A COMPUTATIONAL METHOD FOR THE ANALYSIS OF PAIN PATTERNS AND PROGRESSION OF PANCREATITIS WITH A LARGE NUMBER OF PREDICTOR VARIABLES

by

Ye Tian

BS, Huazhong University of Science and Techonology, China, 2008

MPH, University of Pittsburgh, 2009

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Ye Tian

It was defended on

March 26, 2014

and approved by

Dissertation Advisor:
Stephen Wisniewski, PhD, Professor, Department of Epidemiology
Associate Dean, Graduate School of Public Health, University of Pittsburgh

Committee Member:
Marianne Bertolet, PhD, Assistant Professor, Department of Epidemiology
Graduate School of Public Health, University of Pittsburgh

Committee Member:
Joseph Zmuda, PhD, Associate Professor, Department of Epidemiology
Graduate School of Public Health, University of Pittsburgh

Committee Member:
David Whitcomb, MD PhD, Professor
Chief, Division of Gastroenterology, Hepatology and Nutrition
School of Medicine, University of Pittsburgh

Stephen R. Wisniewski, PhD

A COMPUTATIONAL METHOD FOR THE ANALYSIS OF PAIN PATTERNS AND PROGRESSION OF PANCREATITIS WITH A LARGE NUMBER OF PREDICTOR VARIABLES

Ye Tian, PhD

University of Pittsburgh, 2014

ABSTRACT

Chronic pancreatitis (CP) is a major burden of gastrointestinal disease in the United States accounts for significant healthcare costs to the society. Abdominal pain is the most common symptom in CP patients and the development of CP is challenging medical practice. It has been proposed that a combination of genetic, environmental, and metabolic risk factors contribute to the pain patterns in CP patients and development of CP. This research aimed to introduce a new data analytic strategy Random Forests (RF) to support big data analysis in studying CP and epidemiological researches.

RF has been becoming a popular non-parametric algorithm in computational method and used in many scientific areas in the context of big data era. RF is an ensemble of individual decision trees to help explore data structure and hidden information in high dimensional data. RF could deal with correlated predictor variables and integrates complex interaction effects during modeling process to evaluate the entire effects of all predictor variables on outcome variable and produce estimates of importance scores for all predictor variables.

In this work, a framework of combining RF analyses with traditional statistical analyses was developed to investigate important risk factors associated with different pain patterns in patients with CP and disease progression from recurrent acute pancreatitis (RAP) to CP. The public health significance of this novel analytic method is that it successfully examined a large amount of predictor variables in a multivariable way and would help researchers to better understand complex mechanisms in CP.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 PANCREAS

The pancreas is an organ about six inches long and is located deep in the abdomen, behind the stomach. It is made up of glandular tissue and a system of ducts. The main duct is the pancreatic duct which runs the length of the pancreas and has many small side branches. It drains the pancreatic fluid from the gland and carries it to the duodenum, the first part of the small intestine. The pancreatic duct is merged with the bile duct in the head of the pancreas to form a widening of the duct just before it empties into the duodenum.[1]

The pancreas is both an endocrine and exocrine gland that has a critical role in energy balance and digestion.[2] The endocrine cells (islets of Langerhans) of the pancreas produce and secrete hormones (e.g. insulin and glucagon) into the bloodstream to maintain the proper level of sugar in the blood used for energy. The exocrine cells (acinar cells) of the pancreas produce and transport inactive digestive enzymes that go through pancreatic duct. These cells are secreted in the small intestine where they are activated to assist in the digestion of food. However, when pancreatic enzymes (especially trypsin) are activated in the pancreas instead of the small intestine, it will cause pancreatic injury, inflammation of the pancreas, and immune responses that are recognized as pancreatitis.[2]

1.2 PANCREATITIS

Pancreatitis is a major contributor to the burden of gastrointestinal disease in the United States.[3-5] In 2004, there were 475,000 hospital visits and 277,000 hospitalizations due to pancreatitis as the primary diagnosis in the United States. The number of hospitalization for pancreatitis increased 62% from 1988 to 2004, with the $3.7 billion total estimated health care cost in 2004.[4, 5] Defined as an inflammation of the pancreas, it can be acute or chronic, requiring medical treatment ranging from minor outpatient management to intensive care for organ failures.[6, 7] The most common symptom in both acute and chronic pancreatitis is the upper abdominal pain radiating to the back with other symptoms, such as nausea, vomiting, and abdominal distension, depending on severity and etiology of the disease.[8, 9]

1.2.1 Acute Pancreatitis

Acute pancreatitis (AP) is a sudden inflammation of the pancreas and commonly result in hospital admissions.[8] In the United States, the annual incidence of AP is 42/100,000 population.[10] In 2009, AP accounts for about 274,119 hospitalizations in the United States, which is the most common single gastrointestinal disease of hospital admissions.[11] The mortality rate is less than 5% for those younger than 40 years old but increased to 28% or higher in those older than 60 years old.[12] Most AP is associated with minor organ dysfunction, and commonly resolves within seven days with supportive medical treatment including pain medications, fasting, and intravenous nutritional support. [13] Alcohol abuse and gallstones are the most common etiological factors for AP, accounting for over 80% of cases in Western countries.[8] AP could be confirmed by a battery of diagnostic tools such as physical examinations, laboratory tests, and imaging studies, to determine severity and presence of bleeding in or around the pancreas.[8, 13] While most of AP patients are often successfully relieved by supportive medical treatment, severe cases of AP may require admission to the intensive care unit, endoscopic therapy, or surgery to deal with complications of the disease process because of the development of a systemic inflammatory response (SIRS) and multiple organ failure (MOF).[13-15], which result in majority of death in patients with AP.[16]

1.2.2 Recurrent Acute Pancreatitis

For majority of patients who have a first episode of AP, physicians can usually determine and treat the underlying cause of disease so that AP never returns. However, any established risk factor that is associated with a patient having the first episode of AP has the potential to initiate the subsequent attacks within a few years.[17, 18] Recurrent acute pancreatitis (RAP) is defined as patients having two or more attacks of AP without morphological changes to the pancreas detected by imaging studies including CT scan, endoscopic retrograde cholangiopancreatography (ERCP), magnetic resonance cholangiopancreatography (MRCP), or endoscopic ultrasound (EUS), etc.[19] The median time of the first readmission for AP is 7.2 month and the proportion of patients developing RAP after the first event of AP ranges from 4.2% to 14.4% and the mortality rate of RAP is lower than AP (<1% -3.2%).[3, 12] Even though the mortality rate of RAP is lower than the first episode of AP, RAP may be associated with impairment in quality of life.[17]

Similar to AP, 70% to 80% of cases of RAP are associated with either alcohol abuse or gallstone. Other risk factors include hypertriglyceridemia, smoking, pain medications for first attack of AP, etc.[17, 18] Because of the most causes of AP can lead to recurrent attacks if the underlying risk factor remains uncorrected, patients having recurrent attacks of AP need a more extensive examination to determine the underlying cause.[2] Patients with RAP should be treated with the same supportive treatment such as pain medications, fasting, and intravenous nutritional support as those with AP. The need for any other specific treatment, such as endoscopic therapy or surgery depends on the underlying causes (e.g. gallstone pancreatitis) for prevention of more attacks.[17]

1.2.3 Chronic Pancreatitis

Chronic pancreatitis (CP) is defined by pathological evidence of progressive pancreatic damage with inflammation, fibrosis, anatomic features (e.g. calcification) and loss of function.[2] There are few epidemiological data in the world describing the incidence, hospitalization rate, and prevalence of CP probably due to the non-consensus of diagnosis criteria of CP.[20] Yadav et al reviewed few data and showed that the annual incidence of CP ranges from 5 to 12/ 100,000 population, and the prevalence of CP is about 50/100,000 popualtion.[11] Although the incidence of CP is significantly lower than that of AP, CP is significantly associated with impairment of quality of life since patients diagnosed with CP usually have repeated intermittent or continuous abdominal pain, maldigestion, steatorrhea, hospitalization, or diabetes mellitus.[21] In addition, patients with CP has been shown to have an increased risk of pancreatic cancer, which typically has an extremely poor prognosis.[21-23]

The etiology and mechanisms of CP is still under investigation and only partially known.[2] In 20th century, alcohol abuse has long been characterized as the primary cause of CP. From last decade, however, recent multicenter studies showed that only about 34% to 44% of CP cases having alcohol abuse as the primary single cause considered by physicians.[24, 25]

Other important risk factors include hyperlipidemia/hypertriglyceridemia, pancreas divisum, sphincter of oddi, pancreatic duct obstruction, early or late idiopathic etiologies, and other toxic causes, such as smoking, medication (e.g. valproate, phenacitin, thiazide, oestrogen, and azathioprine) etc.[22] In addition, a number of genetic susceptibility factors have been identified with CP.[2, 21, 22, 26-30] Recent advances in genetics, such as genome-wide association studies (GWAS) and next generation sequencing (NGS) technology provide new possibilities to accurately identify risk factors leading to CP.[26, 28, 29, 31] The Midwest Multicenter Pancreatic Study Group drafted a "TIGAR-O" classification system to categorize major predisposing risk factors to CP including toxic-metabolic, idiopathic, genetic, autoimmune, recurrent and severe AP associated, and obstructive risk factors (Figure 1).[2, 21] Through recent studies, an important conceptual change in understanding the development of CP is that there is not a single etiology for CP, but rather it is a complex, multi-factorial disease with different pathways and interactions. Moreover, it is possible that other yet-to-be-identified genetic and environmental risk factors may play in the complex mechanisms.[2, 22, 24, 26, 28-30]

There are no definite criteria and limited consensuses on the diagnosis of CP.[21] As in other diseases, tissue diagnosis should be the gold standard to diagnose CP to identify chronic inflammation and irregularly placed fibrosis in the pancreas. However, using pancreatic biopsy or resected specimen of pancreas from patients is impractical because of the likelihood of triggering AP or other complications.[2, 21, 22] Currently, a combination of clinical features including imaging studies (e.g. CT, MRI, MRCP, ERCP, EUS) and functional tests (e.g. exocrine pancreatic function test, liver function test) is widely used for diagnosis of CP. The selection of the appropriate diagnostic tests depends on resources and individual circumstances. For example, use a single imaging study to diagnose long-standing severe CP with extensive calcifications and ductal dilation in the pancreas is simple whereas the detection of early CP with minimal morphological changes by any single imaging study or function test is sometimes difficult.[21-23]

The goals of treatment for CP are to relieve acute or chronic pain, slow down the disease process to prevent future painful attacks, correct metabolic consequences such as diabetes or malnutrition, use endoscopic therapy or surgery to manage complications such as bile duct stricture, pseudocyst, portal hypertension, etc.[22]

1.3 PAIN IN CHRONIC PANCREATITIS

Abdominal pain in CP is the most common symptom and the largest clinical challenge for both the patients and clinicians. In a recent large multicenter cohort study with 540 CP patients from the North American Pancreatitis Study-2 (NAPS2), pain is present in 77% of patients with different patient self-identified pain patterns.[32] Some other studies showed that the prevalence of pain is up to 90% and the pain is the primary cause of hospital admissions in CP patients.[5] Adversely effects, poor quality of life and increased medical costs are associated with managing pain in CP patients.[5] The estimated annual cost to deal with pain in CP patients in the United States is over $638 million.[4, 32]

1.3.1 Mechanisms of Pain

The mechanisms of pain in CP patients is still under investigation and is very complex. Many theories have been proposed over the years and the current consensus is that the pain is multi-factorial and heterogeneous among CP patients. [5, 9, 32-36] The multi-factorial causes of pain in CP may also explain why the patterns of pain are highly variable and the poor performance in current pain management in CP patients.[33] Therefore, understanding the mechanisms of pain in CP and risk factors that trigger different types of pain in patients are very important in studying pain in chronic pancreatitis.

There are a variety of hypotheses for the cuase of pain in patients with CP. Originally, pain was hypothesized to be related to pancreatic duct hypertension, pancreatic tissue pressure and/or neural alterations focused on inflammation and morphologic abnormalities.[5, 9, 33, 35, 36] Recent researches indicate genetics may play an important role in different phenotypes of pain, tolerance, and pain therapy effects. The indication may help physicians to understand the question of CP patients who have similar amounts of pancreatic injury but have variant expression of pain. However, the evidence of important candidate gene in CP pain is still insufficient.[9]

1.3.2 Pain Patterns

The characteristics of abdominal pain in CP patients is highly variable among patients. Commonly, based on temporality and severity, the pain may be demonstrated as mild or severe, intermittent or continuous. The data from NAPS2 study showed that in patients with CP, constant pain produces higher rates of disability, hospitalization, and negative impact on quality of life compared with intermittent pain, even if the intermittent pain is much more severe in intensity.[32] On the other hand, the pain patterns in CP patients may be altered, stabilized, worsened, lightened, or disappeared over time, but the progression of pain is not able to predict in an individual patient and a specific treatment for pain may be effective, inadequate, or unpredictable.[32, 34] Therefore, these complexities of pain patterns in CP patients may be due multiple etiologies.[32, 33, 36]

1.3.3 Pain Management

Because the cause of abdominal pain in CP is multi-factorial, and the underlying mechanisms are still not completely known, no effective medical therapy or intervention provides reliable relief of pain in all patients. Neuropathic pain medications are the most used to conservatively control pain in CP patients. The drugs include paracetamol, dextropropoxyphene, prednisolone, non-steroidal anti-inflammatory drugs, tricyclic antidepressants or narcotic analgesics, etc.[35, 37] However, a problem with the use of analgesic drugs is that patients often become addicted on heavy dosage.[9, 33, 35]

Pancreatic enzyme may be used as an alternative for pain control in CP patients because of its ability to diminish stimulation of the exocrine pancreas, thereby reduce pain. However, the results from a meta-analysis showed that there was no significant benefit of pancreatic enzyme therapy to relieve pain associated with CP.[38] In addition, octreotide and antioxidant therapy have also been evaluated from a few of studies but showed no statistically significant benefits for pain control in patients with CP.[9] Therefore, further studies are needed before they can become widely used.

Endoscopic therapy is focused on relieving pain in CP patients due to pancreatic duct obstruction from strictures or gallstones and surgery are primarily used in those with a dilated pancreatic duct. Surgical therapy, such as ductal drainage or resection is used in CP patients who require long time use of pain medications, or are unable to maintain normal daily life and quality of life because of chronic or recurrent pain.[9]

1.4 GENETICS IN PANCREATITIS

In the last two decades, genetics is becoming an important area in exploring pancreatic diseases. A number of pancreatic-targeting genetic factors have been identified in influencing trypsinogen activation and are associated with changes in susceptibility to pancreatic diseases, including cationic trypsinogen (PRSS1), anionic trypsinogen (PRSS2), serine protease inhibitor Kazal 1 (SPINK1), cystic fibrosis transmembrane conductance regulator (CFTR), chymotrypsinogen C (CTRC), calcium-sensing receptor (CASR), and claudin 2 (CLDN2).[2, 21, 26, 28, 31, 39, 40] These genetic factors are associated with premature activation of trypsinogen to trypsin or the failure to eliminate active trypsin within the pancreas that cause pancreatitis.[28] In addition, it is now widely recognized that pancreatitis is a group of complex inflammatory syndromes that appear to involve different combinations of genetic, environmental, and metabolic factors, etc.[2, 26, 39, 41]

A breakthrough in genetics to understand RAP and CP was discovered in 1996. Mutations in the gene that encodes PRSS1 was linked to hereditary pancreatitis, a syndrome characterized by RAP and later CP.[42] Later, mutations in gene that encode SPINK1 and CFTR were identified that related to pancreatitis. Mutations in CTRC and CASR were also reported with small risk with pancreatitis. These genes have different roles in the disease pathways and the different mutations within each gene have different functional effects to determine each gene’s relative contribution to the various forms of CP.[26, 28, 30, 31] However, all these susceptible genetic factors have been found using simple hypothesis-driven candidate-gene approaches on the trypsin-activity model which may be incomplete compared to most recent genomewide association study (GWAS), which provides the opportunity to perform an complete search for genetic risk factors and discover new genetic variants in thousands of genes,[26] The recent NAPS2 project was successfully organized and aimed to perform GWAS to provide new insights into the etiology and pathogenesis of RAP and CP.[24] A recent breakthrough from NAPS2 was reported that the mutations in PRSS1-PRSS1 and CLDN2 was related to pancreatitis which could help to understand the complex genetic variants on effects of RAP and CP.[40]

1.5 ANALYSIS OF GENETIC DATA

1.5.1 Genetic Epidemiology

Genetic epidemiology is an interdisciplinary field to explore both genetic and environmental factors and their interactions in determining the distribution of traits and diseases in human populations. Previous researches have utilized a simple gene environment model incorporating a particular gene variant and an environmental exposure as the potential cause of a chronic disease in a regression model to identify association with a gene and disease. In these models, the gene environment interaction is defined as their co-participation in mechanism of disease development.

Many chronic diseases are increasingly considered as complex, multi-factorial diseases, meaning they are likely associated with the interactive effects of multiple genetic factors in combination with numerous environmental factors. Although various chronic diseases, such as cardiovascular disease, diabetes and cancers, etc., have been identified that they are substantially clustered in families indicating important effect by genetic factors, they do not have a definite pattern of inheritance in a family. Thus, it is difficult to predict a person's risk of developing these diseases simply based on genetic aspects. These complex diseases are difficult to treat because some underlying risk factors that cause these diseases have not yet been identified.

Currently, the field of human genetics research has advanced to include techniques to generate high-throughput data.[43] GWAS approaches, which are used to help discover thousands of each single-nucleotide polymorphisms (SNP) within genes that are associated with hundreds of common, complex human traits, have been proven to be successful in the identification of association between new genetic risk factors and the risk of a wide range of complex diseases.[44-46] The most common approach of GWAS is the case-control study which compares two large groups of individuals, one case group affected by a disease and one healthy control group. All individuals in each group are genotyped for the majority of commonly known SNPs. The exact number of SNPs depends on the genotyping technology, but are typically one million or more.

1.5.2 Challenges in Analyzing Genetic Data with Large Number of Variables

Most GWAS focus on the detection of main effects by using an allele or genotype based test separately. However, because genetic factors in combination with environmental factors or other genetic factors are expected to contribute to susceptibility to disease, multiple SNPs and interaction effects should be analyzed simultaneously.[46, 47] As a result, there are several challenges in studying the comprehensive effects of multiple genetic and environmental variables. First, in typical GWAS, investigation performed in several thousand subjects with genotypes of millions of SNPs, leading to high dimensional data (large p, small n problem, many more predictor variables (e.g. SNPs) than study subjects).[43, 48, 49] Second, the large volume of risk factors leads to exponentially increase in the number of interactions with different levels that requiring assessment of multiple complex comparisons.[46, 50-53] Third, there is a large number of correlated predictor variables (SNPs) genotyped in GWAS needs to be taken into account.[50, 54]

Because of these challenges, standard regression models (e.g. linear regression models, Cox proportional hazard regression models, and mixed effects regression models) that have been commonly used in assessing the association between the development of a disease and risk factors are not sufficient. For example, because of the large number of variables, the variable selection and interaction effects for fitting a model is hard to determine. In addition, model interpretation can be problematic in the presence of higher-order interactions among hundreds or thousands of influential predictor variables.

