benefit from specific treatment [132]	Outcome	Categorical nominal

Table 3.2 Description of relevant attributes

3.2 Data Collection

The secondary use of clinical data in support of medical practice, research and predictive analytics is becoming increasingly popular. Potential sources of existing data for secondary use were identified, including the National Center for Health Statistics (NCHS) [133] and Surveillance, Epidemiology, and End Results (SEER) Program, cBIO cancer genomics portal [134], Cancer genetics network (CGN) [135], Database of genotypes and phenotypes (dbGaP) [136], and DNA-mutation Inventory to Refine and Enhance Cancer Treatment (DIRECT) [137]. Several of these databases offered public use of microarray data but did not report on EGFR and KRAS DNA mutations which were critical to the research analysis. DIRECT collects many of the same attributes [138], [139], but has restricted access due to copyrights on the data. Although it was tempting to rely on one source of data which was easy to acquire and analyze, it was not possible to identify a single source that could provide data to answer the current research question. As an alternative, the decision was made to use freely available, multiple data sources to create a dataset. These sources include PubMed, Catalogue of somatic mutations in cancer (COSMIC) [140], and EGFR Mutations database (SM-EGFR-DB) [141].

As described previously, secondary sources of evidence, such as those higher up in the evidence pyramid, were used to identify well-studied and appraised predictor and outcome *attributes* involved in determining the treatment response to EGFR-TKIs in advanced NSCLC. However, the secondary sources were a synthesis of evidence and did not provide patient-level data for each of the attributes. Consequently, primary sources and mutation databases were used to identify specific *instances having those attributes*. The systematic search strategy for publications from these sources is described below.

PubMed

The MeSH terms ("Carcinoma, Non-Small-Cell Lung" AND "Receptor, Epidermal Growth Factor") were used to search for case series, case reports, and research publications between the years 2000-2012. Only articles on Human species and published in English were selected. The bibliographies of these articles also pointed to relevant literature containing patient-level data.

Catalogue of Somatic Mutations in Cancer

The Catalogue of somatic mutations in cancer (COSMIC) provides a distribution of somatic mutations for EGFR and a list of publication references for each of these mutations. Articles which provided details of mutated and non-mutated samples are assigned COSMIC identification numbers. Each mutated sample may have additional information associated with it.

Somatic Mutations in EGFR Database

The Somatic Mutation in EGFR Database is a collection of EGFR mutations in NSCLC and other cancers. It provides links to original articles and a table of representative sample origins, response to treatment, EGFR mutation site, and mutations detection method, pathology of NSCLC, gender and smoking status amongst other attributes.

This search resulted in the identification of 1962 papers from PubMed, 4681 unique mutated samples (each associated with a PubMed ID) from the COSMIC database and 167 articles from SM-EGFR-DB. There were two main types of publications; original

articles, including reports from both prospective and retrospective studies as well as case reports.

A manual review was performed for articles that met the inclusion and exclusion criteria (Table 3.3). Some authors reported patients and rare mutations in multiple papers and care was taken to delete duplicate entries for individual cases. A summary table of data sources with author names, article and journal title, year of publication is provided in the Appendix A.

Inclusion criteria	Exclusion criteria
Full text available	Articles in foreign language
English language	In-vitro studies
Human studies	Aggregate patient data
Individual patient-level data	Studies reporting acquired EGFR-TKI resistance
Minimum inclusion variables: gender, smoking status, confirmed diagnosis, advanced stage, EGFR mutation status, treatment with erlotinib, gefitinib, or a combination of these with other drugs, and tumor response	Treatment with second-generation EGFR-TKI such afatinib
Additional variables: age, ethnicity, performance status, KRAS mutation status, line of treatment, PFS, and OS	

Table 3.3 Inclusion and exclusion criteria for data collection

3.3 Data Extraction

To demonstrate the data extraction process, one example from an original research article and one case report is presented.

Original research article

Studies have been reported using selected patient-level information in addition to aggregate statistics of the study population. An example is shown Figure 3.4 from [142], where the authors evaluated the efficacy of gefitinib in patients with NSCLC and the correlation of EGFR mutations with the response. In their methods, Zhang et al clearly define the patient eligibly criteria, drug administration and assessment of response using RECIST. EGFR gene sequencing was performed on 30 patients of which 12 had mutations. Clinical features and mutation characteristics of mutation positive patients are provided in the table. Each row of the table represents an individual case and the columns provide the corresponding values for each variable. For example, Case 1 is a male patient who was diagnosed with adenocarcinoma. This patient had stable disease (SD) upon assessment of tumor response after treatment with gefitinib. His PFS was 3.2 months and OS was 24.8 months. EGFR gene sequencing revealed amino acid sequence change E746-A750 in exon 19. Patient-level data presented in tables from 34 articles was transferred to a data spreadsheet.

Table 6. Clinical features and mutation, p-EGFR status ($n = 12$)

No.	Gender	Smoking status	Pathol.	Response	PFS (months)	OS (months)	Survival status	Site (exon)	Nucleotide sequence	Amino acid sequence	p-EGFR status
1	M	Former	ADC	SD	3.2	24.8	D	19	2235–2249 del	E746–A750	–
2	F	Never	ADC	PR	9.6	14.0	A	19	2236–2250 del	E746–A750	+
3	M	Never	SCC	SD	3.1	4.6	D	19	2236–2250 del	E746–A750	–
4	F	Never	ADC	PR	9.8	18.7	D	19	2240–2257 del	L747–P753, insS	+
5	F	Never	ADC	PR	9.1	17.9	A	21	2819T → G	L858R	+
6	F	Never	ADC	PR	12.4	12.4	A	21	2819T→G	L858R	+
7	F	Never	ADC	SD	5.8	16.5	D	21	2819T→G	L858R	–
8	M	Never	BAC	PR	11.2	25.3	A	21	2819T → G	L858R	–
9	F	Never	ADC	PR	8.0	13.7	A	21	2819T → G	L858R	+
10	F	Never	BAC	PR	13.4	18.5	A	21	2819T→G	L858R	+
11	F	Never	ADC	PR	9.5	9.5	A	21	2819T→G	L858R	–
12	F	Never	ADC	SD	3.0	9.8	D	21	2819T → G	L858R	–

F, female; M, male; Pathol., pathological diagnosis; ADC, adenocarcinoma; BAC, brochioloalveolar carcinoma; SCC, squamous cell carcinoma; SD, stable disease; PR, partial response; PFS, progress-free survival; OS, overall survival; D, dead; A, alive; p-EGFR, phosphorylation epidermal growth factor receptor.

Figure 3.3 Data extraction from sample research article

From X. T. Zhang, L. Y. Li, X. L. Mu, Q. C. Cui, X. Y. Chang, W. Song, S. L. Wang, M. Z. Wang, W. Zhong, and L. Zhang, "The EGFR mutation and its correlation with response of gefitinib in previously treated Chinese patients with advanced non-small-cell lung cancer," *Ann Oncol,* vol. 16, no. 8, pp. 1334-42, Aug, 2005. by permission of Oxford University Press.

Case report

Case reports are descriptions of an individual case or up to three cases [143], providing details of patient history, clinical presentation, evaluation, diagnostic testing, and follow-up care. Often times, they are used to communicate unique associations between patient signs and disease, novel therapeutic approaches, or unusual events in the course of treatment. Although there is much argument about the usefulness of case reporting, especially in a world of evidence-based medicine and practice, such reports provide substantial contribution to the understanding of known diseases and the recognition of the unexpected [144]. For this research, 14 case reports were identified which related to the research question and provided the key elements noted in Table 3.3. An example of a typical case report with highlighted variables and its data representation is shown in Figure 3.5.

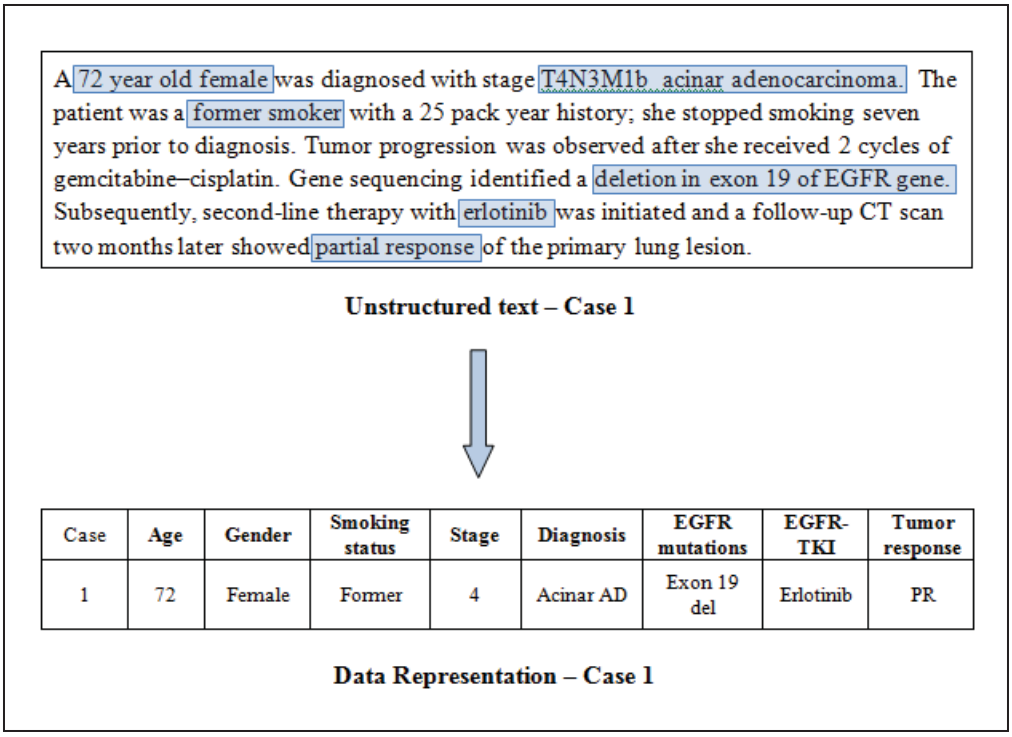

Figure 3.4 Data extraction from sample case report

In the above example, patient demographics of gender and age are stated immediately. Further details of patient history reveal that this 72 year old female was diagnosed with Stage 4 acinar adenocarcinoma (diagnosis). After she demonstrated disease progression with first-line chemotherapy, gene sequencing revealed an exon 19 deletion (EGFR mutation). Erlotinib (EGFR-TKI) was chosen as the second-line therapy and disease assessment two months later showed partial tumor response (PR).

A dataset was created by combining the tabular individual-level participant data reported by 34 research articles and 14 case reports. Final data representation was in the form a table where the individual columns represented the attributes and rows represented instances.

3.4 Data Pre-processing

Integration of data from multiple sources often results in incomplete and noisy data. The quality of the data affects any subsequent analytic method, thereby making data pre-processing a critical stage of data preparation. The data pre-processing methodology is represented in Figure 3.5.

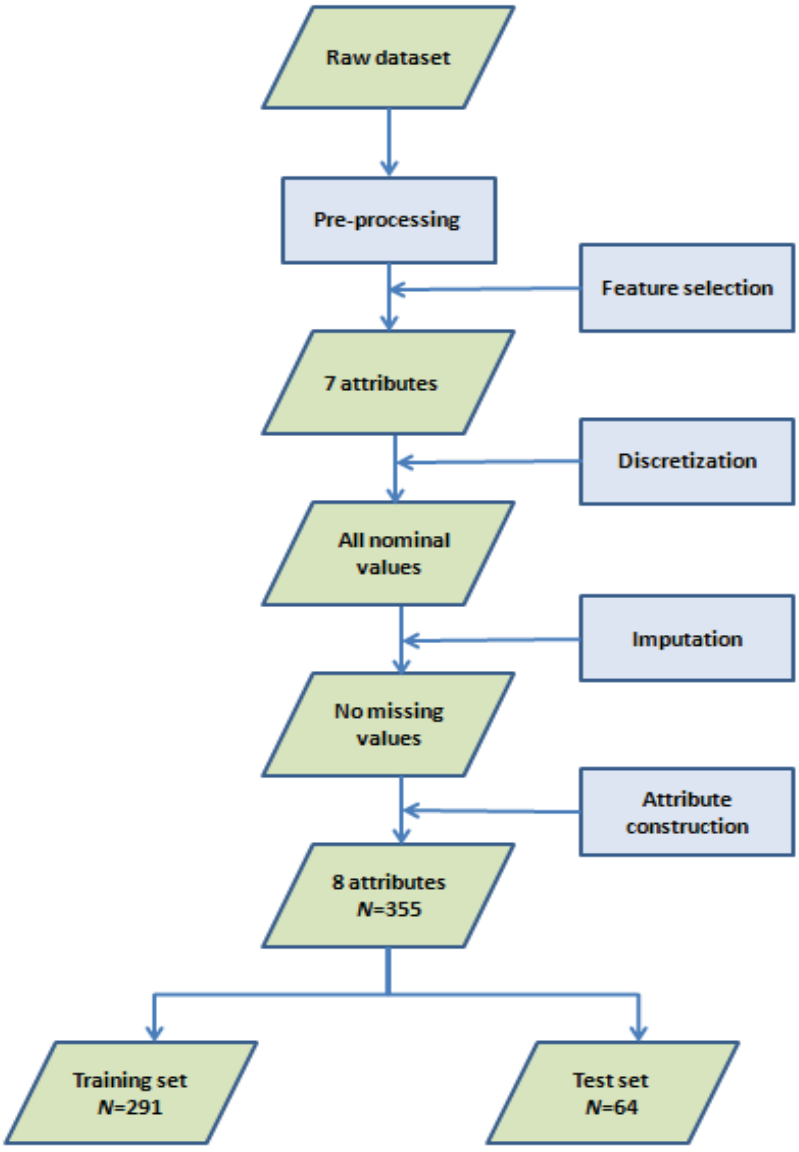

Figure 3.5 Data pre-processing methodology

Attribute	Missing (%)	Type
Age	25	Numeric
Gender	5	Nominal
Smoking status	24	Nominal
Ethnicity	67	Nominal
Diagnosis	6	Nominal
Stage	47	Nominal
EGFR mutation type	0	Nominal
KRAS mutation type	48	Nominal
1st line therapy	64	Nominal
2nd line therapy	68	Nominal
3rd line therapy	78	Nominal
Target drug or combination	0	Nominal
RECIST Response	3	Nominal
PFS	53	Numeric
OS	49	Numeric

Table 3.4 Initial attributes with missing value frequency counts

Manual feature selection

Upon reviewing the initial dataset, it was observed that there was consistency in the reporting of some values over others. Manual feature selection was performed; a list of the removed attributes and rationale for elimination are explained below.

Ethnicity: The estimated frequency of EGFR mutations in East Asians is 30-40% [145], compared to 10-15% for non-Asians [105], [146]. However, sensitizing mutations in EGFR confer sensitivity to EGFR-TKIs such as gefitinib, irrespective of ethnicity [147].

Case reports often provided complete demographic and clinical profile, but research studies rarely reported sample ethnicity.

Stage: The research question was limited to determining the efficacy of erlotinib and gefitinib in *advanced* NSCLC. As described in Chapter 1, locally advanced and advanced NSCLC include Stages 3 and 4. Given that the selection of cases was restricted to advanced NSCLC, the attribute for stage did not provide additional information.

KRAS mutation status: EGFR and KRAS mutations are generally mutually exclusive; the presence of one indicates the absence of the other [148]. Studies have investigated the predictive power of KRAS mutations for chemotherapy benefit [149], [150], however results have not been definitive because of the small sample size. Evidence does not support KRAS mutational analysis for routine clinical use [151].

Lines of treatment: Erlotinib is approved for first-line treatment in patients with metastatic NSCLC harboring EGFR exon 19 deletions or exon 21 L858R substitute mutations. It is also approved for maintenance therapy in locally advanced or advanced NSCLC patients who do not show progression after first-line treatment with a platinum-based chemotherapy. Gefitinib is approved in the European Union for locally advanced or metastatic NSCLC with activating mutations of epidermal growth factor receptor-tyrosine kinase across all lines of therapy. Data for prior chemotherapy was not typically collected by most studies and there was no direct evidence of its predictive value for determining tumor response.

PFS and OS: A survey of reviews retrieved from DARE [Table 3.1] revealed that the endpoints of therapeutic efficacy included objective response rate, PFS and OS. From the three choices of outcome, tumor response was consistently reported by studies and case reports in our data sources. Tumor response, when assessed by standard criteria such as RECIST, remains a validated endpoint to measure anti-tumor activity, especially in phase II clinical trials [152]. Nevertheless, response is not a validated surrogate marker for increased survival benefit. Given the scope of the thesis, tumor response was selected as the outcome measure and it is acknowledged that additional information from PFS and OS would provide a broader picture of the associations amongst clinical endpoints.

Discretization

Age was the only continuous numeric attribute in the dataset with a mean value of 60.2 ± 12.1 (minimum=24, maximum=94).

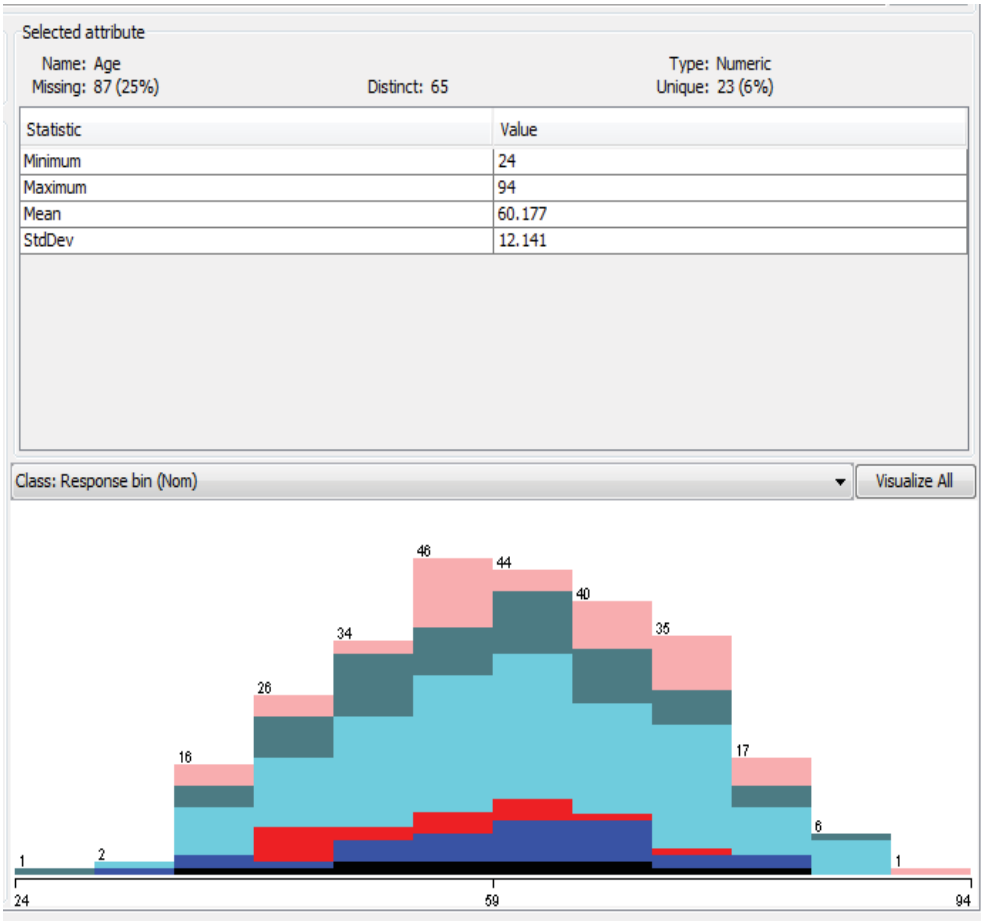

Figure 3.6 Distribution of age

Stable disease Partial response Progressive disease

Complete response Miscellaneous response

In order to divide the continuous data into finite values, discritization was used. There were no informative cut-off points in the literature to define the data ranges or number of bins; instead parameter tuning was used to optimize the range and number of equal-width bins. The process resulted in five bins for the attribute of age. Partition of the original

continuous attributes, maintained the distribution patterns of age, where approximately 60% patients were ≥ 49 years and < 0.3% were <30 years old.

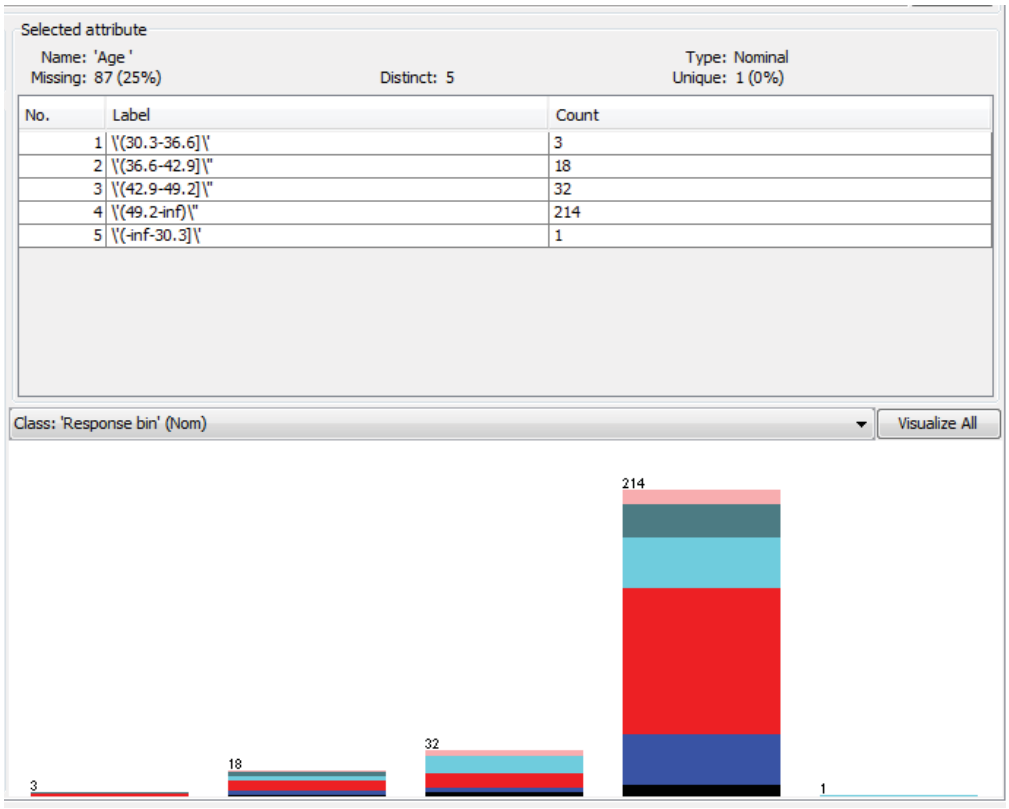

Figure 3.7 Discrete values of age

Stable disease Partial response Progressive disease

Complete response Miscellaneous response

Missing value imputation

Five attributes in the dataset contained missing values: age, gender, smoking, diagnosis, and RECIST response. One of the most drastic methods to deal with missing values is to remove the instances with incomplete data; however this approach would lead to a severe reduction of the already limited data size. Other common strategies are to replace missing values with the most common attribute value or its central tendency [153], but such

simplistic approaches can often bias the data [45]. Su et al demonstrated that classifier-based nominal imputation (CNI) [154] improves classification performance for learning algorithms. Thus, CNI was used to impute the missing values in the dataset. Table 3.5 shows the results of imputation on the counts for each attribute.

Attribute	Label	Count	
		Before imputation	After imputation
Age	≥30.3	1	1
	30.3-36.6	3	3
	36.6-42.9	18	18
	42.9-49.2	32	32
	≥ 49.2	214	301
Gender	Male	139	143
	Female	198	212
Smoking status	Never	174	176
	Former	72	145
	Current	23	34
Diagnosis	Adenocarcinoma	255	276
	SCC	21	21
	BAC	21	21
	AWBF	12	12
	LC	10	10
RECIST response	MR	44	44
	CR	15	15
	PR	143	151
	PD	80	80
	SD	61	65

Table 3.5 Missing value imputation results

Attribute construction

The original attribute of EGFR mutation status included 70 distinct mutations, some of which are rare and thus occurred infrequently in the dataset. Some authors have studied mutations and their relationship to EGFR-TKI response according to the mutation's physical location in the EGFR gene sequence (exon 18-21), type of mutation (point mutation, insertion, deletion, or duplication), and complexity (single mutation, double mutation [155], classical mutation or complex mutation [156]. Using this domain knowledge, a new attribute called EGFR class was constructed from the existing attribute of EGFR mutation [157]. The addition of EGFR class to the original dataset can improve the representation of the problem and aims to help predictive classifiers discern patterns that may otherwise be difficult to recognize.

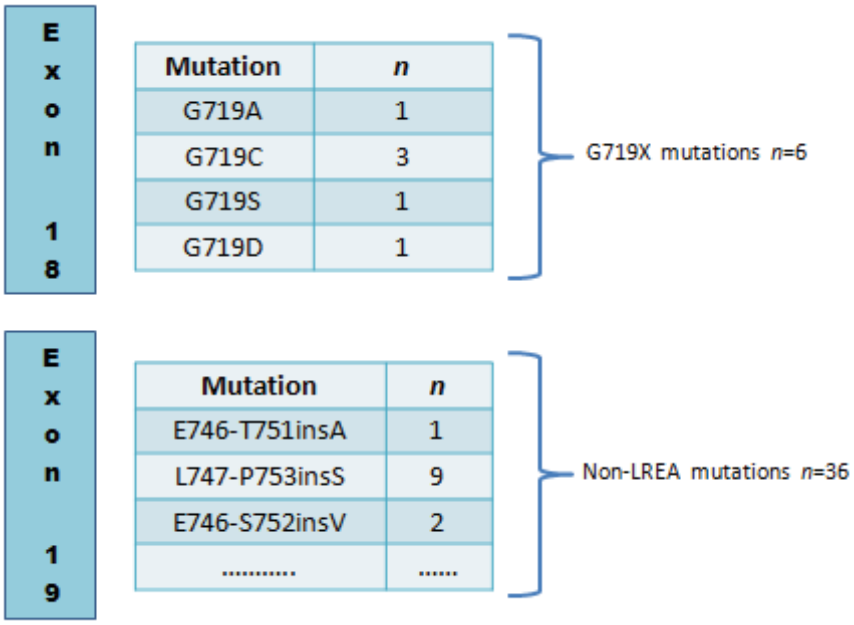

Figure 3.8 Example of attribute construction

3.5 Description of Final Dataset

The final dataset contained 355 data instances and eight attributes. The complete dataset was divided into a training set with 291 instances and a test set with 64 instances. The counts and labels of each attribute for both sets are shown in Tables 3.8 and 3.9.

Ref no	ID	Age	Gender	Smoking	Dx	EGFR mutations	EGFR class	Target drug	RECIST
1	1	'(49.2-inf)\"	f	never	AD	'exon 19 del'	exon 19 LREA del	'erlotinib therapy'	MR
2	2	'(42.9-49.2]\"	f	former	AD	'exon 19 del L747-T751'	exon 19 non LREA del	'erlotinib therapy'	CR
2	3	'(49.2-inf)\"	f	never	AD	'exon 19 E746-T751insI'	exon 19 non LREA del	'erlotinib therapy'	CR
3	4	'(49.2-inf)\"	m	never	AD	'exon 19 del'	exon 19 LREA del	'gefitinib or erlotinib'	PR
3	5	'(49.2-inf)\"	f	never	AD	'exon 19 del'	exon 19 LREA del	'gefitinib or erlotinib'	PR
3	6	'(42.9-49.2]\"	f	never	AD	wt	wildtype	'gefitinib or erlotinib'	PD
3	7	'(42.9-49.2]\"	m	never	AD	wt	wildtype	'gefitinib or erlotinib'	PD
3	8	'(49.2-inf)\"	m	former	NOS	wt	wildtype	'gefitinib or erlotinib'	PD
3	9	'(42.9-49.2]\"	m	former	NOS	wt	wildtype	'gefitinib or erlotinib'	PD
3	10	'(49.2-inf)\"	f	never	AD	wt	wildtype	'gefitinib or erlotinib'	PR
4	11	'(49.2-inf)\"	m	never	'acinar AD'	'exon 19 del'	exon 19 LREA del	'erlotinib therapy'	MR
4	12	'(49.2-inf)\"	f	never	'acinar AD'	'exon 19 del'	exon 19 LREA del	'erlotinib therapy'	MR
4	13	'(36.6-42.9]\"	f	never	'acinar AD'	'exon 19 del E746-A750'	exon 19 LREA del	'erlotinib therapy'	MR
4	14	'(36.6-42.9]\"	f	never	'acinar AD'	'exon 19 del E746-A750'	exon 19 LREA del	'erlotinib therapy'	MR
4	15	'(49.2-inf)\"	m	never	'acinar AD'	'exon 19 del E746-A750'	exon 19 LREA del	'erlotinib therapy'	MR
5	16	'(49.2-inf)\"	f	never	AD	'exon 20 A767-V769dupASV'	exon 20 ins/del/dup	gefitinib	PD
5	17	'(49.2-inf)\"	f	never	AD	'exon 20 D770-N771insD'	exon 20 ins/del/dup	gefitinib	PD
5	18	'(42.9-49.2]\"	f	never	AD	'exon 20 P772-H773insYNP + H773Y'	exon 20 ins/del/dup	gefitinib	PD
5	19	'(49.2-inf)\"	f	never	AD	'exon 20 R776G + exon 21 L858R'	exon 20 21 complex mutation	gefitinib	PD
5	20	'(49.2-inf)\"	f	never	AD	'exon 20 S768-D770dupSVD'	exon 20 ins/del/dup	gefitinib	PR
5	21	'(49.2-inf)\"	f	never	AD	'exon 20 S768-D770dupSVD'	exon 20 ins/del/dup	gefitinib	PD
5	22	'(49.2-inf)\"	f	never	AD	'exon 20 S768-D770dupSVD'	exon 20 ins/del/dup	gefitinib	PD
5	23	'(36.6-42.9]\"	f	never	AD	'exon 20 S768-D770dupSVD'	exon 20 ins/del/dup	gefitinib	PD
5	24	'(49.2-inf)\"	f	never	AD	'exon 20 T790M + exon 21 L858R'	T790M complex mutation	gefitinib	PD

Figure 3.9 Final dataset with eight attributes

Age: 85% of patients were > 50 years old. Discrete bins of age retained the original distribution of age in the data.