1.5.3 Classification and Regression Tree used in Analyzing Large Number of Variables

Decision tree methods, also called recursive partitioning methods, are a group of computational analysis of machine learning algorithms using non-parametric tree structure techniques to handle high dimensional data with complex interactions. These methods provide a useful alternative to parametric regression methods without the need to specify hypothesis and a particular model.[44, 49]

The most common decision tree algorithm is the classification and regression tree (CART) introduced by Breiman in 1984,[55] while other methods, such as ID 3 and C4.5 both developed by Quinlan in 1986 and 1992 are also used.[56, 57] Decision trees are broadly classified into classification trees that used for categorical outcomes and regression trees that used for continuous outcomes. Decision tree methods are attractive in genetic epidemiology because they allow non-parametric analyses of large numbers of genetic variables in small sets of data (large p, small n problem) and can help to identify interaction effects even though genetic variable only have small marginal effects on the health status.[44, 49]

In the computational learning theory, a training dataset is a set of data consisting one or more predictor variables and outcome variable to construct a model for discovering potentially predictive relationships and/or learning complex structure in the data. A model learned by training datatset is based on empirical relationships tend to overfit the data, meaning that they can identify relationships in the training dataset but do not guarantee the future use in general. A testing dataset is a set of data that is independent of the training data but follows the same probability distribution as the training data. Using the testing dataset could evaluate the overfitting in a model trained by the training dataset.

Using the CART algorithm, the whole sample is divided into two datasets, a training dataset and a testing dataset. Using the training dataset, a decision tree begins to build by first splitting a root node containing all sample to one of two child nodes, one left and one right as illustrated in Figure 2.[58] This split process is based on a selected splitting rule, such as decrease of Gini Impurity for categorical outcome or minimizing residual sum of squares for continuous outcome, by testing every predictor variables and find the best one to split the dataset into two subsets of the data that are the most different with respect to the outcome of interest. To build a complete tree, this process is recursively repeated to split the child nodes until some specified stopping criterion (e.g., minimum number of observations in a node for split) is met, or to child nodes are completely homogeneous with respect to the outcome of interest so that there is no further possible split. After that, a tree is usually “pruned” back by sequentially removing the weakest split in the tree to prevent overfitting, because building a large and complex tree that closely fits the training dataset tends to have poor fit on a new dataset. This procedure is called "pruning" and helps us to find the optimal tree by evaluating model fit in testing dataset and improve its generalizability. After the recursive partition is completed, a predictive value of the outcome variable is calculated by weighting the class frequencies in terminal nodes relative to the class frequencies in the root node (for classification trees) or the mean of the outcomes (for regression trees) from the terminal node. Then the predicted outcome of each observation in testing dataset is estimated by sending the predictor values down the tree and taking the predictive value from the terminal node into which the observation falls. This process is applied for all subjects in testing dataset and then could provide the prediction error by this tree: misclassification error for classification trees or mean squared error for regression trees.[55, 59, 60]

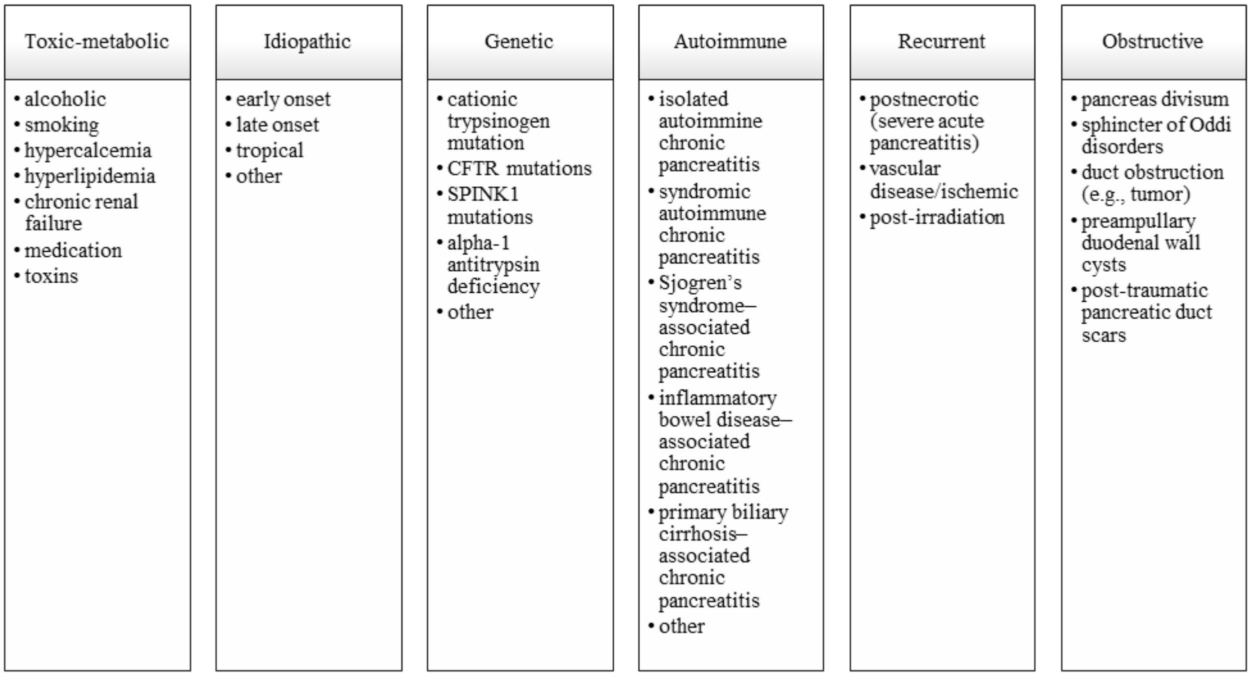
CART algorithm has several advantages. It is not based on a particular model and easy to interpret. The computation runs fast and could incorporate different types of variables simultaneously. However, decision trees tend to have a high instability in node split due to small changes in data. Even though it has ways to optimize tree for generalizability, it cannot guarantee avoiding overfitting and thus do not always have good performance on future dataset. [49, 52]

One approach to overcome these limitations from a single decision tree is to use an ensemble of individual trees. In 2001, Breiman extended his CART algorithm to forest based approaches, named as Random Forests (RF),[61] which have become a popular non-parametric algorithm in computational method and used in many scientific areas. A RF consists of hundreds or thousands of trees which built from a random sub-sample of the original dataset. The predicted outcome for each observation is estimated by selecting the most frequently predicted category (for classification) or average of predictive value tree (for regression) from each individual tree. Based on some previously published studies, researchers found results from RF showed better performance compared to single CART.[52, 59, 61]

1.6 AIMS

The aims of this dissertation include: 1) Introduce the principles of RF, review the applications, recent development of RF, describe its advantages and limitations, discuss and provide epidemiological examples. 2) Use RF to identify and rank important risk factors associated different temporal pain patterns in patients who have chronic pancreatitis. 3) Apply RF to identify risk factors associated with the risk of having disease progression from recurrent acute pancreatitis to chronic pancreatitis.

1.7 FIGURES



Toxic-metabolic	Idiopathic	Genetic	Autoimmune	Recurrent	Obstructive
• alcoholic • smoking • hypercalcemia • hyperlipidemia • chronic renal failure • medication • toxins	• early onset • late onset • tropical • other	• cationic trypsinogen mutation • CFTR mutations • SPINK1 mutations • alpha-1 antitrypsin deficiency • other	• isolated autoimmine chronic pancreatitis • syndromic autoimmune chronic pancreatitis • Sjogren's syndrome–associated chronic pancreatitis • inflammatory bowel disease–associated chronic pancreatitis • primary biliary cirrhosis–associated chronic pancreatitis • other	• postnecrotic (severe acute pancreatitis) • vascular disease/ischemic • post-irradiation	• pancreas divisum • sphincter of Oddi disorders • duct obstruction (e.g., tumor) • preampullary duodenal wall cysts • post-traumatic pancreatic duct scars

Figure 1. Risk factors of CP: TIGAR-O classification system

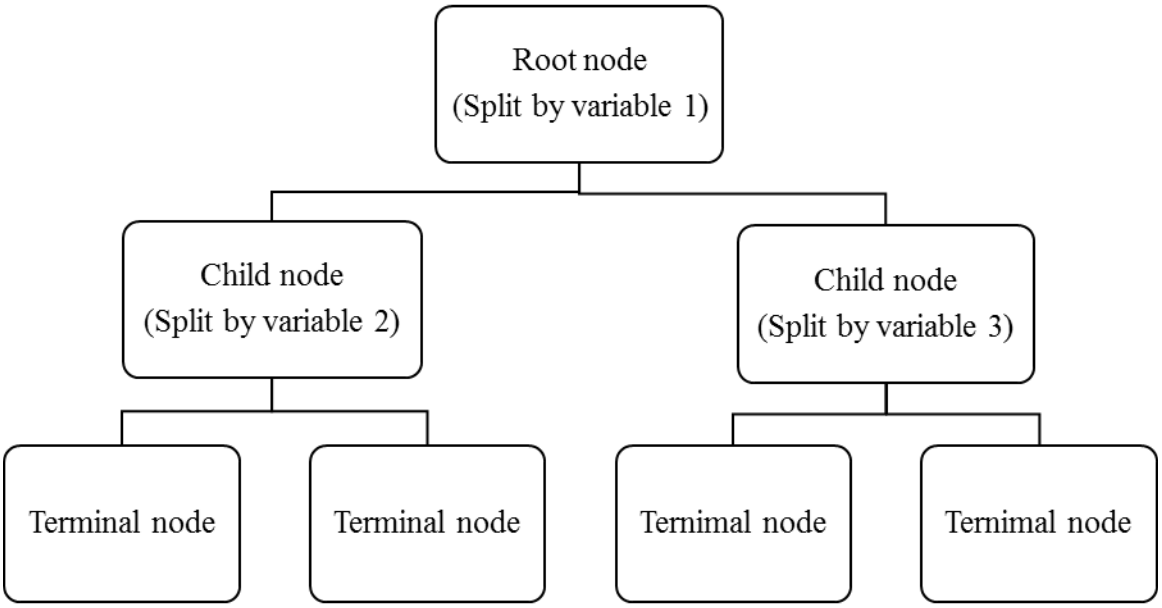


Figure 2. Basic Structure of CART

2.0 RANDOM FORESTS FOR EPIDEMIOLOGY IN THE ERA OF BIG DATA

2.1 ABSTRACT

The Random Forests (RF) algorithm has become a very useful computational analytic method in a variety of research areas for both classification and regression problems. RF is an ensemble decision trees featuring by a two-way randomness through random bootstrap sampling and random variable selection in tree-building process. RF is a powerful analytic framework to analyze large-scale dataset, e.g. variable selection, variable importance estimation, and outcome prediction, etc. It has been applied in a number of epidemiological researches especially in genetic epidemiology due to its capability to explore data structure and hidden information in high dimensional data. This paper is aimed to review RF methodology, provide examples of applications to demonstrate how RF works for epidemiological studies in the context of future researches, and discuss its advantages and limitations.

2.2 BACKGROUND

In epidemiology, researchers seek to understand the association between risk factors and health outcomes, assess the importance of risk factors, find the pattern of disease or health status in population, and sometimes build models to predict or classify different outcomes of interest (e.g.

prognosis, death, treatment effect, etc.). There are a variety of types of studies could be used in epidemiological researches depending on study questions that helps to achieve research objectives such as cohort studies, case-control studies, and cross-sectional studies, etc. In these studies, descriptive measures of frequency, some relative measures as well as absolute measures could be used to quantify the association between specific risk factor and health outcome. When investigating several candidate risk factors, estimating interaction effects and/or adjusting confounding factors, regression modeling is the most common method for the estimation.[62] Traditionally, statistical methods including linear regression for continuous outcome, logistic regression for binary outcome, Cox proportional hazard regression for survival models, and mixed effects regression or generalized estimating equations for repeated outcomes design (longitudinal study), have been well developed and are frequently used for data analysis in epidemiological studies.

Advances in technology and computing, accelerated information accrual and exchange have led to the “big data era” and is contributing in the development of epidemiological research, such as genetic and clinical epidemiology, etc.[63] Today, data from epidemiological studies are much more complex and high dimensional with large number of predictor variable that sometimes result in violation of assumptions for traditional statistical models, such as distribution of parameters, correlation among predictors variables, and many missing values. The large number of predictor variables lead to more complex modeling including complex interactions among predictor variables that are difficult to evaluate and interpret from traditional statistical analysis.

Technological advances in genetic epidemiology, such as genomewide association studies (GWAS), DNA resequencing, or DNA microarray are among the most revolutionary developed in last two decades that challenging data analysis because they simultaneously analyze large amount of genetic markers.[43, 44, 46, 49, 52] As a result, researchers are facing significant challenges from these large-scale data in advanced genetic studies, where thousands of genes are considered as potential risk factors of a disease, makes traditional statistical methods no longer feasible. The highly correlated structure of genetic variables violates the assumption required by traditional models. In addition, many undetected mechanisms that involve gene–gene interactions and gene-environmental interactions are difficult to pre-specify in traditional models especially for higher order interactions. Only a small set of genetic markers are expected to be associated with a particular disease while performing variable selection for high dimensional, correlated, and interactive genetic variables are challenging for traditional statistical methods.

2.2.1 Classification and Regression Trees

Decision tree methods, also called recursive partitioning, are a group of machine learning algorithm and have become popular non-parametric analytic approaches for multivariable analysis. Recursive partitioning helps to identify important risk factors and their interactions in influencing disease and other health outcome in epidemiological areas, particularly widely used in genetic studies.[49, 59]

One of the most popular algorithm of recursive partitioning is Classification and Regression Trees (CART) introduced by Leo Breiman et al in 1984.[55] CART are commonly used as a predictive model to predict health outcomes, to identify important predictor variables associated with the outcome, and/or to visualize the way of predictor variables interact with each other, etc. The outcome variable could be either categorical or continuous, such as disease and no disease, severity of disease, dead and alive, etc.[58, 64, 65]

2.2.2 CART Advantages and Disadvantages

CART analysis could be used in a wide range of epidemiological studies because it is non-parametric computational method without statistical assumptions compared to traditional statistical models. CART does not require pre-specified variable selections and can incorporate many nuisance predictor variables. It can detect complex interactions among predictor variables and has specialized methodology to work with missing values and outliers. However, the largest disadvantage of individual decision trees is that they are unstable and overfit to the data. A minor change in the dataset, such as removing a small set of observations may result in dramatic change in decision tree structure, that is, increase or decrease of tree complexity, changes in splitting variables and cutting values in a split, and impair the tree performance.[49, 59, 66]

To overcome the limitations in a single decision tree, an ensemble of single trees, or use of forest, could improve the model performance and stability while maintaining advantages of individual trees.[59, 61]

2.3 RANDOM FOREST

The Random Forests (RF) algorithm is tree-based ensembles machine learning method for classification or regression by constructing a great quantity of CART-like decision trees featuring by a two-way randomness through random sub-sampling and random variable selection in tree-building process.[59, 61] RF was introduced by Leo Breiman in 2001 and is a popular non parametric analysis method in areas such as genetic[48, 51, 53, 67-70], biological[71, 72], clinical[73, 74], psychological researches[75]. RF is an extension of Breiman's previous CART algorithm and offers high prediction accuracy and variable importance estimation.[55, 61] It could handle high dimensional data with highly correlated predictor variables to identify key subsets of variables related to the outcome and integrate complex interaction effects. In this paper, we will review the RF methodology and discuss how RF works for epidemiological studies. We will review some examples of applications using RF in epidemiological studies and its advantages and limitations.

2.3.1 Random Forests Methodology

To build a forest of trees, the RF algorithm has a number of steps: First, for each individual tree, a different training dataset is created by randomly select two-third of total subjects from the original sample resulting in a particular training dataset to build each tree. This process is called bootstrapping and these training datasets are called bootstrap samples. After each bootstrapping, the remaining subjects from the original sample are called "out-of-bag" (OOB) samples for each individual tree that contains one-third of total subjects from the original sample. Each OOB sample is used as a testing dataset for that tree and also help to estimate variable importance. Using bootstrap sample, each tree is grown from the root node by evaluating ability to split a node using a random subset of all predictor variables. The split criteria are same as the criteria in CART algorithm (e.g. decrease of Gini Impurity, residuals sum of squares). After the bootstrap sample has been split at the top node, the splitting process is repeated and each tree is grown to its largest extent, for classification, the trees are grown until each terminal node contains members of only one class, while for regression they are grown until each terminal node contains same outcome value. These procedure is repeated many times to generate specified number of trees (e.g. 100, 500, 1000, 5000, etc.). After all trees have been fully grown, training process is finished and the forest is formed (Figure 3). [59, 61, 76, 77]

Each subject goes through each individual tree in the forest and get prediction result from each tree. Then the predictions of all trees are aggregated by majority voting to determine the class for classification problems or averaging each predictive value for regression problems.[43, 59, 61, 68, 76, 77] The OOB samples also go through the individual trees they belong to and aggregate prediction results by majority voting or averaging. Then the RF prediction error could be assessed by OOB sample. The key to the accuracy of RF predictions is low bias and low correlation. Low bias is caused by averaging or voting prediction from large number of individual trees. During individual tree building process, the use of different bootstrap sample and randomly selection of a subset of predictor variables for consideration of split at each node result in low correlation among trees. If trees are not built by this two-way randomness, all individual trees will be similar with significant overfitting due to the largest extent in tree building process and therefore lead to poor future prediction while aggregating results from all trees.[59, 61]

2.3.2 Prediction Accuracy

In machine learning, a predictive model tends to overfit since the model is built by the training dataset but may result in low prediction accuracy in future use. Due to empirical relationship from training dataset, overfitting will be more apparent in complex model. A common way to assess overfitting is to use a separate testing dataset to estimate the model prediction accuracy. This may result in an undesirable small training dataset especially when the amount of whole sample is small. The second option is called cross-validation, which provides an assessment of model performance without reducing the training data set and no need of an independent testing dataset. For example, in CART a 10-fold cross-validation is widely used. The original dataset is randomly assigned into 10 sub-datasets with equal sample size. Then one sub-dataset is used as the validation dataset for testing a tree trained by the other 9 sub-dataset. The process is then repeated 10 times, with each of the 10 sub-dataset used once as the validation dataset to produce the overall prediction error.

RF has an inherent cross-validation process that internally validate model through OOB samples and average OOB error to yield overall prediction error for future prediction. When two-thirds of the sample is randomly selected for building each individual tree, the remaining one-third OOB sample is a validation dataset to estimate prediction accuracy for each tree, which is called the OOB error. The overall prediction accuracy of RF is estimated by averaging OOB error from all trees in the forest providing an overall estimate of the prediction accuracy.[49, 59, 61, 76]

In addition, in binary classification problems, the Receiver Operating Characteristic (ROC) curve is generated by plotting the false positive rate against true positive rate and area under the curve (AUC) values provide an indicator of the prediction accuracy and robustness of the model, which is similar to logistic regression.[78-80]

2.3.3 Splitting Criteria

The splitting rules in RF are: 1) decrease of Gini impurity for classification trees and, 2) minimizing residual sum of squares for regression trees, the most commonly used methods in CART.

In classification problem, to create a single tree in a RF, each split in the tree building process is started by measuring the impurity of the root node which defined as pi/t (probability that the outcome variable as class i) in Node t. A node with zero impurity consists of members belonging to in one class (e.g., all yes, or all death). The decrease of impurity function for a split is measured by the difference between the impurity in the parent node and the average impurity in the two child nodes.

For a binary outcome (e.g. 0 or 1), decrease of Gini impurity can be calculated as follows:

1. Calculate the Gini impurity function for the parent node (t): Gini impurity = 2pi/t(1 – pi/t).

2. Calculate the Gini impurity for each of the two child nodes into which the parent node splits: Gini impurity for left child node = 2pi/l(1 – pi/l), Gini impurity for right child node = 2pi/r(1 – pi/r).

3. Calculate the weighted Gini impurity for two child nodes, according to the proportion of the parent node that is included in each child node (pl and pr):

Weighted Gini impurity = (pl)(2pi/l(1 – pi/l)) + (pr)(2pi/r(1 – pi/r)), pl and pr refer to the proportions of the parent node that are included in the left and right child nodes.

4. Calculate the decrease of Gini impurity, which is equal to the following: Decrease of Gini impurity = Gini impurity in parent node – weighted Gini impurity in two child nodes.

Larger values of the decrease of Gini impurity indicate greater difference with respect to the class distribution of outcome in the two child nodes. The predictor variable whose split provides the largest value of the decrease of Gini impurity is selected for splitting at each node.

To demonstrate, consider a hypothetical example in which a parent node includes 100 people from a case control study with a 50 cases of chronic pancreatitis (Y), which is the outcome variable. (Figure 4) Therefore, the impurity of the root node which defined as pi/t equals 0.5. Alcohol abuse (Yes or No) and smoking (Yes or No) are two candidate predictor variables to be evaluated to split the node. The alcohol abuse (X1) is first to split the parent node into two child nodes. The left child node which includes 60 people with alcohol abuse (X1=Yes) has 45 people with chronic pancreatitis. The right child node also includes 40 people without alcohol abuse (X1=No) has 5 people with chronic pancreatitis. In this situation, the Gini impurity is 0.5 for the parent node (Step 1: Gini impurity = 2*0.5*(1-0.5)). The Gini impurity is 0.375 for the left child node and 0.219 for right child node. (Step 2: Gini impurity for left child node = 2*0.75(1 – 0.75), Gini impurity for right child node = 2*0.125(1 – 0.125)). Since the proportion of subjects included from the parent node is 0.6 in each child node and 0.4 in right child node, the weighted Gini impurity for two child nodes also equals to 0.313 (Step 3: Weighted Gini impurity = (0.6*0.375+0.5*0.219)). Then the decrease of Gini impurity is equal to 0.50 - 0.313 = 0.187 (Step 4) using alcohol abuse to split parent node into two child nodes.

Meanwhile, another predictor variable smoking (X2) is also tested to split the parent node. In the same way, the left child node which includes 70 smokers has 40 patients with chronic pancreatitis, and the right child node which also includes 30 non-smokers has 10 patients with chronic pancreatitis. Follow the same formula above, this time the Gini impurity is 0.49 for left child node and 0.44 for right child node. The weighted Gini impurity for two child nodes equals 0.475. Then the decrease of Gini impurity would be equal to 0.50 - 0.475 = 0.025 (Step 4), which is much smaller compared to result using alcohol abuse to split (0.187). As a result, this parent node will split using alcohol abuse (X1) into two child nodes because it has larger value for decrease of Gini impurity.

For continuous predictor variable, the decrease of Gini impurity was evaluated for all possible values in this variable to find the optimal cut off value for split by this continuous predictor variable and then compare to other predictor variables including categorical and/or continuous variables.

In regression problem, as we did for the linear model, minimizing residual sum of squares is used to decrease the impurity of node. Using this criterion, the best split at a node is the split on one of all variables which most successfully separates the high outcome value from the low outcome value in the parent node and therefore minimize residual sum of squares in two child nodes.[55]

2.3.4 Variable Importance

One of the key results from RF is that it measures the entire effect of all predictor variables because RF is able to integrate not only the main effects but also the interaction effects even though the variables with weak marginal effects.

The RF methodology ranks the candidate variables with respect to their importance in predicting the outcome throughout the RF. It is determined by the mean difference of prediction accuracies observed for each tree using OOB error before and after random permutation of a predictor variable.[59, 61, 81] To illustrate, supposing a dataset of chronic pancreatitis study, those patients who drank alcohol more frequently per month (e.g. more than 15 days per month) are more likely to develop to chronic pancreatitis. However, randomly permuting the values of drinking frequency per month in all subjects should destroy this association. Subsequently, using new OOB samples consisting permuted values for drinking frequency to predict whether a patient has chronic pancreatitis or not, the prediction accuracy will decrease since the more influential variables will have more occupancies of splits throughout the forest and the nodes split by drinking frequency will lead to wrong partitions after permutation. Therefore, a large decrease in prediction accuracy indicates a strong association between a predictor variable and the outcome because random permutation destroys their original relationship in OOB sample and lead to worse prediction error. Values around zero or even negative indicate that a variable is not important and no association with outcome because they are weak predictors with small chance to have a spot for split throughout the forest. The permutation process is repeated for all predictor variables and then their importance to the outcome can be ranked in terms of the difference in prediction accuracy between permutation OOB error and the original OOB error.[59, 61]

As mentioned above, the variable importance measures in RF have ability to detect complex interactions between predictor variables because variables involve in specific interactions are likely to stand out as ‘important’. Random permutation in one variable should also destroy its interaction effects with other predictor variables to the outcome and as a result, the prediction accuracy decreases.[59] In addition, RF could capture important predictor variable correlated with other predictor variables. Because of random selection of small number of variables in split process, RF sometimes splits on one and sometimes on another correlated variables. Therefore, RF tends to identify all of the correlated predictors as important if any one of them are important.[59]

There is a second measure available in RF to compute variable importance called Gini importance which is only used in classification problems.[49, 61, 77] The Gini importance of a variable is simply calculated by the sum of decrease of Gini impurity of the variable used to split a node throughout the entire forest. A more important predictor always has more splits in the forest with larger amount of decrease of Gini impurity, leading to a high Gini importance. However, a selection bias may occurs in Gini importance because of the randomly selected variables at each node in which Gini importance calculated with its occurrence of a predictor in the trees.[49, 76, 77] This selection bias does not affect the permutation importance because it is based on the decrease of prediction accuracy resulting from OOB sample. Even if unimportant predictors are selected for split due to the selection bias in tree building process, they do not have ability to significantly improve the overall OOB prediction accuracy, thus do not have higher permutation importance. Practically, however, Breiman said the Gini importance is often very consistent with the permutation importance method.[61]

2.3.5 Proximities

A special feature of RF is the calculation of proximities between each pair of subjects among samples.[43, 59, 82] After the forest is built, the proximity between two subjects could be calculated as the number of times the two subjects end up in the same terminal node of a tree in the forest, divided by the number of trees in the forest. Therefore, a proximity equals to 1 means two subjects always lie in the same terminal node across all trees while a proximity equals to 0 means two subjects are never in the same terminal node. Repeated for all pairs of subjects over all terminal nodes in the forest, the proximity scores could generate a proximity matrix to demonstrate the degree of dissimilarity among subjects as well as to produce multidimensional scaling plots to visualize the distance between subjects typically in a two-dimensional plot.[83] The proximity plot could be used to reveal more hidden data structure, e.g. sub-clusters, outliers, and mislabeled cases that could be of interest to the researchers. In addition, the proximity scores could be used to impute missing values.[59, 82]

2.3.6 Missing Values

RF could handle missing values in the predictor variables. During the tree building process, RF could simply impute missing values using the most frequent non-missing value for categorical variables or computing the median for continuous variables.[59, 61]

Moreover, RF has an advanced method to give better imputation performance for missing values by integrating proximities mentioned above. After the forest is generated with simple missing value imputation, an adjusted missing value imputations are computed using a proximity-weighted majority frequency for categorical variables or a proximity-weighted average for continuous variables and then replace simply imputed values. After that, a new forest is built with new proximities and missing value imputations. It is suggested that repeat this process 4 – 6 times to optimize the imputations but however, this method is computationally expensive and seldom used.[59]

2.3.7 Parameters

RF analysis has two major parameters that are pre-specified: the number of variables to sample at each node (mtry), and the number of trees to build (ntree).[59, 61, 77] In addition, the class weight for classification trees and the tree size could also be adjusted but less commonly specified in practice. There is no optimal values for these parameters because they are data dependent. Since minimizing prediction error is the most important creation to judge the performance of a RF model, the result of OOB error, which is the unbiased estimation of generalization error in RF, could be used to find optimal value of these parameters by running sensitivity analyses.[59, 61, 77] In practice, it has been shown that the RF are not very sensitive to these parameters.[59, 77]

mtry

At each node of an individual tree building process in RF, a small number of different subsets of variables is randomly selected from all predictor variables to find an optimal split (mtry).[77] Smaller values of the mtry may decreases the accuracy of each tree when there are a large number of noise predictor variables because a small subset of all variables may randomly select all noise predictors only thus leading to poor split and increasing the bias. However, when there are large number of less or moderate important predictor variables, small mtry may give them more opportunities to split and higher ranking in variable importance. The large mtry tends to decrease tree complexity because fewer variables will be used in trees and produce more correlated trees. There is no definite value of mtry as the optimal choice. The best choice is reflected by the prediction error calculated by OOB error. However, Breiman recommended a default value of mtry equals square root of total number of variables (√p).[61, 68, 76]

ntree

Another important parameter is how many trees to grow (ntree). The number of trees in the forest should increase with the increase of number of candidate predictor variables, so that each predictor variable has more opportunities to be randomly selected to consider to split a node. But increasing ntree will lead to extra computation required. Breiman demonstrated that increasing the number of trees does not result in overfitting and a larger value of ntree always produces more reliable prediction ability compared to a smaller value.[59, 61] Thus, it is recommended to increase the value of ntree and to stop increasing till OOB error is stable. It is a recommended to default value of ntree = 500 - 1000, which is large enough to produce a stable result with respect to OOB error. [49, 59, 61, 77]

Class weight

In some classification problems, the proportion of minority class in the sample may be extremely small, e.g., only 5% of population have a disease versus 95% of population healthy in a population. Generally, RF tends to minimize overall error rate, keep the error rate relatively low on the majority class while allowing the minority class to have a large error rate. However, the misclassification cost may be very high due to large proportion misclassified minority cases.

RF offers a way to adjust the weights for each class to help to generate more balanced results in classification error for all classes for unbalanced data. The larger weight could be assigned to minority class to penalize its misclassification error rate. However, while getting this balance, the overall classification error rate will go up.[59] The best value of weight could be tested by trying different weights to accommodate the relatively balanced error rate among classes.

There is another technique to deal with unbalanced data called "balanced" RF introduced by Chen et al, by down-sampling majority class in bootstrapping in each tree building process, drawing a bootstrap sample from the minority class and then randomly draw the same number of cases from the majority class.[84]

Tree Size

Limiting the size of individual trees does not often occur because in RF theoretically builds the largest extent unpruned trees. However, RF has the options to limit size of a tree, such as the maximum number of splits, the minimum number of cases in the terminal nodes. It may save computation time and may or may not provide more reliable individual trees to provide lower overall errors. As same as other parameters, there are no optimal values and the overall OOB error is used to see if there are necessary to adjust on the model.[59, 77]

2.4 APPLICATIONS OF RANDOM FORESTS IN EPIDEMIOLOGY

RF has been successfully applied in genetic epidemiology especially with GWAS data.[48, 50, 69, 70, 77]

A study by Goldstein et al used RF analysis in a multiple sclerosis case-control study comprised of over 300,000 SNP genotypes in 931 patients and 2,431 controls.[68] Their results showed that a group of SNPs from a region in chromosome 6p were ranked as top important variables which were consistent with marginal chi-square statistics and the OOB error of RF was 35%. After that, they removed all SNPs on chromosome 6p since association between these SNPs and the disease has been well established from previous findings. Therefore, RF analysis was performed again to search weaker effects after removing chromosome 6p SNPs. The new RF results were compared to findings from the previous study and new top 25 important variables on the ranking list were supported by them. In addition, four new interesting candidate genes were identified that strongly deserve further investigation.[68]

Xu et al applied RF to the prediction of exacerbations in a population of childhood asthmatics participating in the Childhood Asthma Management Program.[70] The outcome of this study was an emergency room visit or a hospitalization for asthma symptoms during a four year follow-up period. There were 417 children enrolled in the study and 127 (~30%) of them experienced at least one severe asthma case. They used age, sex, pre-bronchodilator FEV1%, and treatment group, and/or SNPs as predictors to predict severe asthma exacerbations. To reduce time expense due to implementing all SNPs at a time in RF analysis, they first computed RF variable importance scores for all SNPs, 4,000 at a time in chromosomal order. Base on variable importance sores for all SNPs, then they selected the top 4,000 SNPs, and reran RF with these selected SNPs to re-rank them to generate a candidate gene ranking list. Then they ran RF analysis and repeated it with 4 clinical characteristics only, and clinical characteristics plus different numbers of SNPs selected based on variable importance score from previous candidate gene SNPs list as predictors. Instead of using OOB error, they used an independent testing sample and selected AUC as the indicator of prediction accuracy. With just the 4 clinical characteristics as predictors, the RF model had an AUC = 0.56. The AUCs were 0.57, 0.62, 0.66, and 0.66 in repeated RF models with clinical characteristics plus 10, 40, 160 and 320 SNPs, respectively, indicating that the severe asthma exacerbation in children is affected by genetic as well as environmental factors. They concluded that a reasonable prediction model of asthma exacerbations in children can be achieved through the combination of SNPs and clinical charactertics in a RF model and results improved the understanding of the biologic mechanisms behind why only certain individuals with asthma are at risk for exacerbations.[70]

RF has been also used in other areas of epidemiology such as molecular epidemiology.[67, 71, 85] Barrett et al applied RF to classify samples from 76 breast cancer patients and 77 controls whose proteomic profile had been obtained using mass spectrometry and 365 biomarkers detected from candidate predictor variables.[71] Based on the prediction from OOB samples in RF analysis, the overall OOB error rate was 16.3% and the importance of biomarkers were ranked. They used an independent sample to test and showed consistency in terms of the performance and they concluded that the RF provides a high-performance classification system for proteomic data.

Gurm et al evaluated prediction performance using RF analysis on risk of contrast-induced nephropathy in patients undergoing contemporary percutaneous coronary procedures.[86] Since they had a very large study cohort with 68000 patients, they randomly selected approximately 48,000 patients to run RF analysis and 20,000 patients used as separate validation sample. The study had 46 baseline clinical variables used in RF analysis. All 46 baseline variables were ranked by variable importance in RF and they selected 15 variables with the largest importance scores to build a reduced RF model. The full and reduced RF models were evaluated regarding to prediction accuracy in the validation sample. Predictive accuracy for full and reduced models were higher than 0.8. They concluded that this risk prediction model may prove useful for both clinical decision making and risk assessment.[86]

2.5 PROBLEMS IN RANDOM FORESTS

The primary limitation of RF is that the rank of variable importance does not specify the actual variable interactions, for example whether predictors have an effect in combination with other predictors and if yes with which.[76] Therefore, it is difficult to interpret since there is no information about the splits in each individual tree in the output of RF.

Decision tree methods are well suited for non-linear modeling that help to identify conditional interactions, e.g. if a predictor variable is used as a split in a child node on left branch of the tree but not on the other branch which indicates an interaction effect between that variable and the variable in its parent node. As each individual tree is different from others in RF, the overall variable importance score only provides the importance ranking of all predictor variables but does not specify the actual variable interactions. Moreover, in a situation where two interacted variables have no main effect, it is harder to interpret the interaction effect due to the lack of a marginally detectable main effect.[76]

Second, variable importance provides a ranking of important predictor variables. However, it does not show the significance values of these predictors or a threshold to define which predictor should be selected for further interpretation. Variable importance always provides a ranking - even if all predictors are useless to the prediction problem. Some researches have investigated this issue but no formal approach has been adopted. Therefore, sometimes variable importance from RF analysis is hard to be interpreted due to the lack of inference threshold such as the p-value used in statistical analysis.

In an exploratory study using RF, Strobl et al suggested that excluding variables whose importance is negative, zero or has a small positive value that lies in the same range as the negative values from further exploration by the rationale of random variation of the importance score around zero for unimportant variables. Therefore, positive values of importance score that exceed this range may indicate that a predictor variable is informative and worth for further investigation. [49] Díaz-Uriarte et al suggested to iteratively remove those variables with the smallest variable importance, typically the bottom 20% and re-run RF until prediction accuracy significantly decreases.[87] In a study by Goldstein et al, they used scree plot to visualize variable importance and chose the ‘elbow’ as the threshold to determine the important variables.[68]

2.6 UTILIZATIONS OF RANDOM FORESTS WITH OTHER APPROACHES

While RF can successfully be used by itself, one of its greatest utilities comes in combination with other modeling approaches. Many authors have performed multistage analyses using RF as a first stage screening step and then followed up with other statistical analysis.

In a study by Zyriax et al, baseline characteristics, and 41 SNPs that have previously been found to be associated with type 2 diabetes were analyzed by RF first to select the most important risk factors for contributing pre-diabetes. They found 3 baseline characteristics and 6 SNPs are relatively important variables using the variable selection criteria suggested by Strobl. Then they performed logistic regression and found 3 SNPs was significantly associated with higher risk to pre-diabetes and 1 SNP was significantly associated with lower risk to pre-diabetes, while other 2 SNPs showed a tendency towards a higher risk. However, In this paper, they did not evaluate interaction effects in the logistic regression model.[88]

Jiang et al studied case-control data by first running a RF with all SNPs to obtain variable importance and then designing a sliding window sequential forward feature selection algorithm that could select a small group of candidate SNPs to minimize the classification error and then used a B statistic to test up to three-way interactions of the candidate SNPs.[51, 89] They named it as epiforest for the detection of epistatic interactions using RF.

2.7 CONCLUSION

The RF has become a very useful computational analytic algorithm in various research areas for both classification and regression problems. It has been applied in a number of epidemiological researches especially in genetic epidemiology due to its capability to explore data structure and hidden information in large-scale data. The RF is a powerful analytic framework that help to analyze large dataset, e.g. variable importance evaluation, missing value imputation, outcome prediction, variable selection, etc. RF already has generated many successful applications in a variety of fields and contributed in many publications in last decades. However, RF still under investigation and expansion to be better understood by scientific researchers. Regarding to epidemiological researches, there are still many questions to utilize RF for analysis. The results from RF analysis is dramatically different from traditional statistical models such as the lack of p-values, confidence interval and statistical inference. The new perspective may hard to be interpreted in an epidemiological research. How stable the result is and how to combine RF and other analytic models to initiate more sophisticated data analysis are still questioned. Therefore, RF is a promising data analytic tool that need more exploration and its application in epidemiology is expected to be much more popular in the big data era.