Gender: 40% patients were male and 60% were female.

Smoking status: Patient history of smoking was recorded as never smoker, former smoker or current smoker. 50% patients reported never smoking, 40% were former smokers and 10% were current smokers.

Diagnosis: Eleven distinct sub-histologies for NSCLC were documented. The details are shown in Table 3.6

Histology	Count (*n*)
Adenocarcinoma (AD)	276
Squamous cell cancer (SCC)	21
Bronchoalveolar (BAC)	21
AD with BAC (AWBF)	14
Large cell cancer (LCC)	10
Acinar Adenocarcinoma	5
Not otherwise specified (NOS)	3
Large cell neuroendocrine cancer (LCNEC)	2
Undifferentiated	1
BAC with focal invasion (BWFI)	1
Other	1

Table 3.6 Histopathological subtypes in NSCLC

EGFR mutations: Seventy distinct EGFR mutations spanning exons 18, 19, 20 and 21 were included in the dataset. These included point mutation, insertions, deletions, duplications, classical mutations, and complex mutations.

EGFR class: 18 EGFR classes were constructed from the exiting EGFR mutations. In general, a *complex mutation* represented point mutations from two different exons in a single patient, for example exon 20 21 complex mutations include 'exon 20 R776G + exon 21 L858R', 'exon 20 G779S + exon 21 L858R ', and 'exon 20 R776H + exon 21 L858R'. *T790M complex mutations* included the exon 20 T790M point mutation in combination with another EGFR mutation. There were three cases of such complex mutations and a review of the articles' methodology revealed that mutational analysis

was done on surgical tissue specimens obtained before the administration of TKI therapy. Thus, it was concluded that the presence of T790M was a baseline mutation and not the more common acquired resistance mutation that often develops after treatment with erlotinib and gefitinib. A *classical complex mutation* was defined as exon 19 deletion and exon 21 L858R point deletion co-existing. The term *double mutation* denoted two different point mutations on the same exon of EGFR. For example, exon 18 double mutations included 'exon 18 G719A + exon 18 S720F', 'exon 18 E709A + exon 18 G719C' and 'exon 18 G719A + exon 18 E709A'.If a mutation could not be grouped into the aforementioned classes, it was made into its own class, for example, the commonly occurring L858R mutation, exon 19 double mutation, and exon 20 T790M had their own class.

EGFR Class	**EGFR mutations**	**Count (*n*)***
Exon 19 LREA deletions	Any deletion in residues 746-750 of exon 19	94
Exon 21 L858R	Only exon 21 L858R point mutation	62
Exon 19 non LREA deletions	Any deletion outside of 746-750 of exon 19	36
Exon 20 insertion/deletion/duplication	Any insertion, deletion or duplication in exon 20	11
Classical complex mutation	Coexisting exon 19 deletion and L858R	10
Exon 21 double mutation	Two point mutation in exon 21	10
Exon 20 21 complex mutations	Coexisting exon 20 and exon 21 mutations	9
Exon 18 G719X	G719 A,C,S,D	6
Exon 21 L861Q	Only exon 21 L861Q point mutation	4
Exon 19 21 complex mutation	Coexisting exon 19 and exon 21 mutations	4

T790M complex mutation	T790M coexisting with one other mutation	3
Exon 18 21 complex mutation	Coexisting exon 18 and exon 21 mutations	3
Exon 18 double mutation	Two coexisting exon 18 mutations	3
Exon 21 point mutations	Rare point mutations in exon 21 exclusive of L858R and L861Q	2
Exon 19 double mutation	Two coexisting exon 19 mutations	1
Exon 18 deletion 719G	Only exon 18 719 deletion	1
Exon 20 T790M	Only exon 20 T790M	1

*the total number of EGFR classes does not add up to the individual mutations since more than one type of mutation was included in a class

Table 3.7 EGFR classes

Target drug: All patients were treated with an EGFR-TKI, either erlotinib or gefitinib. Three levels were created for target therapy: erlotinib therapy included those patients who received either erlotinib alone or in combination with another platinum-based chemotherapeutic agent, gefitinib therapy included patients who only received this agent, and erlotinib/gefitinib therapy was reserved for patients who had received either or both drugs during the course of treatment.

RECIST response: Most studies reported the treatment response as complete response (CR), partial response (PR), stable disease (SD) or progressive disease (PD) as defined by RECIST [18]. If the study did not specify the use of RECIST and response was classified using terms such as partial regression, complete regression, partial remission, complete remission, major response or minor response, it was grouped into a new category labelled Miscellaneous response (MR).

Feature	Label	Mutated	Wildtype	Total
Age	≤49 years	29	11	40
	≥50 years	188	62	250
Gender	Female	127	39	166
	Male	91	34	125
Smoking	Current	14	13	27
	Former	93	42	135
	Never	111	18	129
Diagnosis	Adenocarcinoma	165	52	217
	Large cell carcinoma	4	5	9
	Squamous cell carcinoma	10	7	17
	Other	39	9	48
Drug	Erlotinib therapy	66	3	69
	Gefitinib	138	41	179
	Gefitinib or Erlotinib	14	29	43
Response	Complete response	11	2	13
	Miscellaneous response	32	5	37
	Partial response	113	12	125
	Stable disease	39	15	54
	Progressive disease	23	39	62

Table 3.8 Training set

Feature	Label	Mutated	Wildtype	Total
Age	≤49 years	5	8	13
	≥50 years	37	14	51
Gender	Female	31	15	46
	Male	11	7	18
Smoking	Current	3	4	7
	Former	6	4	10
	Never	33	14	47
Diagnosis	Adenocarcinoma	40	19	59
	Large cell carcinoma	0	1	1
	Squamous cell carcinoma	2	2	4
Drug	Erlotinib therapy	1	2	3
	Gefitinib	33	14	47
	Gefitinib or Erlotinib	8	6	14
Response	Complete response	2	0	2
	Miscellaneous response	5	2	7
	Partial response	23	3	26
	Stable disease	7	4	11
	Progressive disease	5	13	18

Table 3. 9 Test set

3.6 Chapter Summary

This chapter discussed an evidence-based approach to determine variables that are needed to answer the research question. The PICO framework was used to isolate the essential components of the research question. The highest quality of evidence was assessed in sourcing the variables that are to be used subsequently to answer the research question. After determining the variables of interest, an appropriate data collection strategy was devised for the utilization of secondary heath care data from freely available sources to generate a dataset. An analysis of the final dataset revealed a mixture of numeric and nominal data with missing values. Data pre-processing steps such as feature selection, discretization, missing value imputation and attribute construction were performed to improve the quality of data before the application of analytical techniques.

CHAPTER 4: PATTERN DISCOVERY OF PATIENT CHARACTERISTICS AND TUMOR RESPONSE

4.1 Patient and Tumor Response Patterns in Advanced NSCLC

The discovery of EGFR mutations in NSCLC patients and the subsequent development of EGFR tyrosine kinase inhibitors such as erlotinib and gefitinib have revolutionized treatment options for patients with advanced stage NSCLC. EGFR-TKIs not only increase survival, but also improve tumor-related symptoms and quality of life in these patients [158]. The history of clinical trials in this area reveals the frequent patterns that have been uncovered in patients who are responders to EGFR-TKIs. This chapter will start by providing an overview of these trials and their findings in unselected and selected populations. Tables 4.1 and 4.2 provide details of both types of trials. Studies reported response rate (RR), PFS and OS, however for the purposes of this thesis, studies were compared only using RR since this endpoint is similar to the current research question's outcome measure of tumor response.

4.1.1 Unselected patient trials

Unselected trials are designed to test all patients enrolled regardless of their clinical risk factors or marker status [159]. In the case of advanced NSCLC, unselected trials did not test for EGFR mutation (marker) status as part of the study design although consent for tumor biopsy testing as part of retrospective analysis may have been obtained. The Iressa Dose Evaluation in Advanced Lung cancer (IDEAL) 1 trial was designed to compare the efficacy and safety of 250 mg versus 500 mg dose of gefitinib in previously treated advanced NSCLC patients [31]. Population analysis revealed that Japanese patients had significantly higher response rate than non-Japanese and female gender. Adenocarcinoma histology and performance status of 0-1 were baseline characteristics that may account for these differences. The IDEAL 2 study [33] confirmed the previous association of response and female gender reported by IDEAL 1. IDEAL 2 also demonstrated that adenocarcinoma histology, including cases with bronchoalveolar features and never smoking history, were strong predictors of tumor response in a NSCLC population

treated with gefitinib. BR.21 was a phase 3 clinical trial of erlotinib versus placebo in patients who had failed first or second-line chemotherapy which revealed a response rate of 8.9% in the erlotinib group and <1% in the placebo group [161]. A sub-group analysis in the same study population showed that women, non-smokers, Asian ethnicity, and adenocarcinoma were all factors that predicted response.

Author	Trial	Study phase	Treatment arms	RR (%)
Fukuka et al [31]	IDEAL-1	Phase 2	Gefitinib 250mg/500mg	18.4/19.0
Kris et al [32]	IDEAL-2	Phase 2	Gefitinib 250mg/500mg	12.0/9.0
Cufer et al [160]	SIGN	Phase 2	Gefitinib/Docetaxel	13.2/13.7
Shepherd et al [161]	BR.21	Phase 3	Erlotinib/Placebo	8.9/<1.0
Thatcher et al [67]	IRESSA	Phase 3	Gefitinib/Placebo	8.0/1.3
Kim et al [162]	INTEREST	Phase 3	Gefitinib/Docetaxel	9.1/7.6
Kelly et al [163]	SWOG S0023	Phase 3	Gefitinib/Placebo	8.3/11.7

Table 4.1 Unselected clinical trials for NSCLC

Although the response rates for erlotinib and gefitinib in unselected trials were not favorably higher that alternative treatment arms, the molecular characterization of lung tumors on the basis of EGFR mutation status was an emergent finding. Retrospective analysis has consistently established clinical and molecular predictors of response to EGFR-TKI therapy. Clinical predictors are Asian ethnicity, female gender, adenocarcinoma including bronchoalveolar histology, and non-smoking history [164], [165].

4.1.2 Selected patient trials

When compelling evidence suggests that certain subgroups of patients may benefit from treatment, a selected or enriched design strategy is adopted [159]. Such trials select patients based on risk factors and integrate molecular testing into the study design, thereby increasing the probability of improved clinical outcomes by offering tailored treatment selection. In 2004, Lynch [14] and Paez [13] demonstrated that somatic mutations in the tyrosine kinase domain of EGFR gene were positively associated with response to tyrosine kinase inhibitors. This landmark discovery led to a number of Phase 2 and 3 selected trials, some of which are summarized in Table 4.2.

Author	Trial	Study phase	Treatment	RR (%)
Sequist et al [96]	iTARGET	Phase 2	Gefitinib 250 mg	55.0
Mok et al [166]	IPASS	Phase 3	Gefitinib/Carboplatin +Paclitaxel	43.0/32.3
Lee et al [167]	FIRST SIGNAL	Phase 3	Gefitinib/Cisplatin + Gemcitabine	53.5/45.3
Mitsudomi et al [168]	WJTOG 3405	Phase 3	Gefitinib/ Cisplatin + Docetaxel	62.1/32.2
Maemondo et al [169]	NEJ002	Phase 3	Gefitinib/Carboplatin +Paclitaxel	73.7/30.7
Rosell et al [170]	EURTAC	Phase 3	Erlotinib/ Cisplatin + gemcitabine or docetaxel	58.0/15.0
Zhou et al [171]	OPTIMAL	Phase 3	Erlotininb/ Gemcitabine and Carboplatin	83.0/36.0

Table 4.2 Selected clinical trials for NSCLC

The iTARGET trial enrolled chemotherapy-naïve patients with advanced NSCLC and ≥1 clinical characteristic associated with EGFR mutations. The overall response rate (ORR) was 55%, but in patients with L858R and exon 19 deletions, the response was 78% and 59% respectively. IPASS (IRESSA Pan Asia Study) was a phase 3 randomized trial

comparing gefitinib with carboplatin and paclitaxel in never or former smokers with pulmonary adenocarcinoma. The overall RR was 43 vs 32.3%. In a sub-group analysis of mutation-positive patients, the ORR was 71.2% with gefitinib versus 47.3% with carboplatin–paclitaxel. WJTOG3405 and NEJ002 were trials performed in Japanese patients. In WJTOG3405, EGFR mutation-positive patients were randomly assigned to receive gefitinib or a combination of cisplatin and docetaxel. The RR was 62.1% and 32.2% respectively in the gefitinib and chemotherapy group respectively. The NEJ002 Trial compared the efficacy of gefitinib with carboplatin and paclitaxel for first-line treatment of EGFR mutation positive patients. RR was 73.7% in the gefitinib arm and 30.7% in the alternative treatment arm. Patients with sensitizing EGFR mutations (exon 19 deletion and L858R) were selected in the EURTAC study. The majority of the study population were females, never-smokers and had adenocarcinomas. The RR in the erlotinib arm was significantly higher than the chemotherapy arm. The Phase 3 OPTMAL trial compared erlotinib to gemcitabine and carboplatin in EGFR mutation-positive. The overall RR was 83% in the erlotinib arm and subgroup analyses for females, adenocarcinoma and never smoking history demonstrated that the RR for erlotinib versus chemotherapy was higher for these clinical risk factors.

For this thesis, the research dataset was created using multiple sources, each having its individual design, setting, intervention and outcome measurement protocol. Given the atypical nature of our dataset, it was hypothesized that patterns from these sources were carried over into the current research dataset and as such we expect to find many of the same patient and response behaviors. The identification of frequent patterns that corroborate with previous findings from the literature will facilitate the validation of the research dataset [172], which in turn will ensure that data being used are appropriate for the pursued outcome.

Several authors have reported the success of using frequent patterns, expressed as IF-THEN rules, in medical databases to confirm previous biomedical knowledge and reveal novel associations between variables. Agarwal et al performed an analysis of lung cancer data from the Surveillance, Epidemiology, and End Results (SEER) Program

to identify patients segments where survival time was higher or lower than the overall average survival time [173]. Using the HotSpot algorithm, the authors presented non-redundant survival patterns from lung cancer patients that conformed with existing medical knowledge. Ordonez et al performed frequent pattern mining to predict the presence or absence of heart disease [174]. Their results demonstrated that age, gender, diabetes and cholesterol levels were frequently occurring risk factors, corroborating current understanding of cardiac disease. Using a nephrology database, Elfangary et al found useful patterns of tests associated with the diagnosis of IgA glomerulonephritis. These patterns were validated by medical specialists and deemed useful and understandable [175]. Following a similar approach, it is proposed that the identification of frequent patterns in the research dataset can serve two purposes:

1. The discovery of associations that are similar to findings reported in the literature will validate the derived research data. For example, evidence suggests that female gender and never-smoking history are strong predictive factors for higher response rate to erlotinib or gefitinib. If this finding is converted to an IF-THEN rule, it may take the form: IF gender=female AND smoking history=never THEN responder.

2. The discovery of new association patterns that make sense biologically may further suggest new hypotheses that warrant additional investigation. Given that the research dataset consists of both frequent and rare mutations, it may be possible that previously unknown associations between EGFR mutations and other attributes could be discovered.

Association Rules

Frequent sets of items are often presented as IF-THEN rules, also known as association rules. Let $I = \{\ \}$ be a set of literals call items. Let D be a set of all transactions where each transaction T is a set of items such that $T \subseteq I$. Let X, Y be a set of items such that $X, Y \subseteq I$. An association rule is an implication in the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, $X \cap Y = \emptyset$ [176].

The left hand side (LHS) of the rule is called the antecedent and the right hand side (RHS) is called the consequent. An association rule, $X \Rightarrow Y$, implies that transactions which contain X, also contain Y.

Support

Support is the number of items in the dataset which correspond to the rule's antecedent and consequent. For any rule $X \Rightarrow Y$, the support s can be defined as

$Support(X=>Y) = P(X \cap Y) = \{X \cap Y\}/D$

Frequent itemset

Any itemset I is said to be frequent if its support is s or more.

Confidence

Confidence is the proportion of the examples covered by the antecedent that are also covered by the consequent. Both general and class association rules can only be mined using confidence. According to the support-confidence framework, support prunes the search space using its downward closure property and thereafter, confidence generates rules from frequently occurring itemsets.

$Confidence(X=>Y) = P(Y \mid X) = support(X \cup Y)/support(X)$

Lift

Lift is confidence divided by the proportion of all examples that are covered by the consequent.

$Lift(X=>Y) = P(X \cap Y)/P(X)\ P(Y)$

Since lift is a measure that is independent of support, it remains an important measure of evaluation of an association rule. A lift value of 1 is suggestive that the antecedent and the consequent are independent of each other and no correlation rule can be drawn. Positive lift indicates that the occurrence of the antecedent strongly affects the occurrence of the consequent.

Classical association rule learning

Classical algorithms use a two step process to produce association rules. In the first phase, the algorithm finds all large item sets whose support > minimum support. This is based on the downward closure property of the Apriori principle: a k-itemset is frequent only if all of its sub-itemsets are frequent. The first scan of the transaction database yields all frequent 1-itemsets which can then be used to generate frequent 2-itemsets. This process is repeated until no more *k*-itemsets can be found for a given value of *k*. In the second phase, for a given large itemset, the algorithm generates all rules whose confidence > minimum confidence.

Constrained association rule mining

If the LHS or RHS of an association rule is restricted to user-specified attributes, a subset of association rules will be generated [177]. This allows the user to bypass an exhaustive list of rules and focus on the presence of certain interesting items in the database. Using the market-basket example, constraining items to those that interest retailers would allow them to identify frequently purchased itemsets and recognize consumer behavior patterns. The systematic method is similar to the classical approach, where the algorithm first finds frequent itemsets that satisfy user constraints from the transaction database. Using these frequent itemsets, the algorithm constructs association rules in the form $X \Rightarrow Y$, where X and Y are of specific interest to the user [178].

Pattern mining algorithms

The original frequent itemset mining and association rule learning algorithm, Apriori, was proposed by Agrawal and Srikant [179]. This classical algorithm generates association rules by mining all frequent itemsets and returns rules that conform to the user defined minimum confidence threshold. Despite its historical significance, the algorithm suffers from drawbacks [180]. The primary limitation of this approach is the large number of frequent itemsets and redundant association rules that are generated, many of which are not interesting for the user. In addition, decisions about minimum support threshold and confidence values pose difficulties. Since then, many variations of the Apriori algorithm have been proposed, many of which improve upon these

shortcomings. For a comparative study and evaluation of popular association rule algorithms on real world and artificial datasets, see [181]. The introduction of constrained association mining [182], [183] limits the output of frequent itemsets and association rules to only those which contain user-specified items. Predictive Apriori was proposed [184] as an alternative to Apriori algorithm as it does not require the specification of support and confidence values. It performs a Bayesian calculation for the exact expected predictive accuracy which Scheffer [185] defines as:

Let D be a database whose individual records r are generated by a static process P, let X ⇒ Y be an association rule. The predictive accuracy c(X ⇒Y) =Pr(r satisfies Y |r satisfies X) is the conditional probability of Y ⊆ r given that X ⊆ r when the distribution of r is governed by P.

4.2 Our Pattern Discovery Approach

The objective in pursuing pattern discovery is to compare the regularities in the research dataset with previously well-understood patterns and trends in order to validate the research dataset, and even to discover some interesting new patterns. Two types of patterns relating to patient characteristics and treatment efficacy in advanced NSCLC are recognized in the literature and summarized from the unselected and selected trials. Studies report the co-occurrence of clinical patient attributes such as gender and smoking status and gender and histology. These associations may be referred to as *patient characteristic patterns*. Evidence also strongly suggests the association of sensitizing mutations in EGFR with treatment response after erlotinib or gefitinib. These associations are *tumor response patterns*. The frequent pattern mining approach is to identify both patient characteristic and tumor response patterns using classical and constrained association rule learning. In the classical approach, the algorithm will discover frequent patterns and association rules in the dataset. In constrained mining, the consequent will be constrained to include only the response attribute and this will specifically support the algorithm to detect tumor response patterns. The choice of algorithms include Apriori and Predictive Apriori which have been used with success in risk factor extraction [186], breast cancer [187], [188], lung cancer [189], and other clinical databases [190]. The

algorithms are implemented in the Waikato Environment for Knowledge Analysis (WEKA). Post-processing will reduce the large number of generated rules by selecting only those with that are potentially useful and interesting. An understanding of the data structure and domain knowledge is integrated into post-processing to prune and filter irrelevant rules. Finally, objective and subjective measures are applied to evaluate the interestingness of rules. Objective measures include support, confidence, lift and predictive accuracy. Subjective measures consist of unexpectedness and actionability. A schematic representation of the pattern discovery approach is shown in Figure 4.1.

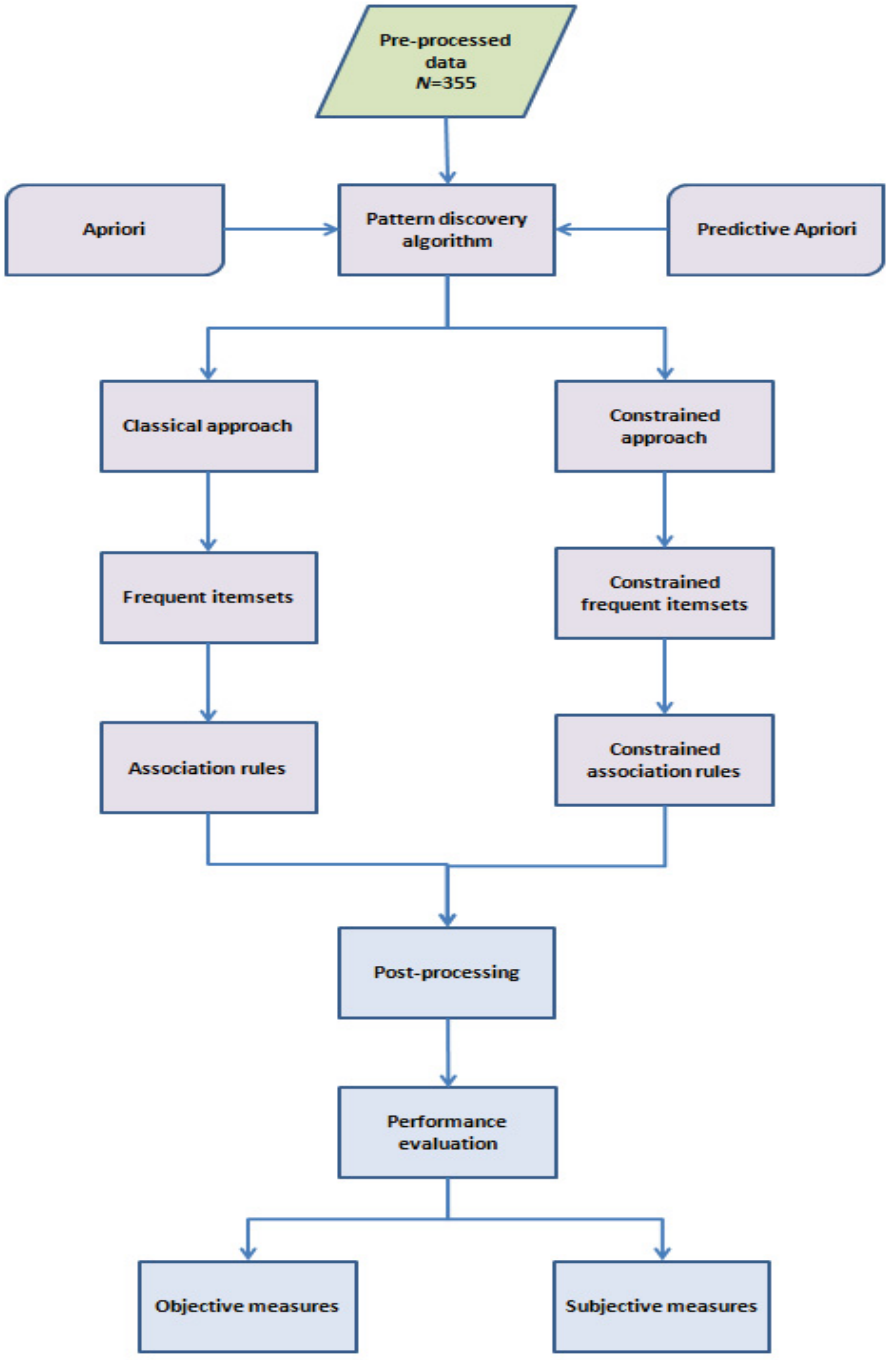

Figure 4.1 Schematic representation of pattern discovery approach

4.2.1 Experiment 1: Association Rules for Patient Characteristics and Tumor Response Using Apriori Algorithm

Three parameters are specified by the user in the Apriori algorithm: support, confidence or lift and number of rules. The output can be sorted according to confidence, lift, leverage or conviction. In the current research dataset, some classes occur more frequently than others. Since calculation of confidence is relies on frequency of the consequent, it often produces high confidence rules for the most common class [191]. To overcome this problem, lift was used to rank the rule interestingness.

The input data consisted of 355 instances and six attributes (age, gender, smoking status, diagnosis, EGFR mutation and RECIST response). EGFR mutation class was not included in the attributes because this attribute was constructed from EGFR mutation. Upper bound minimum support was set to 100% and the lower bound to 10%. Starting with the upper bound support, the algorithm incrementally decreased support by 5% and stopped when the lower bound for minimum support was reached or when a minimum of 50 rules with a minimum lift of 1.0 were discovered.

The output consisted of 50 association rules with lift values ranging from 1.01 to 1.29 and confidence values from 40-80%. Table 4.3 presents selected rules which are numbered sequentially and given the prefix “A” to denote the Apriori algorithm. The parameter settings for class pattern mining using the Apriori algorithm are shown in Figure 4.2.

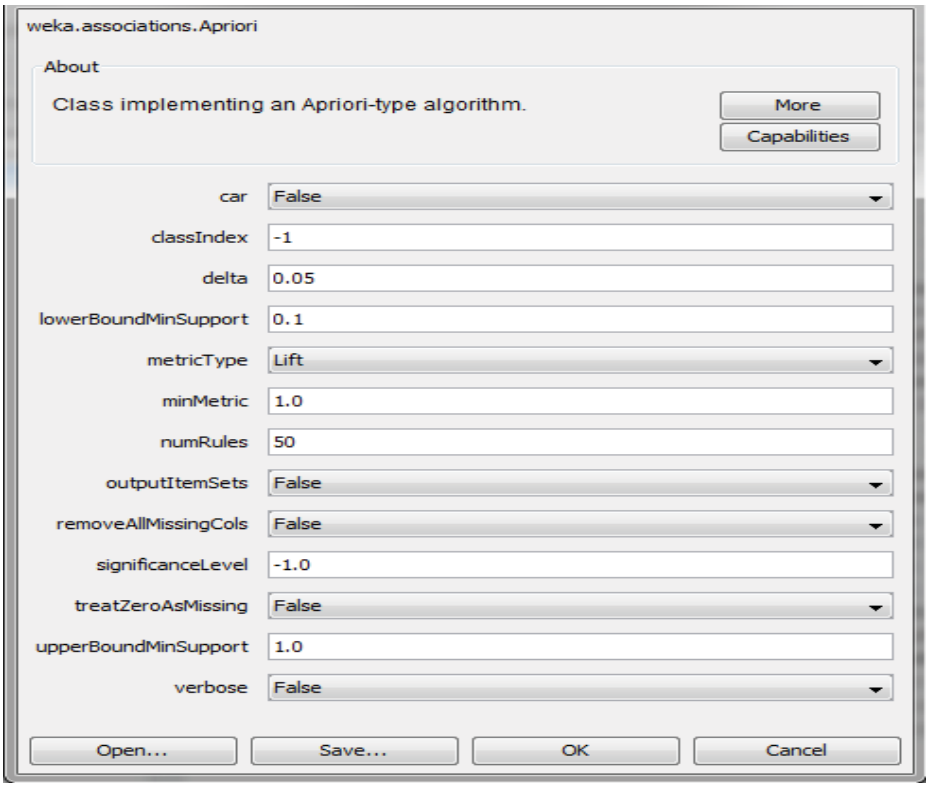

Figure 4.2 Apriori algorithm parameter settings

Using constrained association rule mining, tumor response patterns in the dataset were explored and compared with previously reported findings. Constrained association rule mining is performed by constraining the rule consequent to the class value (RECIST response). The resulting rules will be limited to only those that contain RECIST response in the RHS. The parameter settings for constrained pattern mining using the Apriori algorithm are shown in Figure 4.3. The input dataset contained 355 attributes and seven attributes (age, gender, smoking status, diagnosis, EGFR mutation and RECIST response). If *car* (class association rules) is set to true, then the algorithm only finds rules with the class in the consequent. Upper bound minimum support was set to 100% and the lower bound to 10%. In constrained mining, the algorithm only allows the use of confidence metric for ranking. Initially, the confidence is set to 90% and incrementally dropped by 10% until constrained rules are obtained. The minimum confidence at which 43 constrained rules are produced is 40%. Lift values are calculated using the rule's confidence as the numerator and the conditional probability of the consequent as the denominator.

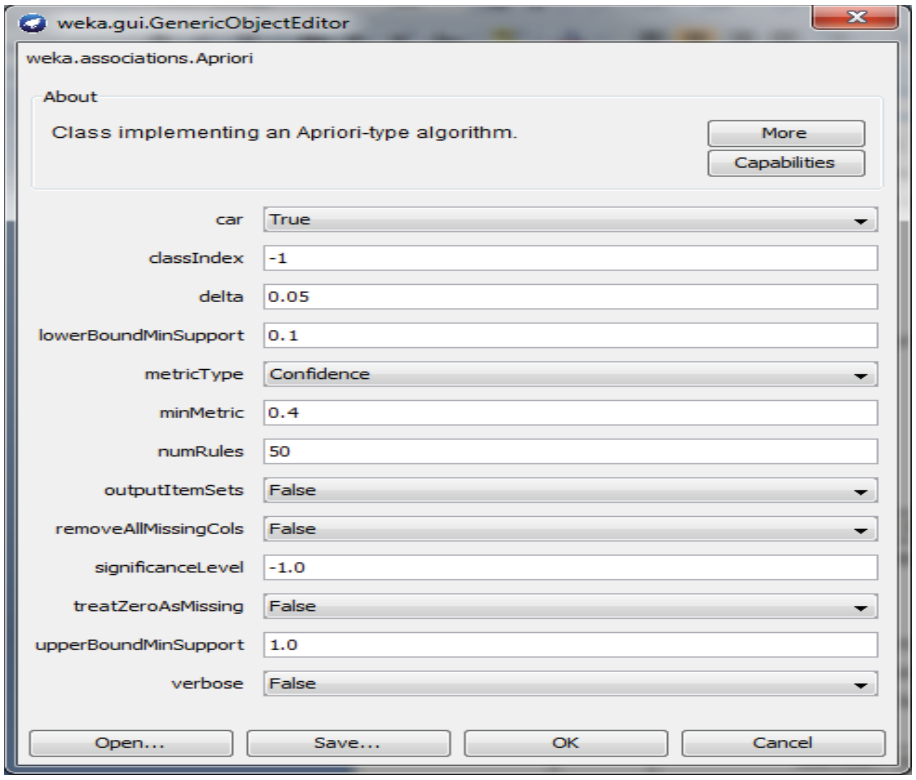

Figure 4.3 Apriori algorithm constrained mining parameter settings

4.2.1.1 Results from Experiment 1

Tables 4.3 and 4.4 present selected association rules that were discovered based on the specified thresholds for support, confidence and lift by the Apriori algorithm. A total of 93 rules were generated, many of which were redundant. These tables illustrate three rules obtained from general mining (Table 4.3) and three from constrained mining (Table 4.4). All rules concur with what is already known about patient characteristic and tumor response patterns in advanced NSCLC.

No.	Association Rule	Confidence	Lift
A1	*Smoking=never Diagnosis=adenocarcinoma ==> Gender=female* IF the patient reports a history of never smoking and the lung cancer diagnosis is adenocarcinoma, THEN the patient's gender is female.	0.77	1.29
A2	*Gender=female ==> Smoking=never* IF the patient's gender is female, THEN the patient reports history of never smoking.	0.60	1.21
A3	*Diagnosis=adenocarcinoma ==> RECIST response=partial response* IF the patient's lung cancer diagnosis is adenocarcinoma, THEN the tumor response recorded by RECIST is partial response to erlotinib or gefitinib.	0.47	1.11

Table 4.3 Selected association rules from Apriori algorithm. Rules are numbered with Prefix of A to denote Apriori algorithm

Rules A1 and A2 show the relationship between gender, smoking status, and histology. Females often report a history of never-smoking and the most common histology for NSCLC in never-smokers is adenocarcinoma [32]. In addition, similar to what many of the unselected clinical trials found, Rule A3 states that patient's with a diagnosis of adenocarcinoma achieve partial response after treatment with erlotinib and gefitinib [32].

No.	Constrained Association Rule	Confidence	Lift
A4	*Diagnosis=adenocarcinoma EGFR mutation=wildtype ==> RECIST response=progressive disease* IF the patient's lung cancer diagnosis is adenocarcinoma and EGFR mutation status is wildtype THEN the RECIST response is progressive disease.	0.58	2.58
A5	*EGFR mutation=exon 19 del E746-A750 ==> RECIST response =partial response* IF the patient's EGFR mutation status is exon 19 del E746-A750, THEN the RECIST response is partial response.	0.55	1.29
A6	*Gender=female Smoking=never Diagnosis=Adenocarcinoma ==> RECIST response =partial response* IF the patient's gender is female and the patient reports history of never smoking and the diagnosis is adenocarcinoma THEN the RECIST response is partial response.	0.44	1.03

Table 4.4 Selected constrained association rules from Apriori algorithm. Rules are numbered with Prefix of A to denote Apriori algorithm

Constrained patterns that were limited to the tumor response in the consequent also agreed with what the literature has reported. Rule A4 demonstrates the interaction of histology, mutation and tumor response. Although adenocarcinoma is considered a favorable histology [13], [14], the presence of EGFR wildtype status leads to progressive disease after treatment with an EGFR-TKI [192]. Sensitizing mutations in the tyrosine kinase domain of EGFR include deletions in exon 19 and L858R in exon 21 [193]. The presence of sensitizing mutations increases the efficacy of the EGFR-TKI and leads to an improved response as shown by Rule A5. According to Rule A6, elderly female patients

with diagnosed adenocarcinoma achieve a partial response when treated with gefitinib. This finding has been confirmed by the response rate in [31], [32]. Comparing Rule A3 and A4, we observe that a diagnosis of adenocarcinoma is frequently seen in patients who achieve partial response. However, when the antecedent contains the additional information of wildtype EGFR status with adenocarcinoma histology, there is a drastic change in the tumor response to progressive disease. This agrees with the current understanding of clinical and molecular predictors of response to erlotinib and gefitinib. The combination of clinical and molecular characteristics is more informative of treatment response in advanced NSCLC than clinical predictors alone [118].

4.2.1.2 Performance Evaluation of Experiment 1

In the objective performance evaluation we focus on the two metrics of lift and confidence. All association rules from Apriori algorithm had a lift value >1 demonstrating that the frequency of the antecedent and consequent occurring together was higher than the frequency of either occurring independently. Rule A3 suggests that EGFR-TKIs are effective in adenocarcinoma and frequently produce a partial tumor response. This rule has a confidence of 42% and a lift ratio of 1.11. When EGFR mutation status is included in the LHS, the improved confidence is 58% and lift ratio reaches 2.59 in Rule A4. This underlies the significance of genotyping for patient selection in personalized therapy of advanced NSCLC.

The lift ratio for general association rules (A1-A3) was slightly lower than that of the constrained association rules (A4-A6) establishing that tumor response patterns were more interesting than general patient characteristic patterns. Confidence levels varied with the frequency of the consequent in the dataset. Rules A1 and A2 had higher confidence levels, but when the consequent contained a value for tumor response, the confidence was low.

4.2.2 Experiment 2: Association Rules for Patient Characteristics and Tumor Response Using Predictive Apriori

Predictive Apriori has been proposed as an improvement on the classical Apriori algorithm because it eliminates the need to indicate a minimum support threshold and minimum confidence value. In WEKA's implementation of Predictive Apriori, the algorithm searches with an increasing support threshold for the best *n* rules. A rule is added to the output if the expected predictive accuracy of this rule is among the *n* best and it is not subsumed by a rule with at least the same expected predictive accuracy.

There are two user-defined parameters in the Predictive Apriori algorithm. If class association rules (car), is set to "false", no constraints are applied and the user only specifies the number of rules to be generated. The maximum number of rules is set to 100. The input data consisted of 355 instances and seven attributes. The output was 50 rules with accuracies ranging from 95%-99%. Table 4.5 presents selected rules which are numbered sequentially and given the prefix "P" to denote the Predictive Apriori algorithm. The parameter settings for the Predictive Apriori algorithm are shown in Figure 4.4.

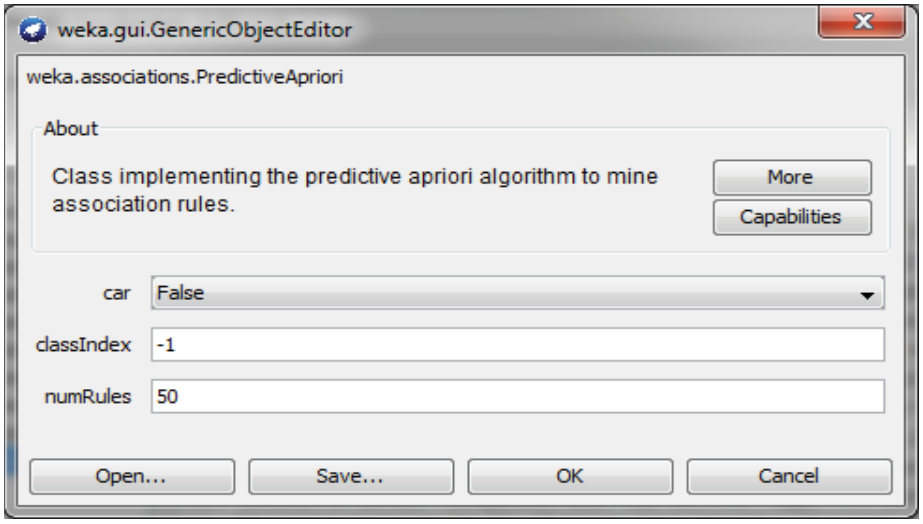

Figure 4.4 Predictive Apriori algorithm parameter settings

To constrain the rule consequent to class, the parameter settings were changed by setting class association rules (car) to “true” as shown in Figure.4.5.

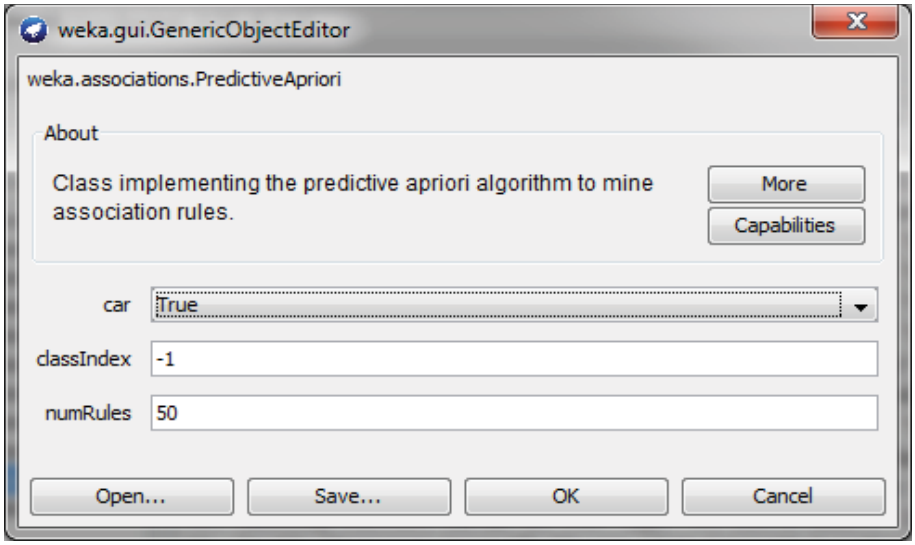

Figure 4.5 Predictive Apriori algorithm constrained mining parameter settings

4.2.2.1 Results from Experiment 2

A total of 100 rules were generated, many of which were redundant and some which were similar to the results from the Apriori. Tables 4.5 and 4.6 present the association rules that were discovered with and without constraints by the Predictive Apriori algorithm. The predictive accuracy range for all rules was 68%-99%. Table 4.5 highlights three rules obtained without any constraints and Table 4.6 shows six rules from constrained mining. Rules P1-P3 are similar to the patterns observed from Apriori’s result, so these are not discussed a second time. Constrained patterns uncovered by Predictive Apriori reveal the relationships of histologies and complex combinations of EGFR mutations with tumor response. From Rules P4 and P5, the diagnosis of adenocarcinoma with bronchoalveolar features and acinar adenocarcinoma occur frequently with a miscellaneous response. In Chapter 3, the levels of tumor response recorded after treatment with erlotinib or gefitinib were outlined. If the observed response did not follow RECIST, it was termed a miscellaneous response. Examples of miscellaneous response include partial regression, complete regression, partial metabolic remission, complete metabolic remission, major

response or minor response. Studies have shown that adenocarcinoma with bronchoalveolar features are responsive to gefitinib [33].

No.	Association Rule	Accuracy
P1	*Age = 42.9-49.2 Smoking=never ==> Diagnosis=adenocarcinoma* IF the patient's age is between 42.9-49.2 and patient has a history of never smoking THEN diagnosis is adenocarcinoma	0.99
P2	*Target drug =gefitinib or erlotinib RECIST response=partial response ==> Diagnosis=adenocarcinoma* IF targeted drug is either gefitinib or erlotinib and RECIST response is partial response, THEN patient's lung cancer diagnosis is adenocarcinoma	0.97
P3	*Gender=female Target drug =gefitinib or erlotinib RECIST response= progressive disease ==> EGFR mutation=wildtype* IF patient's gender is female and targeted drug treatment is with either gefitinib or erlotinib and RECIST response is progressive disease, THEN EGFR mutation status is wildtype	0.95

Table 4.5 Selected association rules from Predictive Apriori algorithm
(Rules are numbered with Prefix of P to denote Predictive Apriori algorithm)

No.	Constrained Association Rule	Accuracy
P4	*Diagnosis=adenocarcinoma with bronchoalveolar features ==> Response=miscellaneous response* IF patient's diagnosis is adenocarcinoma with bronchoalveolar features, THEN patient has response (not specified by RECIST)	0.99
P5	*Diagnosis=acinar adenocarcinoma ==> Response=miscellaneous response* IF patient's diagnosis is acinar adenocarcinoma, THEN patient has response (not specified by RECIST)	0.98
P6	*Gender=female Diagnosis=bronchoalveolar carcinoma EGFR mutation=exon 21 L858R ==> RECIST response=partial response* IF patient's gender is female and diagnosis is bronchoalveolar carcinoma and EGFR mutation status is L858R,THEN RECIST response is partial response	0.95
P7	*EGFR mutation=exon 20 G779S + exon 21 L858R ==> RECIST response=partial response* IF patient's EGFR mutation status is exon 20 G779S + exon 21 L858R, THEN RECIST response is partial response	0.92
P8	*EGFR mutation=exon 20 T790M + exon 21 L858R ==> RECIST response=progressive disease* IF patient's EGFR mutation status is exon 20 T790M + exon 21 L858R, THEN RECIST response is progressive disease	0.92
P9	*Gender=female EGFR mutation=exon 18 G719C ==> RECIST response=partial response* IF patient's gender is female and EGFR mutation status is exon 18 G719C, THEN RECIST response is partial response	0.92

Table 4.6 Selected constrained association rules from Predictive Apriori algorithm
(Rules are numbered with Prefix of P to denote Predictive Apriori algorithm)

According to Rule P6, female patients with bronchoalveolar carcinoma who have the sensitizing L858R mutation achieve partial response to EGFR-TKIs. The sensitizing role of L858R in EGFR mutated tumors treated with gefitinib is well documented [194], [195]. Similarly, the presence of exon 18 G719C mutation confers increased sensitivity (Rule P9) as shown by [196], [197],[106]. Rules P7 and P8 represent the unusual effect of co-occurring sensitizing and resistant EGFR mutations. The combination of exon 20 G779S + L858R results in partial response, whereas exon 20 T790M + L858R leads to progressive disease. T790M is strongly associated with resistance to drug susceptibility [198] especially after treatment with gefitinib [199], leading to stable or progressive tumor response. Exon 20 mutations are also associated with poor gefitinib response [197], however in combination with the classical L858R mutation, the response is favorable. This has been confirmed in [200] and [156]. Overall, the frequent patterns observed from Predictive Apriori corroborate with previously reported studies.

4.2.2.2 Performance evaluation of Experiment 2

The predictive accuracy of all rules without applying any constraints was 95-99%. When the consequent was constrained to tumor response, rule accuracy was 68-99%. The dynamic pruning process of Predictive Apriori produced a higher quality set of rules with reduced redundancy compared to Apriori. However the computational performance time is compromised-runtime for Predictive Apriori was 10 seconds without constraints and 3 seconds with constraints compared to <1 second for Apriori.

Liu et al proposed two subjective measures of interestingness [201]: unexpectedness and actionability, which state that rules are useful if they contradict what is previously known and can be acted upon. A complex mutation is the presence of more than one EGFR mutation in a single tumor sample [202] and the frequency of complex mutations is estimated to be 3-7% [156]. Although few associations of complex mutation patterns and tumor response have been reported, these mutations are not rare [202]. Further work is required to understand the interaction of sensitizing and resistant mutations. At present, the molecular mechanisms of tumor response in the presence of concurrent mutations are not clear.

4.3 Chapter Summary

The objective behind using pattern discovery was to uncover the inherent relationships of the research dataset and compare them to the patient characteristics and tumor response patterns that have emerged in the last ten years. Pattern discovery was pursued using both classical and constrained mining. Classical mining allowed all possible frequent itemsets to be discovered while constraining the consequent to tumor response allowed the algorithm to specially construct rules containing tumor response. Apriori and Predictive Apriori confirmed the relationship between the demographic and clinical factors in NSCLC. Gender, smoking status and histology were highly correlated. Sensitizing mutations in EGFR including exon 19 deletions and L858R, produced at least a partial response after treatment with erlotinib or gefitinib. Tumor response patterns for rare combinations of mutations were also discovered which is in part due to the method in which the dataset was created, i.e., using articles and case series which reported both common and uncommon EGFR mutations. The lift ratio and predictive accuracy of reported rules is high and they validate previous findings from the literature. More rules with higher performance metrics could be obtained given a larger dataset. In conclusion, the results suggest that patient characteristics and tumor response patterns found in the dataset are similar to previous research findings. Strong associations exist among female gender, never smoking history and adenocarcinoma histology. EGFR wildtype status frequently leads to progressive disease after EGFR-TKI therapy, whereas sensitizing mutations in exon 19 and exon 21 confer increased sensitivity.

The rules highlighted in this chapter correspond to combinatorial patterns that have been reported previously and helped to support the collection of the research dataset. These association rules contain the itemsets of female=gender, smoking status= never, and diagnosis= adenocarcinoma in the antecedent or consequent. As in many experimental procedures, there were a small number of unexpected results, which have not been elaborated upon. The association rules generated using Apriori algorithm yielded a total of 78 rules. Of these, 66 can be explained using analogous reasoning and previous reports and 12 were unexpected. Predictive Apriori generated a total of 100 rules from both experiments; of these, 98 corroborate the findings of previous work in personalized

therapy of NSCLC and 8 were unexpected. Unexpected rules associated male gender and former smoking status with adenocarcinoma diagnosis and partial response to targeted therapy. Additionally, some rules expressed a non-favorable response to erlotinib or gefitinib in the presence of sensitizing EGFR mutations. These rules may represent anomalies of this specific research dataset or require scrutiny and further experimentation with larger datasets. It is entirely possible that a number of interesting rules may be generated if variations of attributes and constraints are used.

CHAPTER 5: DATA-DRIVEN DECISION SUPPORT

5.1 Clinical Decision Support for NSCLC

An abundance of healthcare data exists in electronic health records, clinical trial reports, disease registries, and pharmaceutical records, yet, often remains unused [203]. Disparate sources, variable collection methods and complexity of medical data are only some of the challenging considerations that must be taken into account when trying to leverage this data to provide intelligent and actionable information. As we enter the generation of personalized medicine, it is becoming increasingly important to harness the power of health data and transform it to improve both individual patient care and healthcare systems [204].

Clinical decision support (CDS) tools can utilize streams of medical data to generate preventative, diagnostic and therapeutic decisions. These tools facilitate health care professionals to make clinical assessments, diagnoses, recommendations and decisions about individual patients at the point of care [205]. CDS presents intelligently filtered information at appropriate times for effective and efficient patient management using computerized alerts and reminders, clinical guidelines, and order sets, among other tools [206]. There are two main types of CDS tools: knowledge based and non-knowledge based [207]. Three components that are common to knowledge-based CDS tools (or expert systems), include a knowledge-base, inference engine and communication interface. The knowledge base is often in the form of guidelines, IF-THEN rules or probabilistic associations. The inference engine uses a reasoning mechanism to combine knowledge base rules with patient data. Finally, the communication component transmits the system output to the end user [207].

Although evidence-based medicine is practiced by adhering to clinical practice guidelines and reference to expert opinion, an alternative approach is to use retrospective patient data to derive evidence which can in turn support the decision-making process [208]. Such non-knowledge-based decision support models eliminate the need for rules or direct expert input in the knowledge-base; they use artificial intelligence techniques to learn

from clinical data and use this as knowledge to provide decision support [209]. This type of data-driven decision support can complement clinical experience and *a priori* knowledge using the accumulation of past patient cases that are digitally generated and widely documented [210], [211].

Clinical decision-making involves the interpretation of complex factors and this has encouraged the development of computerized models for aiding decision support in an attempt to reduce medical errors and improve patient outcomes [212]. In NSCLC, a wide array of factors such as patient demographics, pathological diagnosis, staging and genetic mutations influence the decision-making process [213]. Tools which can combine retrospective data to improve the treatment choice accuracy and clinical outcomes of patients have been successfully employed in NSCLC. Machine learning and statistical pattern recognition are two of the most popular techniques in artificial intelligence that have gained popularity in the biomedical community and review of both supervised and unsupervised approaches is provided in [214]. Many clinical prediction models for NSCLC exist and have been reviewed in Chapter 2; the use of machine learning techniques for clinical decision support in lung cancer, however, is limited. A PubMed search using the terms (("Artificial Intelligence"[Mesh]) AND ("Decision Support Systems, Clinical"[Mesh] OR "Decision Support Techniques"[Mesh])) AND "Lung Neoplasms"[Mesh] yielded 14 results. Using the Query Builder in EMBASE, the Emtree terms ('machine learning'/exp OR 'machine learning' AND ('decision support'/exp OR 'decision support') AND ('lung cancer'/exp OR 'lung cancer') resulted in seven publications, with some overlapping studies. Relevant studies, authors, year of publication and application descriptions are summarized in Table 5.1

Author/ year	Clinical application	Machine learning method
Lee et al/ 2012 [215]	Early diagnosis of NSCLC	Random forest
Zhao et al/ 2011 [216]	Treatment regimen for NSCLC	Q-learning
van den Branden et al/ 2011 [217]	Case based reasoning for lung cancer	Genetic algorithm & k-nearest neighbor
Lee et al/ 2010 [218]	Diagnosis of pulmonary nodules	Genetic algorithm & random subspace method
Sacile et al/ 2003 [219]	Malignancy associated changes (MAC) in lung cancer	Artificial neural network
Wolfe et al/ 2004 [220]	Nuclear morphometry for diagnosis of NSCLC	Logistic regression and CART
Biganzoli et al/ 1998 [221]	Censored survival times for lung cancer	Feed forward neural network
Schweiger et al/ 1993 [222]	Evaluation of lung cancer tumor markers	Back propagation neural network
Gutte et al/2007 [223]	Lung cancer PET/CT staging	Neural network
Polak et al/ 2004 [224]	Pharmacoeconomics	Neural network
Campadelli et al/ 2006 [225]	Lung cancer nodules digital radiographs	Support vector machine

Table 5.1 Reports of clinical decision support in lung cancer

Individualized therapy in advanced NSCLC depends on the accurate classification of patients into groups based on their responsiveness to targeted therapy. Although a large amount of literature is dedicated to the role of clinical and molecular predictors of response to EGFR-TKI in NSCLC [226-228], to the author's knowledge, there are no studies that have demonstrated the use of individual-level data to develop data-driven clinical decision support for predicting tumor response to erlotinib or gefitinib. This

chapter will demonstrate that data-driven models can predict the therapeutic response based on a composition of clinical, histological and molecular factors.

5.2 Measuring Tumor Response

The focus of this chapter is to investigate data-driven decision support in determining EGFR-TKI responsiveness for advanced NSCLC. Before outlining the solution approach, the **concept and significance** of assessing and measuring tumor response will be defined. Clinical assessment of tumor burden is an important feature of both cancer therapeutics and molecular medicine. Patients with identical clinical symptoms and histopathological characteristics may respond differently to the same drug and the measurement of response helps describe this complex biologic phenomenon [229]. Although the ultimate gold standard for treatment efficacy is the improvement in overall survival, tumor response is a surrogate marker for measuring tumor shrinkage following chemotherapy. Therapeutic effectiveness of anticancer agents is evaluated by the reduction in tumor size as assessed by anatomic and diagnostic imaging modalities which objectively quantify tumor dimensions. Currently, RECIST is the most commonly used standard set of rules for evaluating tumor response [230]. According to these criteria, linear measurement of tumor target lesions is a sufficient proxy for tumor mass. RECIST recommends the following four categories of tumor response to anticancer drug treatment: complete response (CR) is the complete disappearance of all target lesions; partial response (PR) is 30% decrease in the sum of the longest diameter of target lesions; progressive disease (PD) is 20% increase in the sum of the longest diameter of target lesions; and stable disease (SD) consists of small changes that do not meet the preceding criteria [18].

Assessment of anticancer tumor response serves at least two distinct purposes [17]. Firstly, it functions as a surrogate endpoint in Phase 2 and 3 clinical trials, by evaluating the clinical benefit and benefit-to-risk profile of intervention. Chemotherapeutic agents that provide significant clinical benefit may be eligible for accelerated approval [231]. Secondly, the evaluation of tumor response in the context of clinical, histological and molecular factors enables healthcare providers to target patients who will benefit from treatment before the initiation of therapy. Timely assessment of tumor response avoids

adverse effects from needless chemotherapy and allows the prompt transition to alternative treatment options [232]. For patients undergoing treatment, healthcare decisions to continue, modify or withdraw chemotherapy also rely on the periodic evaluation of tumor response.

5.3 Solution approach for Developing Decision Support Model

There are two major observations which guided the solution approach. First, there is a lack of data-driven decision support in the context of determining the responsiveness to EGFR-TKIs such as erlotinib and gefitinib, for the treatment of advanced NSCLC. Secondly, the integration of clinical and molecular predictors to the outcome of tumor response is not well established. Using the studies reported in Table 5.1 as a guide, it was determined that the classification algorithms could be employed to predict tumor response to targeted therapy in advanced NSCLC. The six steps of the solution approach are described in detail in the following sections.

1. Determining the feature vector

Classification seeks to determine if learning algorithms can be used to accurately predict patients as responders or non-responders from treatment with erlotinib or gefitinib in advanced NSCLC. Using domain knowledge and literature as a guide, a number of attributes were defined that contribute to the responsiveness to EGFR-TKIs in advanced NSCLC. As described in Chapter 3, age, gender, ethnicity, smoking status, and genetic polymorphisms in EGFR, KRAS and EML4-ALK were identified as potential attributes to describe the research problem. Using these attributes, a feature vector was constructed consisting of eight attributes (age, gender, smoking status, diagnosis, EGFR mutation, EGFR class, target therapy, and response). The dataset generated from Chapter 3 was used for classification. The attributes selected for classification are shown in Figure 5.1.

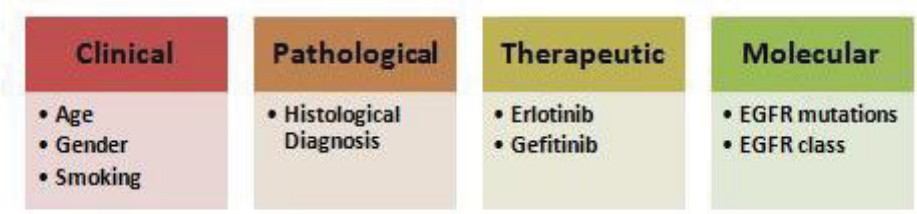

Figure 5.1 Attributes for classification

2. Determining the output of interest

As discussed previously, tumor response was selected as the outcome of interest. The original tumor responses to erlotinib and gefitinib were recorded according to RECIST (CR, PR, PD or SD) or as miscellaneous response (MR), if RECIST was not followed in the data sources. The distribution of instances for each of these five classes was: MR 12.7%, CR 4.5%, PR 43%, PD 21.3%, and SD 18.6%. Recognizing that the class imbalance would cause learning algorithms to be dominated by the majority class [233], a binary outcome using the multiclass outcome was also constructed.

In the binary class model, the original response attributes were recoded into two classes. The criteria used to group the response attributes is based on two main approaches cited in the literature. Traditionally, objective response rate is calculated from CR + PR [234] but recently the development of disease control rate (DCR) [235] has been proposed, where CR + PR + SD were shown to be stronger predictors of survival. Using a modification of the former approach, we categorized the multiclass attributes so that CR, PR and MR acquired the label "responder", whereas SD and PD were labeled "non-responder" [236]. In the binary model, the distribution of the classes was 60% and 40%, for responders and non-responders respectively.

The task for classification was performed over two models, using multiclass and binary response outcomes. In the remainder of the chapter, these are referred to as the multiclass model and the binary class model.

3. Establishing the role of clinical, molecular and integrative predictors

As described in Chapter 4, associations of varying strength exist between different attributes describing NSCLC. The following categories of attributes are defined:

Clinical predictors of response include patient variables and the histological diagnosis.