2.8 FIGURES

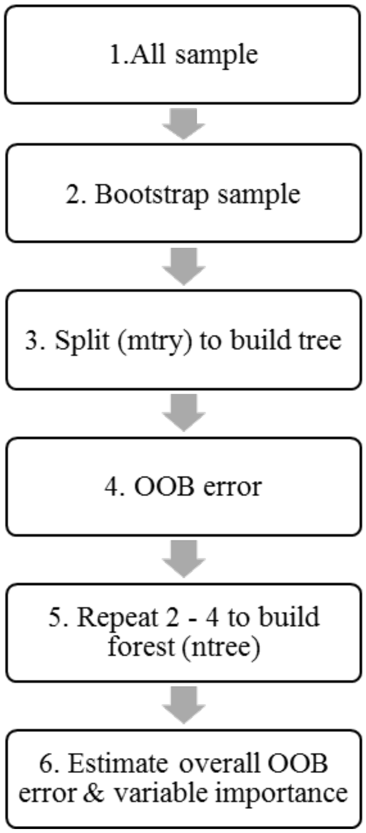
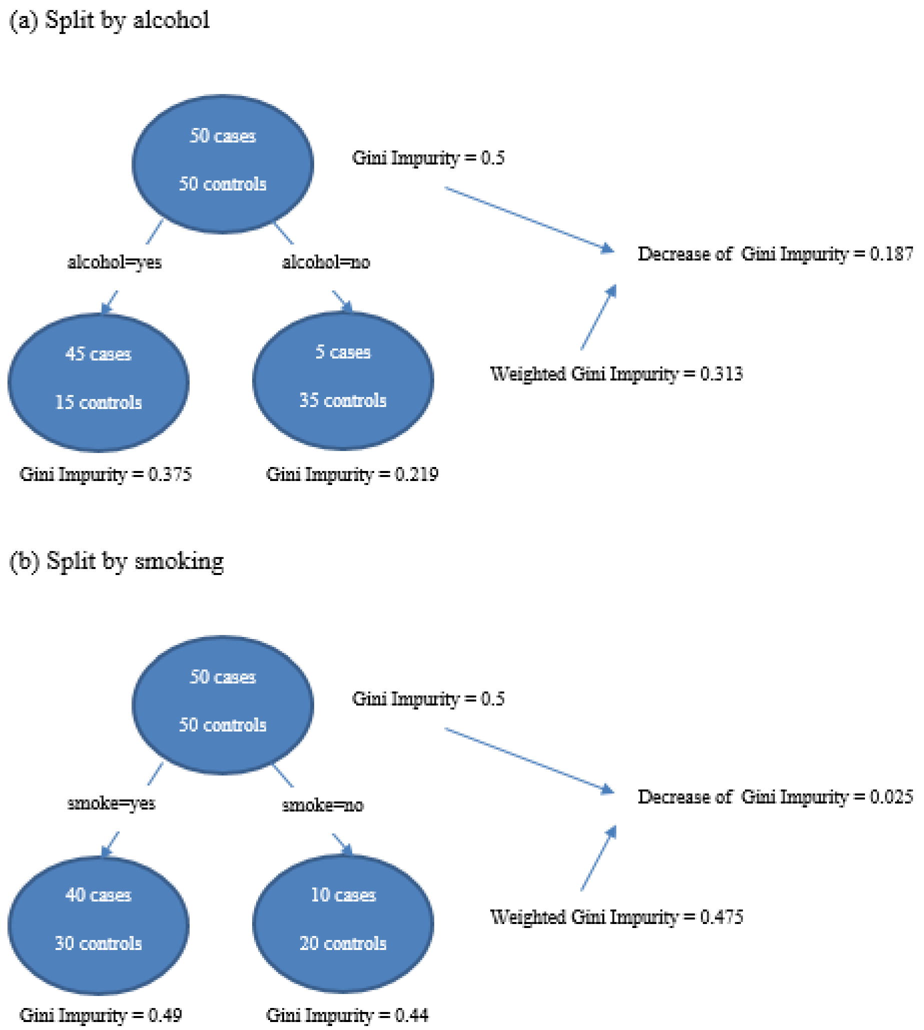

1.All sample
2. Bootstrap sample
3. Split (mtry) to build tree
4. OOB error
5. Repeat 2 - 4 to build forest (ntree)
6. Estimate overall OOB error & variable importance

Figure 3. Random Forests Algorithm



(a) Split by alcohol

(b) Split by smoking

Figure 4. Example of Decrease of Gini Impurity

3.0 APPLICATION OF THE RANDOM FORESTS METHOD TO IDENTIFY IMPORTANT RISK FACTORS ASSOCIATED WITH DIFFERENT PAIN PATTERNS IN PATIENTS WITH CHRONIC PANCREATITIS

3.1 ABSTRACT

The abdominal pain is the most common clinical symptom in chronic pancreatitis (CP) patients. The expressions of pain in CP patients are highly variable and the causes are still under investigation. The mechanism of pain in CP patients is thought to be multifactorial and interactive. The present study is aimed to detect important factors that associated with developing different pain patterns (constant pain vs intermittent pain) in CP patients. Random Forests (RF) analyses were performed to analyze North American Pancreatic Study 2 - Continuation and Validation (NAPS2-CV) study, and was able to identify important risk factors associated with development of different temporal pain patterns in CP patients. Patients' age at CP diagnosis, and two genetic variants rs10818187 and rs7894089 are the most important predictor variables to determine intermittent or constant pain in CP patients. The results confirmed that the complex mechanisms for development of pain in CP. The identification of these factors may be useful in clinical practice to identify individuals at risk for the constant pain at the early stage in CP.

3.2 INTRODUCTION

Chronic pancreatitis (CP) is characterized pathological evidence of destruction of pancreatic tissue, ductal abnormality, fibrosis, inflammation, and loss of both endocrine and exocrine function.[2, 33] Abdominal pain is the principle symptom of CP because pain is the most common clinical symptom occuring in up to 90% of patients and accounts for repeated hospitalizations, interventions, narcotics addiction, detrimental effects on quality of life and medical costs associated with caring for these patients.[5, 32, 90]

Pain in CP is most often located in the upper abdomen. It is sometimes extended to a other parts of the abdomen and radiated to the back.[35] However, the expressions of pain in CP patients are highly variable in severity, frequency and features, and have been challenging to physicians.[32, 91] Temporality and severity of pain are the most two common ways to characterize the pain patterns in CP.

Pain is CP patients may be intermittent and/or constant, mild, moderate and/or severe. Pain may express in different forms, ranging from those patients with no or little mild pain to those with continuous severe pain.[34] There is no universal pain pattern in CP even in similar status of the destruction of pancreas. But it is usual that pain is deep, penetrating and debilitating, and the extent may increase after eating.[35] Moreover, the pattern of pain may change over time in the same patient.

Mullady et al. analyzed the relationship between pain pattern and the disease burden associated with pain using a national cohort study of North American Pancreatitis Study-2 (NAPS2).[24, 32] They classified a total of 414 CP patients with pain regarding to their pain status with respect to temporality pattern (intermittent vs. constant pain), and severity pattern (mild/moderate vs. severe pain). They found that patients who experienced constant pain had higher rates of disability, hospitalization, use of pain medication and lower quality of life compared to patients with intermittent pain patterns. In contrast, there were no significant associations between the quality of life and the difference in severity of pain.[32] Therefore, it has been suggested that management of relapse of pain is more important than relieving severe pain in CP patients which indicates that the evaluating determinants associated with the constant pain might be much more meaningful to clinical practice in management of pain in CP.

The mechanisms of pain in CP continue to be investigated. There are no significant associations between imaging findings or function testing and different pain patterns from previous studies.[35] Most current theories are linked to peripheral and central nociceptive nerve sensitization, that may be motivated by repeated episodes of inflammation and pancreatic injury.[34] But the cause of pain in CP patients is thought to be multifactorial and interactive because the mechanism of development of CP is very complex. As was mentioned earlier, pain patterns are variable in CP patients with similar injury to the pancreas, which indicates that genetic factors may contribute to different pain patterns in CP patients. Based on existing literatures, there are no evidence of the role of genetics on pain patterns.

The present study is aimed to detect important factors that associated with developing different pain patterns (constant pain vs intermittent pain) in CP patients.

3.3 MATERIALS AND METHODS

Study Population

Study subjects and data were derived from the North American Pancreatic Study 2 - Continuation and Validation (NAPS2-CV) study, a cross-sectional and observational follow-up study. The original North American Pancreatic Study 2 (NAPS2) was a multicenter, molecular epidemiology study designed to understand the underlying environmental, metabolic, and genetic factors associated pancreatitis. Between August 2000 and September 2006, 1,000 human subjects with recurrent acute pancreatitis (RAP) or CP, plus 695 controls were ascertained and the study data has been used to conduct many genetic and gene-environment studies for CP and RAP.[24] In 2008, the NAPS2-CV was funded to continue the work started by original NAPS2. The goal of NAPS2-CV was to collect a validation group of European ancestry subjects for potential genome-wide association studies (GWAS) and there were 521 CP patients recruited.

In NAPS2-CV, all patients described their pain features of two kinds of abdominal pain - constant and/or intermittent pain patterns in two separate pain questionnaires using newly developed McGill pain short form.[92] Some patients having one of two pain patterns filled out one of two questionnaires, while some patients filled out neither or both. Since the study objective was to identify important predictor variables that contribute to different temporal pattern of abdominal pain. We included CP patients with either constant pain or intermittent pain only to extract eligible subjects from all CP patients in NAPS2-CV. Therefore, the outcome of interest was pain pattern status classified as constant pain or intermittent pain. In NAPS2-CV, study subjects were also asked to assess the temporality and severity of their pain based on recommendations of the American Gastroenterological Association (AGA) which was used in original NAPS2 study and successfully evaluated in previous study.[32] As a result, the classification of temporality of pain in CP patients in NAPS2-CV could be summarized by two separate ways and we used sample of patients whose pain patterns classified by AGA recommendations as validation for current study.

Demographic and Clinical Variables

The following candidate predictor variables were selected or calculated from NAPS2-CV questionnaires for evaluation. Patient's gender, race, ethnicity (Hispanic or Latino), Ashkenazi Jewish heritage, family history of any pancreatic disease, family history of AP, family history of CP, family history of pancreatic cancer, patient history of AP, additional attacks of AP, age of CP diagnosis, alcohol consumption during the period of maximum drinking in patient's lifetime (abstainer, light drinker, moderate drinker, heavy drinker or very heavy drinker), alcohol consumption in the months before getting pancreatitis, smoking (non-smoker, former smoker, current smoker), amount of smoking (none, less than 1 pack per day, more than 1 pack per day), exocrine insufficiency, endocrine insufficiency, TIGAR-O etiologic risk factors including any toxic-metabolic factor, alcohol, tobacco, hyperlipidemia, hypercalcemia, medications, chronic renal failure, toxins, idiopathic factors, any genetic factor, Cationic Trypsinogen mutation, CFTR mutations, SPINK1 mutations, Alpha 1-antitrypsin deficiency, other genetics, any autoimmune pancreatitis (AIP) factor, Sjogren's disease, rheumatoid arthritis, primary sclerosing cholangitis (PSC), retroperitoneal fibrosis, other AIP, any autoimmune disease-associated factor, Crohn's disease, ulcerative colitis, autoimmune hepatitis, other autoimmune disease-associated factors, any recurrent and severe acute pancreatitis associated chronic pancreatitis factor, postnecrotic, postirradiation, vascular diseases/ischemic, any obstructive factor, pancreas divisum, sphincter of oddi disorders, posttraumatic pancreatic stricture, preampullary duodenal diverticulum, duct obstruction, pancreatic cancer, intraductal papillary mucinous neoplasm (IPMN), other obstructive factors, gallstones, and other risk factors.

Genetic variables

DNA was isolated from patient blood samples using the Qiagen FlexiGene DNA Kit (Qiagen Inc, Valencia, CA, USA).[93] Genotyping was performed at the University of Pittsburgh Genomics and Proteomics Core Laboratories using the MassARRAY iPLEX GOLD (Sequenom, Inc, San Diego, CA, USA). The gene polymorphisms of 58 candidate SNPs (Table 1) that may associated with chronic pain in CP identified from the first stage GWAS in NAPS2 study were included in analysis.[40]

3.4 ANALYSIS

Descriptive analyses for all predictor variables between two pain pattern groups were reported as means and standard deviation for quantitative variables or count and proportions for categorical variables. All genetic variables were analyzed as binary categorical variables with common allele homozygotes treated as reference group. Bivariate comparisons for continuous variables were performed using Student’s t-test or Mann–Whitney U test and for categorical variables using chi-squared tests or Fisher’s exact tests, as applicable.

Random Forests (RF) analysis was used to identify important risk factors in all 117 candidate predictor variables associated with the temporal pain patterns in patients with CP. RF is an ensemble of large amount of individual decision trees by randomly selecting a bootstrap subset sample of two-thirds of whole sample per tree and randomly selecting a subset of all predictor variables at each node of the tree. At each node, RF selects the predictor variable that best splits data into two child nodes.[61] This process allows for all demographic, clinical and genetic factors to work simultaneously in predicting the pain patterns in CP patients. RF determines classification error using the out of bag sample (OOB), those one-thirds of subjects not randomly selected to build a given individual tree. RF works through this process of selecting bootstrap samples to build the tree and using the OOB samples to determine error and variable importance. The variable importance was evaluated using permutation importance. The given variable was randomly permuted in the OOB sample for the tree and new estimate of OOB error was calculated. All candidate predictor variables in the model were ranked based on the difference between this estimate and the original OOB error. The larger increase in classification error indicates a stronger association between a given predictor variable and the pain patterns because random permutation destroys their original relationship in OOB sample and lead to worse classification error.

To address clinical question that help physicians to determine if a CP patient at risk of having constant pain patterns, a set of predictor variables and preferably small set of predictor variables were expected to be identified using RF analysis. An iterative RF analysis framework was used to identify and select important variables for further examination. To build the 1st RF, we used all variable without pre-selection to get the ranking of variable importance. Based on the recommendation by Díaz-Uriarte et al,[87] the bottom 20% of variables with the lowest importance score were dropped and a new RF was built. This process was repeated till there were two variables left to evaluate in the final RF analysis. After building all RF, The OOB error from all RF were evaluated. In steading of selecting the best model, our purpose is to use RF analysis to help us to select a relatively less complex and parsimonious model with minimum number of predictor variables which performs not significantly worse than the best model regarding to the classification accuracy (OOB error). Therefore, if there is a RF with smaller number of variables and the OOB error is not significant higher than the RF with the lowest

OOB error, its set of variables was selected for further investigation. The important variables found in these best models were analyzed in a multivariable logistic regression model to estimate the association among the set of identified predictor variables and the pain patterns. The area under the receiver operation characteristic (ROC) curve (AUC) from logistic regression model was evaluated to examine model discrimination abilities. Typically, an AUC value of 0.5 means a model accuracy of 50% in predicting positives and is no better than the random assignment of positive or negative status. An AUC value of 1.0 shows the model accurately classified 100%. If AUC exceeds the critical value of 0.7, the model is considered with high predictive power.[94] To compare the model using the optimal combination of predictor variables selected from RF analysis, a logistic regression using stepwise variable selection procedure for all 117 candidate variable was performed to identify important predictor variables associated with the difference in temporal pain patterns in CP.

To validate the results using new adopted McGill short pain form for classifying intermittent and constant pain groups, the sample of CP patients in NAPS2-CV whose temporal pain patterns classified by AGA recommendations was used as validation dataset. The samples were not mutually exclusive because patients might filled out both pain pattern questions in NAPS2-CV. The validation sample was scored in the logistic regression model concluded from RF analysis to compare the difference in classification ability using AUC statistic. If classification ability were similar then it could be shown that the pain pattern classification using McGill short pain form are parallel to original AGA recommendation pain pattern classification system and the results from the current study were valid.

RF analyses were implemented in Salford Predictive Modeler, version 7.0 (Salford Systems, San Diego, CA). The RF analyses were implemented with the recommended parameter settings: the number of trees was set to ntree=1000; the number of variables to test at each node was set to mtry=√N (N=Total number of variables used in each RF). The OOB error was generated for each RF analysis and the variable importance score was calculated by permutation importance method.

Multivariate logistic regression models were built using the most promising variables found in RF analysis and odds ratio (OR) with 95% confidence interval (CI) were calculated. Descriptive and logistic regression analysis were performed using SAS, version 9.3 (SAS, Inc., Cary, NC).

3.5 RESULTS

In NAPS2-CV study, intermittent pain characters were reported by 323 of the 521 CP patients enrolled (62.0%) and constant pain characters were reported by 260 of the 521 CP patients enrolled (50.0%) based on Short Form McGill Pain Questionnaire. Among them, 173 were patients with intermittent pain only, and 110 were patients with constant pain only. 150 patients with both pain patterns were excluded from the current study. Due to availability of genetic variants data, a total of 181 patients were enrolled for the analysis with 112 patients classified as intermittent pain group and 69 patients classified as constant pain group.

The distribution of all demographic and clinical candidate variables were presented in Table 2. The initial RF analysis was implemented using all these predictor variables to build up the forest. The OOB error rate equals 0.3702 in classifying two pain patterns in CP patients, which means the overall classification accuracy of the 1st RF is 63.0%. All variables were assigned an importance score with 58 of 117 variables have score higher than 0. The rest 59 variables got score of 0 or lower.

The second RF was analyzed by dropping 20% (n=23) of all variables with lowest importance score in the first iteration of RF. A total of 94 variables was kept to run the second RF analysis and a new ranking of variable importance was generated and the new OOB error rate equals 0.3812. Iteratively, a total of 18 RFs was built by removing the weakest 20% of variables from the previous ranking of variable importance. The 18th RF only contained 2 variables: age at diagnosis of CP and a SNP rs10818187. All OOB error rates and number of variable used for 18 RF analyses are listed in Table 3 and the OOB error rate curve from the 1st RF to the 18th RF is presented in Figure 5. The 8th RF model with 24 variables has the relatively lowest OOB error 0.2541 which means the highest classification accuracy generated by the 8th RF. The ranking of variable importance to classify two pain patterns in CP patients from the 8th RF model is listed in Table 4. Among 24 variables, age at CP diagnosis and SNP rs10818187 have relatively higher importance scores (16.52 and 16.45, respectively) compared to other 22 variables. The distributions of genetic variants in the 8th RF model were presented in Table 5. 9 of 24 variables, Age at CP diagnosis, rs10818187, rs10108543, rs7438388, rs7338234, rs7894089, rs465189, rs7389298, and Sphincter of Oddi disorders risk factor are univariately associated with two pain patterns in CP patients. rs10818187, rs10108543, rs7438388, rs465189, rs7389298, and Sphincter of Oddi disorders risk factor are significantly associated with higher risk of having constant pain in CP patients. Increased age at CP diagnosis, rs7338234, and rs7894089 are significantly associated with lower risk having constant pain in CP patients.

From 8th to 17th RF analysis, the OOB error rate is fluctuated between 25.41% and 28.73%. However, the OOB error rate of the 18th RF sharply increased from 28.18% in the 17th

RF analysis to 32.60% after dropping the 3rd as well as the weakest variable rs7894089 from the 17th RF analysis, which means rs7894089 is an informative predictor variable that should be included in the model to help to correctly classify or predict pain patterns in CP patients. The variable importance score in the 17th are listed in Table 6. According to the 17th RF model, age at CP diagnosis, rs10818187, and rs7894089 represent the most powerful risk factors to determine the temporal pain patterns in CP patients. A multivariable logistic regression was performed including three variables. There was no interaction existing amongst three predictors.(Table 7) The results showed that age at CP diagnosis (OR for 10 years increase=0.776, CI 0.651-0.961, p=0.02) and rs7894089 (OR=0.394, CI 0.174-0.896, p=0.03) are significantly associated with the lower risk for constant pain, whereas rs10818187 (OR=2.603, CI 1.139-5.951, p=0.02) is a significant risk factor associated with the increased risk of having constant pain in CP patients (Table 8). The AUC equaled to 0.706 which indicates that the logistic regression model has good discrimination ability.