Molecular predictors of response are reviewed in Chapter 2 and include EGFR, KRAS, and EML4-ALK; this research focuses on EGFR mutations.

Integrated predictors are the sum of clinical and molecular predictors.

In order to determine the independent contribution of clinical, molecular and combined predictors to the response outcome, the dataset was divided as shown in Figure 5.2.

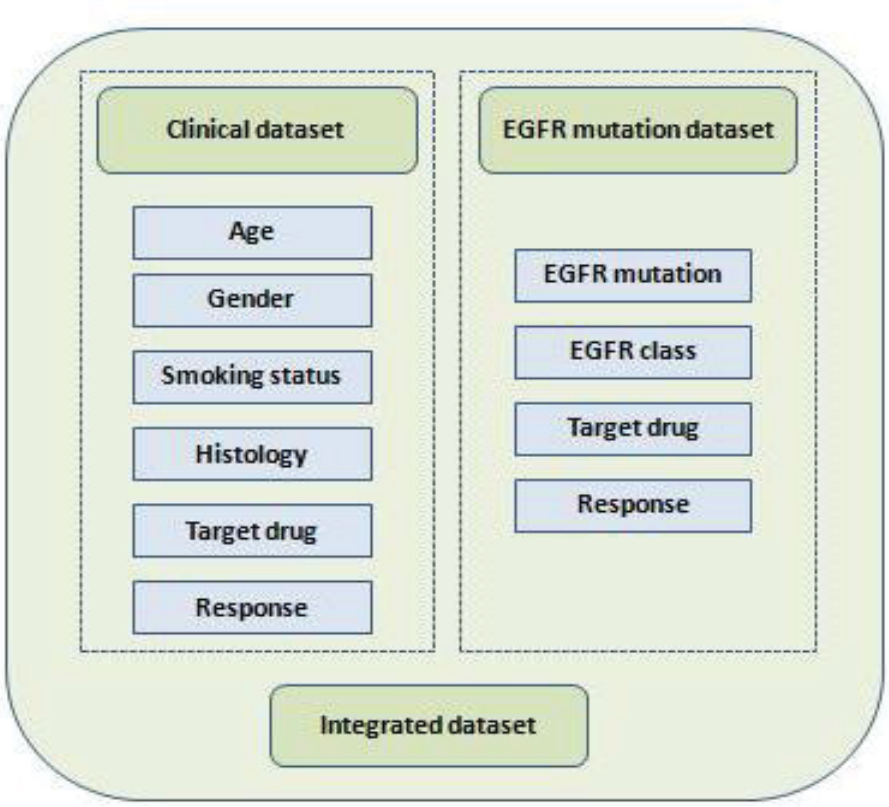

Figure 5.2 Clinical, EGFR mutation and Integrated datasets

Clinical dataset: included the attributes for age, gender, smoking, diagnosis, target drug, and response.

EGFR mutation dataset: comprised four attributes, namely, EGFR mutation, EGFR class, target drug, and response.

Integrated dataset: consisted of a combination of all attributes from the dataset, specifically, age, gender, smoking, diagnosis, EGFR mutation, EGFR class, target drug, and response.

The objective was to investigate how separate subsets of features correlate to the targeted drug response and whether the integration of clinical data with EGFR mutations enhances the predictive power of classification. A similar approach was adopted by Berrar et al [237], where the authors demonstrated that a combination of microarray data with clinical patient data improved the predictive accuracy for 5-year lung cancer survival.

4. Choice and comparison of classification algorithms

The selection of learning algorithms for a specific problem has been the subject of much debate [238- 240], and the *No Free Lunch Theorem* (NFL) recognizes that there is no optimal classifier for a given problem [241]. The choice of classifiers is based on studies which have used machine learning algorithms for predictive modeling in NSCLC. The selection of algorithms included Naïve Bayes [242], neural network [243], [244], Support vector machine [245-247], and classifier trees [248-250].

All learning algorithms were trained and tested on the Clinical, EGFR mutation and Integrated datasets. The results of the top four performing classifiers, support vector machine, J48, Random forest and classification and regression tree (CART), are reported here.

5. Evaluating classifier performance

Caruana et al discuss the value of evaluating algorithms over a variety of performance metrics, as individual algorithms are designed to enhance different performance measures [251]. Classifier performance was evaluated using the metrics of accuracy, error,

precision, recall and the receiver operator curve (ROC) or the area under the curve (AUC). The 2x2 confusion matrix describes the performance of a classifier with its predictions (Predicted) against the actual target classes (Actual). The confusion matrix of a binary class problem and resulting definitions of metrics used in this work, are defined as shown in Figure 5.3.

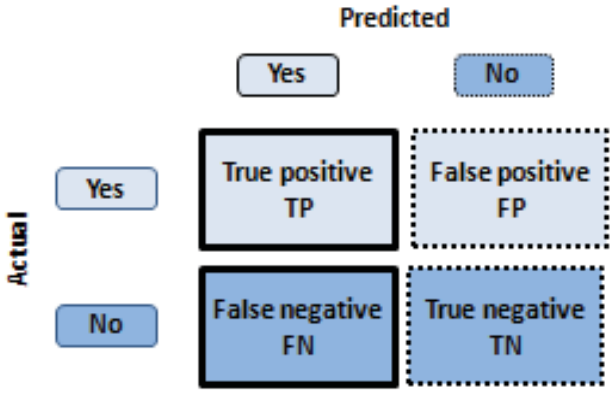

Figure 5.3 Confusion matrix of a binary class problem

Accuracy= (TP + TN) / (TP + TN + FP + FN)

Error=1-Accuracy

Precision= TP / (TP + FP)

Recall= TP / (TP + FN)

Accuracy and error are the simplest and most intuitive measures of classifier performance; having said that, the simplicity comes at a price. Accuracy does not distinguish between FP and FN and in clinical classification problems, this poses a tremendous disadvantage. This can be further explained in the context of the current research question where classification was used to predict patients as responders or non-responders from treatment with erlotinib or gefitinib. In this scenario, if patients with advanced NSCLC who are responders to EGFR-TKI therapy are predicted as non-responders by the learning algorithm, they fail to receive targeted therapy. Extensive

studies have demonstrated that patients harboring characteristics such as female gender, adenocarcinoma and EGFR mutation show a significant response to EGFR-TKIs and this in turn confers a distinct survival advantage [252], [253], [169]. Therefore, misclassifying patients as non-responders has the detrimental consequence of withholding treatment that could potentially increase survival.

Since accuracy does not capture these details, it was necessary to introduce two measure commonly used to test the performance of medical screening tests-sensitivity and specify. Sensitivity and specificity are defined as follows:

Sensitivity (or recall) = TP / (TP + FN)

Where sensitivity is the proportion of patients predicted as responders who are in fact responders to EGFR-TKI therapy.

Specificity = TN / (TN + FP)

Where specificity identifies the proportion of patients predicted as non-responders, who are truly non-responders.

ROC is fundamental performance tool for predictive diagnostic testing in clinical sciences that combines the two measures of sensitivity and specificity [254]. ROC is defined by plotting sensitivity against '1 – specificity'. A learning model that has perfect prediction will generate an ROC curve that follows the y-axis along the upper left quadrant of the ROC plot, while a model with random predictions will generate a ROC curve that follows the 1:1 line.

AUC is derived from the ROC and expressed as a proportion of the total area of the square defined by the axes [255]. The maximum AUC value is thus 1.0 for perfect predictive ability and 0.5 if the model's predictive value is no better than random.

Domain validation using expert knowledge and biological information reported in previous studies was also used to validate performance of the integrated model [256].

6. Interpreting the results

In addition to evaluating the classification accuracy of a model, it is crucial to understand and recognize the reasons behind each class decision. Decision trees produce human-readable descriptions of the underlying data structure and as such present an intelligible and powerful technique in medical decision-making. Following the completion of each experiment, the output from tree-based classification was analyzed for its readability and comprehensibility of representation.

5.4 Overview of Classification Experiments

All classification techniques were carried out using the open source Waikato Environment for Knowledge Analysis (WEKA) [153]. The systematic approach used for both experiments included dividing the Clinical, EGFR mutation and Integrated datasets into training (n=291) and test sets (n=64). The training set was used to build several models using different algorithms. Feature selection and parameter tuning were performed to optimize the models. Subsequently, these models were applied to the test sets and those that best fit the relationship between the input features and class labels were identified using performance evaluation. Figure 5.4 summarizes the methodology for the classification experiments.

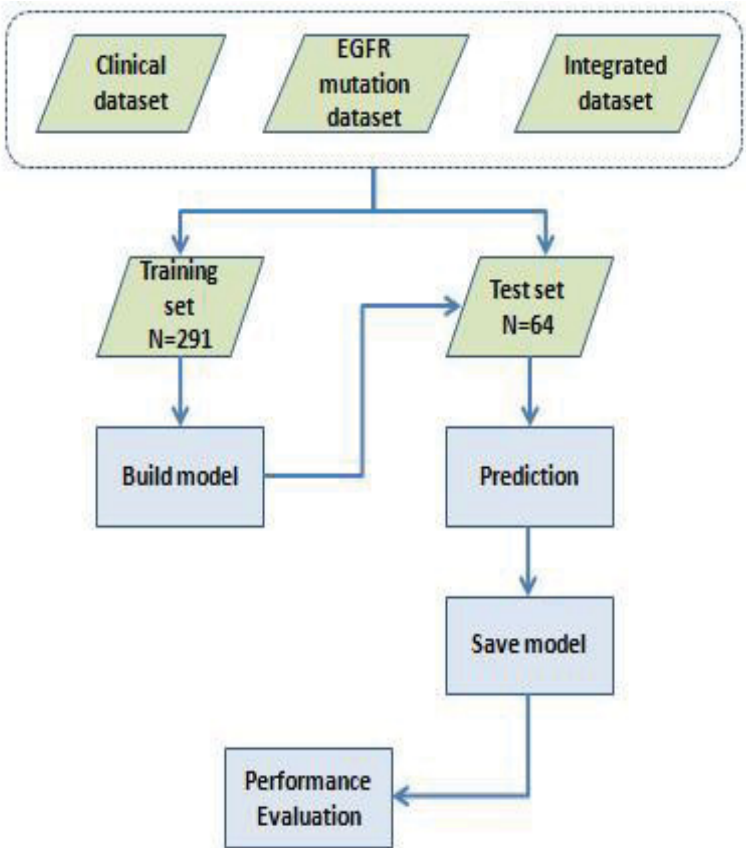

Figure 5.4 Overview of classification methodology

The classification task consisted of two experiments: Experiment 1 was a multiclass model (complete response, partial response, miscellaneous response, stable disease and progressive disease) and Experiment 2 was a binary class model (responder and non-responder).

Feature selection

Both experiments were performed with and without feature selection. Attribute selection was performed using the Correlation Feature-based Attribute evaluator (CfsSubset evaluator in WEKA), which considers the redundancy between features as well as the correlation to the class in order to produce a subset of features with minimum redundancy and maximized predictive ability [257].

Classifiers

Experiments were performed using support vector machine and three tree-based classifiers, J48, Random forest and CART.

Support vector machine

WEKA uses Platt's sequential minimal optimization (SMO) algorithm to train a support vector classifier, where multiclass problems are solved using pairwise classification [258], [259]. The optimal parameters for training the support vector were polykernal using a complexity parameter of 1.0 and a tolerance parameter of 0.0010.

J48

The classic C4.5 is re-implemented as J48 [260] in WEKA. The decision tree was created using a confidence threshold of 25%, a minimum of two instances per leaf, and three instances were used for error pruning. Confidence-based post-pruning and sub-tree raising was used to prune the decision tree.

Random forest

The number of attributes for random selection was set to log_2(number_of_attributes)+1. Ten trees were generated and one execution slot was used to construct the ensemble.

Simple CART

Simple CART uses minimal cost complexity pruning to produce a classification tree [261]. A heuristic search was used for binary split for nominal attributes and fivefold cross validation was used in the minimal cost complexity pruning.

5.4.1 Experiment 1: Classification for Multiclass Model

Results of Correlation Feature-based Attribute evaluator

Clinical dataset: Using best first search method, 18 subsets were evaluated and the merit of the best subset was 0.177. Only the attribute Diagnosis was selected.

EGFR mutation dataset: 6 subsets were evaluated and the best merit was 0.24. Only the attribute EGFR mutation was selected.

Integrated dataset: Using the best first search method on the full training set, 30 subsets were evaluated and the merit of the best subset was 0.276 which included the attributes of Diagnosis and EGFR mutation.

Performance Evaluation

The evaluation of classification is based on the test set performance (n=64). ROC area weighted averages for multiclass labels are shown in the Table 5.2.

Overall, algorithm performance on the Clinical dataset was poor when compared to the EGFR mutation and Integrated dataset. In the multiclass model, the accuracy of all algorithms on the Clinical dataset was at least 15.62% lower than the other two sets. Accuracies were comparable between the EGFR mutation and Integrated datasets, but precision for SMO and J48 increased when combined features were used. Also, the weighted ROC area for SMO, J48, and CART showed slight improvement when Integrated data was used instead of simply EGFR mutations. In the Integrative dataset, Random forest outperformed all algorithms with an accuracy of 56.69%, while the level of accuracy was equivalent for SMO, J48, and CART (56.25%). The weighted ROC was the highest for SMO, followed by CART, J48 and Random forest.

Dataset	Classifier*	Accuracy	Error	Precision	Recall	ROC area
Clinical	SMO	37.50	62.50	0.261	0.375	0.51
	J48	40.63	59.38	0.278	0.406	0.51
	Random Forest	40.63	59.38	0.278	0.406	0.51
	CART	40.63	59.38	0.278	0.406	0.51
EGFR	SMO	56.25	43.75	0.389	0.563	0.66
	J48	56.25	43.75	0.389	0.563	0.66
	Random Forest	56.25	43.75	0.389	0.563	0.66
	CART	56.25	43.75	0.389	0.563	0.66
Integrated	SMO	56.25	43.75	0.4	0.563	**0.69**
	J48	56.25	43.75	0.394	0.563	0.67
	Random Forest	54.69	45.31	0.389	0.547	0.65
	CART	56.25	43.75	0.452	0.563	0.68

*results for tree based classifiers are with CFS and results for SMO are without CFS

Table 5.2 Performance evaluation of classifiers from Experiment 1

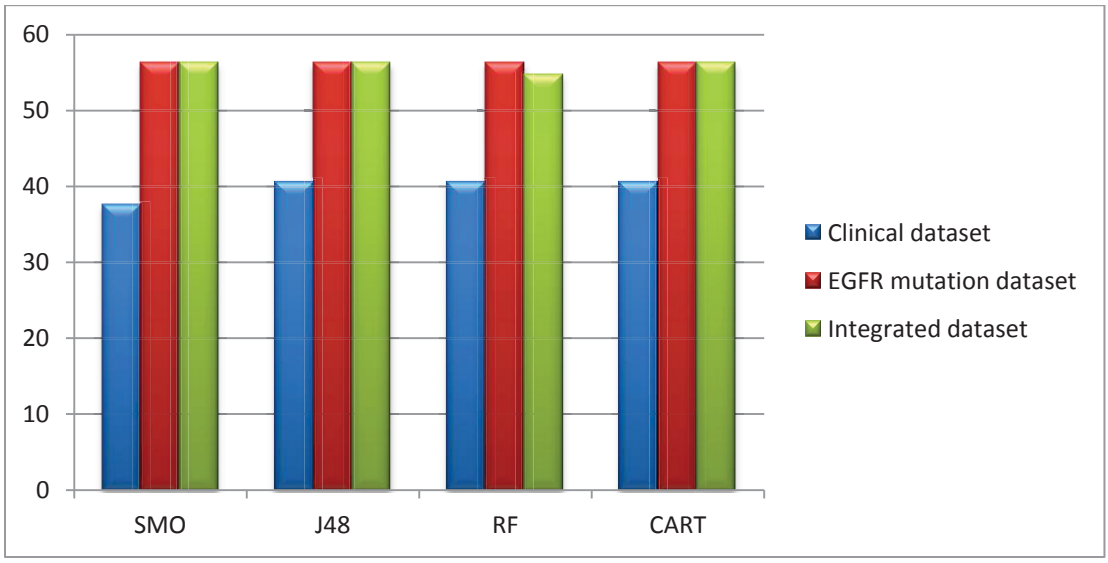

Figure 5.5 Accuracy comparison of four classifiers for three models in Experiment 1
SMO:support vector machine, J48:implementation of C4.5, RF: random forest, and CART: classification and regression tree

Analysis of Decision tree

Two sections of the decision tree output from the Integrated dataset in Experiment 1 are highlighted in Figure 5.6. The decision tree selected EGFR mutations as the root node. Because the dataset had 67 distinct mutations, the resultant tree is too large to display. For the purposes of illustration, Figure 5.6 shows EGFR classes instead. Numbers at the end of classes in brackets indicate the number of instances that follow the formula, followed by the number of misclassified instances.

Within the most commonly detected mutations of exon 19 deletion E746-A750 and exon 21 L858R, NSCLC sub-histologies differed in their individual responses. Adencarcinoma (AD), adenocarcinoma with bronchoalveolar features (AWBF), and bronchoalveolar carcinoma (BAC) were all responsive to EGFR-TKI therapy. However, squamous cell carcinoma (SCC) and large cell carcinomas (LC) had stable or progressive disease. Adenocarcinoma and its variants have been well studied in NSCLC and EGFR mutations frequently occur in this pathological subtype, leading it to be the most sensitive histology. Little information is available for the specific response to TKIs in SCC and

LC, however data from the SEER program revealed that one year survival in LC patients was less than 12.8%; the lowest observed in all histologic subtypes [262]. It is possible that even in the presence of sensitizing EGFR mutations, specific tumor histologies remain non responsive to targeted treatments due to the presence of coexisting primary resistance mutations.

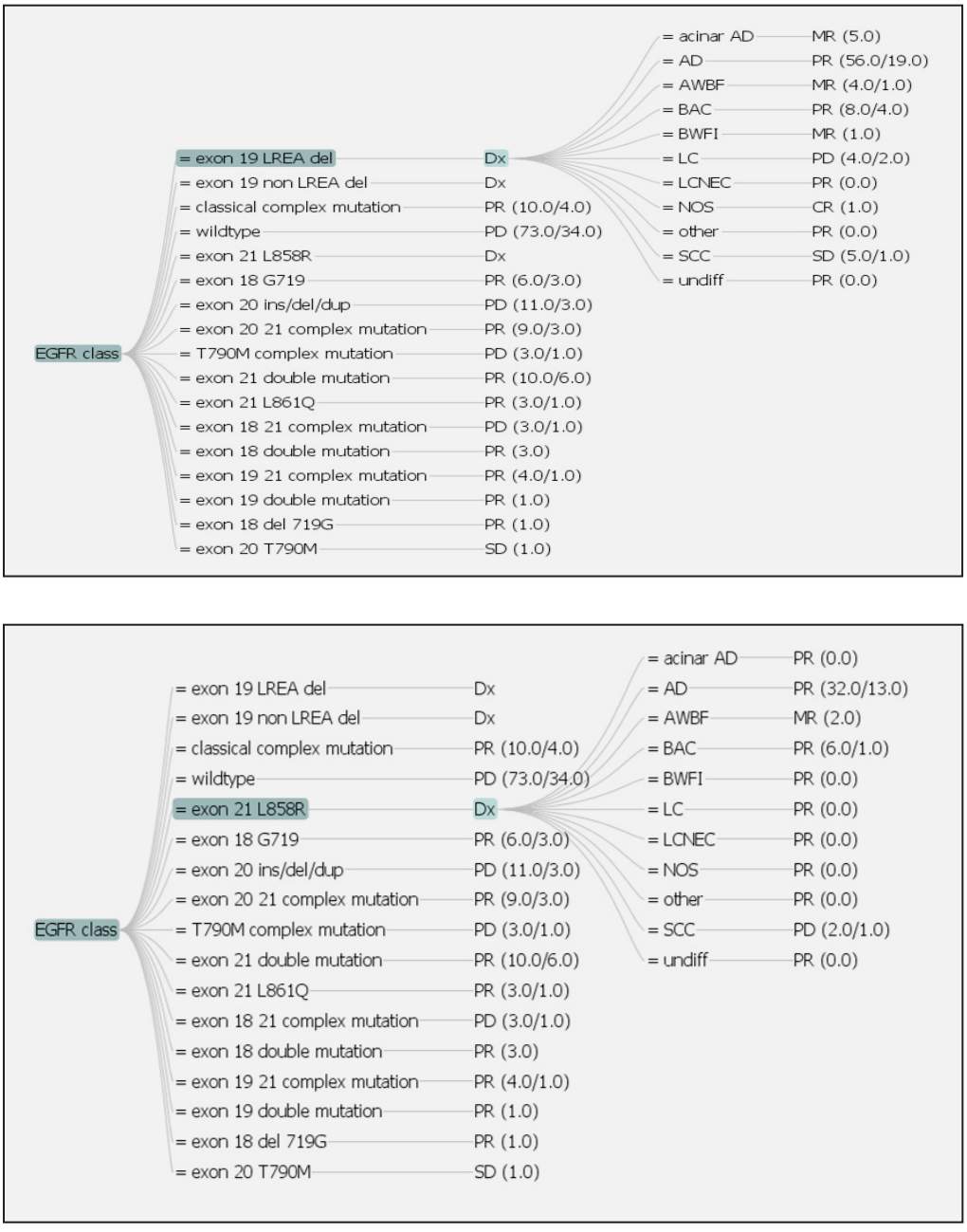

Figure 5.6 Highlights of decision tree from Experiment 1

5.4.2 Experiment 2 Classification for Binary Model

In Experiment 2, the multiclass problem was changed to a binary one, where the response to treatment was recorded as "responder" or "nonrepsonder". The aim was to determine if the predictive power of the classifiers could be increased by reducing the unbalanced classes in Experiment 1.

Results of Correlation Feature-based Attribute Evaluator

Clinical dataset: 17 subsets were evaluated and the merit of the best subset was 0.082. Smoking and Diagnosis were chosen for their predictive ability.

EGFR mutation dataset: 8 subsets were evaluated and the best merit was 0.158. Only EGFR mutation was selected.

Integrated dataset: Using the best first search method on the full training set, 31 subsets were evaluated and the merit of the best subset was 0.162 which included the attributes of Diagnosis, EGFR mutations, and EGFR class.

Performance Evaluation

The evaluation of classification is based on the test set performance (n=64) and results are shown in Table 5.3.

Dataset	Classifier*	Accuracy	Error	Precision	Recall	ROC area
Clinical	SMO	**57.81**	42.19	0.572	0.578	**0.56**
	J48	51.56	48.44	0.443	0.516	0.46
	Random Forest	56.25	43.75	0.557	0.563	0.53
	CART	54.69	45.31	0.53	0.547	0.51
EGFR	SMO	73.44	26.56	0.741	0.734	0.72
	J48	**76.56**	23.44	0.783	0.766	**0.75**
	Random Forest	73.44	26.56	0.741	0.734	0.69
	CART	73.44	26.56	0.741	0.734	0.70
Integrated	SMO	76.56	23.44	0.769	0.766	**0.76**
	J48	76.56	23.44	0.783	0.766	0.75
	Random Forest	75.00	25.00	0.755	0.75	0.69
	CART	76.56	23.44	0.769	0.766	0.74

*results for SMO and J48 are without attribute selection and results for Random forest and CART are with attribute selection

Table 5.3 Performance evaluation of classifiers from Experiment 2

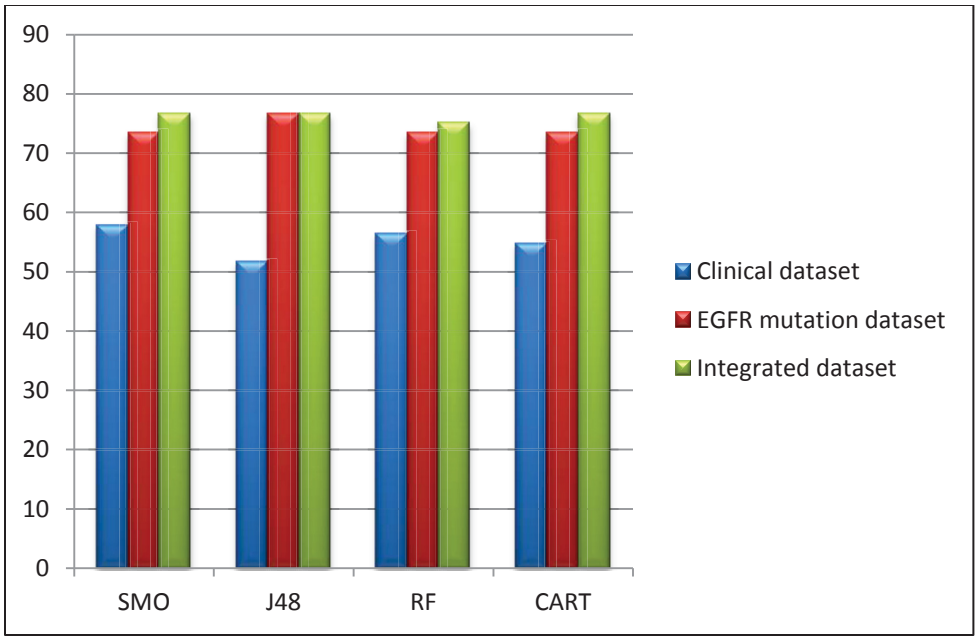

Figure 5.7 Experiment 2 Accuracy comparison of four classifiers

(SMO:support vector machine, J48:implementation of C4.5, RF: random forest, and CART: classification and regression tree)

An accuracy comparison of the four classifiers used in Experiment 2 is shown in Figure 5.7. In the binary class model, the accuracy of all algorithms on the Clinical dataset was at least 17.19% lower than that of algorithms on the EGFR mutation and Integrative dataset. In the Integrative dataset, SMO, J48, and CART achieved an accuracy of 76.56%, however, SMO had the highest AUC at 0.76, while that for J48, CART and Random forest was 0.75, 0.74, and 0.69 respectively.

All performance metrics improved when the response was binary instead of multiclass. The pattern of change between the two models was similar- models created on the Integrated dataset performed better than those for only clinical or EGFR mutation features.

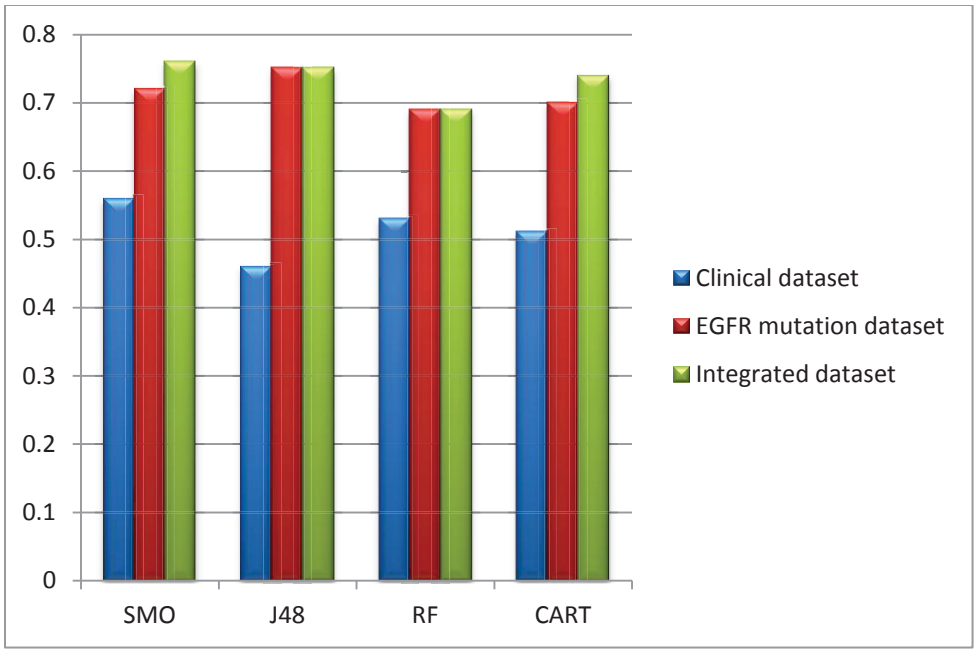

Figure 5.8 Experiment 2 ROC comparison of four classifiers

(SMO:support vector machine, J48:implementation of C4.5, RF: random forest, and CART: classification and regression tree)

In order to test whether the integration of clinical data and EGFR mutations performed significantly different than clinical data and EGFR mutations alone, the AUC was compared between the three datasets (Figure 5.8). Figure 5.9 illustrates the three ROCs generated from SMO. Using DeLong's method for comparing the area under multiple correlated curves [263], the three AUC values for support vector machine were evaluated. Table 5.4 describes the summary for AUC values, standard errors and 95% confidence intervals for all three datasets produced using support vector machine. Table 5.5 provides the pair-wise comparison of the AUC values between datasets. The model created using the integration of patient clinical data with EGFR mutations performed significantly better than the model created from clinical data alone (p=0.0363**).** However, the AUC of the integrated model was not significantly different from the EGFR model (p=0.1498).