The result from logistic regression using stepwise variable selection procedure (P < 0.30 for entry in the model and P > 0.35 for removing the variables) for all 117 variables showed that no variable should be included and therefore no model was built through stepwise selection.

The validation sample contained 105 intermittent pain CP patients and 166 constant pain patients. The validation sample was scored using 3-predictor logistic regression model and the AUC was 0.685 which indicated the two pain classification system are parallel.

3.6 DISCUSSION

Abdominal pain is the most common symptom in CP and still remained as a major clinical challenge. Different pain patterns have been described previously and the constant pain is significantly associated with poor quality of life in CP patients.[32] There are some neurobiological theories proposed to understand pain in CP. Also, it was hypothesized that differences in pain patterns are related with different etiology of CP but has not been confirmed.[95] The etiology of pain in CP is likely involves multiple mechanisms. However, the studies on the influence of multiple factors including demographic, clinical, and genetic variables to the determination of pain patterns in CP patients is still limited.

Because previous findings showed that patients who experienced constant pain had higher disease burden and lower quality of life compared to patients with intermittent pain patterns, we developed a backward elimination RF analysis framework combined with follow-up logistic regression model to identify important factors associated with different temporal pain patterns in CP patients. In high dimensional data, there might be a large number of uninformative predictor variables. Therefore, it is possible to see that in RF analyses if number of candidate predictor variables decreases, the OOB error will be lower because in tree building process with a large amount of candidate predictor variables, the random selection of a small subset of variables maybe all uninformative variables that lead to poor performance in some node split of the tree. Without variable pre-selection, results from iterations of RF analyses identified 24 predictor variables are potentially related to the determination of different temporal pain patterns in CP patients regarding to the lowest OOB error using these variables together. These variables include patients' age at diagnosis, their smoking status, alcohol drinking categories, family history of AP, and Sphincter of Oddi disorders, and a group of genetic variants. Alcohol drinking and smoking are well established risk factors for development of CP but their relationships with pain patterns were not described previously. Sphincter of Oddi disorders is an obstructive etiology of CP. These result could be used as an evidence that the variant manifestations of different pain patterns in CP patients are extremely complex and likely involved multiple mechanisms and predisposing factors suggested by many studies focusing on pain in CP. Among 24 variables, Age at CP diagnosis, rs10818187, rs10108543, rs7438388, rs7338234, rs7894089, rs465189, rs7389298, and Sphincter of Oddi disorders are significantly associated with the difference in pain patterns using traditional univariate statistical analysis. All these 9 significant variables are ranked the on the top in the variable importance list generated in RF. Sphincter of Oddi disorders risk factor is the lowest among these factors with the 15th ranking apparently due to its rare occurrence in the sample. Other 13 non-significant factors may involve in conditional interaction or serve as mediators with other factors in the context of complex mechanisms of pain in CP.

A reduced RF model incorporating the most three important factors, the patients' age at CP diagnosis, rs10818187, and rs7894089, has an OOB error of 0.28, which is slightly higher than the RF model with lowest OOB error (0.25). Regarding to the simplicity consideration, the results from this small RF analysis could be more meaningful for clinical purpose. The age at CP diagnosis is ranked as the 1st with importance score of 41.64 which indicated age is the most important predictor variable to classify CP patients with different temporal pain patterns. The 2nd is rs10818187 and the 3rd is rs7894089 (importance score: 36.58, 21.77, respectively).

In a logistic regression model, the 3 selected risk factors variables were all significantly associated with different pain patterns. The larger the age of CP diagnosis, the probability of intermittent pain increases in patients. Specifically, we estimated that the odds for having constant pain in CP patients decrease 23% for each 10 years increase of age (10-year OR = 1-0.97510 = 0.23) after adjusting for rs10818187 and rs7894089. The rs10818187 is a risk genetic variant (OR=2.603) of having constant pain whereas the rs7894089 is a protective genetic variant (OR=0.394) to having constant pain in CP patients. The SNP rs10818187 is inter-genic region between LOC389787 and DBC1 gene and rs7894089 located in FAM53B gene. To our knowledge, identification of these genetic variants has not been previously reported to relate to pain patterns in CP. In addition, other SNPs identified in the RF model with the lowest OOB error should also be investigated for the association with pain patterns because they may have complex interaction effects with other environmental, metabolic, and genetic factors. Therefore, further investigation is required to confirm the findings from current study.

In addition, it was notably that performing stepwise logistic regression failed to build a model with one or more predictor variables to classify pain patterns in CP. The reason was that there were a number of variables with large number of missing value and the listwise deletion of incomplete subjects was used in stepwise regression modeling which resulted in small number of sample size and the loss of power to detect associations. However, in RF analysis, the missing values were automatically imputed using the most frequent –non-missing value or median. Missing value is frequently happened in high-dimensional data. The current study demonstrated the advantage of RF to handling missing value and therefore successfully identify important predictor variables for modeling purpose.

The key findings from this study showed that using RF analyses, we successfully identified risk factor associated with having constant pain vs intermittent pain in CP patients. Analysis of the NAPS2-CV data indicates that using SNPs rs10818187, rs7894089, and the patients' age at CP diagnosis could be reliably used to make a distinction between intermittent and constant pain patterns in patients who suffer abdominal pain after diagnosis of the CP.

3.7 LIMITATIONS

A number of limitations of this study should be addressed. First, RF did not provide visual information about variable interactions even though it takes into account complex interaction effects in decision tree building process. The variables in the ranking of variable importance may have interaction effects with others but they were not specified. Second, the variable importance ranking in RF does not have a threshold to show evaluate the significance of importance which made them hard to interpret from a statistical perspective and make the decision on variable selection difficultly. Third, there was no formal way to select the optimal RF model, we selected a parsimonious RF model as the optimal model followed by statistical analysis based on its simplicity for clinical purpose. Finally, the sample size was relatively small and therefore the association between all candidate predictor variables and pain patterns may not be truly reflected.

3.8 CONCLUSIONS

We used RF analyses to analyze NAPS2-CV data, and was able to identify important risk factors associated with temporal pain patterns in CP patients. The results confirmed that the complex mechanism for development of pain in CP. Patients' age at CP diagnosis, and two genetic variants rs10818187 and rs7894089 are the most important predictor variables to determine intermittent or constant pain in CP patients. The identification of these factors may be useful in clinical practice to identify individuals at risk for the constant pain at the early stage in CP. In the future, more research is required to confirm these findings to establish a clinically applicable tool.

3.9 TABLES AND FIGURES

Table 1. 58 Candidate pain SNPs

SNP	Chromosome	Minor Allele	SNP	Chromosome	Minor Allele
rs4143111	1	A	rs757323	7	G
rs4927113	1	G	rs2129557	7	A
rs7540125	1	A	rs10108543	8	G
rs887958	2	A	rs382796	8	A
rs1001763	3	A	rs1010587	9	A
rs16861588	3	G	rs10818187	9	A
rs4698390	4	A	rs7389298	9	G
rs7438388	4	G	rs7894089	10	G
rs10017798	4	C	rs4757031	11	A
rs10009455	4	G	rs17132911	11	A
rs13127102	4	A	rs7949201	11	G
rs10461324	4	A	rs516226	11	A
rs7293455	5	G	rs528431	11	A
rs172139	5	G	rs10492094	12	A
rs10042680	5	A	rs2302604	12	A
rs17318106	5	A	rs11055087	12	A
rs13182765	5	A	rs17394079	12	A
rs1796520	6	G	rs10506053	12	G
rs1796521	6	A	rs12582707	12	G
rs1624440	6	G	rs3741658	12	G
rs9295689	6	G	rs2050500	13	A
rs2172007	6	C	rs7338234	13	A
rs2498399	6	G	rs7317522	13	A
rs1778296	6	C	rs10146989	14	A
rs2504284	6	A	rs8029816	15	A
rs1542650	6	G	rs1551355	17	A
rs721025	6	A	rs4804524	19	G
rs13214367	6	G	rs465189	21	A
rs803411	6	A	rs5983020	23	A

Table 2. Baseline characteristics of study population

Variables	Intermittent N=112 (61.88%)	Constant N=69 (38.12%)	p-value
Age at CP diagnosis, mean (SD)	50.07 (17.83)	43.37 (13.58)	0.0022
Gender, n (%)			
Male	55 (56.12)	43 (43.88)	0.0832
Female	57 (68.67)	26 (31.33)	
Ethnicity, n (%)			
Not Hispanic or Latino	111 (62.01)	68 (37.99)	1
Hispanic or Latino	1 (50.00)	1 (50.00)	
Race, n (%)			
White	112 (62.22)	68 (37.78)	0.38
Unknown	0 (0)	1 (100)	
Jewish heritage, n (%)			
No	2 (33.33)	4 (66.67)	0.1627
One parent	108 (62.79)	64 (37.21)	
Both Parents	2 (100)	0 (0)	
Unknown	0 (0)	1 (100)	
Drinking category in maximum period, n (%)			
Abstainer	20 (64.52)	11 (35.48)	0.3256
Light	23 (60.53)	15 (39.47)	
Moderate	17 (80.95)	4 (19.05)	
Heavy	25 (65.79)	13 (34.21)	
Very heavy	25 (54.54)	13 (34.21)	
Drinking category in period before pancreatitis, n (%)			
Abstainer	11 (84.62)	2 (15.38)	0.5093
Light	21 (77.78)	6 (22.22)	
Moderate	10 (65.20)	6 (37.50)	
Heavy	11 (68.75)	5 (31.25)	
Very heavy	20 (62.50)	12 (37.50)	
Smoking, n (%)			
Never	38 (66.67)	19 (33.33)	0.0943
Past	32 (71.11)	13 (28.89)	
Current	42 (53.16)	37 (46.84)	
Amount of smoking, n (%)			
Never	38 (66.67)	19 (33.33)	0.1757
<1 pack/day	36 (67.92)	17 (32.08)	
≥1 pack/day	38 (53.52)	33 (46.48)	
Diagnosis of acute pancreatitis, n (%)			

Table 2 Continued

Yes	75 (67.57)	36 (32.43)	0.6471
No	26 (60.47)	17 (39.53)	
Unknown	4 (57.14)	3 (42.86)	
Family history of any pancreatitis, n (%)			
No	85 (64.39)	47 (35.61)	0.3864
Yes	15 (55.56)	12 (44.44)	
Family history of chronic pancreatitis, n (%)			
No	97 (60.63)	63 (39.38)	0.3379
Yes	15 (71.43)	6 (28.57)	
Family history of acute pancreatitis, n (%)			
No	103 (64.38)	57 (35.63)	0.0563
Yes	9 (42.86)	12 (57.14)	
Family history of pancreatic cancer, n (%)			
No	102 (60.71)	66 (39.29)	0.1283
Yes	10 (76.92)	3 (23.08)	
Age at AP diagnosis, mean (SD)	43.96 (20.09)	41.13 (15.11)	0.3357
Additional attacks of AP, n (%)			
No	47 (55.29)	38 (44.71)	0.0962
Yes	64 (67.37)	31 (32.63)	
Exocrine insufficiency, n (%)			
No	61 (63.54)	35 (36.46)	06244
Yes	51 (60.00)	34 (40.00)	
Endocrine insufficiency, n (%)			
No	66 (58.93)	46 (41.07)	0.2979
Yes	46 (66.67)	23 (33.33)	
TIGAR-O etiologies			
Toxic-metabolic factors, n (%)			
No	38 (67.86)	18 (32.14)	0.2677
Yes	74 (59.20)	51 (40.80)	
Alcohol, n (%)			
No	70 (64.81)	38 (35.19)	0.3225
Yes	42 (57.53)	31 (44.93)	
Smoking, n (%)			
No	55 (67.90)	26 (32.10)	0.0953
Yes	54 (55.67)	43 (44.33)	
Hyperlipidemia, n (%)			
No	81 (57.04)	61 (42.96)	0.1565
Yes	20 (71.43)	8 (28.57)	
Hypercalcemia, n (%)			
No	97 (58.43)	69 (41.57)	0.5868
Yes	1 (100)	0 (0)	

Table 2 Continued

Medications, n (%)				
	No	97 (59.15)	67 (40.85)	0.3682
	Yes	1 (33.33)	2 (66.67)	
Chronic Renal Failure, n (%)				
	No	97 (58.79)	68 (41.21)	0.4312
	Yes	2 (66.67)	1 (33.33)	
Toxins, n (%)				
	No	98 (58.68)	69 (41.42)	.
	Yes	0 (0)	0 (0)	
Idiopathic factors, n (%)				
	No	76 (58.91)	53 (41.09)	0.2625
	Yes	34 (68.00)	16 (32.00)	
Genetic factors, n (%)				
	No	94 (60.65)	61 (39.35)	0.5728
	Yes	16 (66.67)	8 (33.33)	
Cationic Trypsinogen mutation, n (%)				
	No	101 (61.59)	63 (38.41)	0.2879
	Yes	5 (62.50)	3 (37.50)	
CFTR mutation, n (%)				
	No	98 (60.49)	64 (39.51)	0.1575
	Yes	9 (75.00)	3 (25.00)	
SPINK1 mutation, n (%)				
	No	103 (60.95)	66 (39.05)	0.5685
	Yes	3 (75.00)	1 (1.49)	
Alpha 1-antitrypsin deficiency, n (%)				
	No	103 (60.95)	66 (39.05)	0.6118
	Yes	1 (100)	0 (0)	
Other genetic factors, n (%)				
	No	103 (60.95)	66 (39.05)	0.4794
	Yes	1 (50.00)	1 (50.00)	
Autoimmune Pancreatitis, n (%)				
	No	106 (61.27)	67 (38.73)	0.3226
	Yes	4 (66.67)	2 (33.33)	
Sjogren’s disease, n (%)				
	No	108 (61.02)	69 (38.98)	0.6124
	Yes	1 (100)	0 (0)	
Rheumatoid Arthritis, n (%)				
	No	106 (60.57)	69 (39.43)	0.2272
	Yes	3 (100)	0 (0)	
PSC, n (%)				
	No	109 (61.24)	69 (38.76)	.

Table 2 Continued

Yes	0 (0)	0 (0)	
Retroperitoneal fibrosis, n (%)			
No	109 (61.24)	69 (38.76)	.
Yes	0 (0)	0 (0)	
Other AIP, n (%)			
No	109 (61.24)	69 (38.76)	0.6145
Yes	1 (100)	0 (0)	
Autoimmune disease-associated factors?, n (%)			
No	106 (61.27)	67 (38.73)	0.3226
Yes	4 (66.67)	2 (33.33)	
Crohn’s disease, n (%)			
No	108 (61.63)	68 (38.64)	0.4774
Yes	1 (50.00)	1 (50.00)	
Ulcerative Colitis, n (%)			
No	109 (61.24)	69 (38.76)	0.6145
Yes	1 (100)	0 (0)	
Autoimmune hepatitis, n (%)			
No	109 (61.24)	69 (38.76)	.
Yes	0 (0)	0 (0)	
Other AI disease, n (%)			
No	107 (61.14)	68 (38.86)	0.4395
Yes	2 (66.67)	1 (33.33)	
Recurrent and Severe Acute Pancreatitis Associated, n (%)			
No	103 (62.80)	61 (37.20)	0.1414
Yes	6 (42.86)	8 (57.14)	
Postnecrotic, n (%)			
No	104 (62.28)	63 (37.72)	0.1348
Yes	5 (45.45)	6 (54.55)	
Postirradiation, n (%)			
No	108 (61.02)	69 (38.98)	.
Yes	0 (0)	0 (0)	
Vascular Diseases / ischemic, n (%)			
No	108 (61.02)	69 (38.98)	.
Yes	0 (0)	0 (0)	
Obstructive factors, n (%)			
No	84 (60.43)	55 (39.57)	0.6009
Yes	26 (65.00)	14 (35.00)	
Pancreas Divisum, n (%)			
No	95 (60.90)	61 (39.10)	0.8004
Yes	11 (57.89)	8 (42.11)	
Sphincter of Oddi disorders, n (%)			

Table 2 Continued

	No	105 (62.13)	64 (37.87)	0.0090
	Yes	0 (0)	5 (100)	
Posttraumatic pancreatic stricture, n (%)				
	No	103 (59.88)	69 (40.12)	0.3628
	Yes	2 (100)	(0)	
Preampullary duodenal diverticulum, n (%)				
	No	105 (60.34)	69 (39.66)	.
	Yes	0 (0)	0 (0)	
Duct obstruction, n (%)				
	No	100 (59.88)	67 (40.12)	0.2672
	Yes	5 (71.43)	2 (29.57)	
Pancreatic cancer, n (%)				
	No	105 (60.34)	69 (39.66)	.
	Yes	0 (0)	0 (0)	
IPMN, n (%)				
	No	105 (60.34)	69 (39.66)	0.6057
	Yes	1 (100)	0 (0)	
Other obstructive, n (%)				
	No	96 (58.90)	67 (41.10)	0.0485
	Yes	12 (85.71)	2 (14.29)	
Gallstones, n (%)				
	No	108 (61.02)	69 (38.98)	0.1436
	Yes	4 (100)	0 (0)	
Miscellaneous factors, n (%)				
	No	95 (60.51)	62 (39.49)	0.4887
	Yes	15 (68.18)	7 (31.82)	

Table 3. Number of variables and OOB error rate of each RF analysis

RF analysis	Number of variables	OOB error (%)
1	117	37.02
2	94	38.12
3	75	35.36
4	60	29.28
5	48	29.28
6	38	30.39
7	30	27.07
8	24	25.41
9	19	27.07
10	15	28.73
11	12	28.73
12	10	27.62
13	8	28.73
14	6	28.18
15	5	27.07
16	4	28.18
17	3	28.18
18	2	32.60