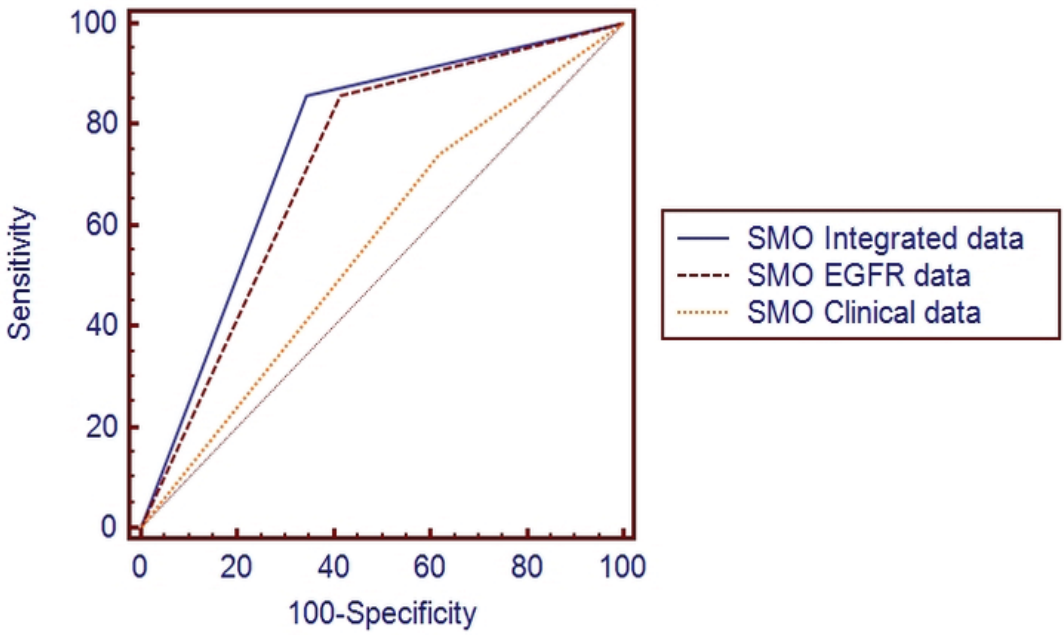

Figure 5.9 ROCs generated from support vector machine

Dataset	AUC	Standard error	95% CI
Clinical	0.561	0.0592	0.431 to 0.685
EGFR	0.722	0.0554	0.596 to 0.826
Integrated	0.756	0.0540	0.633 to 0.855

Table 5.4 AUC for all datasets using support vector machine

Dataset pair	z statistic	*p*-value
Integrated and EGFR	1.440	0.1498
Integrated and Clinical	2.093	**0.0363**
EGFR and Clinical	1.694	0.0902

Table 5.5 Pair-wise comparison of AUC for support vector machine

Analysis of Decision tree

Because J48 has excellent readability and offers a powerful way to express the underlying structure in datasets, it was unpruned and the output of a section of the tree is shown in Figure 5.10. Similar to Experiment 1, the attribute for EGFR mutation formed the root node in Experiment 2. For illustrative purposes we show EGFR class as the root node. A comparison of a similar leaf from Experiment 1 is shown in Figure 5.11. In the multiclass model, only the attribute of Diagnosis was used to determine the class. However, in the binary model, age, gender and smoking were used to determine the class in adenocarcinoma (AD).

Figure 5.10 Experiment 2 Exon 19 LREA deletion decision branch

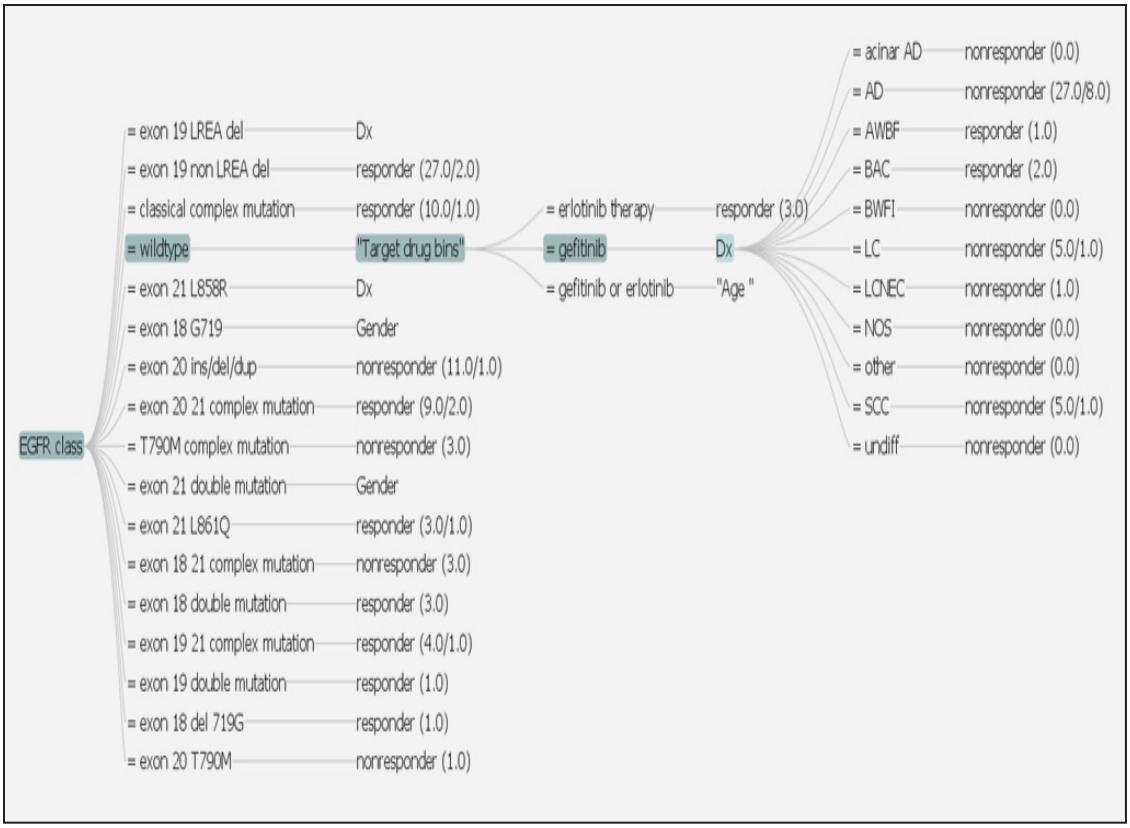

Figure 5.11 Experiment 2 wildtype decision branch

There was a high misclassification rate for EGFR wildtype cases in Experiment 1's decision tree. Although both erlotinib and gefitinib are tyrosine kinase inhibitors, it remains uncertain whether they display different clinical activities in advanced NSCLC. This is in part due to the lack of direct comparison studies of both agents under similar conditions. Several clinical trials have shown that the response rate of gefitinib in EGFR wildtype patients is between 0-6.6% [264-267], whereas erlotinib has achieved a 7% response rate [268]. Recently, it has been illustrated that erlotinib may be more effective than gefitinib in wildtype patients owing to its protein interaction profile [269]. A study conducted at Memorial Sloan Kettering in 2008 demonstrated that some EGFR wildtype patients with bronchoalveolar carcinoma or adenocarcinoma with bronchoalveolar features show minor or partial response to TKIs [270]. A small number of studies have reported EGFR wildtype response to gefitinib and this may explain the histological differences in response seen in gefitinib branch above [271].

5.5 Chapter Summary

Two experiments were set up using different levels of the class variable. Experiment 1 used five levels of response; four of these directly correspond with RECIST (complete response, partial response, stable disease and progressive disease) and the fifth level was created to denote a miscellaneous response as described in Chapter 3. The distribution of the class variables varied greatly, where complete response was found in only 4% of the dataset while the frequency of partial response was 43%. Unbalanced classes along with the small size of the dataset resulted in low accuracy of multiclass model. In Experiment 2, the multiclass problem was transformed to a binary one, by changing the outcome to responder and non-responder. Although this provided lesser detail to the degree of response, the accuracy of binary classification improved dramatically.

Classification was performed on three datasets. In the Clinical dataset, only demographics, pathological diagnosis and target drug were included, whereas the EGFR mutation dataset contained the individual mutations, EGFR class, and target drug. The collection of attributes from these two sets produced the Integrated dataset.

In Experiment 1, SMO achieved an accuracy of 37.5% on the clinical dataset and tree-based classifiers classified 40.6% of instances correctly. In the Integrated dataset, the accuracy of SMO increased to 56.3% and the average accuracy for tree-based classifiers was 55.7%. In Experiment 2, the accuracy of SMO and J48 was 57.8% and 51.7% respectively on the Clinical dataset. By integrating all attributes, SMO, J48 and CART achieved an accuracy of 76.6%. In addition, a comparison of AUC values generated for all three datasets using support vector machine was performed. The AUC for the integrated dataset was significantly different from the clinical dataset. The pair-wise comparison for the integrated dataset and the EGFR dataset did not reach statistical significance. The results of both experiments demonstrate that a set of only clinical attributes has a weak predictive power to classify the response achieved from EGFR-TKI therapy. Conversely, the integration of clinical and EGFR mutations was strongly predictive of the class variable.

Both quantitative and qualitative criteria were used to evaluate the performance of the models. The quantitative measures of classification accuracy, error rate, prediction, recall and ROC area were compared for all classifiers. In addition, the output interpretability of the decision tree was used as qualitative measure of J48's performance.

The decision tree of the multiclass and binary model had 17 decision nodes. Given that exons 19 and 21 of EGFR are most commonly mutated in NSCLC, these branches were emphasized and explained in the results of decision modeling. As shown in Figures 5.6, 5.10 and 5.11, many of the rarer EGFR mutations corresponded directly to a specific tumor response. Although, the number of recorded observations for rare mutations and mutation combinations was limited in the current research dataset, it was noted that the misclassification for rare mutations was unexpectedly low. Since majority of patient data in this project was derived from clinical trials that are highly selective, caution must be exercised when attempting to generalize the tumor response of rare and complex EGFR mutations in to clinical practice.

CHAPTER 6: DISCUSSION

Molecular targeted therapies for NSCLC promise superior clinical outcomes for patients with advanced stage disease. The post-genomic era has witnessed an explosion of biomedical data coupled with plateau of personalized patient selection. With increasing costs of modern drugs, it is critical to identify subsets of patient populations that are likely to benefit from treatments and ensure the best health outcomes. A significant challenge lies in the integration of multifaceted healthcare data to provide clinical predictions for patient outcomes. This thesis presents the integration of clinical and molecular data to guide the personalized treatment selection for targeted therapy in advanced NSCLC.

6.1 Pattern Discovery in NSCLC

In the clinical environment, frequent pattern mining of healthcare data can uncover biological relationships and help us understand genotype-phenotype associations (diagnosis, prognosis, and therapeutic response). In this thesis, two methods of pattern discovery were used to discover meaningful relationships in the research dataset. With the Classical approach, also known as the unsupervised approach, we did not predetermine a target variable. The results of Apriori consistently revealed frequent itemsets that have been previously described from large randomized controlled trials. Unselected trials for advanced NSCLC indicated that clinical characteristics, such as, female gender, a history of never smoking and the diagnosis of adenocarcinoma were all highly correlated to a favorable response to erlotinib or gefitinib. Constrained pattern mining is a form of subset association mining. Intuitively, the advantage of constrained mining is the reduction in the number of rules generated. The subset of rules is limited to specified attributes in the antecedent or consequent. For this thesis, the consequent was constrained to tumor response in order to generate rules demonstrating the relationship between itemsets and responsiveness to EGFR-TKI. Results from classical and constrained approaches confirmed the patient characteristics and tumor response patterns reported in the literature, in turn, validating the dataset.

Pattern mining was a descriptive and exploratory approach; the generated rules expressed the inherent regularities in the dataset and the frequency of attributes occurring together. Despite the fact that pattern mining was not performed sequentially with classification, many of the association rules were found to correspond with classification rules. For example, a constrained rule using Predictive Apriori algorithm demonstrated that exon 20 Q787Q + exon 21 L858R EGFR mutation pattern was likely to be associated with partial response with 92% accuracy. In the multiclass classification model, visualization of the decision tree established that this same mutation pattern resulted in partial response. This observation forms the basis of associative classification, a popular task which combines the power of association rules with classification to enhance prediction [272]. It also indicates the existence of a strong relationship between complex EGFR mutations and tumor response.

6.2 Decision Support for NSCLC

In order to isolate the relationship between tumor response and distinct variables, the predictive ability of subsets of attributes was tested. The clinical dataset consisted of conventional clinical variables such as age, gender, smoking status and histology. In the EGFR mutation dataset, the biomarker status and its class (as defined in Chapter 3) were included, while the integrated dataset combined both clinical and biomarker data. A common observation for both the multiclass and binary models was the significant improvement in performance of all classifiers between the clinical dataset and EGFR dataset. The accuracy and ROC using only clinical characteristics of age, gender, smoking status and histology was weak when compared to the EGFR mutation dataset In the multiclass model the highest accuracy was 40.63% and the weighted ROC was 0.51 using the clinical data characteristics, whereas in the EGFR mutation dataset the highest accuracy was 56.25% and weighted ROC was 0.66. When the clinical and molecular data was integrated into a combined dataset, the classifier ROC further improved (0.69 for support vector machine). A similar pattern was observed in the binary model, where support vector machine achieved the best performance using integrated data

characteristics; 76.56% accuracy and 0.76 AUC. Pair-wise comparison of the AUC for three datasets generated from support vector machine was performed. This assessment revealed that the AUC of the integrated dataset was significantly different from the clinical dataset. This can in part be explained by the nature of the domain- erlotinib and gefitinib target the tyrosine kinase domain of EGFR. If mutational testing data is included in the feature vector, the prediction of tumor response is more accurate. These results are in agreement with the existing understanding that EGFR mutations describe a subset of NSCLC which can guide targeted therapy [95]. Previous research has also established that the combination of clinical and genetic or genomic variables has enhanced predictive ability for risk stratification in lung cancer. In order to identify patients at risk for recurrence after surgical resection in lung cancer, [102] constructed three predictive models. The clinical model consisted of seven clinical variables; stage, cell type, differentiation, smoking history, tumor size, gender, age; the genomic model included QPCR-assayable genes, and a combination of both produced the clinicogenomic model. The authors demonstrated that the clinicogenomic model had superior performance (AUC >0.75) within the validation cohorts compared with the clinical or genomic models alone in predicting recurrence risk for lung cancer patients. In [103] the authors observed that the addition of clinical covariates improved the hazard ratio of gene expression to produce a robust predictor of overall survival in lung cancer patients. Various studies have assessed the AUC of predictive models using a variety of outcomes in lung cancer. In this thesis, validation of trained classifiers was performed on independent test data. The AUC of 0.76 achieved by support vector machine in the integrated model is comparable to the limits reported by [36] and [103].

Given a limited size of training data, it is essential to carefully choose the class distributions as this directly affects classification. Some authors recommend that classifier learning should be based on the natural underlying distribution of classes while others advocate an increase in the minority class examples [273]. An analysis of the confusion matrices for the multiclass model demonstrated that the misclassification for miscellaneous response (MR), complete response (CR) and stable disease (SD) was much higher than partial response (PR) or progressive disease (PD). Because MR, CR and SD were under-represented, the induced classifiers had poor classification for the minority

examples. In [273], the authors provide guidelines for imbalanced datasets and recommend that when accuracy is chosen as the performance measure, natural class distribution is preferred, whereas if AUC is the performance metric, balance class distribution is desirable. In this thesis, to overcome the imbalance class distribution problem the multiclass model was transformed to a binary model using the criteria used to calculate response rate. PR, CR and MR were grouped together as responders and SD and PD were grouped into non-responders. In the binary model, the class distributions were 60% for responders and 40% for non-responders. Classification performance improved for all learners in the binary class model. This can in part be explained by the measurement of tumor response by RECIST and its four response classes. The widely accepted guideline proposed by RECIST defines the framework for converting CT measurements into tumor response categories. Although RECIST is standardized for consistent use across studies and clinical trials, it is highly dependent on individual subjective judgement of tumor margins and increases inter and intra observer variability [274]. The observer variability is multiplied in our dataset, as it draws on tumor assessment for patient samples from several locations. The research community acknowledges that there may be differences among readers in the assignment of tumors into RECIST-defined response categories [274]. Accordingly, by categorizing lung tumor response into two classes; responders and non-responders [229], [275], a marked improvement was observed in the discriminatory ability of all classifiers.

EGFR mutations are most commonly associated with adenocarcinoma histology and cancers with an adenocarcinoma component [13]. Molecular aberrations of EGFR occur infrequently in squamous cell carcinomas and large cell carcinomas that lack an adenocarcinoma component [276], [252] and these subtypes are not currently recommended for EGFR testing. In the multiclass model, the analysis of J48 decision tree output revealed that in the presence of sensitizing mutations in exons 19 and 21, in all histologies except squamous cell carcinoma resulted in response to EGFR-TKI. Although, patients with non-adenocarcinoma histology testing positive for EGFR mutations are rarely evaluated, a pooled analysis of published reports found that gefitinib was less effective in non-adenocarcinoma cancers harboring EGFR mutations [277]. The response rate and PFS was significantly lower in non-adenocarcinoma histologies

compared to adenocarcinomas. One possible explanation of this phenomenon is an alteration in the phosphatidylinositol 3-kinases)/Akt pathway downstream of EGFR in non-adenocarcinomas [278]. The gene encoding the p110α subunit of phosphoinositide 3-kinase (PI3K) α or PIK3CA is not mutually exclusive to EGFR or KRAS mutations and may cause resistance to treatment with erlotinib or gefitinib in squamous cell carcinomas [277].

In the binary class model, a similar response pattern was observed for all histologies. However, in this model, the decision tree provided two additional decision nodes for adenocarcinoma. Patients younger than 49.2 years were likely to be responders, but in older patients, gender and history of smoking further differentiated tumor response. Analogous to our current understanding of clinical characteristics, female patients respond better than males, and never or former smokers respond better than current smokers to targeted therapy.

The binary class model also revealed that some EGFR wildtype tumors may also respond to erlotinib or gefitinib. This was most pronounced in bronchoalveolar carcinoma or adenocarcinoma with bronchoalveolar features. The Sequential Tarceva in Unresectable NSCLC (SATURN) trial investigated the role of erlotinib as maintenance therapy in patients with non-progressing disease after first-line platinum-doublet chemotherapy [279]. Results demonstrated a PFS benefit in patients with both EGFR mutation positive and wildtype tumors. Therefore it may be possible that EGFR-TKIs benefits patients regardless of mutation status or clinical profile in the second-line, third-line or maintenance settings.

The work presented in this thesis focused on clinical decision support for predicting tumor response to the two EGFR-TKIs erlotinib and gefitinib. Projecting forward to the future implementation of this work online, there would likely arise a number of interesting issues that would prove challenging. To begin with, ongoing pharmacogenomic cancer research will inevitably produce new and improved targeted therapies for NSCLC. Before reaching the marketplace, prospective clinical trials will collect longitudinal patient data for a number of years before efficacy of the new treatments can be assessed. Another important issue for consideration would be whether

these new drugs have any interaction with any other cancer treatments for NSCLC. Any such interactions that are reported and documented in the literature would have to be incorporated into the clinical decision support model. A third challenging issue would be to account for the acquisition of drug resistance in the decision support model.

6.3 Limitations

Limitations of this work include the potential omission of rare patterns and low discriminative capacity of decision model. As described previously, the current research dataset was created from diverse sources, each with a slightly different experimental approach, DNA sequencing methodology and assessment of tumor response. The limited sample size (n=355) included both common and rare mutations. Pattern mining techniques, applied to validate the data, discovered many of the well-established clinical and molecular associations. However, traditional pattern mining algorithms are designed to discover interesting patterns in potentially large datasets. Rare features, including several insertions, duplications, deletions and point mutations, had low frequency counts and would not have occurred in frequent itemsets. The highest predictive accuracy of the data-driven decision support model reached 76.53%, implying that there is still room for improvement. Additional clinical features such as performance status, ethnicity, line of treatment, and genetic abnormalities may help explain tumor response to erlotinib or gefitinib [36].

6.4 Future work

The current research demonstrates data-driven decision modeling for predicting tumor response to targeted therapies in advanced stage NSCLC. This proof of concept can be extended and further validated using alternative data sources. Longitudinal electronic health records contain comprehensive patient data, through which learning models can decipher the complex interaction between clinical and molecular characteristics that predict tumor response. Alternatively, data warehouses which are not tied to specific institutes or organizations offer increased scope and utility for mining healthcare data. Predictive ability and reliability of decision models relies heavily on the quality and

volume of training data; this underlies the importance of data aggregation initiatives from both public and private provider organizations.

6.5 Conclusion

Results from the decision support modeling experiments follow the recently released guideline [62] on molecular testing in lung cancer. The College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology suggest that clinical characteristics of age, gender, smoking, are insufficient to guide testing and treatment in patients with advanced NSCLC. The strong association of clinical characteristics and EGFR mutation frequency has been established in population studies, however, these features are inadequate to guide personalized treatment in NSCLC. Treatment decisions with erlotinib or gefitinib and the assessment of their efficacy in individual patients is best established by molecular testing of EGFR mutations [62].

Recommended EGFR testing in NSCLC is only the beginning of individualized therapy in NSCLC. As genome-wide association studies continue to identify additional molecular targets, scientific advancements will outpace the development of consensus guideline and expert opinions. Clinicians will be constantly challenged by the rapid changes driven by experimental research and clinical trials. Even so, data alone is insufficient to drive healthcare research. New and improved computational methods and techniques are the true drivers of data-intensive research and medical practice [208]. The value of the data-driven decision support model is two-fold: the evidence is based on real-world patient cases and models can "learn" personalized data-driven principles for individual cases in the absence of clear guidelines [280]. As an example, the 2013 National Cancer Comprehensive Network guidelines for NSCLC recommend EGFR mutation testing for adenocarcinoma, large cell carcinoma, and NSCLC NOS. For EGFR- mutation positive patients, erlotinib is the drug of choice. Recommendations from guidelines and consensus are based on population studies and say nothing about individual variations that may be seen in daily practice [62]. As an increasing amount of new mutations are reported, medical practitioners are faced with the challenge of optimizing treatment selection for

individual patients. Computer-assisted data extraction and evidence synthesis offered by decision support models can help clinicians improve targeted treatment options for patients based on their individual clinical, histological and molecular features [281].

In summary the achievements from this research are:

1. Construction of research dataset in the absence of structured data

An important contribution of this research is the construction of a dataset from multiple freely available sources, including original research articles and case reports. Data extracted from such sources can be computationally analyzed to provide intelligent and accurate decision support to clinicians.

2. Development of a data-driven clinical decision support model for NSCLC

The first research objective was to categorize patients who are responsive to EGFR-TKI chemotherapy in advanced NSCLC based on personalized clinicopathological data and EGFR mutation status. Multiple models were constructed on the dataset to determine optimal performance using feature selection and parameter tuning. Among several models tested, support vector machine and decision tree-based learners had the best performance. Furthermore, the decision tree learners modeled an easily interpretable relationship of input variables, using EGFR mutation status as the root node, which is very much in accordance with both current research and clinical practice guidelines for advanced NSCLC.

3. Comparison of attribute sets to tumor response

In addition to creating decision support models, the classification performance of three patient feature subsets was compared. Results suggest that the predictive performance of the model is significantly better with the integration of both clinical and genetic data, than when clinical features alone are used. This highlights the importance of the personalization of patient health in advanced NSCLC.

The findings from this research indicate that data-driven decision support is a promising avenue of research in fostering personalized medical decision-making for patients with advanced NSCLC. Potential next steps include learning models on large volumes of patient data to improve prediction and testing these models in pilot trials to validate results and performance.

REFERENCES

[1] C. C. Bennett, T. W. Doub, and R. Selove, "EHRs connect research and practice: Where predictive modeling, artificial intelligence, and clinical decision support intersect," *Health Policy and Technology, v*ol. 1, no. 2, pp. 105-114, Jun 2012.

[2] L. Chouchane, R. Mamtani, A. Dallol, and J. I. Sheikh, "Personalized medicine: a patient-centered paradigm," *J Transl Med,* vol. 9, p. 206, 2011.

[3] L. Mancinelli, M. Cronin, and W. Sadee, "Pharmacogenomics: the promise of personalized medicine," *AAPS PharmSci,* vol. 2, no. 1, pp. E4, 2000.

[4] R. K. Ito and L. M. Demers, "Pharmacogenomics and pharmacogenetics: future role of molecular diagnostics in the clinical diagnostic laboratory," *Clin Chem,* vol. 50, no. 9, pp. 1526-7, Sep, 2004.

[5] E. S. Vesell, "Advances in pharmacogenetics and pharmacogenomics," *J Clin Pharmacol,* vol. 40, no. 9, pp. 930-8, Sep, 2000.

[6] M. Diamandis, N. M. A. White, and G. M. Yousef, "Personalized medicine: marking a new epoch in cancer patient management," *Molecular Cancer Research,* vol. 8, no. 9, pp. 1175-1187, 2010.

[7] M. A. Socinski, "The Emerging Role of Biomarkers in Advanced Non–Small-Cell Lung Cancer," *Clinical lung cancer,* vol. 11, no. 3, pp. 149-159, 2010.

[8] S. Matsui, "Genomic Biomarkers for Personalized Medicine: Development and Validation in Clinical Studies," *Computational and mathematical methods in medicine,* vol. 2013, 2013.

[9] N. B. La Thangue and D. J. Kerr, "Predictive biomarkers: a paradigm shift towards personalized cancer medicine," *Nature Reviews Clinical Oncology,* vol. 8, no. 10, pp. 587-596, 2011.

[10] S. Navada, P. Lai, A.G. Schwartz, G.P. Kalemkerian. “Temporal trends in small cell lung cancer: analysis of the national Surveillance Epidemiology and End-Results (SEER) database,” *J Clin Oncol,* vol. 24, no. 18 suppl, pp. 7082, 2006.

[11] Z. Zhang, A. L. Stiegler, T. J. Boggon, S. Kobayashi, and B. Halmos, "EGFR-mutated lung cancer: a paradigm of molecular oncology," *Oncotarget,* vol. 1, no. 7, pp. 497-514, Nov, 2010.

[12] R.S. Herbst, M. Fukuoka, and J. Baselga,"Gefitinib—a novel targeted approach to treating cancer," *Nature Reviews Cancer*, vol. 4, no. 12, pp. 956-965, Dec, 2004.

[13] J. G. Paez, P. A. Janne, J. C. Lee, S. Tracy, H. Greulich, S. Gabriel*, et al.*, "EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy," *Science,* vol. 304, no. 5676, pp. 1497-500, Jun 2004.

[14] T. J. Lynch, D. W. Bell, R. Sordella, S. Gurubhagavatula, R. A. Okimoto, B. W. Brannigan, P. L. Harris, S. M. Haserlat, J. G. Supko, F. G. Haluska, D. N. Louis, D. C. Christiani, J. Settleman, and D. A. Haber, "Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib," *N Engl J Med,* vol. 350, no. 21, pp. 2129-39, May 2004.

[15] National Cancer Institute, "FDA approval for erlotinib hydrochloride," July 7, 2013. [Online]. Available: http://www.cancer.gov/cancertopics/druginfo/fda-erlotinib-hydrochloride [Accessed: March 1 2013].

[16] R. B. Natale, S. Thongprasert, F. A. Greco, M. Thomas, C. M. Tsai, P. Sunpaweravong, D. Ferry, C. Mulatero, R. Whorf, J. Thompson, F. Barlesi, P. Langmuir, S. Gogov, J. A. Rowbottom, and G. D. Goss, "Phase III trial of vandetanib compared with erlotinib in patients with previously treated advanced non-small-cell lung cancer," *J Clin Oncol,* vol. 29, no. 8, pp. 1059-66, Mar 10, 2011.

[17] P. Therasse, S. G. Arbuck, E. A. Eisenhauer, J. Wanders, R. S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A. T. van Oosterom, M. C. Christian, and S. G. Gwyther, "New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada," *J Natl Cancer Inst,* vol. 92, no. 3, pp. 205-16, Feb 2, 2000.

[18] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij, "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)," *Eur J Cancer,* vol. 45, no. 2, pp. 228-47, Jan, 2009.

[19] M. Markman, H. L. Kaufman, K. Antman, and S. Wadler, *Molecular targeting in oncology*: Springer, 2008.

[20] S. D. Ramsey, L. Clarke, T. V. Kamath, and D. Lubeck, "Evaluation of erlotinib in advanced non-small-cell lung cancer: impact on the budget of a US health insurance plan," *Journal of Managed Care Pharmacy,* vol. 12, no. 6, 2006.

[21] P. A. Bradbury, D. Tu, L. Seymour, P. K. Isogai, L. Zhu, R. Ng*, et al.*, "Economic analysis: Randomized placebo-controlled clinical trial of erlotinib in advanced non–small cell lung cancer," *Journal of the National Cancer Institute,* vol. 102, no. 5, pp. 298-306, 2010.

[22] K. Kawamoto, D. F. Lobach, H. F. Willard, and G. S. Ginsburg, “A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine,” *Bmc Medical Informatics and Decision Making,* vol. 9, Mar , 2009.

[23] G. S. Ginsburg, and H. F. Willard, “Genomic and personalized medicine: foundations and applications,” *Transl Res,* vol. 154, no. 6, pp. 277-87, Dec, 2009.

[24] J. Gardner, and L. Xiong, “An integrated framework for de-identifying unstructured medical data,” *Data & Knowledge Engineering,* vol. 68, no. 12, pp. 1441-1451, Dec, 2009.

[25] G. S. Ginsburg, and J. J. McCarthy, “Personalized medicine: revolutionizing drug discovery and patient care,” *Trends Biotechnol,* vol. 19, no. 12, pp. 491-6, Dec, 2001.

[26] E. A. Balas and S. A. Boren, "Managing clinical knowledge for health care improvement," *Yearbook of medical informatics,* vol. 2000, pp. 65-70, 2000.

[27] F. Meric-Bernstam, C. Farhangfar, J. Mendelsohn, and G. B. Mills, "Building a personalized medicine infrastructure at a major cancer center," *Journal of Clinical Oncology,* vol. 31, no. 15, pp. 1849-1857, 2013.

[28] G. S. Ginsburg, R. P. Konstance, J. S. Allsbrook, and K. A. Schulman, “Implications of pharmacogenomics for drug development and clinical practice,” *Arch Intern Med,* vol. 165, no. 20, pp. 2331-6, Nov 14, 2005.

[29] Department of Health and Human Services, *Realizing the potential of pharmacogenomics: opportunities and challenges. Report of the Secretary's Advisory Committee on Genetics, Health, and Society*. Bethesda: United States; 2008. [Online]. Available: http://oba.od.nih.gov/oba/sacghs/reports/sacghs_pgx_report.pdf [Accessed May 2, 2013].

[30] National Comprehensive Cancer Network (NCCN), *“*NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines). Nonsmall cell lung cancer version 2.2013,” [Online]. Available at: http://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf [Accessed June 3, 2013].

[31]M. Fukuoka, S. Yano, G. Giaccone, T. Tamura, K. Nakagawa, J. Y. Douillard, Y. Nishiwaki, J. Vansteenkiste, S. Kudoh, D. Rischin, R. Eek, T. Horai, K. Noda, I. Takata, E. Smit, S. Averbuch, A. Macleod, A. Feyereislova, R. P. Dong, and J. Baselga, “Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small-cell lung cancer (The IDEAL 1 Trial) [corrected],” *J Clin Oncol,* vol. 21, no. 12, pp. 2237-46, Jun 15, 2003.

[32] M. G. Kris, R. B. Natale, R. S. Herbst, T. J. Lynch Jr, D. Prager, C. P. Belani, et al., "Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non–small cell lung cancer," JAMA: the journal of the American Medical Association, vol. 290, no. 16, pp. 2149-2158, 2003.

[33]V. A. Miller, M. G. Kris, N. Shah, J. Patel, C. Azzoli, J. Gomez, et al., "Bronchioloalveolar pathologic subtype and smoking history predict sensitivity to gefitinib in advanced non–small-cell lung cancer," *Journal of Clinical Oncology,* vol. 22, no. 6, pp. 1103-1109, 2004.

[34] X. Zhang and A. Chang, "Molecular predictors of EGFR-TKI sensitivity in advanced non–small cell lung cancer," *International journal of medical sciences,* vol. 5, no. 4, pp. 209, 2008.

[35] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *New England Journal of Medicine,* vol. 363, no. 4, pp. 301-304, 2010.

[36] A. Lopez-Encuentra, F. Lopez-Rios, E. Conde, R. Garcia-Lujan, A. Suarez-Gauthier, N. Manes, G. Renedo, J. L. Duque-Medina, E. Garcia-Lagarto, R. Rami-Porta, G. Gonzalez-Pont, J. Astudillo-Pombo, J. L. Mate-Sanz, J. Freixinet, T. Romero-Saavedra, M. Sanchez-Cespedes, A. Gomez de la Camara, P. Bronchogenic Carcinoma Cooperative Group of the Spanish Society of, and S. Thoracic, "Composite anatomical-clinical-molecular prognostic model in non-small cell lung cancer," *Eur Respir J,* vol. 37, no. 1, pp. 136-42, Jan, 2011.

[37] A. Spira, J. E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y. M. Dumas, P. Calner, P. Sebastiani, S. Sridhar, J. Beamis, C. Lamb, T. Anderson, N. Gerry, J. Keane, M. E. Lenburg, and J. S. Brody, "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer," *Nat Med,* vol. 13, no. 3, pp. 361-6, Mar, 2007.

[38] J. Beane, P. Sebastiani, T. H. Whitfield, K. Steiling, Y. M. Dumas, M. E. Lenburg, and A. Spira, "A prediction model for lung cancer diagnosis that integrates genomic and clinical features," *Cancer Prev Res (Phila),* vol. 1, no. 1, pp. 56-64, Jun, 2008.

[39] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *Int J Cancer,* vol. 127, no. 12, pp. 2893-917, Dec 15, 2010.

[40] Canadian Cancer Society's Steering Committee on Cancer Statistics, *Canadian cancer Statistics, 2012*, Toronto, ON: Canadian Cancer Society, 2012. [Online]. Available: http://www.cancer.ca/~/media/cancer.ca/CW/cancer%20information/cancer%20101/Canadian%20cancer%20statistics/Canadian-Cancer-Statistics-2012---English.pdf. [Accessed: June 29, 2013].

[41] T. Sher, G. K. Dy, and A. A. Adjei, "Small cell lung cancer," *Mayo Clin Proc,* vol. 83, no. 3, pp. 355-67, Mar, 2008.

[42**]** H. Kantarjian, C. A. Koller, R. A. Wolff, H. Amin, E. K. Abdalla, J. L. Ater, T. P. Avery, R. Avritscher, A. Y. Bedikian, and N. A. Bhadkamkar, *The MD Anderson manual of medical oncology*: McGraw-Hill, 2011.

[43] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2009," *CA Cancer J Clin,* vol. 59, no. 4, pp. 225-49, Jul-Aug, 2009.

[44] American Cancer Society,"Non-small cell lung cancer survival rates by stage," 2013.[Online].Available: http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-survival-rates.[Accessed: Jan 10, 2013].

[45] A. G. Shaper, S. G. Wannamethee, and M. Walker, "Pipe and cigar smoking and major cardiovascular events, cancer incidence and all-cause mortality in middle-aged British men," *Int J Epidemiol,* vol. 32, no. 5, pp. 802-8, Oct, 2003.

[46] S. J. Henley, M. J. Thun, A. Chao, and E. E. Calle, "Association between exclusive pipe smoking and mortality from cancer and other diseases," *Journal of the National Cancer Institute,* vol. 96, no. 11, pp. 853-861, Jun 2, 2004.

[47] B. D. Smith, G. L. Smith, A. Hurria, G. N. Hortobagyi, and T. A. Buchholz, "Future of Cancer Incidence in the United States: Burdens Upon an Aging, Changing Nation," Journal of Clinical Oncology, vol. 27, no. 17, pp. 2758-2765, Jun 10, 2009.

[48] J. Subramanian, D. Morgensztern, B. Goodgame, M. Q. Baggstrom, F. Gao, J. Piccirillo, and R. Govindan, "Distinctive characteristics of non-small cell lung cancer (NSCLC) in the young: a surveillance, epidemiology, and end results (SEER) analysis," *J Thorac Oncol,* vol. 5, no. 1, pp. 23-8, Jan, 2010.

[49] D. Cerny, T. Cerny, S. Ess, G. D'Addario, and M. Fruh, "Lung cancer in the Canton of St. Gallen, Eastern Switzerland: sex-associated differences in smoking habits, disease presentation and survival," *Onkologie,* vol. 32, no. 10, pp. 569-73, Oct, 2009.

[50] H. A. Wakelee, E. T. Chang, S. L. Gomez, T. H. Keegan, D. Feskanich, C. A. Clarke, L. Holmberg, L. C. Yong, L. N. Kolonel, M. K. Gould, and D. W. West, "Lung cancer incidence in never smokers," *J Clin Oncol,* vol. 25, no. 5, pp. 472-8, Feb, 2007..

[51] J. Gasperino, "Gender is a risk factor for lung cancer," *Medical Hypotheses,* vol. 76, no. 3, pp. 328-331, Mar, 2011.

[52] G. D. Santos, F. A. Shepherd, and M. S. Tsao, "EGFR Mutations and Lung Cancer," *Annual Review of Pathology: Mechanisms of Disease, Vol 6,* vol. 6, pp. 49-69, 2011.

[53] A. F. Gazdar, "Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors," *Oncogene,* vol. 28, pp. S24-S31, Aug, 2009.

[54] J. L. Johnson, S. Pillai, and S. P. Chellappan, "Genetic and biochemical alterations in non-small cell lung cancer," *Biochem Res Int,* vol. 2012, pp. 940405, 2012.

[55] A. T. Shaw, B. Y. Yeap, M. Mino-Kenudson, S. R. Digumarthy, D. B. Costa, R. S. Heist, B. Solomon, H. Stubbs, S. Admane, U. McDermott, J. Settleman, S. Kobayashi, E. J. Mark, S. J. Rodig, L. R. Chirieac, E. L. Kwak, T. J. Lynch, and A. J. Iafrate, "Clinical features and outcome of patients with non-small-cell lung cancer who harbor EML4-ALK," *J Clin Oncol,* vol. 27, no. 26, pp. 4247-53, Sep, 2009.

[56] M. Soda, Y. L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, M. Bando, S. Ohno, Y. Ishikawa, H. Aburatani, T. Niki, Y. Sohara, Y. Sugiyama, and H. Mano, "Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer," *Nature,* vol. 448, no. 7153, pp. 561-6, Aug, 2007.

[57] J. P. Koivunen, C. Mermel, K. Zejnullahu, C. Murphy, E. Lifshits, A. J. Holmes, H. G. Choi, J. Kim, D. Chiang, R. Thomas, J. Lee, W. G. Richards, D. J. Sugarbaker, C. Ducko, N. Lindeman, J. P. Marcoux, J. A. Engelman, N. S. Gray, C. Lee, M. Meyerson, and P. A. Janne, "EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer," *Clin Cancer Res,* vol. 14, no. 13, pp. 4275-83, Jul, 2008.

[58] K. Inamura, K. Takeuchi, Y. Togashi, S. Hatano, H. Ninomiya, N. Motoi, M. Y. Mun, Y. Sakao, S. Okumura, K. Nakagawa, M. Soda, Y. L. Choi, H. Mano, and Y. Ishikawa, "EML4-ALK lung cancers are characterized by rare other mutations, a TTF-1 cell lineage, an acinar histology, and young onset," *Mod Pathol,* vol. 22, no. 4, pp. 508-15, Apr, 2009.

[59] Y. Pylayeva-Gupta, E. Grabocka, and D. Bar-Sagi, "RAS oncogenes: weaving a tumorigenic web," *Nat Rev Cancer,* vol. 11, no. 11, pp. 761-74, Nov, 2011.

[60] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun, "Cancer statistics, 2008," *CA Cancer J Clin,* vol. 58, no. 2, pp. 71-96, Mar-Apr, 2008.

[61] American Cancer Society, "Non-small cell lung cancer," Lung Cancer (Non-Small Cell). [Online]. Available: http://www.cancer.org/acs/groups/cid/documents/webcontent/003115-pdf.pdf. [Accessed: Jan 10, 2013].

[62] N. I. Lindeman, P. T. Cagle, M. B. Beasley, D. A. Chitale, S. Dacic, G. Giaccone, R. B. Jenkins, D. J. Kwiatkowski, J. S. Saldivar, J. Squire, E. Thunnissen, and M. Ladanyi, "Molecular Testing Guideline for Selection of Lung Cancer Patients for EGFR and ALK Tyrosine Kinase Inhibitors: Guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology," *J Thorac Oncol,* vol. 8, no. 7, pp. 823-859, Jul, 2013.

[63] L. West, S. J. Vidwans, N. P. Campbell, J. Shrager, G. R. Simon, R. Bueno, P. A. Dennis, G. A. Otterson, and R. Salgia, "A Novel Classification of Lung Cancer into Molecular Subtypes," *Plos One,* vol. 7, no. 2, Feb, 2012.

[64] R. Perez-Soler, A. Chachoua, L. A. Hammond, E. K. Rowinsky, M. Huberman, D. Karp, J. Rigas, G. M. Clark, P. Santabarbara, and P. Bonomi, "Determinants of tumor response and survival with erlotinib in patients with non--small-cell lung cancer," *J Clin Oncol,* vol. 22, no. 16, pp. 3238-47, Aug, 2004

[65] D. M. Jackman, B. Y. Yeap, N. I. Lindeman, P. Fidias, M. S. Rabin, J. Temel, A. T. Skarin, M. Meyerson, A. J. Holmes, A. M. Borras, B. Freidlin, P. A. Ostler, J. Lucca, T. J. Lynch, B. E. Johnson, and P. A. Janne, "Phase II clinical trial of chemotherapy-naive patients > or = 70 years of age treated with erlotinib for advanced non-small-cell lung cancer," *J Clin Oncol,* vol. 25, no. 7, pp. 760-6, Mar, 2007.

[66] W. Brugger, N. Triller, M. Blasinska-Morawiec, S. Curescu, R. Sakalauskas, G. Manikhas, J. Mazieres, R. Whittom, K. Rohr, F. Cappuzzo, and S. Investigators, "Biomarker analyses from the phase III placebo-controlled SATURN study of maintenance erlotinib following first-line chemotherapy for advanced NSCLC," *Journal of Clinical Oncology,* vol. 27, no. 15, May, 2009.

[67] N. Thatcher, A. Chang, P. Parikh, J. Rodrigues Pereira, T. Ciuleanu, J. von Pawel, S. Thongprasert, E. H. Tan, K. Pemberton, V. Archer, and K. Carroll, "Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer)," *Lancet,* vol. 366, no. 9496, pp. 1527-37, Oct 29-Nov, 2005.

[68] D. L. Aisner, and C. B. Marshall, "Molecular pathology of non-small cell lung cancer: a practical guide," *Am J Clin Pathol,* vol. 138, no. 3, pp. 332-46, Sep, 2012.

[69] J.A. Lynch and C.S. Lathan, "Disparities in access to the EGFR assay," *Cancer Epidemiol Biomarkers Prev,* vol. 20, 10 Suppl:B79, Sep, 2011.

[70] E.Li, "NCCN Trends™ Survey and Data: *EGFR* Mutation Testing Practices," National comprehensive Cancer Network. [Online]. Available: http://www.nccn.org/about/news/ebulletin/2010-11-01/patient_advocacy.asp. [Accessed: June 20,2013].

[71] H. J. Lee, J. Jo, D. S. Son, J. Lee, Y. S. Choi, K. Kim, Y. M. Shim, and J. Kim, "Predicting Recurrence Using the Clinical Factors of Patients with Non-small Cell Lung Cancer After Curative Resection," *Journal of Korean Medical Science,* vol. 24, no. 5, pp. 824-830, Oct, 2009.

[72] T. Hoang, R. Xu, J. H. Schiller, P. Bonomi, and D. H. Johnson, "Clinical model to predict survival in chemonaive patients with advanced non-small-cell lung cancer treated with third-generation chemotherapy regimens based on eastern cooperative oncology group data," *J Clin Oncol,* vol. 23, no. 1, pp. 175-83, Jan, 2005.

[73] B. Jeremic, B. Milicic, A. Dagovic, J. Aleksandrovic, and N. Nikolic, "Pretreatment clinical prognostic factors in patients with stage IV non-small cell lung cancer (NSCLC) treated with chemotherapy," *J Cancer Res Clin Oncol,* vol. 129, no. 2, pp. 114-22, Feb, 2003.

[74] S. J. Mandrekar, S. E. Schild, S. L. Hillman, K. L. Allen, R. S. Marks, J. A. Mailliard, J. E. Krook, A. W. Maksymiuk, K. Chansky, K. Kelly, A. A. Adjei, and J. R. Jett, "A prognostic model for advanced stage nonsmall cell lung cancer. Pooled analysis of North Central Cancer Treatment Group trials," *Cancer,* vol. 107, no. 4, pp. 781-92, Aug, 2006.

[75] A. S. Tsao, D. Liu, J. J. Lee, M. Spitz, and W. K. Hong, "Smoking affects treatment outcome in patients with advanced nonsmall cell lung cancer," *Cancer,* vol. 106, no. 11, pp. 2428-36, Jun, 2006.

[76] T. E. Strand, H. Rostad, R. A. M. Damhuis, and J. Norstein, "Risk factors for 30-day mortality after resection of lung cancer and prediction of their magnitude," *Thorax,* vol. 62, no. 11, pp. 991-997, Nov, 2007.

[77] Y. Sakuma, N. Okamoto, H. Saito, K. Yamada, T. Yokose, M. Kiyoshima, Y. Asato, R. Amemiya, H. Saitoh, S. Matsukuma, M. Yoshihara, Y. Nakamura, F. Oshita, H. Ito, H. Nakayama, Y. Kameda, E. Tsuchiya, and Y. Miyagi, "A logistic regression predictive model and the outcome of patients with resected lung adenocarcinoma of 2 cm or less in size," *Lung Cancer,* vol. 65, no. 1, pp. 85-90, Jul, 2009

[78] W. Kossler, A. Fiebeler, A. Willms, T. ElAidi, B. Klosterhalfen, and U. Klinge, "Formation of translational risk score based on correlation coefficients as an alternative to Cox regression models for predicting outcome in patients with NSCLC," *Theoretical Biology and Medical Modelling,* vol. 8, Jul, 2011.

[79] J. Putila, S. C. Remick, and N. L. Guo, "Combining Clinical, Pathological, and Demographic Factors Refines Prognosis of Lung Cancer: A Population-Based Study," *Plos One,* vol. 6, no. 2, Feb, 2011.

[80] T. Hanai, Y. Yatabe, Y. Nakayama, T. Takahashi, H. Honda, T. Mitsudomi, and T. Kobayashi, "Prognostic models in patients with non-small-cell lung cancer using artificial neural networks in comparison with logistic regression," *Cancer Sci,* vol. 94, no. 5, pp. 473-7, May, 2003.

[81] C. Dehing-Oberije, S. Yu, D. De Ruysscher, S. Meersschout, K. Van Beek, Y. Lievens, J. Van Meerbeeck, W. De Neve, B. Rao, H. van der Weide, and P. Lambin, "Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy," *Int J Radiat Oncol Biol Phys,* vol. 74, no. 2, pp. 355-62, Jun, 2009.

[82] A. C. Borczuk, L. Shah, G. D. Pearson, K. L. Walter, L. Wang, J. H. Austin, R. A. Friedman, and C. A. Powell, "Molecular signatures in biopsy specimens of lung cancer," *Am J Respir Crit Care Med,* vol. 170, no. 2, pp. 167-74, Jul, 2004.

[83] H. Y. Chen, S. L. Yu, C. H. Chen, G. C. Chang, C. Y. Chen, A. Yuan, C. L. Cheng, C. H. Wang, H. J. Terng, S. F. Kao, W. K. Chan, H. N. Li, C. C. Liu, S. Singh, W. J. Chen, J. J. Chen, and P. C. Yang, "A five-gene signature and clinical outcome in non-small-cell lung cancer," *N Engl J Med,* vol. 356, no. 1, pp. 11-20, Jan, 2007.

[84] Z. F. Sun, P. Yang, M. C. Aubry, F. Kosari, C. Endo, J. Molina, and G. Vasmatzis, "Can gene expression profiling predict survival for patients with squamous cell carcinoma of the lung?," *Molecular Cancer,* vol. 3, 2004.

[85] H. Endoh, S. Tomida, Y. Yatabe, H. Konishi, H. Osada, K. Tajima, T. Kuwano, T. Takahashi, and T. Mitsudomi, "Prognostic model of pulmonary adenocarcinoma by expression profiling of eight genes as determined by quantitative real-time reverse transcriptase polymerase chain reaction," *Journal of Clinical Oncology,* vol. 22, no. 5, pp. 811-819, Mar, 2004.

[86] H. Y. Jiang, Y. P. Deng, H. S. Chen, L. Tao, Q. Y. Sha, J. Chen, C. J. Tsai, and S. L. Zhang, "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes," *Bmc Bioinformatics,* vol. 5, Jun, 2004.

[87] Y. Lu, W. Lemon, P. Y. Liu, Y. J. Yi, C. Morrison, P. Yang, Z. F. Sun, J. Szoke, W. L. Gerald, M. Watson, R. Govindan, and M. You, "A gene expression signature predicts survival of patients with stage I non-small cell lung cancer," *Plos Medicine,* vol. 3, no. 12, pp. 2229-2243, Dec, 2006.

[88] J. Hou, J. Aerts, B. den Hamer, W. van IJcken, M. den Bakker, P. Riegman, C. van der Leest, P. van der Spek, J. A. Foekens, H. C. Hoogsteden, F. Grosveld, and S. Philipsen, "Gene Expression-Based Classification of Non-Small Cell Lung Carcinomas and Survival Prediction," *Plos One,* vol. 5, no. 4, Apr, 2010.

[89] R. Mitra, J. Lee, J. Jo, M. Milani, J. N. McClintick, H. J. Edenberg, K. A. Kesler, K. M. Rieger, S. Badve, O. W. Cummings, A. Mohiuddin, D. G. Thomas, X. Luo, B. E. Juliar, L. Li, C. Mesaros, I. A. Blair, A. Srirangam, R. A. Kratzke, C. J. McDonald, J. Kim, and D. A. Potter, "Prediction of postoperative recurrence-free survival in non-small cell lung cancer by using an internationally validated gene expression model," *Clin Cancer Res,* vol. 17, no. 9, pp. 2934-46, May 1, 2011.

[90] F. Baty, M. Facompre, S. Kaiser, M. Schumacher, M. Pless, L. Bubendorf, S. Savic, E. Marrer, W. Budach, M. Buess, J. Kehren, M. Tamm, and M. H. Brutsche, "Gene profiling of clinical routine biopsies and prediction of survival in non-small cell lung cancer," *Am J Respir Crit Care Med,* vol. 181, no. 2, pp. 181-8, Jan 15, 2010.

[91] T. Mitsudomi, T. Kosaka, H. Endoh, Y. Horio, T. Hida, S. Mori, S. Hatooka, M. Shinoda, T. Takahashi, and Y. Yatabe, "Mutations of the epidermal growth factor receptor gene predict prolonged survival after gefitinib treatment in patients with non-small-cell lung cancer with postoperative recurrence," *J Clin Oncol,* vol. 23, no. 11, pp. 2513-20, Apr, 2005.

[92] S. W. Han, T. Y. Kim, P. G. Hwang, S. Jeong, J. Kim, I. S. Choi, D. Y. Oh, J. H. Kim, D. W. Kim, D. H. Chung, S. A. Im, Y. T. Kim, J. S. Lee, D. S. Heo, Y. J. Bang, and N. K. Kim, "Predictive and prognostic impact of epidermal growth factor receptor mutation in non-small-cell lung cancer patients treated with gefitinib," *J Clin Oncol,* vol. 23, no. 11, pp. 2493-501, Apr, 2005.

[93] T. Kosaka, Y. Yatabe, H. Endoh, H. Kuwano, T. Takahashi, and T. Mitsudomi, "Mutations of the epidermal growth factor receptor gene in lung cancer: biological and clinical implications," *Cancer Res,* vol. 64, no. 24, pp. 8919-23, Dec, 2004.

[94] M. Taron, Y. Ichinose, R. Rosell, T. Mok, B. Massuti, L. Zamora, J. L. Mate, C. Manegold, M. Ono, C. Queralt, T. Jahan, J. J. Sanchez, M. Sanchez-Ronco, V. Hsue, D. Jablons, J. M. Sanchez, and T. Moran, "Activating mutations in the tyrosine kinase domain of the epidermal growth factor receptor are associated with improved survival in gefitinib-treated chemorefractory lung adenocarcinomas," *Clin Cancer Res,* vol. 11, no. 16, pp. 5878-85, Aug, 2005.

[95] L. V. Sequist, V. A. Joshi, P. A. Janne, A. Muzikansky, P. Fidias, M. Meyerson, D. A. Haber, R. Kucherlapati, B. E. Johnson, and T. J. Lynch, "Response to treatment and survival of patients with non-small cell lung cancer undergoing somatic EGFR mutation testing," *Oncologist,* vol. 12, no. 1, pp. 90-8, Jan, 2007.

[96] L. V. Sequist, R. G. Martins, D. Spigel, S. M. Grunberg, A. Spira, P. A. Janne, V. A. Joshi, D. McCollum, T. L. Evans, A. Muzikansky, G. L. Kuhlmann, M. Han, J. S. Goldberg, J. Settleman, A. J. Iafrate, J. A. Engelman, D. A. Haber, B. E. Johnson, and T. J. Lynch, "First-line gefitinib in patients with advanced non-small-cell lung cancer harboring somatic EGFR mutations," *J Clin Oncol,* vol. 26, no. 15, pp. 2442-9, May, 2008.

[97] J. Pittman, E. Huang, H. Dressman, C. F. Horng, S. H. Cheng, M. H. Tsou, C. M. Chen, A. Bild, E. S. Iversen, A. T. Huang, J. R. Nevins, and M. West, "Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes," *Proc Natl Acad Sci U S A,* vol. 101, no. 22, pp. 8431-6, Jun, 2004.

[98] L. X. Li, "Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information," *Bioinformatics,* vol. 22, no. 4, pp. 466-471, Feb, 2006.

[99] A. J. Stephenson, A. Smith, M. W. Kattan, J. Satagopan, V. E. Reuter, P. T. Scardino, and W. L. Gerald, "Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy," *Cancer,* vol. 104, no. 2, pp. 290-298, Jul, 2005.

[100] S. K. Lau, P. C. Boutros, M. Pintilie, F. H. Blackhall, C. Q. Zhu, D. Strumpf, M. R. Johnston, G. Darling, S. Keshavjee, T. K. Waddell, N. Liu, D. Lau, L. Z. Penn, F. A. Shepherd, I. Jurisica, S. D. Der, and M. S. Tsao, "Three-gene prognostic classifier for early-stage non-small-cell lung cancer," *Journal of Clinical Oncology,* vol. 25, no. 35, pp. 5562-5569, Dec, 2007.

[101] A. Potti, S. Mukherjee, R. Petersen, H. K. Dressman, A. Bild, J. Koontz, R. Kratzke, M. A. Watson, M. Kelley, G. S. Ginsburg, M. West, D. H. Harpole, and J. R. Nevins, "A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer," *New England Journal of Medicine,* vol. 355, no. 6, pp. 570-580, Aug, 2006.

[102] E. S. Lee, D. S. Son, S. H. Kim, J. Lee, J. Jo, J. Han, H. Kim, H. J. Lee, H. Y. Choi, Y. Jung, M. Park, Y. S. Lim, K. Kim, Y. M. Shim, B. C. Kim, K. Lee, N. Huh, C. Ko, K. Park, J. W. Lee, Y. S. Choi, and J. Kim, "Prediction of Recurrence-Free Survival in Postoperative Non-Small Cell Lung Cancer Patients by Using an Integrated Model of Clinical Information and Gene Expression," *Clinical Cancer Research,* vol. 14, no. 22, pp. 7397-7404, Nov, 2008.

[103] K. Shedden, J. M. G. Taylor, S. A. Enkemann, M. S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, A. C. Chang, C. Q. Zhu, D. Strumpf, S. Hanash, F. A. Shepherd, K. Ding, L. Seymour, K. Naoki, N. Pennell, B. Weir, R. Verhaak, C. Ladd-Acosta, T. Golub, M. Gruidl, A. Sharma, J. Szoke, M. Zakowski, V. Rusch, M. Kris, A. Viale, N. Motoi, W. Travis, B. Conley, V. E. Seshan, M. Meyerson, R. Kuick, K. K. Dobbin, T. Lively, J. W. Jacobson, D. G. Beer, and D. s. C. C. Mo, "Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," *Nature Medicine,* vol. 14, no. 8, pp. 822-827, Aug, 2008.

[104] M. Tokumo, S. Toyooka, K. Kiura, H. Shigematsu, K. Tomii, M. Aoe, K. Ichimura, T. Tsuda, M. Yano, K. Tsukuda, M. Tabata, H. Ueoka, M. Tanimoto, H. Date, A. F. Gazdar, and N. Shimizu, "The relationship between epidermal growth factor receptor mutations and clinicopathologic features in non-small cell lung cancers," *Clinical Cancer Research,* vol. 11, no. 3, pp. 1167-1173, Feb, 2005.

[105] H. Cortes-Funes, C. Gomez, R. Rosell, P. Valero, C. Garcia-Giron, A. Velasco, A. Izquierdo, P. Diz, C. Camps, D. Castellanos, V. Alberola, F. Cardenal, J. L. Gonzalez-Larriba, J. M. Vieitez, I. Maeztu, J. J. Sanchez, C. Queralt, C. Mayo, P. Mendez, T. Moran, M. Taron, and S. L. C. Grp, "Epidermal growth factor receptor activating mutations in Spanish gefitinib-treated non-small-cell lung cancer patients," *Annals of Oncology,* vol. 16, no. 7, pp. 1081-1086, Jul, 2005.

[106] T. Y. Chou, C. H. Chiu, L. H. Li, C. Y. Hsiao, C. Y. Tzen, K. T. Chang, Y. M. Chen, R. P. Perng, S. F. Tsai, and C. M. Tsai, "Mutation in the tyrosine kinase domain of epidermal growth factor receptor is a predictive and prognostic factor for gefitinib treatment in patients with non-small cell lung cancer," *Clinical Cancer Research,* vol. 11, no. 10, pp. 3750-3757, May, 2005.

[107] M. A. Levy, C. M. Lovly, and W. Pao, "Translating genomic information into clinical medicine: Lung cancer as a paradigm," *Genome Research,* vol. 22, no. 11, pp. 2101-2108, Nov, 2012.

[108] P. Cerrato, "IBM Watson Finally Graduates Medical School," *Information Week.* 2012. [Online]. Available: http://www.informationweek.com/healthcare/clinical-systems/ibm-watson-finally-graduates-medical-sch/240009562. [Accessed: Jan 10, 2013].

[109] Personalized Medicine Coalition, "The Case for Personalized Medicine," 2011. [Online].Available: http://www.ageofpersonalizedmedicine.org/objects/pdfs/Case_for_PM_3rd_edition.pdf. [Accessed: Jan 10, 2013].

[110] A. K. Smith, J. Z. Ayanian, K. E. Covinsky, B. E. Landon, E. P. McCarthy, C. C. Wee, and M. A. Steinman, "Conducting High-Value Secondary Dataset Analysis: An Introductory Guide and Resources," *Journal of General Internal Medicine,* vol. 26, no. 8, pp. 920-929, Aug, 2011.