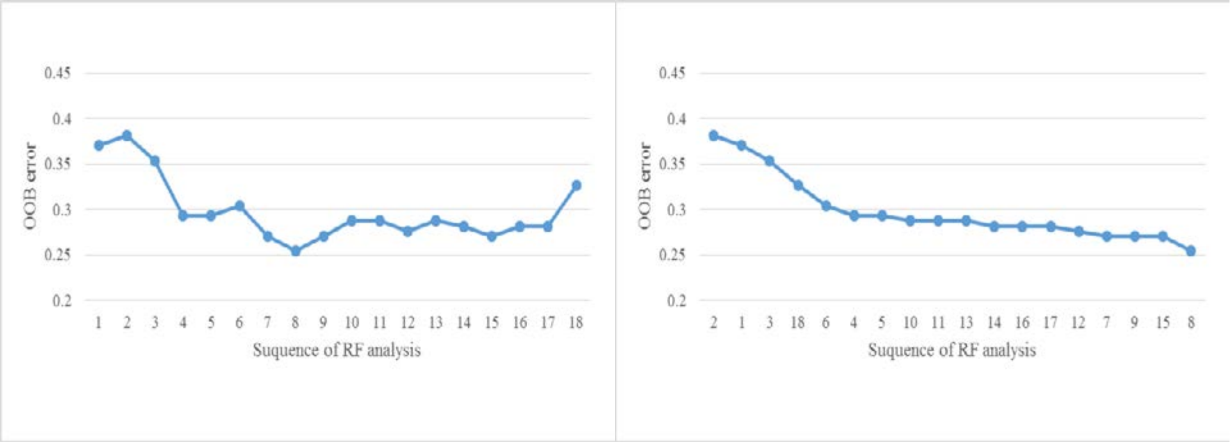


Figure 5. OOB error rate curve for all RF analyses

Table 4. Variable importance score in the 8th RF

Rank	Variables	VI
1	Age at CP diagnosis, mean (SD)	16.52
2	rs10818187, n (%)	16.45
3	Age at AP diagnosis, mean (SD)	7.28
4	rs10108543, n (%)	6.72
5	Drinking category in period before pancreatitis, n (%)	6.55
6	rs7438388, n (%)	6.18
7	rs7338234, n (%)	6.01
8	Smoking, n (%)	4.19
9	rs7894089, n (%)	4.13
10	Smoking risk factor identified by Physician, n (%)	3.97
11	Amount of smoking, n (%)	3.32
12	rs465189, n (%)	2.52
13	rs7389298, n (%)	2.25
14	rs17394079, n (%)	2.01
15	Sphincter of Oddi disorders risk factor identified by Physician, n (%)	1.86
16	rs2129557, n (%)	1.76
17	Family history of AP, n (%)	1.72
18	rs10017798, n (%)	1.66
19	rs2050500, n (%)	1.25
20	rs2302604, n (%)	1.23
21	rs10009455, n (%)	1.11
22	rs12582707, n (%)	0.9
23	rs9295689, n (%)	0.33
24	rs528431, n (%)	0.08

Table 5. Distribution of genetic variants in the 8th RF model

Variables	Intermittent N=112 (61.88%)	Constant N=69 (38.12%)	p-value
rs10818187, n (%)			
Without rare allele	39 (79.59)	10 (20.41)	0.0033
With rare allele	73 (55.73)	58 (44.27)	
rs10108543, n (%)			
Without rare allele	52 (72.22)	20 (27.78)	0.0199
With rare allele	60 (55.05)	49 (44.95)	
rs7438388, n (%)			
Without rare allele	70 (68.63)	32 (31.37)	0.0336
With rare allele	42 (53.16)	37 (46.84)	
rs7338234, n (%)			
Without rare allele	86 (56.95)	65 (43.05)	0.0022
With rare allele	26 (86.67)	4 (13.33)	
rs7894089, n (%)			
Without rare allele	76 (56.30)	59 (43.70)	0.0081
With rare allele	36 (78.26)	10 (21.74)	
rs465189, n (%)			
Without rare allele	86 (67.19)	42 (32.81)	0.0223
With rare allele	26 (49.06)	27 (50.94)	
rs7389298, n (%)			
Without rare allele	70 (67.96)	33 (32.04)	0.0395
With rare allele	39 (52.70)	35 (47.30)	
rs17394079, n (%)			
Without rare allele	71 (59.17)	49 (40.83)	0.2921
With rare allele	41 (67.21)	20 (32.79)	
rs2129557, n (%)			
Without rare allele	104 (63.80)	59 (36.20)	0.1085
With rare allele	8 (44.44)	10 (55.56)	
rs10017798, n (%)			
Without rare allele	78 (60.00)	52 (40.00)	0.4061
With rare allele	34 (66.67)	17 (33.33)	
rs2050500, n (%)			
Without rare allele	77 (58.78)	54 (41.22)	0.1646
With rare allele	35 (70.00)	15 (30.00)	
rs2302604, n (%)			
Without rare allele	70 (58.82)	49 (41.18)	0.2732
With rare allele	41 (67.21)	20 (32.79)	
rs10009455, n (%)			
Without rare allele	97 (59.88)	65 (40.21)	0.1054

Table 5 Continued

With rare allele	15 (78.95)	4 (21.05)	
rs12582707, n (%)			
Without rare allele	93 (60.30)	61 (39.61)	0.3247
With rare allele	19 (70.37)	8 (29.63)	
rs9295689, n (%)			
Without rare allele	33 (73.33)	12 (26.67)	0.0680
With rare allele	79 (58.09)	57 (41.91)	
rs528431, n (%)			
Without rare allele	35 (55.56)	28 (44.44)	0.0996
With rare allele	67 (68.37)	31 (31.63)	

Table 6. Variable importance score in the 17th RF

Rank in the 17th RF	**Variable**	**VI**
1	Age at CP diagnosis	41.64
2	rs10818187	36.58
3	rs7894089	21.77

Table 7. Parameter estimates in logistic regression model for age at CP diagnosis, rs10818187, rs7894089 and interaction effects with the risk of having constant pain

Parameters in logistic regression	β	p-value
Intercept	-0.3822	0.7515
Age at CP diagnosis	-0.0142	0.5607
RS10818187	2.5436	0.0785
RS7894089	-2.6042	0.3456
Age at CP diagnosis*RS10818187	-0.0335	0.2532
Age at CP diagnosis*RS7894089	0.0405	0.4316
RS10818187*RS7894089	-0.8860	0.7839
Age at CP diagnosis* RS10818187*RS7894089	0.0116	0.8480

Table 8. Odd ratio in Logistic regression model for 10 years increase in age at CP diagnosis, rs10818187, rs7894089 with the risk of having constant pain

Parameters in logistic regression	**OR**	**95% CI**		**p-value**
10 years increase Age at CP diagnosis	0.776	0.631	0.961	0.0192
RS10818187	2.603	1.139	5.951	0.0233
RS7894089	0.394	0.174	0.896	0.0262

4.0 APPLICATION OF RANDOM FORESTS TO IDENTIFY RISK FACTORS ASSOCIATED WITH PROGRESSION FROM RECURRENT ACUTE PANCREATITIS TO CHRONIC PANCREATITIS

4.1 ABSTRACT

It remains to be understood why some patients with recurrent acute pancreatitis (RAP) develop chronic pancreatitis with progressive morphological changes of the pancreas, exocrine and endocrine dysfunction in a short time, whereas others have recurrent attacks of AP without development of structural changes or dysfunction of the pancreas over a long period of time. The underlying risk factors contributing to the progression from RAP to CP are still rarely described. It has been hypothesized that the mechanisms of progression from RAP to CP should include genetic, environmental, and other risk factors. A computational method - Random Forests (RF) analyses were used to identify important risk factors contributing to the progression from RAP to CP in patients from North American Pancreatic Study 2 (NAPS2). The results showed that smoking is the strongest risk factor associated with progression. CFTR gene is a potential risk factor for the progression in patients. The identification of these factors may be useful in clinical practice to predict early stage pancreatitis patients at risk for developing CP.

4.2 INTRODUCTION

Recurrent acute pancreatitis (RAP) and chronic Pancreatitis (CP) pancreatitis are two form of pancreatic diseases. RAP is defined as the presence of at least two episodes of acute pancreatitis (AP) in a patient without morphological damage of the pancreas. CP is characterized by persistent inflammation and irreversible morphological changes of the pancreas that often accompanied by variable pain, calcifications, necrosis, fatty replacement, fibrosis, scarring, exocrine and/or endocrine dysfunction, and other complications.[96]

Although AP, RAP, CP are defined as three different diseases, they are related because CP is often considered as the next step after RAP and RAP sometimes is a prerequisite for the development of CP which means RAP may be in the pathway from AP, RAP to CP.[97, 98]

In a study by Yadav et al, they found that the median time to the second attack for AP was 7.2 months after the first attack of AP and the proportion of patients with RAP who received a diagnosis of CP was 32.3% in their population after a median of 40 months following the first attack of AP. [3] Nojgaard et al in a study showed that CP developed within a mean interval of 3.5 years from the first attack of AP and the mortality rate for patients with CP with progression from AP was 2.7 times higher than in AP patients without chronic progression.[7]

It remains to be established why some patients with RAP have morphological changes of the pancreas, exocrine and endocrine dysfunction in a short time, e.g. during a period of 40 months after the first episode of AP, whereas others have recurrent attacks of AP without development of structural changes or dysfunction of the pancreas over a long period of time. The underlying risk factors contributing to the progression from RAP to CP are still rarely described. It has been hypothesized that the mechanisms of progression from RAP to CP should include genetic, environmental, and other risk factors.[2, 28]

Ammann et al conducted a sequential histological study of the pancreas in patients with alcoholic pancreatitis and concluded that there was pathway from AP to CP according to a suggested 'necrosis-fibrosis sequence' theory.[98, 99] The progression of alcoholic AP to advanced CP is determined primarily by two factors: the incidence and severity (mild or severe) of acute attacks of pancreatitis and the location of necrosis within the head of the pancreas.[98] However, This study was in alcoholic pancreatitis only without any evidence for other types of pancreatitis. In 1996, Whitcomb et al. identified gain-of-function mutations in the gene that encodes cationic trypsinogen (PRSS1) cause hereditary pancreatitis, a syndrome characterized by RAP and later CP.[42] They also showed that the majority of patients with hereditary pancreatitis recovered from AP through the normal healing process, whereas the rest of patients failed to recover and progressed to CP.[2, 39] Hereditary pancreatitis then has been emerged as an important pancreatic disease, which was a relatively rare but a breakthrough in understanding of RAP and CP.[39]

Currently, the knowledge of the relationship among AP, RAP and CP is not completely understood and previous findings stimulates researchers to pay more interested in identifying which patients with AP and RAP develop CP, understanding the underlying mechanisms involved so that the disease progression to CP could be prevented or early detected.

Whitcomb et al. defined a "sentinel acute pancreatitis event" (SAPE) hypothesis model that organizing factors associated with AP according to a hypothetical pathway that leads from the first attack of AP towards CP.[2, 30, 41] The first event of AP comes from factors that cause pancreatic injury and then trigger a series of events that potentially lead to CP. They proposed that a SAPE must occur for the development of CP. The events from the second attack may involve a variety of risk factors such as alcohol, smoking, and genetic mutations or other undiscovered factors affect the pancreas to cause various inflammation and inflammation-associated complications.[30] Therefore, current research efforts are focusing on identifying risk factors, mechanisms, and biomarkers of such a pathways progression to CP. The goal of this study was to identify important risk factors contributing to the progression from RAP to CP.

4.3 MATERIALS AND METOHDS

Study Population

Study subjects and data were derived from the North American Pancreatic Study 2 (NAPS2). The NAPS2 Program was designed to help researchers to understand the mechanisms leading to RAP, CP and their complications in human subjects. The overall objective was to ascertain several thousand patients with RAP, CP and controls with detailed risk assessment and deep phenotyping linked to biomarker and genetic information for cross sectional and observational follow-up studies. Enrollment was complete in September 2006 and there were a total of 1000 patients (540 subjects with CP, 460 with RAP) and 695 controls who completed consent forms and questionnaires and donated blood samples comprised the final dataset.[24]

In NAPS2, CP was confirmed by imaging studies and RAP was defined by the presence of two or more attacks of documented AP but with no imaging evidence of CP. The age of the diagnosis of first attack of AP and the age of diagnosis of CP were collected. The number of additional attacks of AP was also collected in RAP and CP patients.

In the current study, we defined two groups of patients. The RAP patient group was defined as participants who had documented the first AP attack more than 4 years before their age at enrollment of NAPS2 and had one or more additional episodes of AP attacks but had never been diagnosed as CP. The CP patient group was defined as participants who had been diagnosed as CP at enrollment of NAPS2 and had documented first AP attack with additional episodes of AP attacks within 4 years before their age at diagnosis of CP.

Demographic and Clinical Variables

A total of 36 demographic and clinical variables were extracted or computed from two comprehensive patient and physician questionnaires in NAPS2 study for the use for current study. Patient's gender, race, Ashkenazi Jewish heritage, family history of any pancreatic disease, age of the first AP diagnosis, the severity of the documented first AP attacks, drinking history (never drink, ever drink), alcohol consumption in the months before getting pancreatitis (abstainer, light drinker, moderate drinker, heavy drinker or very heavy drinker), smoking (non-smoker, former smoker, current smoker), amount of smoking (none, less than 1 pack per day, more than 1 pack per day), TIGAR-O etiologic risk factors including hyperlipidemia, hypercalcemia, chronic renal failure, medications, toxins, idiopathic factors, hereditary factors, Cationic Trypsinogen mutation, CFTR mutations, SPINK1 mutations, Alpha 1-antitrypsin deficiency, other genetic factors, isolated autoimmune pancreatitis (AIP) factor, Inflammatory bowel disease AIP factor, other AIP factors, postnecrotic, vascular diseases/ischemic, postirradiation, pancreas divisum, sphincter of oddi disorders, duct obstruction, preampullary duodenal wall cysts, posttraumatic pancreatic duct scar, intraductal papillary mucinous neoplasm (IPMN), pancreatic cancer, gallstones, other risk factors.

Genetic Variables

Genetic variants was extracted from patient blood samples using the Qiagen FlexiGene DNA Kit (Qiagen Inc, Valencia, CA, USA).[93] Genotyping was performed at the University of Pittsburgh Genomics and Proteomics Core Laboratories using the MassARRAY iPLEX GOLD

(Sequenom, Inc, San Diego, CA, USA). A total of 71 SNPs from a group of gene loci including PRSS1, PRSS2, SPINK1, CFTR, CASR, CTRC, CLDN2 were genotyped because they were found to be associated with the risk of pancreatitis reported from a number of previous studies.[30]

4.4 ANALYSIS

In this study, all genetic variants were analyzed as binary categorical variables with common allele homozygotes in SNPs were treated as reference and hetetozygote and rare allele homozygote genotypes were assigned as risk genetic variants. Distribution of all 107 candidate predictor variables were reported as means and standard deviation for quantitative data and were compared between two study groups using Student's t-test or Mann–Whitney U test. Qualitative scales were reported as counts and proportions and compared using chi-squared tests or Fisher's exact tests.

Random Forest (RF) algorithm, an ensemble of hundreds to thousands of decision tree, was used to help to address the study purpose by its ability of dealing with large number of candidate predictor variables. RF is an ensemble of large amount of individual decision trees by randomly selecting a bootstrap subset sample of two-thirds of whole sample per tree and randomly selecting a subset of all predictor variables at each node of the tree. At each node, RF selects the predictor variable that best splits data into two child nodes.[61] This process allows for all demographic, clinical and genetic factors to work simultaneously in predicting the progression from RAP to CP. RF determines prediction error using the out of bag sample (OOB), those one-thirds of subjects not randomly selected to build a given individual tree. RF works through this process of selecting bootstrap samples to build the tree and using the OOB samples to determine error and variable importance. The variable importance was evaluated using permutation importance. The given variable was randomly permuted in the OOB sample for the tree and new estimate of OOB error was calculated. All candidate predictor variables in the model were ranked based on the difference between this estimate and the original OOB error. The larger increase in prediction error indicates a stronger association between a given predictor variable and the progression from RAP to CP because random permutation destroys their original relationship in OOB sample and lead to worse prediction error.

To identify important risk factors associated with the disease progression from RAP to CP, we used an iterative RF analysis framework to identify important risk factors for further examination regarding to the predictive ability for the disease progression outcome in those patients with RAP. To build the 1st RF, all candidate demographic, clinical, and genetic variable were inputted simultaneously in the 1st RF analysis without any variable pre-selection. The RF analysis then generated a ranking of variable importance for all predictors according to the importance score. After evaluation of all variables in the ranking, 20% of variables with the lowest importance score was dropped. A second RF was built using the rest of 80% variables follow the same algorithm to create new variable importance score and ranking.[87] This iteration process was repeated to build up a series of RF models and the last RF analysis was implemented using two variables left. After building all RF, The OOB error from all RF were evaluated. In steading of selecting the best model, our purpose is to use RF analysis to help us to select a relatively less complex and parsimonious model with minimum number of predictor variables which performs not significantly worse than the best model regarding to the classification accuracy (OOB error). Therefore, if there is a RF with smaller number of variables and the OOB error is not significant higher than the RF with the lowest OOB error, the variables in this RF model were analyzed using logistic regression for further evaluation. The area under the receiver operation characteristic (ROC) curves (AUC) from logistic regression model were also generated to examine model prediction abilities. If AUC exceeds the critical value of 0.7, the model is considered with high predictive power. To compare the variables obtained from RF analysis to the variables obtained from the traditional approach of multivariable modeling, a logistic regression based on stepwise variable selection procedure for all 107 candidate variable was performed.

RF analyses were carried out using Salford Predictive Modeler, version 7.0 (Salford Systems, San Diego, CA) with the recommended parameter settings: the number of trees was set to ntree=1000; the number of variables to test at each node was set to mtry=√N (N=Total number of variables used in each RF). The OOB error were generated for each RF analysis and the variable importance score was computed by permutation importance method.

Multivariable logistic regression model was built using the most promising variables found in RF analysis and odds ratio (OR) with 95% confidence interval (CI) were calculated. All descriptive and logistic fregression analyses were carried out using SAS, version 9.3 (SAS, Inc., Cary, NC).

4.5 RESULTS

In NAPS2 study, there were 460 RAP and 540 CP patients enrolled.[24] Among them, 95 RAP patients and 92 CP patients were eligible for analysis for the current study regarding to criteria described above. Due to availability of genetic variants data, there were 80 patients enrolled into

RAP group and 65 patients enrolled into CP group for analysis. Total population in the current study is N=145.

The distribution of all demographic and clinical candidate variables were presented in Table 9. A total of 107 candidate variables were selected to train the 1st RF model including 36 demographic or clinical predictor variables, and 71 pancreatitis-related genetic variants. The OOB error rate of the 1st RF was 28.97% which means it had an overall accuracy of 71.0% in predicting the patient disease progression from RAP to CP. Among all 107 variables, 53 of them had zero of negative variable importance score and 54 of them had variable importance score greater than zero. The ranking of all 107 variables importance was generated based on their importance score. After the 1st iteration of RF, the 2nd RF model was built by dropping 20% (n=21) of all variables with lowest importance score in the 1st RF. The 2nd RF model included 86 candidate variables and a new variable importance score was computed. The OOB error rate of the 2nd RF also equaled to 28.97%. Using the same iteration procedure, a total of 18 RF models were analyzed and the last model was built by 2 variables only. All OOB error rates and number of variable used in each RF analysis are listed in Table 10 and the OOB error rate curve from the 1st RF to the 18th RF is graphed in Figure 6.

From the summary of all RF analysis, both the 8th RF model with 22 variables and the 9th RF model with 18 variables has the lowest OOB error rate (22.76%). The rankings of variable importance score for these two models are listed in Table 11. Smoking status which categorized as non-smoker, past smoker, and current smoker, was the most important risk factor influencing the progression consequence from RAP to CP. The variable importance scores for smoking status were dramatically higher in both models (48.02, 42.79) compared to the second important risk factor the patient drinking category before getting pancreatitis (12.8, 14.67). Other top important risk factors with variable importance score greater than 3 calculated by permutation imputation in both RF models included amount of smoking, patient age at the 1st AP attack, and two genetic variants rs213938 and rs213945. The results indicated that smoking status was the most important predictor varaibles

The OOB error rate from 9th to 16th RF analyses were fluctuated between 22.76% and 28.28% and then the OOB error rate of the 17th RF built by 3 variables significantly increased to 33.13% and continuously increased to 33.79% in the 18th model, the last RF model built by 2 variables, smoking status and drinking category. The results indicated that the 4th important predictor variable, which was ranked in the last in the 16th RF model should be included. The variable importance score and distribution in the 16th RF analysis are listed in Table 12. From the 16th RF model, smoking status, drinking category before getting pancreatitis, amount of smoking, and genetic variants rs213938 were considered to be a set of important risk factors associated with the development from RAP to CP.