[111] D. Pyle, *Data preparation for data mining*: Morgan Kaufmann, 1999.

[112] A. K. Akobeng, "Principles of evidence based medicine," *Archives of Disease in Childhood,* vol. 90, no. 8, pp. 837-840, Aug, 2005.

[113] P. W. Stone, “Popping the (PICO) question in research and evidence-based practice,” *Applied Nursing Research,* vol. 15, no. 3, pp. 197-198, Aug, 2002.

[114] J.Howick, I.Chalmers, P. Glasziou, T. Greenhalgh, C. Heneghan, A. Liberati, I.Moschetti, B.Phillips, and H.Thornton, "The 2011 Oxford CEBM Evidence Levels of Evidence (Introductory Document)," *Oxford Centre for Evidence-Based Medicine.* [Online]. Available: http://www.cebm.net/index.aspx?o=5653. [Accessed: June 2, 2013].

[115] The Cochrane Collaboration, “About us,” 2013. [Online]. Available: http://www.cochrane.org. [Accessed: June 2, 2013].

[116] S. Green, and S. McDonald, “Cochrane Collaboration: more than systematic reviews?” *Internal Medicine Journal,* vol. 35, no. 1, pp. 4-5, Jan, 2005.

[117] H. Gao, X. Ding, D. Wei, P. Cheng, X. Su, H. Liu, F. Aziz, D. Wang, and T. Zhang, “Efficacy of erlotinib in patients with advanced non-small cell lung cancer: a pooled analysis of randomized trials,” *Anticancer Drugs,* vol. 22, no. 9, pp. 842-52, Oct, 2011.

[118] M. S. Tsao, A. Sakurada, J. C. Cutz, C. Q. Zhu, S. Kamel-Reid, J. Squire, I. Lorimer, T. Zhang, N. Liu, M. Daneshmand, P. Marrano, G. D. Santos, A. Lagarde, F. Richardson, L. Seymour, M. Whitehead, K. Y. Ding, J. Pater, and F. A. Shepherd, “Erlotinib in lung cancer - Molecular and clinical predictors of outcome,” *New England Journal of Medicine,* vol. 353, no. 2, pp. 133-144, Jul 14, 2005.

[119] K. Ohtsuka, H. Ohnishi, G. Furuyashiki, H. Nogami, Y. Koshiishi, A. Ooide, S. Matsushima, T. Watanabe, and T. Goya, “Clinico-pathological and biological significance of tyrosine kinase domain gene mutations and overexpression of epidermal growth factor receptor for lung adenocarcinoma,” *Journal of Thoracic Oncology,* vol. 1, no. 8, pp. 787-795, Oct, 2006.

[120] S. B. Hulley, S. R. Cummings, W. S. Browner, D. G. Grady, and T. B. Newman, *Designing clinical research*: LWW, 2013.

[121] National Cancer Institute, “Age (CUI C0001779),” *NCI Metathesaurus*. [Online]. Available: http://ncim.nci.nih.gov/ncimbrowser/ConceptReport.jsp?dictionary=NCI%20Thesaurus&code=C0001779','NCImBrowser%20C0001779. [Accessed: June 2, 2013].

[122] National Cancer Institute, “Gender (CUI C0079399),” *NCI Metathesaurus*.MSH Definition [Online]. Available: http://ncim.nci.nih.gov/ncimbrowser/ConceptReport.jsp?dictionary=NCI%20MetaThesaurus&code=C0079399. [Accessed: June 2, 2013].

[123] Y. L. Choi, J. M. Sun, J. Cho, S. Rampal, J. Han, B. Parasuraman, E. Guallar, G. Lee, J. Lee, and Y. M. Shim, "EGFR Mutation Testing in Patients with Advanced Non-Small Cell Lung Cancer: A Comprehensive Evaluation of Real-World Practice in an East Asian Tertiary Hospital," *Plos One,* vol. 8, no. 2, Feb 28, 2013.

[124] National Cancer Institute, "Performance status," *NCI Dictionary of Cancer Terms*.[Online]. Available: http://www.cancer.gov/dictionary?CdrID=44174. [Accessed: June 2, 2013].

[125] National Cancer Institute, "Diagnosis," *NCI Dictionary of Cancer Terms*.[Online]. Available: http://www.cancer.gov/dictionary?CdrID=46450. [Accessed: June 2, 2013].

[126] National Cancer Institute, "Stage," *NCI Dictionary of Cancer Terms*.[Online]. Available: http://www.cancer.gov/dictionary?CdrID=45885. [Accessed: June 2, 2013].

[127] National Cancer Institute, "First-line therapy," *NCI Dictionary of Cancer Terms*.[Online]. Available: http://www.cancer.gov/dictionary?CdrID=346494. [Accessed: June 2, 2013].

[128] National Cancer Institute, "Second-line therapy," *NCI Dictionary of Cancer Terms*.[Online]. Available: http://www.cancer.gov/dictionary?CdrID=346513. [Accessed: June 2, 2013].

[129] National Cancer Institute, "Performance status," *NCI Dictionary of Cancer Terms.*[Online]. Available: http://www.cancer.gov/dictionary?CdrID=346514. [Accessed: June 2, 2013].

[130] National Cancer Institute, "Progression-free survival," *NCI Dictionary of Cancer Terms*.[Online]. Available: http://www.cancer.gov/dictionary?CdrID=44782. [Accessed: June 2, 2013].

[131] National Cancer Institute, "Overall survival," NCI Dictionary of Cancer Terms.[Online]. Available: http://www.cancer.gov/dictionary?CdrID=655245. [Accessed: June 2, 2013].

[132] W. A. Weber, "Assessing Tumor Response to Therapy," *Journal of Nuclear Medicine,* vol. 50, pp. 1S-10S, May 1, 2009.

[133] Centers for Disease Control and Prevention, "Public-Use Data Files and Documentation," *Data Access*, 2012. [Online]. Available: http://www.cdc.gov/nchs/data_access/ftp_data.htm. [Accessed: June 2, 2013].

[134] Memorial Sloan Kettering Cancer Center, "cBioPortal for Cancer Genomics," [Online]. Available: http://cbioportal.org. [Accessed: June 2, 2013].

[135] Massachusetts General Hospital Biostatistics Center, "The Cancer Genetics Networks," 2010. [Online]. Available: http://www.cancergen.org. [Accessed: June 2, 2013].

[136] National Center for Biotechnology, "Database of Genotypes and Phenotypes (dbGaP)," [Online]. Available: http://www.ncbi.nlm.nih.gov/gap. [Accessed: June 2, 2013].

[137] Vanderbilt-Ingram Cancer Center, "DNA-mutation Inventory to Refine and Enhance Cancer Treatment (DIRECT)," My Cancer Genome. [Online]. Available**:** http://www.mycancergenome.org/about/direct.[Accessed: June 2, 2013].

[138] L. Horn, H. Chen, C. Lovly, J. Andrews, P. Yeh, M. Levy, and W. Pao, "DIRECT: DNA-mutation Inventory to Refine and Enhance Cancer Treatment—A catalogue of clinically relevant somatic mutations in lung cancer," *J Clin Oncol,* vol. 29, no. suppl, pp. 7575, 2011.

[139] P. Yeh, H. Chen, J. Andrews, R. Naser, W. Pao, and L. Horn, "DNA-Mutation Inventory to Refine and Enhance Cancer Treatment (DIRECT): A Catalog of Clinically Relevant Cancer Mutations to Enable Genome-Directed Anticancer Therapy," *Clinical Cancer Research,* vol. 19, no. 7, pp. 1894-1901, Apr, 2013.

[140] Wellcome Trust Sanger Institute, "Catalogue of Somatic Mutations in Cancer (COSMIC)," [Online]. Available: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic.[Accessed: June 2, 2013].

[141] "Somatic Mutations in Epidermal Growth Factor Receptor DataBase (SMEGFR-DB)," *EGFR Mutations Database,* 2008. [Online]. Available: http://somaticmutations-egfr.org. [Accessed: June 2, 2013].

[142] X. T. Zhang, L. Y. Li, X. L. Mu, Q. C. Cui, X. Y. Chang, W. Song, S. L. Wang, M. Z. Wang, W. Zhong, and L. Zhang, "The EGFR mutation and its correlation with response of gefitinib in previously treated Chinese patients with advanced non-small-cell lung cancer," *Ann Oncol,* vol. 16, no. 8, pp. 1334-42, Aug, 2005.

[143] J. Albrecht, A. Meves, and M. Bigby, "Case reports and case series from Lancet had significant impact on medical literature," *J Clin Epidemiol,* vol. 58, no. 12, pp. 1227-32, Dec, 2005.

[144] J. P. Vandenbroucke, "In defense of case reports and case series," *Ann Intern Med,* vol. 134, no. 4, pp. 330-4, Feb 20, 2001.

[145] K. Yoshida, Y. Yatabe, J. Y. Park, J. Shimizu, Y. Horio, K. Matsuo, T. Kosaka, T. Mitsudomi, and T. Hida, "Prospective validation for prediction of gefitinib sensitivity by epidermal growth factor receptor gene mutation in patients with non-small cell lung cancer," *J Thorac Oncol,* vol. 2, no. 1, pp. 22-8, Jan, 2007.

[146] R. Rosell, T. Moran, C. Queralt, R. Porta, F. Cardenal, C. Camps, M. Majem, G. Lopez-Vivanco, D. Isla, M. Provencio, A. Insa, B. Massuti, J. L. Gonzalez-Larriba, L. Paz-Ares, I. Bover, R. Garcia-Campelo, M. A. Moreno, S. Catot, C. Rolfo, N. Reguart, R. Palmero, J. M. Sanchez, R. Bastus, C. Mayo, J. Bertran-Alamillo, M. A. Molina, J. J. Sanchez, M. Taron, and G. Spanish Lung Cancer, "Screening for epidermal growth factor receptor mutations in lung cancer," *N Engl J Med,* vol. 361, no. 10, pp. 958-67, Sep 3, 2009.

[147] M. Kris , T. Mok , E. Kim , et al, "Response and progression-free survival in 1006 patients with known EGFR mutation status in phase III randomized trials of gefitinib in individuals with non-small cell lung cancer," *EJC Supplements*, abstract O-9003, 2009.

[148] A. Marchetti, C. Martella, L. Felicioni, F. Barassi, S. Salvatore, A. Chella, P. P. Camplese, T. Iarussi, F. Mucilli, A. Mezzetti, F. Cuccurullo, R. Sacco, and F. Buttitta, "EGFR mutations in non-small-cell lung cancer: analysis of a large series of cases and development of a rapid and sensitive method for diagnostic screening with potential implications on pharmacologic treatment," *J Clin Oncol,* vol. 23, no. 4, pp. 857-65, Feb 1, 2005.

[149] H. Linardou, I. J. Dahabreh, D. Kanaloupiti, F. Siannis, D. Bafaloukos, P. Kosmidis, C. A. Papadimitriou, and S. Murray, "Assessment of somatic k-RAS mutations as a mechanism associated with resistance to EGFR-targeted agents: a systematic review and meta-analysis of studies in advanced non-small-cell lung cancer and metastatic colorectal cancer," *Lancet Oncol,* vol. 9, no. 10, pp. 962-72, Oct, 2008.

[150] E. Felip, F. Rojo, M. Reck, A. Heller, B. Klughammer, G. Sala, S. Cedres, S. Peralta, H. Maacke, D. Foernzler, M. Parera, J. Mocks, C. Saura, U. Gatzemeier, and J. Baselga, "A phase II pharmacodynamic study of erlotinib in patients with advanced non-small cell lung cancer previously treated with platinum-based chemotherapy," *Clin Cancer Res,* vol. 14, no. 12, pp. 3867-74, Jun 15, 2008.

[151] P. J. Roberts, and T. E. Stinchcombe, "KRAS mutation: should we test for it, and does it matter?," *J Clin Oncol,* vol. 31, no. 8, pp. 1112-21, Mar 10, 2013.
[152] Therasse, P. "Measuring the clinical response. What does it mean?." European Journal of Cancer 38, no. 14 (2002): 1817-1823.

[153] I. H. Witten, and E. Frank, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2005.

[154] X. Su, R. Greiner, T. M. Khoshgoftaar, and A. Napolitano, "Using classifier-based nominal imputation to improve machine learning," *Advances in Knowledge Discovery and Data Mining*, pp. 124-135: Springer, 2011.

[155] I. Y. Tam, E. L. Leung, V. P. Tin, D. T. Chua, A. D. Sihoe, L. C. Cheng, L. P. Chung, and M. P. Wong, "Double EGFR mutants containing rare EGFR mutant types show reduced in vitro response to gefitinib compared with common activating missense mutations," *Mol Cancer Ther,* vol. 8, no. 8, pp. 2142-51, Aug, 2009.

[156] S. G. Wu, Y. L. Chang, Y. C. Hsu, J. Y. Wu, C. H. Yang, C. J. Yu, M. F. Tsai, J. Y. Shih, and P. C. Yang, "Good response to gefitinib in lung adenocarcinoma of complex epidermal growth factor receptor (EGFR) mutations with the classical mutation pattern," *Oncologist,* vol. 13, no. 12, pp. 1276-84, Dec, 2008.

[157] L. H. Liu, and H. Motoda, *Feature extraction, construction and selection: A data mining perspective*: Springer, 1998.

[158] A. Bezjak, D. Tu, L. Seymour, G. Clark, A. Trajkovic, M. Zukin, J. Ayoub, S. Lago, R. de Albuquerque Ribeiro, A. Gerogianni, A. Cyjon, J. Noble, F. Laberge, R. T. Chan, D. Fenton, J. von Pawel, M. Reck, F. A. Shepherd, and B. R. National Cancer Institute of Canada Clinical Trials Group Study, "Symptom improvement in lung cancer patients treated with erlotinib: quality of life analysis of the National Cancer Institute of Canada Clinical Trials Group Study BR.21," *J Clin Oncol,* vol. 24, no. 24, pp. 3831-7, Aug 20, 2006.

[159] S. J. Mandrekar, and D. J. Sargent, "Clinical trial designs for predictive biomarker validation: one size does not fit all," *Journal of biopharmaceutical statistics,* vol. 19, no. 3, pp. 530-542, 2009.

[160] T. Cufer, E. Vrdoljak, R. Gaafar, I. Erensoy, K. Pemberton, and S. S. Group, "Phase II, open-label, randomized study (SIGN) of single-agent gefitinib (IRESSA) or docetaxel as second-line therapy in patients with advanced (stage IIIb or IV) non-small-cell lung cancer," *Anticancer Drugs,* vol. 17, no. 4, pp. 401-9, Apr, 2006.

[161] F. A. Shepherd, J. Rodrigues Pereira, T. Ciuleanu, E. H. Tan, V. Hirsh, S. Thongprasert, D. Campos, S. Maoleekoonpiroj, M. Smylie, R. Martins, M. van Kooten, M. Dediu, B. Findlay, D. Tu, D. Johnston, A. Bezjak, G. Clark, P. Santabarbara, L. Seymour, and G. National Cancer Institute of Canada Clinical Trials, "Erlotinib in previously treated non-small-cell lung cancer," *N Engl J Med,* vol. 353, no. 2, pp. 123-32, Jul 14, 2005.

[162] E. S. Kim, V. Hirsh, T. Mok, M. A. Socinski, R. Gervais, Y. L. Wu, L. Y. Li, C. L. Watkins, M. V. Sellers, E. S. Lowe, Y. Sun, M. L. Liao, K. Osterlind, M. Reck, A. A. Armour, F. A. Shepherd, S. M. Lippman, and J. Y. Douillard, "Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial," *Lancet,* vol. 372, no. 9652, pp. 1809-18, Nov 22, 2008.

[163] K. Kelly, K. Chansky, L. E. Gaspar, K. S. Albain, J. Jett, Y. C. Ung, D. H. Lau, J. J. Crowley, and D. R. Gandara, "Phase III trial of maintenance gefitinib or placebo after concurrent chemoradiotherapy and docetaxel consolidation in inoperable stage III non-small-cell lung cancer: SWOG S0023," *J Clin Oncol,* vol. 26, no. 15, pp. 2450-6, May 20, 2008.

[164] V. A. Miller, M. G. Kris, N. Shah, J. Patel, C. Azzoli, J. Gomez, L. M. Krug, W. Pao, N. Rizvi, and B. Pizzo, "Bronchioloalveolar pathologic subtype and smoking history predict sensitivity to gefitinib in advanced non–small-cell lung cancer," *Journal of Clinical Oncology,* vol. 22, no. 6, pp. 1103-1109, 2004.

[165] P. A. Janne, S. Gurubhagavatula, B. Y. Yeap, J. Lucca, P. Ostler, A. T. Skarin, P. Fidias, T. J. Lynch, and B. E. Johnson, "Outcomes of patients with advanced non-small cell lung cancer treated with gefitinib (ZD1839, "Iressa") on an expanded access study," *Lung Cancer,* vol. 44, no. 2, pp. 221-30, May, 2004.

[166] T. S. Mok, Y. L. Wu, S. Thongprasert, C. H. Yang, D. T. Chu, N. Saijo, P. Sunpaweravong, B. Han, B. Margono, Y. Ichinose, Y. Nishiwaki, Y. Ohe, J. J. Yang, B. Chewaskulyong, H. Jiang, E. L. Duffield, C. L. Watkins, A. A. Armour, and M. Fukuoka, "Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma," *N Engl J Med,* vol. 361, no. 10, pp. 947-57, Sep 3, 2009.

[167] J. S. Lee, K. Park, S.-W. Kim, D. H. Lee, H. T. Kim, J.-Y. Han, T. Yun, M.-J. Ahn, J. S. Ahn, and C. Suh, "A randomized phase III study of gefitinib (IRESSATM) versus standard chemotherapy (gemcitabine plus cisplatin) as a first-line treatment for never-smokers with advanced or metastatic adenocarcinoma of the lung." pp. S283-S284.

[168] T. Mitsudomi, S. Morita, Y. Yatabe, S. Negoro, I. Okamoto, J. Tsurutani, T. Seto, M. Satouchi, H. Tada, T. Hirashima, K. Asami, N. Katakami, M. Takada, H. Yoshioka, K. Shibata, S. Kudoh, E. Shimizu, H. Saito, S. Toyooka, K. Nakagawa, M. Fukuoka, and G. West Japan Oncology, "Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial," *Lancet Oncol,* vol. 11, no. 2, pp. 121-8, Feb, 2010.

[169] M. Maemondo, A. Inoue, K. Kobayashi, S. Sugawara, S. Oizumi, H. Isobe, A. Gemma, M. Harada, H. Yoshizawa, I. Kinoshita, Y. Fujita, S. Okinaga, H. Hirano, K. Yoshimori, T. Harada, T. Ogura, M. Ando, H. Miyazawa, T. Tanaka, Y. Saijo, K. Hagiwara, S. Morita, T. Nukiwa, and G. North-East Japan Study, "Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR," *N Engl J Med,* vol. 362, no. 25, pp. 2380-8, Jun 24, 2010.

[170] R. Rosell, E. Carcereny, R. Gervais, A. Vergnenegre, B. Massuti, E. Felip, R. Palmero, R. Garcia-Gomez, C. Pallares, and J. M. Sanchez, “Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial,” *The lancet oncology,* vol. 13, no. 3, pp. 239-246, 2012.

[171] C. Zhou, Y.-L. Wu, G. Chen, J. Feng, X.-Q. Liu, C. Wang, S. Zhang, J. Wang, S. Zhou, and S. Ren, “Erlotinib versus chemotherapy as first-line treatment for patients with advanced< i> EGFR</i> mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study,” *The lancet oncology,* vol. 12, no. 8, pp. 735-742, 2011.

[172] V. Somaraki, D. Broadbent, F. Coenen, and S. Harding, "Finding temporal patterns in noisy longitudinal data: a study in diabetic retinopathy," *Advances in Data Mining. Applications and Theoretical Aspects*, pp. 418-431: Springer, 2010.

[173] A. Agrawal, and A. Choudhary, "Identifying HotSpots in Lung Cancer Data Using Association Rule Mining." pp. 995-1002.

[174]C. Ordonez, N. Ezquerra, and C. A. Santana, “Constraining and summarizing association rules in medical data,” *Knowledge and Information Systems,* vol. 9, no. 3, pp. 1-2, 2006.

[175] L. Elfangary, and W. A. Atteya, "Mining Medical Databases Using Proposed Incremental Association Rules Algorithm (PIA)."in *Digital Society,Second International Conference on the, Feb. 10-15,2008*, Sainte Luce, pp.88-92.

[176] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases." In *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207-216, 1993.

[177] R. Srikant, Q. Vu, and R. Agrawal, "Mining Association Rules with Item Constraints." In *KDD*, vol. 97, pp. 67-73. 1997.

[178] C. K.-S. Leung, "Constraint-Based Association Rule Mining," *Encyclopedia of Data Warehousing and Mining, Second Edition*, pp. 307-312: IGI Global, 2009.

[179] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, “Fast Discovery of Association Rules,” *Advances in knowledge discovery and data mining,* vol. 12, pp. 307-328, 1996.

[180] E. García, C. Romero, S. Ventura, and T. Calders, "Drawbacks and solutions of applying association rule mining in learning management systems." In *Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML 2007), Crete, Greece*, pp. 13-22. 2007.

[181] Z. Zheng, R. Kohavi, and L. Mason, "Real world performance of association rule algorithms," in Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2001. pp. 86-98.

[182] B. Goethals, and J. V. d. Bussche, "Interactive constrained association rule mining," *arXiv preprint cs/0112011*, 2001.

[183] B. Jeudy, and J.-F. Boulicaut, "Optimization of association rule mining queries," *Intelligent Data Analysis,* vol. 6, no. 4, pp. 341-357, 2002.

[184] E. García, C. Romero, S. Ventura, and C. de Castro, "Using rules discovery for the continuous improvement of e-learning courses," *Intelligent Data Engineering and Automated Learning–IDEAL 2006*, pp. 887-895: Springer, 2006.

[185]T. Scheffer, "Finding association rules that trade support optimally against confidence," *Principles of Data Mining and Knowledge Discovery*, pp. 424-435: Springer, 2001.

[186] J. Nahar, and K. S. Tickle, "Significant cancer risk factor extraction: an association rule discovery approach." In Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on, 2008. pp. 108-114.

[187] M. Karabatak, and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert Systems with Applications,* vol. 36, no. 2, pp. 3465-3469, 2009.

[188] Q. Fan, C.-J. Zhu, J.-Y. Xiao, B.-H. Wang, L. Yin, X.-L. Xu, and F. Rong, "An Application of Apriori Algorithm in SEER Breast Cancer Data."In Artificial Intelligence and Computational Intelligence (AICI), 2010 International Conference on, vol. 3, 2010. pp. 114-116.

[189] M. Z. R. Mukti, and F. Ahmed, "Early Detection of Lung Cancer Risk Using Data Mining," *Asian Pacific Journal of Cancer Prevention,* vol. 14, no. 1, pp. 595-598, 2013.

[190] S. Stilou, P. Bamidis, N. Maglaveras, and C. Pappas, "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare," *Studies in health technology and informatics*, no. 2, pp. 1399-1403, 2001.

[191] M.Hahsler, "A Comparison of Commonly Used Interest Measures for Association Rules," 2010. [Online].
Available:http://michael.hahsler.net/research/association_rules/measures.html.
[Accessed: June 2, 2013].

[192] Y. L. Wu, C. Zhou, Y. Cheng, S. Lu, G. Y. Chen, C. Huang, Y. S. Huang, H. H. Yan, S. Ren, Y. Liu, and J. J. Yang, "Erlotinib as second-line treatment in patients with advanced non-small-cell lung cancer and asymptomatic brain metastases: a phase II study (CTONG-0803)," *Ann Oncol,* vol. 24, no. 4, pp. 993-9, Apr, 2013.

[193] K. P. Chung, S. G. Wu, J. Y. Wu, J. C. Yang, C. J. Yu, P. F. Wei, J. Y. Shih, and P. C. Yang, "Clinical outcomes in non-small cell lung cancers harboring different exon 19 deletions in EGFR," *Clin Cancer Res,* vol. 18, no. 12, pp. 3470-7, Jun 15, 2012.

[194] S. Tracy, T. Mukohara, M. Hansen, M. Meyerson, B. E. Johnson, and P. A. Janne, "Gefitinib induces apoptosis in the EGFRL858R non-small-cell lung cancer cell line H3255," *Cancer Res,* vol. 64, no. 20, pp. 7241-4, Oct 15, 2004.

[195] W. Pao, V. Miller, M. Zakowski, J. Doherty, K. Politi, I. Sarkaria, B. Singh, R. Heelan, V. Rusch, L. Fulton, E. Mardis, D. Kupfer, R. Wilson, M. Kris, and H. Varmus, "EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib," *Proc Natl Acad Sci U S A,* vol. 101, no. 36, pp. 13306-11, Sep 7, 2004.

[196] S. W. Han, T. Y. Kim, P. G. Hwang, S. Jeong, J. Kim, I. S. Choi, D. Y. Oh, J. H. Kim, D. W. Kim, D. H. Chung, S. A. Im, Y. T. Kim, J. S. Lee, D. S. Heo, Y. J. Bang, and N. K. Kim, "Predictive and prognostic impact of epidermal growth factor receptor mutation in non-small-cell lung cancer patients treated with gefitinib," *J Clin Oncol,* vol. 23, no. 11, pp. 2493-501, Apr 10, 2005.

[197] J.-Y. Wu, S.-G. Wu, C.-H. Yang, C.-H. Gow, Y.-L. Chang, C.-J. Yu, J.-Y. Shih, and P.-C. Yang, "Lung cancer with epidermal growth factor receptor exon 20 mutations is associated with poor gefitinib treatment response," *Clinical cancer research,* vol. 14, no. 15, pp. 4877-4882, 2008.

[198] D. W. Bell, I. Gore, R. A. Okimoto, N. Godin-Heymann, R. Sordella, R. Mulloy, S. V. Sharma, B. W. Brannigan, G. Mohapatra, and J. Settleman, "Inherited susceptibility to lung cancer may be associated with the T790M drug resistance mutation in EGFR," *Nature genetics,* vol. 37, no. 12, pp. 1315-1316, 2005.

[199] A. Ogino, H. Kitao, S. Hirano, A. Uchida, M. Ishiai, T. Kozuki, N. Takigawa, M. Takata, K. Kiura, and M. Tanimoto, "Emergence of epidermal growth factor receptor T790M mutation during chronic exposure to gefitinib in a non–small cell lung cancer cell line," *Cancer Research,* vol. 67, no. 16, pp. 7807-7814, 2007.

[200] C. H. Gow, C. R. Chien, Y. L. Chang, Y. H. Chiu, S. H. Kuo, J. Y. Shih, Y. C. Chang, C. J. Yu, C. H. Yang, and P. C. Yang, "Radiotherapy in lung adenocarcinoma with brain metastases: effects of activating epidermal growth factor receptor mutations on clinical response," *Clin Cancer Res,* vol. 14, no. 1, pp. 162-8, Jan 1, 2008.

[201] B. Liu, W. Hsu, S. Chen, and Y. Ma, “Analyzing the subjective interestingness of association rules,” *Intelligent Systems and Their Applications, IEEE,* vol. 15, no. 5, pp. 47-55, 2000.

[202] A. Hata, H. Yoshioka, S. Fujita, K. Kunimasa, R. Kaji, Y. Imai, K. Tomii, M. Iwasaku, A. Nishiyama, T. Ishida, and N. Katakami, “Complex mutations in the epidermal growth factor receptor gene in non-small cell lung cancer,” *J Thorac Oncol,* vol. 5, no. 10, pp. 1524-8, Oct, 2010.

[203] D. Nicolini, J. Powell, P. Conville, and L. Martinez-Solano, “Managing knowledge in the healthcare sector. A review,” *International Journal of Management Reviews,* vol. 10, no. 3, pp. 245-263, 2008.

[204] B.Goldman, “Scientists consider potential of abundant biomedical data,” Stanford School of Medicine, 2013. [Online]. Available:http://med.stanford.edu/ism/2013/may/bigdata-052813.html.[Accessed: July 2, 2013].

[205] A. X. Garg, N. K. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. Devereaux, J. Beyene, J. Sam, and R. B. Haynes, “Effects of computerized clinical decision support systems on practitioner performance and patient outcomes,” *JAMA: the journal of the American Medical Association,* vol. 293, no. 10, pp. 1223-1238, 2005.

[206] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, “Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success,” *Bmj,* vol. 330, no. 7494, pp. 765, 2005.

[207] M. Abbasi, and S. Kashiyarndi, “Clinical Decision Support Systems: A discussion on different methodologies used in Health Care,” 2006.

[208] N. H. Shah, and J. D. Tenenbaum, “The coming age of data-driven medicine: translational bioinformatics' next frontier,” *J Am Med Inform Assoc,* vol. 19, no. e1, pp. e2-4, Jun, 2012.
.
[209] K. K. Mane, C. Schmitt, P. Owen, K. Gersing, S. C. Ahalt, and K. Wilhelmsen, "Data-driven approaches to augment clinical decision in EMR Era." In Cognitive Information Processing (CIP), 2012 3rd International Workshop on, 2012. pp. 1-5.

[210] C. P. Friedman, A. K. Wong, and D. Blumenthal, “Achieving a nationwide learning health system,” *Sci Transl Med,* vol. 2, no. 57, pp. 57cm29, Nov 10, 2010.

[211] P. LePendu, M. A. Musen, and N. H. Shah, "The Age of Data-Driven Medicine: Mining the Electronic Health Record." In *ICBO*. 2011.

[212] D. Bates, M. Cohen, L. Leape, J. M. Overhage, M. M. Shabot, and T. Sheridan, "Reducing the frequency of errors in medicine using information technology," *Journal of the American Medical Informatics Association,* vol. 8, no. 4, pp. 299-308, 2001.

[213] M. D. Brundage, P. A. Groome, D. Feldman-Stewart, J. R. Davidson, and W. J. Mackillop, "Decision analysis in locally advanced non-small-cell lung cancer: is it useful?," *J Clin Oncol,* vol. 15, no. 3, pp. 873-83, Mar, 1997.

[214] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu Rev Biomed Eng,* vol. 8, pp. 537-65, 2006.

[215] H. J. Lee, Y. T. Kim, P. J. Park, Y. S. Shin, K. N. Kang, Y. Kim, and C. W. Kim, "A novel detection method of non–small cell lung cancer using multiplexed bead-based serum biomarker profiling," *The Journal of thoracic and cardiovascular surgery,* vol. 143, no. 2, pp. 421-427. e3, 2012.

[216] Y. Zhao, D. Zeng, M. A. Socinski, and M. R. Kosorok, "Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer," *Biometrics,* vol. 67, no. 4, pp. 1422-1433, 2011.

[217] M. van den Branden, N. Wiratunga, D. Burton, and S. Craw, "Integrating case-based reasoning with an electronic patient record system," *Artificial Intelligence in Medicine,* vol. 51, no. 2, pp. 117-123, 2011.

[218] M. C. Lee, L. Boroczky, K. Sungur-Stasik, A. D. Cann, A. C. Borczuk, S. M. Kawut, and C. A. Powell, "Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction," *Artificial intelligence in medicine,* vol. 50, no. 1, pp. 43-53, 2010.

[219] R. Sacile, E. Montaldo, C. Ruggiero, H. E. Nieburgs, and G. Nicolò, "A decision support system to detect morphologic changes of chromatin arrangement in normal-appearing cells," *NanoBioscience, IEEE Transactions on,* vol. 2, no. 2, pp. 118-123, 2003.

[220] P. Wolfe, J. Murphy, J. McGinley, Z. Zhu, W. Jiang, E. B. Gottschall, and H. J. Thompson, "Using nuclear morphometry to discriminate the tumorigenic potential of cells: a comparison of statistical methods," *Cancer Epidemiology Biomarkers & Prevention,* vol. 13, no. 6, pp. 976-988, 2004.

.

[221] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach," *Statistics in Medicine,* vol. 17, no. 10, pp. 1169-1186, 1998.

[222] C. R. Schweiger, G. Soeregi, S. Spitzauer, G. Maenner, and A. Pohl, "Evaluation of laboratory data by conventional statistics and by three types of neural networks," *Clinical chemistry,* vol. 39, no. 9, pp. 1966-1971, 1993.

[223] H. Gutte, D. Jakobsson, F. Olofsson, M. Ohlsson, S. Valind, A. Loft, L. Edenbrandt, and A. Kjaer, "Automated interpretation of PET/CT images in patients with lung cancer," *Nucl Med Commun,* vol. 28, no. 2, pp. 79-84, Feb, 2007.

[224] S. Polak, A. Skowron, A. Mendyk, and J. Brandys, "Artificial neural network in pharmacoeconomics," *Stud Health Technol Inform,* vol. 105, pp. 241-9, 2004.

[225] P. Campadelli, E. Casiraghi, and D. Artioli, "A fully automated method for lung nodule detection from postero-anterior chest radiographs," *IEEE Trans Med Imaging,* vol. 25, no. 12, pp. 1588-603, Dec, 2006.

[226] L. Toschi, and F. Cappuzzo, "Understanding the new genetics of responsiveness to epidermal growth factor receptor tyrosine kinase inhibitors," *Oncologist,* vol. 12, no. 2, pp. 211-20, Feb, 2007.

[227] S. Murray, V. Karavasilis, M. Bobos, E. Razis, S. Papadopoulos, C. Christodoulou, P. Kosmidis, and G. Fountzilas, "Molecular predictors of response to tyrosine kinase inhibitors in patients with Non-Small-Cell Lung Cancer," *J Exp Clin Cancer Res,* vol. 31, pp. 77, 2012.

[228] X. Zhang, and A. Chang, "Molecular predictors of EGFR-TKI sensitivity in advanced non-small cell lung cancer," *Int J Med Sci,* vol. 5, no. 4, pp. 209-17, 2008.

[229] W. A. Weber, V. Petersen, B. Schmidt, L. Tyndale-Hines, T. Link, C. Peschel, and M. Schwaiger, "Positron emission tomography in non-small-cell lung cancer: prediction of response to chemotherapy by quantitative assessment of glucose use," *J Clin Oncol,* vol. 21, no. 14, pp. 2651-7, Jul 15, 2003.

[230] C. Suzuki, H. Jacobsson, T. Hatschek, M. R. Torkzad, K. Boden, Y. Eriksson-Alm, E. Berg, H. Fujii, A. Kubo, and L. Blomqvist, "Radiologic measurements of tumor response to treatment: practical approaches and limitations," *Radiographics,* vol. 28, no. 2, pp. 329-44, Mar-Apr, 2008.

[231] B. Clinton, and A. Gore, *Reinventing the Regulation of Cancer Drugs: Accelerating Approval and Expanding Access*: National Performance Review, 1996.

[232] J. W. Wang, W. Zheng, J. B. Liu, Y. Chen, L. H. Cao, R. Z. Luo, A. H. Li, and J. H. Zhou, "Assessment of early tumor response to cytotoxic chemotherapy with dynamic contrast-enhanced ultrasound in human breast cancer xenografts," *PLoS One,* vol. 8, no. 3, pp. e58274, 2013.

[233] J P. Jeatrakul, K. W. Wong, C. C. Fung, and Y. Takama, "Misclassification analysis for the class imbalance problem." In World Automation Congress (WAC), 2010. pp. 1-6.

[234] R. L. Schilsky, "End points in cancer clinical trials and the drug approval process," *Clin Cancer Res,* vol. 8, no. 4, pp. 935-8, Apr, 2002.

[235] P. N. Lara, M. W. Redman, K. Kelly, M. J. Edelman, S. K. Williamson, J. J. Crowley, and D. R. Gandara, "Disease control rate at 8 weeks predicts clinical benefit in advanced non–small-cell lung cancer: results from Southwest Oncology Group randomized trials," *Journal of Clinical Oncology,* vol. 26, no. 3, pp. 463-467, 2008.

[236] J. Wang, N. Wu, M. D. Cham, and Y. Song, "Tumor response in patients with advanced non-small cell lung cancer: perfusion CT evaluation of chemotherapy and radiation therapy," *AJR Am J Roentgenol,* vol. 193, no. 4, pp. 1090-6, Oct, 2009.

[237] D. P. Berrar, B. Sturgeon, I. Bradbury, and W. Dubitzky, "Microarray data integration and machine learning techniques for lung cancer survival prediction." In *Proceedings of the the International Conference of Critical Assessment of Microarray Data Analysis*, pp. 43-54. 2003.

[238] L. Kotthoff, I. P. Gent, and I. Miguel, "A preliminary evaluation of machine learning in algorithm selection for search problems," In *Fourth Annual Symposium on Combinatorial Search*. 2011.

[239] L. Xu, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "SATzilla: Portfolio-based Algorithm Selection for SAT," *J. Artif. Intell. Res.(JAIR),* vol. 32, pp. 565-606, 2008.

[240] E. O'Mahony, Eoin, Emmanuel Hebrard, Alan Holland, Conor Nugent, and Barry O'Sullivan. " E. O'Mahony, E. Hebrard, A. Holland, C. Nugent, and B. O'Sullivan, "Using case-based reasoning in an algorithm portfolio for constraint solving," In *Irish Conference on Artificial Intelligence and Cognitive Science*. 2008.

[241] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural computation,* vol. 8, no. 7, pp. 1341-1390, 1996.

[242] Y. W. Wan, E. Sabbagh, R. Raese, Y. Qian, D. Luo, J. Denvir, V. Vallyathan, V. Castranova, and N. L. Guo, "Hybrid models identified a 12-gene signature for lung cancer prognosis and chemoresponse prediction," *PLoS One,* vol. 5, no. 8, pp. e12222, 2010.

[243] E. Adetiba and F.A. Ibikunle, "Ensembling of EGFR Mutations' based Artificial Neural Networks for Improved Diagnosis of Non-Small Cell Lung Cancer," *International Journal of Computer Applications*, vol.20, no. 7, pp. 39-47, April 2011.

[244] E. Adetiba, J. Ekeh, V. Matthews, S. Daramola, and M. Eleanya, "Article Title: Estimating An Optimal Backpropagation Algorithm for Training An ANN with the EGFR Exon 19 Nucleotide Sequence: An Electronic Diagnostic Basis for Non-Small Cell Lung Cancer (NSCLC)."

[245] C. D. Coldren, B. A. Helfrich, S. E. Witta, M. Sugita, R. Lapadat, C. Zeng, A. Baron, W. A. Franklin, F. R. Hirsch, M. W. Geraci, and P. A. Bunn, Jr., "Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines," *Mol Cancer Res,* vol. 4, no. 8, pp. 521-8, Aug, 2006.

[246] Z. H. Zhu, B. Y. Sun, Y. Ma, J. Y. Shao, H. Long, X. Zhang, J. H. Fu, L. J. Zhang, X. D. Su, Q. L. Wu, P. Ling, M. Chen, Z. M. Xie, Y. Hu, and T. H. Rong, "Three immunomarker support vector machines-based prognostic classifiers for stage IB non-small-cell lung cancer," *J Clin Oncol,* vol. 27, no. 7, pp. 1091-9, Mar 1, 2009.

[247] L. Wu, W. Chang, J. Zhao, Y. Yu, X. Tan, T. Su, L. Zhao, S. Huang, S. Liu, and G. Cao, "Development of Autoantibody Signatures as Novel Diagnostic Biomarkers of Non–Small Cell Lung Cancer," *Clinical Cancer Research,* vol. 16, no. 14, pp. 3760-3768, 2010.

[248] M. Wislez, M. Antoine, L. Baudrin, V. Poulot, A. Neuville, M. Pradere, E. Longchampt, S. Isaac-Sibille, M. P. Lebitasy, and J. Cadranel, "Non-mucinous and mucinous subtypes of adenocarcinoma with bronchioloalveolar carcinoma features differ by biomarker expression and in the response to gefitinib," Lung Cancer, vol. 68, no. 2, pp. 185-91, May, 2010.

[249] J. S. Kaminker, Y. Zhang, A. Waugh, P. M. Haverty, B. Peters, D. Sebisanovic, J. Stinson, W. F. Forrest, J. F. Bazan, S. Seshagiri, and Z. Zhang, "Distinguishing cancer-associated missense mutations from common polymorphisms," *Cancer Res,* vol. 67, no. 2, pp. 465-73, Jan 15, 2007.

[250] E. C. Farlow, K. Patel, S. Basu, B. S. Lee, A. W. Kim, J. S. Coon, L. P. Faber, P. Bonomi, M. J. Liptay, and J. A. Borgia, "Development of a multiplexed tumor-associated autoantibody-based blood test for the detection of non-small cell lung cancer," *Clin Cancer Res,* vol. 16, no. 13, pp. 3452-62, Jul 1, 2010.

[251] R.Caruana and Alexandru Niculescu-Mizil, "An empirical comparison of supervised learning algorithms." In *Proceedings of the 23rd international conference on Machine learning* 2006. pp. 161-168.

[252] K. Sugio, H. Uramoto, K. Ono, T. Oyama, T. Hanagiri, M. Sugaya, Y. Ichiki, T. So, S. Nakata, M. Morita, and K. Yasumoto, "Mutations within the tyrosine kinase domain of EGFR gene specifically occur in lung adenocarcinoma patients with a low exposure of tobacco smoking," *Br J Cancer,* vol. 94, no. 6, pp. 896-903, Mar 27, 2006.

[253] D. M. Jackman, B. Y. Yeap, L. V. Sequist, N. Lindeman, A. J. Holmes, V. A. Joshi, D. W. Bell, M. S. Huberman, B. Halmos, M. S. Rabin, D. A. Haber, T. J. Lynch, M. Meyerson, B. E. Johnson, and P. A. Janne, "Exon 19 deletion mutations of epidermal growth factor receptor are associated with prolonged survival in non-small cell lung cancer patients treated with gefitinib or erlotinib," *Clin Cancer Res,* vol. 12, no. 13, pp. 3908-14, Jul 1, 2006.

[254] J. Fan, S. Upadhye, and A. Worster, "Understanding receiver operating characteristic (ROC) curves," *CJEM,* vol. 8, no. 1, pp. 19-20, Jan, 2006.

[255] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science,* vol. 240, no. 4857, pp. 1285-93, Jun 3, 1988.

[256] S. Dey, R. Gupta, M. Steinbach, and V. Kumar, "Integration of Clinical and Genomic data: a Methodological Survey," 2013.[Online]. Available: https://wwws.cs.umn.edu/tech_reports_upload/tr2013/old_files/13-005.pdf. [Accessed: June 28, 2013].

[257] M. A. Hall, and L. A. Smith, "Practical feature subset selection for machine learning," Australian Computer Science Conference, 1998. pp.181-191.

[258] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," In Advances in Kernel Methods - Support Vector Learning, 1998.

[259] T. Hastie, and R. Tibshirani, "Classification by pairwise coupling," *The annals of statistics,* vol. 26, no. 2, pp. 451-471, 1998.

[260] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[261] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees. Wadsworth & Brooks," *Monterey, CA*, 1984.

[262] K. Cetin, D. S. Ettinger, Y. J. Hei, and C. D. O'Malley, "Survival by histologic subtype in stage IV nonsmall cell lung cancer based on data from the Surveillance, Epidemiology and End Results Program," *Clin Epidemiol,* vol. 3, pp. 139-48, 2011.

[263] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics,* vol. 44, no. 3, pp. 837-45, Sep, 1988.

[264] J. Y. Douillard, F. A. Shepherd, V. Hirsh, T. Mok, M. A. Socinski, R. Gervais, M. L. Liao, H. Bischoff, M. Reck, M. V. Sellers, C. L. Watkins, G. Speake, A. A. Armour, and E. S. Kim, "Molecular predictors of outcome with gefitinib and docetaxel in previously treated non-small-cell lung cancer: data from the randomized phase III INTEREST trial," *J Clin Oncol,* vol. 28, no. 5, pp. 744-52, Feb 10, 2010.

[265] R. Maruyama, Y. Nishiwaki, T. Tamura, N. Yamamoto, M. Tsuboi, K. Nakagawa, T. Shinkai, S. Negoro, F. Imamura, K. Eguchi, K. Takeda, A. Inoue, K. Tomii, M. Harada, N. Masuda, H. Jiang, Y. Itoh, Y. Ichinose, N. Saijo, and M. Fukuoka, "Phase III study, V-15-32, of gefitinib versus docetaxel in previously treated Japanese patients with non-small-cell lung cancer," *J Clin Oncol,* vol. 26, no. 26, pp. 4244-52, Sep 10, 2008.

[266] T. S. Mok, Y. L. Wu, S. Thongprasert, C. H. Yang, D. T. Chu, N. Saijo, P. Sunpaweravong, B. Han, B. Margono, Y. Ichinose, Y. Nishiwaki, Y. Ohe, J. J. Yang, B. Chewaskulyong, H. Jiang, E. L. Duffield, C. L. Watkins, A. A. Armour, and M. Fukuoka, "Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma," *N Engl J Med,* vol. 361, no. 10, pp. 947-57, Sep 3, 2009.

[267] F. R. Hirsch, M. Varella-Garcia, P. A. Bunn, Jr., W. A. Franklin, R. Dziadziuszko, N. Thatcher, A. Chang, P. Parikh, J. R. Pereira, T. Ciuleanu, J. von Pawel, C. Watkins, A. Flannery, G. Ellison, E. Donald, L. Knight, D. Parums, N. Botwood, and B. Holloway, "Molecular predictors of outcome with gefitinib in a phase III placebo-controlled study in advanced non-small-cell lung cancer," *J Clin Oncol,* vol. 24, no. 31, pp. 5034-42, Nov 1, 2006.

[268] C.-Q. Zhu, G. da Cunha Santos, K. Ding, A. Sakurada, J.-C. Cutz, N. Liu, T. Zhang, P. Marrano, M. Whitehead, and J. A. Squire, "Role of KRAS and EGFR as biomarkers of response to erlotinib in National Cancer Institute of Canada Clinical Trials Group Study BR. 21," *Journal of Clinical Oncology,* vol. 26, no. 26, pp. 4268-4275, 2008.

[269] A. Augustin, J. Lamerz, H. Meistermann, S. Golling, S. Scheiblich, J. C. Hermann, G. Duchateau-Nguyen, M. Tzouros, D. W. Avila, H. Langen, L. Essioux, and B. Klughammer, "Quantitative chemical proteomics profiling differentiates erlotinib from gefitinib in EGFR wild-type non-small cell lung carcinoma cell lines," *Mol Cancer Ther,* vol. 12, no. 4, pp. 520-9, Apr, 2013.

[270]V. A. Miller, G. J. Riely, M. F. Zakowski, A. R. Li, J. D. Patel, R. T. Heelan, M. G. Kris, A. B. Sandler, D. P. Carbone, A. Tsao, R. S. Herbst, G. Heller, M. Ladanyi, W. Pao, and D. H. Johnson, "Molecular characteristics of bronchioloalveolar carcinoma and adenocarcinoma, bronchioloalveolar carcinoma subtype, predict response to erlotinib," *J Clin Oncol,* vol. 26, no. 9, pp. 1472-8, Mar 20, 2008.

[271] C. H. Yang, C. J. Yu, J. Y. Shih, Y. C. Chang, F. C. Hu, M. C. Tsai, K. Y. Chen, Z. Z. Lin, C. J. Huang, C. T. Shun, C. L. Huang, J. Bean, A. L. Cheng, W. Pao, and P. C. Yang, "Specific EGFR mutations predict treatment outcome of stage IIIB/IV patients with chemotherapy-naive non-small-cell lung cancer receiving first-line gefitinib monotherapy," *J Clin Oncol,* vol. 26, no. 16, pp. 2745-53, Jun 1, 2008.

[272] K. Kianmehr, and R. Alhajj, "CARSVM: a class association rule-based classification framework and its application to gene expression data," *Artif Intell Med,* vol. 44, no. 1, pp. 7-25, Sep, 2008.

[273] G. M. Weiss, and F. J. Provost, "Learning when training data are costly: the effect of class distribution on tree induction," *J. Artif. Intell. Res.(JAIR),* vol. 19, pp. 315-354, 2003.

[274] W. Land, D. Margolis, R. Gottlieb, E. Krupinski, and J. Yang, "Improving CT prediction of treatment response in patients with metastatic colorectal carcinoma using statistical learning theory," *BMC genomics,* vol. 11, no. Suppl 3, pp. S15, 2010.

[275] S. O. Kim, J. Y. Jeong, M. R. Kim, H. J. Cho, J. Y. Ju, Y. S. Kwon, I. J. Oh, K. S. Kim, Y. I. Kim, S. C. Lim, and Y. C. Kim, "Efficacy of gemcitabine in patients with non-small cell lung cancer according to promoter polymorphisms of the ribonucleotide reductase M1 gene," *Clin Cancer Res,* vol. 14, no. 10, pp. 3083-8, May 15, 2008.

[276] G. Sartori, A. Cavazza, A. Sgambato, A. Marchioni, F. Barbieri, L. Longo, M. Bavieri, B. Murer, E. Meschiari, S. Tamberi, A. Cadioli, F. Luppi, M. Migaldi, and G. Rossi, "EGFR and K-ras mutations along the spectrum of pulmonary epithelial tumors of the lung and elaboration of a combined clinicopathologic and molecular scoring system to predict clinical responsiveness to EGFR inhibitors," *Am J Clin Pathol,* vol. 131, no. 4, pp. 478-89, Apr, 2009.

[277] T. Shukuya, T. Takahashi, R. Kaira, A. Ono, Y. Nakamura, A. Tsuya, H. Kenmotsu, T. Naito, K. Kaira, H. Murakami, M. Endo, K. Takahashi, and N. Yamamoto, "Efficacy of gefitinib for non-adenocarcinoma non-small-cell lung cancer patients harboring epidermal growth factor receptor mutations: a pooled analysis of published reports," *Cancer Sci,* vol. 102, no. 5, pp. 1032-7, May, 2011.

[278] H. Yamamoto, H. Shigematsu, M. Nomura, W. W. Lockwood, M. Sato, N. Okumura, J. Soh, M. Suzuki, Wistuba, II, K. M. Fong, H. Lee, S. Toyooka, H. Date, W. L. Lam, J. D. Minna, and A. F. Gazdar, "PIK3CA mutations and copy number gains in human lung cancers," *Cancer Res,* vol. 68, no. 17, pp. 6913-21, Sep 1, 2008.

[279] F. Cappuzzo, T. Ciuleanu, L. Stelmakh, S. Cicenas, A. Szczésna, E. Juhász, E. Esteban, O. Molinier, W. Brugger, and I. Melezínek, "Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: a multicentre, randomised, placebo-controlled phase 3 study," *The lancet oncology,* vol. 11, no. 6, pp. 521-529, 2010.

[280] J. Frankovich, C. A. Longhurst, and S. M. Sutherland, "Evidence-based medicine in the EMR era," *N Engl J Med,* vol. 365, no. 19, pp. 1758-9, Nov 10, 2011.

[281] G. Karakülah, A. Suner, K.-P. Adlassnig, and M. Samwald, "A data-driven living review for pharmacogenomic decision support in cancer treatment," *Studies in health technology and informatics,* vol. 180, pp. 688, 2012.

APPENDIX A DATA SOURCES

Ref no.	Title	Type
1	Lung Adenocarcinoma with Concurrent Exon 19 EGFR Mutation and ALK Rearrangement Responding to Erlotinib	Case Report
2	Complete Pathologic Response in Lung Tumors in Two Patients with Metastatic Non-small Cell Lung Cancer Treated with Erlotinib	Case Report
3	Characterization of epidermal growth factor receptor mutations in non-small-cell lung cancer patients of African-American ancestry	Article
4	Miliary Never-Smoking Adenocarcinoma of the Lung Strong Association with Epidermal Growth Factor Receptor Exon 19 Deletion	Case Report
5	Lung Cancer with Epidermal Growth Factor Receptor Exon 20 Mutations Is Associated with Poor Gefitinib Treatment	Article
6	The EGFR mutation and its correlation with response of gefitinib in previously treated Chinese patients with advanced non-small-cell lung cancer	Article
7	Multiplicity of EGFR and KRAS Mutations in Non-small Cell Lung Cancer (NSCLC) Patients Treated with Tyrosine Kinase Inhibitors	Article
8	Good Response to Gefitinib in Lung Adenocarcinoma of Complex Epidermal Growth Factor Receptor (EGFR) Mutations with the Classical Mutation Pattern	Article
9	Benchmarking of Mutation Diagnostics in Clinical Lung Cancer Specimens	Article
10	Near Total Regression of Diffuse Brain Metastases in Adenocarcinoma of the Lung with an EGFR Exon 19 Mutation: A Case Report and Review of the Literature	Case Report
11	Role of cMET expression in non-small-cell lung cancer patients treated with EGFR tyrosine kinase inhibitors	Article
12	De Novo Resistance to Epidermal Growth Factor Receptor-Tyrosine Kinase Inhibitors in EGFR Mutation-PositivePatients with Non-small Cell Lung Cancer	Case Report
13	Favorable Response to Erlotinib in a Lung Adenocarcinoma With Both Epidermal Growth Factor Receptor Exon 19 Deletion and K-ras G13D Mutations	Case Report
14	EGFR Mutations Detected in Plasma Are Associated with Patient Outcomes in Erlotinib Plus Docetaxel-Treated Non-small Cell Lung Cancer	Article
15	Large-Cell Neuroendocrine Carcinoma of the Lung Harboring EGFR Mutation and Responding to Gefitinib	Case Report
16	Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib	Article

17	EGF receptor gene mutations are common in lung cancers from ‘‘never smokers’’ and are associated with sensitivity of tumors to gefitinib and erlotinib	Article
18	EGFR mutations in patients with brain metastases from lung cancer: Association with the efficacy of gefitinib	Article
19	Response to second-line erlotinib in an EGFR mutation-negative patient with non-small-cell lung cancer: make no assumptions	Case Report
20	Erlotinib May be More Effective for Central Nervous Metastasis of Lung Adenocarcinoma Than Gefi tinib Because of the Difference in the Clinical Regimes	Case Report
21	Impact of specific mutant KRAS on clinical outcome of EGFR-TKI-treated advanced non-small cell lung cancer patients with an EGFR wild type genotype	Article
22	EGFR Mutation and Resistance of Non–Small-Cell Lung Cancer to Gefitinib	Case Report
23	Activity of Epidermal Growth Factor Receptor-Tyrosine Kinase Inhibitors in Patients with Non-small Cell Lung Cancer Harboring Rare Epidermal Growth Factor Receptor Mutations	Article
24	Mutations in the Epidermal Growth Factor Receptor and in KRAS Are Predictive and Prognostic Indicators in Patients With Non–Small-Cell Lung Cancer Treated With Chemotherapy Alone and in Combination With Erlotinib	Article
25	Detectionof Epidermal Growth Factor ReceptorMutations in Serum as a Predictor of the Response to Gefitinib in Patients with Non ^Small-Cell LungC ancer	Article
26	Gefitinib-SensitiveMutations of the Epidermal Growth Factor Receptor Tyrosine Kinase Domain in Chinese Patientswith Non^Small Cell Lung Cancer	Article
27	Epidermal growth factor receptor mutations are associated with gefitinib sensitivity in non-small cell lung cancer in Japanese	Article
28	EGFR/KRAS Mutations and Gefitinib Therapy in Chinese NSCLC Patients	Article
29	EGFR and KRAS mutations as criteria for treatment with tyrosine kinase inhibitors: retro- and prospective observations in non-small-cell lung cancer	Article
30	Epidermal growth factor receptor (EGFR) mutations in a series of non-small-cell lung cancer (NSCLC) patients and response rate to EGFR-specifi c tyrosine kinase inhibitors (TKIs)	Article
31	Epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (TKIs) are effective for leptomeningeal metastasis from non-small cell lung cancer patients with sensitive EGFR mutation or other predictive factors of good response for EGFR TKI	Article

32	Response to Erlotinib in First-Line Treatment of Non–Small-Cell Lung Cancer in a White Male Smoker with Squamous-Cell Histology	Case Report
33	Erlotinib for Pretreated Squamous Cell Carcinoma of the Lung in Japanese Patients	Case Report
34	Brain metastasis from non-small cell lung cancer: sustained response with erlotinib	Case Report
35	Early and Complete Response of Bone Metastases, Documented by FDG-PET/CT Scan, in a Patient With NSCLC	Case Report
36	High Frequency of Epidermal Growth Factor Receptor Mutations with Complex Patterns in Non–Small Cell Lung Cancers Related to Gefitinib Responsiveness in Taiwan	Article
37	ActivatingMutations in the Tyrosine Kinase Domain of the Epidermal Growth Factor ReceptorAreAssociatedwith Improved Survival in Gefitinib-Treated Chemorefractory Lung Adenocarcinomas	Article
38	'Classical' but not 'other' mutations of EGFR kinase domain are associated with clinical outcome in gefitinib-treated patients with non-small cell lung cancer	Article
39	A phase II trial of gefitinib as first-line therapy for advanced non-small cell lung cancer with epidermal growth factor receptor mutations	Article
40	Phase II Clinical Trial of Chemotherapy-Naı¨ve Patients 70 Years of Age Treated With Erlotinib for Advanced Non–Small-Cell Lung Cancer	Article
41	EGFR Mutation Status in Primary Lung Adenocarcinomas and Corresponding Metastatic Lesions: Discordance in Pleural Metastases	Article
42	Mutations of epidermal growth factor receptor of non-small cell lung cancer were associated with sensitivity to gefitinib in recurrence after surgery	Article
43	Use of Cetuximab After Failure of Gefitinib in Patients With Advanced Non–Small-Cell Lung Cancer	Article
44	Complex Mutations in the Epidermal Growth Factor Receptor Gene in Non-small Cell Lung Cancer	Article
45	Effects of Erlotinib in EGFR Mutated Non-Small Cell Lung Cancers with Resistance to Gefitinib	Article
46	Erlotinib after Gefitinib failure in female never-smoker Asian patients with pulmonary adenocarcinoma	Article
47	Mutation in the Tyrosine Kinase Domain of Epidermal Growth Factor Receptor Is a Predictive and Prognostic Factor for Gefitinib Treatment in Patients with Non-Small Cell Lung Cancer	Article
48	The Relationship between Epidermal Growth Factor Receptor Mutations and Clinicopathologic Features in Non–Small Cell Lung Cancers	Article

APPENDIX B COPYRIGHT PERMISSION LETTER

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS
Jul 19, 2013

This is a License Agreement between Nelofar Kureshi ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3167810673476
License date	Jun 14, 2013
Licensed content publisher	Oxford University Press
Licensed content publication	Annals of Oncology
Licensed content title	The EGFR mutation and its correlation with response of gefitinib in previously treated Chinese patients with advanced non-small-cell lung cancer
Licensed content author	X.-T. Zhang, L.-Y. Li, X.-L. Mu, Q.-C. Cui, X.-Y. Chang, W. Song, S.-L. Wang, M.-Z. Wang, W. Zhong, L. Zhang
Licensed content date	August 2005
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Personalized Medicine: Development of a Predictive Computational Model for Personalized Therapeutic Interventions
Publisher of your work	n/a
Expected publication date	Aug 2013
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oxfordjournals.org

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never

granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4