A multivariable logistic regression model was analyzed using 4 important predictor variables, smoking status, drinking category before getting pancreatitis, amount of smoking, and rs213938. The result showed that only amount of smoking was not significant in the multivariable regression model. (p=0.38) The correlation between smoking status and amount of smoking was evaluated and they were highly correlated (r=0.83, p<0.001). This result indicates that the ability of RF to identify relevant predictor variables when they have similar impact on outcome of interest and are highly correlated, e.g. genetic variants in same gene loci. From Table 11, the importance scores for SNPs rs213938 and rs213945 could also reflect this scenario in which rs213945 always followed rs213938 in the rankings with similar importance score because they are highly correlated (r=0.98, p<0.001) and in the same gene CFTR.

To identify a small set of predictor variables that could be used for clinical purpose and avoid violation of multicollinearity assumption in statistical regression model, a new multivariable logistic regression model was built using the other three important predictor variables, smoking status, drinking category before getting pancreatitis, and rs213938. The full model with interactions was not fit since there was no interaction existing amongst three variables. The final model with three main effects showed that current smokers (OR=20.78 CI: 6.30-68.53) and genetic variants rs213938 (OR=2.98 CI: 1.22-7.31) were significantly associated with a higher risk for development of CP from RAP. Patients with alcohol consumption history showed a tendency towards a lower risk as to the progression, which was significantly in light and heavy drinkers (Table 13). The AUC of the logistic regression is 0.85 indicate a high prediction accuracy.

In the logistic regression model with a stepwise variable selection procedure (P < 0.30 for entry in the model and P > 0.35 for removing the variables), there were 29 of all 145 subjects deleted in the modeling because of the missing values. The variables selected by the stepwise logistic regression were listed in Table 14.

4.6 DISCUSSION

This study examined a large number of demographic, clinical and genetic variables using a novel computational method of RF followed by traditional statistical regression models to identify combinations among these predictors in the relationship with the disease progression from RAP to CP. Using a sequence of RF models, we successfully identified a group of important risk factors with the most contribution to the disease outcome and examined further by multivariable logistic regression.

The results from this study demonstrated that smoking is the most important risk factor to the development of CP in RAP patients. The variable importance score of smoking status from RF analysis was significantly higher than any other variables. In multivariable logistic regression model, smoking status was also significantly associated with disease progression and current smokers was identified as the most contributor in the progression from RAP to CP (adjusted OR=20.78). However, there was no increased risk in past smokers compared to non-smokers. This result confirmed previous findings that smoking is an important risk factors for CP.[100]

The alcohol consumption is also significantly related to the progression from RAP to CP. Alcohol drinking has been widely recognized as a major risk factor for pancreatic diseases. However, a recent study reported that alcohol intake is not significantly associated with the progression from RAP to CP.[7] In our study, compared to abstainers, light, moderate and heavy drinkers tend to slightly decrease the risk of developing CP from RAP after adjusting for other variables. The higher risk in abstainers might be due to other etiologic causes and the small sample size in each category. Therefore, the effect of alcohol consumption in the disease progression should be further investigated.

The SNP rs213938, which is in CFTR (cystic fibrosis transmembrane conductance regulator) gene was significantly associated with higher risk of developing CP in RAP patients. From the results of RF analyses, a group of SNPs in CFTR gene were also identified that ranked at the top of variable importance score. In the 9th RF model which had the best prediction accuracy, 6 of 10 all candidate SNPs are located in CFTR gene (rs213938, rs213945, rs10487368, rs2237723, rs17449197, rs17451754). The CFTR gene mutations have been suggested as a pancreatitis related genes that cause cystic fibrosis, an autosomal recessive disease characterized by the development of CP because CFTR gene were found to linked with an extra–acinar cell mechanism to eliminate trypsin by flushing it out of the pancreatic duct and into the duodenum. Therefore, mutations in CFTR gene reduce fluid secretion and trypsinogen/trypsin wash-out.[28, 30]

In stepwise logistic regression model, there were 29 incomplete subjects were listwise deleted due to missing values resulted in smaller of sample size and potential of loss of power to detect associations. However, in RF analysis, the missing values were automatically imputed using the most frequent –non-missing value or median. Missing value is frequently happened in high-dimensional data. The current study demonstrated the advantage of RF to handle missing value and therefore successfully identify important predictor variables for modeling purpose.

The key findings from this study indicate that smoking, alcohol drinking, and genetic variants in CFTR are important determinants for the development of CP in patients with RAP in the short period of time. Analysis of the NAPS2 data indicates that using patients smoking status, drinking category, and SNP rs213938 could be potentially used to make a disease progression prediction in RAP patients.

4.7 LIMITATIONS

There are some limitations in current study. First, the data for analysis was from a cross-sectional study thus the patients answering about past AP events may be not accurate. Second, the results from RF analyses did not provide visual information about interaction effects among predictor variables. The variables in the ranking of variable importance may have interaction effects with others but they were not specified. Third, the variable importance ranking in RF does not have a threshold to show evaluate the significance of importance which made them hard to interpret from a statistical perspective and make the decision on variable selection difficultly. Finally, the sample size was relatively small and therefore the association between all candidate predictor variables and disease progression from RAP to CP may not be truly reflected.

4.8 CONCLUSIONS

We combined RF analyses and logistic regression models to analyze NAPS2 data, and was able to identify important risk factors associated with disease progression from RAP to CP within a time frame of 4 years. Smoking is the strongest risk factor associated with progression. CFTR gene is a potential risk factor for the progression in patients. The identification of these factors may be useful in clinical practice to predict early stage pancreatitis patients at risk for developing CP. Certainly, additional research is needed to evaluate the findings from this study.

4.9 TABLES AND FIGURES

Table 9. Baseline characteristics of study population

Variables		RAP N=80 (55.17%)	CP N=65 (44.83%)	p-value
Gender, n (%)				
	Male	35 (50.00)	35 (50.00)	0.2263
	Female	45 (60.00)	30 (40.00)	
Race, n (%)				
	White	80 (55.17)	65 (44.83)	.
	Other	0 (0)	0 (0)	
Jewish heritage, n (%)				
	No	71 (52.99)	63 (47.01)	0.3218
	One parent	2 (100)	0 (0)	
	Both Parents	5 (83.33)	1 (16.67)	
	Unknown	1 (50.00)	1 (50.00)	
Drinking category in period before pancreatitis, n (%)				
	Abstainer	14 (43.75)	18 (56.52)	0.0033
	Light	37 (72.55)	14 (27.45)	
	Moderate	12 (50.00)	12 (50.00)	
	Heavy	10 (62.50)	6 (37.50)	
	Very heavy	5 (25.00)	15 (75.00)	
Ever Drink, n (%)				
	No	14 (43.75)	18 (56.52)	0.1521
	Yes	65 (58.04)	47 (41.96)	
Smoking, n (%)				
	Never	41 (75.93)	13 (24.07)	<0.0001
	Past	30 (65.22)	16 (34.78)	
	Current	8 (18.18)	36 (81.82)	
Amount of smoking, n (%)				
	Never	41 (75.93)	13 (24.07)	0.0011
	<1 pack/day	18 (51.43)	17 (48.57)	
	≥1 pack/day	29 (40.82)	29 (59.18)	
Family history of pancreatitis, n (%)				
	No	62 (56.53)	48 (43.64)	0.7905
	Yes	15 (53.57)	13 (46.43)	
Age at AP diagnosis, mean (SD)		41.45 (16.59)	47.50 (15.63)	0.0266
AP severity, n (%)				

Table 9 Continued

	Mild	37 (48.05)	40 (51.95)	0.1861
	Moderate	8 (47.06)	9 (52.94)	
	Severe	29 (64.44)	16 (35.56)	
Etiologies				
Hyperlipidemia, n (%)				
	No	64 (52.46)	58 (47.54)	0.1302
	Yes	16 (69.57)	7 (30.43)	
Hypercalcemia, n (%)				
	No	79 (56.03)	62 (43.97)	0.1978
	Yes	1 (25.00)	3 (75.00)	
Medications, n (%)				
	No	72 (53.73)	62 (46.27)	0.2233
	Yes	8 (72.73)	3 (27.27)	
Chronic Renal Failure, n (%)				
	No	79 (54.86)	65 (45.14)	0.5517
	Yes	1 (100)	0 (0)	
Toxins, n (%)				
	No	80 (55.17)	65 (44.83)	.
	Yes	0 (0)	0 (0)	
Idiopathic factors, n (%)				
	No	38 (52.05)	35 (47.95)	0.4472
	Yes	42 (58.33)	30 (41.67)	
Genetic factors, n (%)				
	No	74 (54.01)	63 (45.99)	0.1570
	Yes	6 (75.00)	2 (25.00)	
CFTR mutation, n (%)				
	No	73 (53.28)	64 (46.72)	0.0586
	Yes	7 (87.50)	1 (12.50)	
SPINK1 mutation, n (%)				
	No	79 (54.86)	65 (45.14)	0.5517
	Yes	1 (100)	0 (0)	
Alpha 1-antitrypsin deficiency, n (%)				
	No	80 (55.17)	65 (44.83)	.
	Yes	0 (0)	0 (0)	
Other genetic factors, n (%)				
	No	78 (54.93)	64 (45.07)	0.4127
	Yes	2 (66.67)	1 (33.33)	
IBD Autoimmune, n (%)				
	No	78 (54.93)	64 (45.07)	0.4127
	Yes	2 (66.67)	1 (33.33)	
Isolated Autoimmune, n (%)				
	No	80 (55.56)	64 (44.44)	0.4483
	Yes	0 (0)	1 (100)	

Table 9 Continued

Other AIP, n (%)				
	No	80 (55.17)	65 (44.83)	.
	Yes	0 (0)	0 (0)	
Postnecrotic, n (%)				
	No	79 (56.43)	61 (43.57)	0.1087
	Yes	1 (20.00)	4 (80.00)	
Postirradiation, n (%)				
	No	80 (55.56)	64 (44.44)	0.4483
	Yes	0 (0)	1 (100)	
Vascular Diseases / ischemic, n (%)				
	No	80 (55.17)	65 (44.83)	.
	Yes	0 (0)	0 (0)	
Pancreas Divisum, n (%)				
	No	74 (57.81)	54 (42.19)	0.0794
	Yes	6 (35.29)	11 (64.71)	
Sphincter of Oddi disorders, n (%)				
	No	64 (52.89)	57 (47.11)	0.2152
	Yes	16 (66.67)	8 (33.33)	
Posttraumatic pancreatic stricture, n (%)				
	No	80 (55.17)	65 (44.83)	.
	Yes	0 (0)	0 (0)	
Preampullary duodenal diverticulum, n (%)				
	No	80 (55.17)	65 (44.83)	.
	Yes	0 (0)	0 (0)	
Duct obstruction, n (%)				
	No	75 (55.56)	60 (44.44)	0.2409
	Yes	5 (50.00)	5 (50.00)	
Pancreatic cancer, n (%)				
	No	79 (54.86)	65 (45.14)	0.5517
	Yes	1 (100)	0 (0)	
IPMN, n (%)				
	No	78 (54.44)	65 (45.45)	0.3027
	Yes	2 (100)	0 (0)	
Gallstones, n (%)				
	No	75 (56.82)	57 (43.18)	0.2042
	Yes	5 (38.46)	8 (61.54)	
Miscellaneous factors, n (%)				
	No	71 (52.21)	65 (47.79)	0.0038
	Yes	9 (100)	0 (0)	

Table 10. Number of variables and OOB error rate of each RF analysis

RF analysis	Number of variables	OOB error (%)
1	107	28.97
2	86	28.97
3	69	28.28
4	55	28.28
5	44	26.21
6	35	24.14
7	28	25.52
8	22	22.76
9	18	22.76
10	14	24.14
11	11	26.90
12	9	28.28
13	7	26.21
14	6	24.14
15	5	26.21
16	4	24.14
17	3	33.10
18	2	33.80

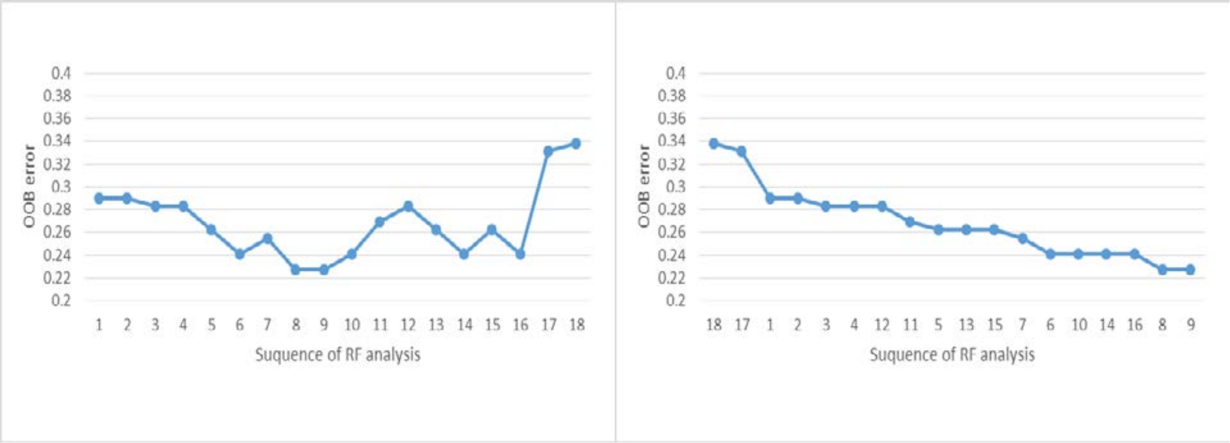


Figure 6. OOB error rate curve for all RF analyses

Table 11. Variable importance score in the 8th and 9th RF model

	8th RF		9th RF	
Rank	**Variables**	**VI score**	**Variable**	**VI score**
1	Smoking	48.02	Smoking	42.79
2	Drinking Category	12.8	Drinking Category	14.67
3	Amount of Smoking	8.64	Amount of Smoking	8.99
4	rs213938	4.45	Age at 1st AP diagnosis	4.18
5	rs213945	3.46	rs213938	3.91
6	Age at 1st AP diagnosis	3.02	rs213945	3.84
7	Pancreas Divisum Risk Factor	2.42	Pancreas Divisum Risk Factor	2.69
8	Hyperlipidemia Risk Factor	2.03	Hyperlipidemia Risk Factor	2.61
9	Drinking History	1.72	rs17250717	2.07
10	rs17250717	1.66	rs10273639	2.02
11	rs10273639	1.59	Other Risk Factor	1.97
12	Other Risk Factor	1.53	rs10487368	1.93
13	rs2237723	1.4	Drinking History	1.89
14	rs1393198	1.38	rs10934578	1.87
15	rs10934578	1.26	rs1393198	1.47
16	rs17451754	1.1	rs2237723	1.27
17	rs10487368	0.93	rs17449197	0.98
18	rs17449197	0.76	rs17451754	0.86
19	rs12008279	0.74		
20	Ashkenazi Jewish heritage	0.63		
21	rs2202127	0.24		
22	Medications Risk Factor	0.23		

Table 12. Importance score and distribution of predictor variables in the 16th RF

Rank	Variables	VI	All N=145	RAP N=80 (55.17%)	CP N=65 (44.83%)	p-value
1	**Smoking, n (%)**	61.48				
	Never		54	41 (75.93)	13 (24.07)	<0.001
	Past		46	30 (65.22)	16 (34.78)	
	Current		44	8 (18.18)	36 (81.82)	
2	**Drinking category in period before pancreatitis, n (%)**	14.22				
	Abstainer		32	14 (43.75)	18 (56.52)	0.003
	Light		51	37 (72.55)	14 (27.45)	
	Moderate		24	12 (50.00)	12 (50.00)	
	Heavy		16	10 (62.50)	6 (37.50)	
	Very heavy		20	5 (25.00)	15 (75.00)	
3	**Amount of smoking, n (%)**	13.91				
	Never		54	41 (75.93)	13 (24.07)	0.001
	<1 pack/day		35	18 (51.43)	17 (48.57)	
	≥1 pack/day		49	29 (40.82)	29 (59.18)	
4	**rs213938, n (%)**	10.39				
	Without rare allele		99	61 (61.62)	38 (38.38)	0.022
	With rare allele		46	19 (41.30)	27 (58.70)	

Table 13. Smoking, Drinking Category, rs213938 and the risk of progression from RAP to CP

Parameters in logistic regression	OR	95% CI		p-value
Smoking (vs Never)				<0.001
Past	2.66	0.93	7.61	
Current	20.78	6.30	68.53	
Drinking Category (vs Abstained)				0.009
Light	0.16	0.05	0.50	
Moderate	0.27	0.07	1.08	
Heavy	0.14	0.03	0.72	
Very heavy	0.80	0.17	3.70	
rs213938	2.98	1.22	7.31	0.017

Table 14. Parameter estimates in logistic regression model using stepwise variable selection

Parameters	β	p-value
Intercept	-0.5735	0.2565
Past smoker	1.2203	0.0772
Current smoker	4.2855	<.0001
Medication risk factors	-3.6258	0.0017
Ever drinker	-2.1132	0.0026
Postnecrotic risk factors	2.9783	0.0207
Hypercalcemia risk factors	4.5751	0.0354
rs13221882	1.9544	0.0134

5.0 DISCUSSION

5.1 SUMMARY OF FINDINGS

This dissertation introduced the principles of random forests (RF), reviewed the RF methodology, described its advantages and limitations, and provided epidemiological examples to explore the use of RF to analyze high dimensional data in epidemiology. By analyzing large scale epidemiological data combining demographical, clinical and genetic variables from NAPS2-CV and NAPS2 study, the dissertation developed a framework of combining RF analyses with traditional statistical analyses to investigate important risk factors associated with different pain patterns in patients with CP and disease progression from RAP to CP. The dissertation provided a novel data analytic strategy to support big data analysis in epidemiological researches.

The first paper of this dissertation was a review paper focused on RF methodology. RF has been becoming a popular non-parametric algorithm in computational method and used in many scientific areas in the context of big data era. RF is an ensemble of CART-like individual decision trees introduced by Leo Breiman in 2001. RF could handle both classification and regression problems to explore data structure and hidden information in high dimensional data.

RF modeling is featured by a two-way randomness through random sub-sampling and random variable selection in tree-building process and has been tested with higher prediction accuracy compared to individual decision tree method such as CART. RF deals with correlated predictor variables and integrates complex interaction effects during modeling process. The most attracting feature from RF is that it could evaluate the entire effects of all predictor variables on outcome variable and produce a ranking of variable importance scores regarding to their discrimination ability on the outcome. RF has been applied in a variety of epidemiological studies especially in genetic epidemiology to help identify unknown relationship in high dimensional data.

The second paper focused on identifying important risk factors that associated with different temporal pain patterns in CP patients. The study population were from NAPS2-CV study, 112 patients with intermittent pain and 69 patients with constant pain. The total number of candidate predictor variables were 117 including demographical, clinical, genetic markers. A framework of RF iterations combined with logistic regression model were used to identify important factors associated with different temporal pain patterns in CP patients.

A reduced RF model identified patients' age at CP diagnosis, rs10818187, and rs7894089 are the most three important factors to determine temporal pain patterns in CP patients. The multivariable logistic regression indicated that the increase of age at CP diagnosis and the SNP rs7894089 are significantly associated with the lower risk for constant pain whereas the SNP rs10818187 are significantly associated with the lower risk for constant pain. No interaction effects were significant among age at CP diagnosis, rs7894089, and rs10818187. The identification of these factors may be useful in clinical practice to identify individuals at risk for the constant pain at the early stage in CP.

A larger RF model with 24 predictor variables yielded the lowest OOB error indicating that all 24 variables were potentially related to the determination of different temporal pain patterns in CP patients. The results indicated that the underlying mechanisms contributing to pain patterns in CP patients were complex.

The third paper focused on using RF iterations and logistic regression model to identify important risk factors contributing to disease progression from RAP to CP in a period of 4 years after the 1st attack of AP. This study found that smoking was the most important risk factor to the progression compared to any other risk factors. Gene mutations in CFTR gene also contributed a lot in the progression. The identification of these factors may be useful in clinical practice to predict early stage pancreatitis patients at risk for developing CP. The result from RF analysis also showed that a group of 18 to 22 risk factors were related to the progression from RAP to CP indicating that the causes of disease progression were complex and need further investigation.

5.2 STRENGTHS

There are several strengths in this current dissertation research. The review of RF described the methodology of RF and its implementation in data analysis for epidemiological researches. The first paper was a beneficiary of expert review, which provide examples of RF applications in epidemiology and possible directions for future researches.

The development of an analytic framework of RF analyses and traditional statistical analyses in large pancreatitis studies allowed for the examination of large amount of predictor variables simultaneously and their influence on different pain patterns in patients with CP and disease progression from RAP to CP in a multivariable way.

5.3 LIMITATIONS

In application of RF analysis, the interaction effects by predictor variables were not specified because RF did not provide information on how they interact with each other. It made the difficulty in interpreting results.

Second, RF analysis did not provide a threshold to define which predictor was significant and/or should be selected for further investigation. Variable importance score always provided a ranking which made them hard to interpret from a statistical perspective and make the decision on variable selection.

Third, there was no formal way to select the optimal RF model, the smaller RF model was selected as the optimal model followed by statistical analysis based on its simplicity for clinical purpose.

Fourth, the number of patients who chose to participate in current studies were small. This small sample limits the power to detect true differences between groups and it calls into question to representativeness of the sample.

5.4 PUBLIC HEALTH SIGNIFICANCE

CP is a major burden of gastrointestinal disease in the United States accounts for significant healthcare costs to the society. Abdominal pain is the most common symptom in CP patients and the relapse or persistence of pain is challenging medical practice. It has been proposed that a combination of genetic, environmental, and metabolic risk factors contribute to the development of CP and pain patterns in CP patients. This dissertation provided a data analysis framework utilizing the novel computational method RF to better understand complex mechanisms in the development of CP and pain patterns in CP patients. RF is suitable for analyzing high dimensional data for epidemiological studies in the context of the big data era.

5.5 FUTURE STUDIES

Future research on should continue to concentrate on understanding complex effects of genetic variations and environmental factors in CP. The capability of RF should be further developed in analyzing large epidemiological studies. RF analysis are needed to fit in the statistical framework to be better understood from a statistical point of view.

BIBLIOGRAPHY

1. Go, V.L., et al., The pancreas. 1993: Raven Press New York.

2. Whitcomb, D.C., Mechanisms of disease: Advances in understanding the mechanisms leading to chronic pancreatitis. Nat Clin Pract Gastroenterol Hepatol, 2004. 1(1): p. 46-52.

3. Yadav, D., M. O'Connell, and G.I. Papachristou, Natural history following the first attack of acute pancreatitis. Am J Gastroenterol, 2012. 107(7): p. 1096-103.

4. Everhart, J.E. and C.E. Ruhl, Burden of Digestive Diseases in the United States Part III: Liver, Biliary Tract, and Pancreas. Gastroenterology, 2009. 136(4): p. 1134-1144.

5. Pasricha, P.J., Unraveling the mystery of pain in chronic pancreatitis. Nat Rev Gastroenterol Hepatol, 2012. 9(3): p. 140-51.

6. Moores, K., et al., A systematic review of validated methods for identifying pancreatitis using administrative data. Pharmacoepidemiol Drug Saf, 2012. 21 Suppl 1: p. 194-202.

7. Nøjgaard, C., et al., Progression from acute to chronic pancreatitis: prognostic factors, mortality, and natural course. Pancreas, 2011. 40(8): p. 1195-1200.

8. Jha, R.K., et al., Acute pancreatitis: a literature review. Med Sci Monit, 2009. 15(7): p. RA147-56.

9. Fasanella, K.E., et al., Pain in chronic pancreatitis and pancreatic cancer. Gastroenterol Clin North Am, 2007. 36(2): p. 335-64, ix.

10. Frey, C.F., et al., The incidence and case-fatality rates of acute biliary, alcoholic, and idiopathic pancreatitis in California, 1994-2001. Pancreas, 2006. 33(4): p. 336-344.

11. Yadav, D. and A.B. Lowenfels, The epidemiology of pancreatitis and pancreatic cancer. Gastroenterology, 2013. 144(6): p. 1252-61.

12. Yadav, D. and A.B. Lowenfels, Trends in the epidemiology of the first attack of acute pancreatitis: a systematic review. Pancreas, 2006. 33(4): p. 323-330.

13. Dzakovic, A. and R. Superina, Acute and chronic pancreatitis: surgical management. Semin Pediatr Surg, 2012. 21(3): p. 266-71.

14. Makhija, R. and A.N. Kingsnorth, Cytokine storm in acute pancreatitis. Journal of hepato-biliary-pancreatic surgery, 2002. 9(4): p. 401-410.

15. Bhatia, M., et al., Pathophysiology of acute pancreatitis. Pancreatology, 2005. 5(2): p. 132-144.

16. Johnson, C. and M. Abu-Hilal, Persistent organ failure during the first week as a marker of fatal outcome in acute pancreatitis. Gut, 2004. 53(9): p. 1340-1344.

17. Somogyi, L., S.P. Martin, and C.D. Ulrich, Recurrent acute pancreatitis. Current Treatment Options in Gastroenterology, 2001. 4: p. 361-368.

18. Sand, J. and I. Nordback, Acute pancreatitis: risk of recurrence and late consequences of the disease. Nature Reviews Gastroenterology and Hepatology, 2009. 6(8): p. 470-477.

19. Levy, M.J. and J.E. Geenen, Idiopathic acute recurrent pancreatitis. The American journal of gastroenterology, 2001. 96(9): p. 2540-2555.

20. Yadav, D. and D.C. Whitcomb, The role of alcohol and smoking in pancreatitis. Nat Rev Gastroenterol Hepatol, 2010. 7(3): p. 131-45.

21. Etemad, B. and D.C. Whitcomb, Chronic Pancreatitis: Diagnosis, Classification, and New Genetic Developments. Gastroenterology, 2001. 120(3): p. 682-707.

22. Braganza, J.M., et al., Chronic pancreatitis. The Lancet, 2011. 377(9772): p. 1184-1197.

23. Otsuki, M., Chronic pancreatitis in Japan: epidemiology, prognosis, diagnostic criteria, and future problems. Journal of gastroenterology, 2003. 38(4): p. 315-326.

24. Whitcomb, D.C., et al., Multicenter approach to recurrent acute and chronic pancreatitis in the United States: the North American Pancreatitis Study 2 (NAPS2). Pancreatology, 2008. 8(4-5): p. 520-31.

25. Frulloni, L., et al., Chronic pancreatitis: report from a multicenter Italian survey (PanCroInfAISP) on 893 patients. Digestive and Liver Disease, 2009. 41(4): p. 311-317.

26. Chen, J.M. and C. Ferec, Chronic pancreatitis: genetics and pathogenesis. Annu Rev Genomics Hum Genet, 2009. 10: p. 63-87.

27. LaRusch, J. and D.C. Whitcomb, Genetics of pancreatitis. Curr Opin Gastroenterol, 2011. 27(5): p. 467-74.

28. Whitcomb, D.C., Genetic aspects of pancreatitis. Annu Rev Med, 2010. 61: p. 413-24.

29. Whitcomb, D.C., Framework for interpretation of genetic variations in pancreatitis patients. Front Physiol, 2012. 3: p. 440.

30. Whitcomb, D.C., Genetic risk factors for pancreatic disorders. Gastroenterology, 2013. 144(6): p. 1292-302.

31. Derikx, M.H. and J.P. Drenth, Genetic factors in chronic pancreatitis; implications for diagnosis, management and prognosis. Best Pract Res Clin Gastroenterol, 2010. 24(3): p. 251-70.

32. Mullady, D.K., et al., Type of pain, pain-associated complications, quality of life, disability and resource utilisation in chronic pancreatitis: a prospective cohort study. Gut, 2011. 60(1): p. 77-84.

33. Sakorafas, G.H., A.G. Tsiotou, and G. Peros, Mechanisms and natural history of pain in chronic pancreatitis: a surgical perspective. Journal of clinical gastroenterology, 2007. 41(7): p. 689-699.

34. Lieb, J.G., 2nd and C.E. Forsmark, Review article: pain and chronic pancreatitis. Aliment Pharmacol Ther, 2009. 29(7): p. 706-19.

35. Andren-Sandberg, A., D. Hoem, and H. Gislason, Pain management in chronic pancreatitis. Eur J Gastroenterol Hepatol, 2002. 14(9): p. 957-70.

36. Demir, I.E., et al., Pain mechanisms in chronic pancreatitis: of a master and his fire. Langenbecks Arch Surg, 2011. 396(2): p. 151-60.

37. Ingemar Ihse, M., M. Kurt Borch, and M. Jörgen Larsson, Chronic pancreatitis: results of operations for relief of pain. World journal of surgery, 1990. 14(1): p. 53-58.

38. Brown, A., et al., Does pancreatic enzyme supplementation reduce pain in patients with chronic pancreatitis: a meta-analysis. The American journal of gastroenterology, 1997. 92(11): p. 2032-2035.

39. Whitcomb, D., Hereditary pancreatitis: new insights into acute and chronic pancreatitis. Gut, 1999. 45(3): p. 317-322.

40. Whitcomb, D.C., et al., Common genetic variants in the CLDN2 and PRSS1-PRSS2 loci alter risk for alcohol-related and sporadic pancreatitis. Nat Genet, 2012. 44(12): p. 1349-54.

41. Whitcomb, D.C., Value of genetic testing in the management of pancreatitis. Gut, 2004. 53(11): p. 1710-7.

42. Whitcomb, D.C., et al., Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. Nat Genet, 1996. 14(2): p. 141-5.

43. Touw, W.G., et al., Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform, 2013. 14(3): p. 315-26.

44. Chen, X., M. Wang, and H. Zhang, The use of classification trees for bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov, 2011. 1(1): p. 55-63.

45. Kawaguchi, A. Variable Ranking by Random Forests Model for Genome-Wide Association Study. in Proceedings of the International MultiConference of Engineers and Computer Scientists. 2012.

46. Szymczak, S., et al., Machine learning in genome-wide association studies. Genet Epidemiol, 2009. 33 Suppl 1: p. S51-7.

47. Holzinger, E., et al. Athena: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. Pacific Symposium on Biocomputing, 2012. World Scientific.

48. Chen, X. and H. Ishwaran, Random forests for genomic data analysis. Genomics, 2012. 99(6): p. 323-9.

49. Strobl, C., J. Malley, and G. Tutz, An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods, 2009. 14(4): p. 323-48.

50. Winham, S., et al., SNP interaction detection with Random Forests in high-dimensional genetic data. BMC bioinformatics, 2012. 13(1): p. 164.

51. Jiang, R., et al., A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinformatics, 2009. 10 Suppl 1: p. S65.

52. Yoo, W., et al., A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions. International journal of applied science and technology, 2012. 2(7): p. 268.

53. De Lobel, L., et al., A screening methodology based on Random Forests to improve the detection of gene-gene interactions. Eur J Hum Genet, 2010. 18(10): p. 1127-32.

54. Hapfelmeier, A. and K. Ulm, A new variable selection approach using Random Forests. Computational Statistics & Data Analysis, 2013. 60: p. 50-69.

55. Olshen, L.B.J.F.R. and C.J. Stone, Classification and regression trees. Wadsworth International Group, 1984.

56. Quinlan, J.R., Induction of decision trees. Machine learning, 1986. 1(1): p. 81-106.

57. Quinlan, J.R., C4. 5: programs for machine learning. Vol. 1. 1993: Morgan kaufmann.

58. Lemon, S.C., et al., Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of Behavioral Medicine, 2003. 26(3): p. 172-181.

59. Cutler, A., D.R. Cutler, and J.R. Stevens, Tree-based methods, in High-Dimensional Data Analysis in Cancer Research. 2009, Springer. p. 1-19.

60. Steinberg, D. and P. Colla, CART: classification and regression trees. The Top Ten Algorithms in Data Mining, 2009: p. 179-201.

61. Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.

62. Selvin, S., Statistical analysis of epidemiologic data. 2004: Oxford University Press.

63. Howe, D., et al., Big data: The future of biocuration. Nature, 2008. 455(7209): p. 47-50.

64. Podgorelec, V., et al., Decision trees: an overview and their use in medicine. Journal of medical systems, 2002. 26(5): p. 445-463.

65. Marshall, R.J., The use of classification and regression trees in clinical epidemiology. Journal of Clinical Epidemiology, 2001. 54(6): p. 603-609.

66. Speybroeck, N., Classification and regression trees. Int J Public Health, 2012. 57(1): p. 243-6.

67. Fan, Y., et al., Applying random forests to identify biomarker panels in serum 2D-DIGE data for the detection and staging of prostate cancer. J Proteome Res, 2011. 10(3): p. 1361-73.

68. Goldstein, B.A., et al., An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC Genet, 2010. 11: p. 49.

69. Maenner, M.J., et al., Detecting gene-by-smoking interactions in a genome-wide association study of early-onset coronary heart disease using random forests. BMC Proceedings, 2009. 3(Suppl 7): p. S88.

70. Xu, M., et al., Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. BMC Medical Genetics, 2011. 12(1): p. 90.

71. Barrett, J.H. and D.A. Cairns, Application of the random forest classification method to peaks detected from mass spectrometric proteomic profiles of cancer patients and controls. Stat Appl Genet Mol Biol, 2008. 7(2): p. Article4.

72. Marino, S.R., et al., Identification by random forest method of HLA class I amino acid substitutions associated with lower survival at day 100 in unrelated donor hematopoietic cell transplantation. Bone Marrow Transplant, 2012. 47(2): p. 217-26.

73. Barco, L., et al., Application of the Random Forest method to analyse epidemiological and phenotypic characteristics of Salmonella 4,[5],12:i:- and Salmonella Typhimurium strains. Zoonoses Public Health, 2012. 59(7): p. 505-12.

74. Peng, S.Y., et al., Random forest can predict 30-day mortality of spontaneous intracerebral hemorrhage with remarkable discrimination. Eur J Neurol, 2010. 17(7): p. 945-50.

75. Scott, S.B., et al., Combinations of stressors in midlife: examining role and domain stressors using regression trees and random forests. J Gerontol B Psychol Sci Soc Sci, 2013. 68(3): p. 464-75.

76. Boulesteix, A.L., et al., Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012. 2(6): p. 493-507.

77. Goldstein, B.A., E.C. Polley, and F.B. Briggs, Random forests for genetic association studies. Stat Appl Genet Mol Biol, 2011. 10(1): p. 32.

78. Hong, W., et al., Prediction of severe acute pancreatitis using classification and regression tree analysis. Dig Dis Sci, 2011. 56(12): p. 3664-71.

79. Lasko, T., et al., The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform, 2005. 38(5): p. 404 - 415.

80. Surnamespratt, G., G. Surnameju, and G.R. Surnamebrasier, A structured approach to predictive modeling of a two-class problem using multidimensional data sets. Methods, 2013. 61(1): p. 73-85.

81. Yoo, W., et al., A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions. International journal of applied science and technology, 2012. 2(7): p. 268.

82. Cutler, A. and J.R. Stevens, Random forests for microarrays. Methods Enzymol, 2006. 411: p. 422-32.

83. Greenstein, D., et al., Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls. Frontiers in Psychiatry, 2012. 3.

84. Chen, C., A. Liaw, and L. Breiman, Using random forest to learn imbalanced data. University of California, Berkeley, 2004.

85. Washam, C.L., et al., Identification of PTHrP(12-48) as a plasma biomarker associated with breast cancer bone metastasis. Cancer Epidemiol Biomarkers Prev, 2013. 22(5): p. 972-83.

86. Gurm, H.S., et al., A novel tool for reliable and accurate prediction of renal complications in patients undergoing percutaneous coronary intervention. J Am Coll Cardiol, 2013. 61(22): p. 2242-8.

87. Díaz-Uriarte, R. and S.A. De Andres, Gene selection and classification of microarray data using random forest. BMC bioinformatics, 2006. 7(1): p. 3.

88. Zyriax, B.C., et al., The Association of Genetic Markers for Type 2 Diabetes with Prediabetic Status - Cross-Sectional Data of a Diabetes Prevention Trial. PLoS One, 2013. 8(9): p. e75807.

89. Zhang, Y. and J.S. Liu, Bayesian inference of epistatic interactions in case-control studies. Nat Genet, 2007. 39(9): p. 1167-1173.

90. Issa, Y., et al., Surgical and Endoscopic Treatment of Pain in Chronic Pancreatitis: A Multidisciplinary Update. Dig Surg, 2013. 30(1): p. 35-50.

91. Lazarev, M., et al., Does the pain-protective GTP cyclohydrolase haplotype significantly alter the pattern or severity of pain in humans with chronic pancreatitis? Mol Pain, 2008. 4: p. 58.

92. Dworkin, R.H., et al., Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). Pain, 2009. 144(1): p. 35-42.

93. Brand, H., et al., Variation in the gamma-glutamyltransferase 1 gene and risk of chronic pancreatitis. Pancreas, 2013. 42(5): p. 836-40.

94. Hosmer Jr, D.W. and S. Lemeshow, Applied logistic regression. 2004: John Wiley & Sons.

95. Poulsen, J.L., et al., Pain and chronic pancreatitis: a complex interplay of multiple mechanisms. World J Gastroenterol, 2013. 19(42): p. 7282-91.

96. Larusch, J. and D.C. Whitcomb, Genetics of pancreatitis with a focus on the pancreatic ducts. Minerva Gastroenterol Dietol, 2012. 58(4): p. 299-308.

97. Ammann, R.W., P.U. Heitz, and G. Kloppel, Course of alcoholic chronic pancreatitis: a prospective clinicomorphological long-term study. Gastroenterology, 1996. 111(1): p. 224-231.

98. Ammann, R.W. and B. Muellhaupt, Progression of alcoholic acute to chronic pancreatitis. Gut, 1994. 35(4): p. 552-6.

99. Klöppel, G. and B. Maillet, Pathology of acute and chronic pancreatitis. Pancreas, 1993. 8(6): p. 659-670.

100. Yadav, D., et al., Alcohol consumption, cigarette smoking, and the risk of recurrent acute and chronic pancreatitis. Arch Intern Med, 2009. 169(11): p. 1035-45.