A BAYESIAN APPROACH TO LEARNING DECISION TREES FOR PATIENT-SPECIFIC MODELS

by

Joyeeta Dutta-Moscato

BA & BSE, University of Pennsylvania, 2000

MS, University of Pittsburgh, 2007

MS, University of Pittsburgh, 2013

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Joyeeta Dutta-Moscato

It was defended on

August 2, 2018

and approved by

Michael J. Becich, MD, PhD, Department of Biomedical Informatics

Xinghua Lu, MD, PhD, Department of Biomedical Informatics

David C. Whitcomb, MD, PhD, Department of Medicine

Dissertation Director: Shyam Visweswaran, MD, PhD, Department of Biomedical Informatics

A BAYESIAN APPROACH TO LEARNING DECISION TREES FOR PATIENT-SPECIFIC MODELS

Joyeeta Dutta-Moscato, PhD

University of Pittsburgh, 2018

A principal goal of precision medicine is to identify genomic factors that are predictive of outcomes in complex diseases, to provide better insight into their molecular mechanisms. Based on our current understanding, there are many genomic factors that are likely to be pathogenic in small subpopulations while being rare in the population as a whole. This research introduces a new machine learning method for discovering single nucleotide variants (SNVs), both common and rare, that in a given person are predictive of that person developing a disease or disease outcome.

The new method described in this research constructs decision tree models, uses a Bayesian score to evaluate the models, and employs a person-specific search strategy to identify SNVs that are predictive in a subpopulation whose members are similar to the person of interest. This method, called the Personalized Decision Tree Algorithm (PDTA), works by constructing a decision tree model from the data and then identifying a path in the tree that has excellent prediction for the person of interest, or constructing a new path if none of the paths in the tree have excellent prediction.

The PDTA was refined iteratively on synthetic data and was experimentally evaluated on five datasets. One of the datasets was synthetic, one was semi-synthetic, and three were biological datasets collected from patients with chronic pancreatitis that included one small genomic dataset, a whole exome dataset, and a whole exome dataset focused on patients with diabetes in chronic pancreatitis. The performance of the method was evaluated using area under the Receiver Operating Characteristic curve and F1 score, as well as the ability to retrieve known and unknown rare SNVs. The PDTA was found to be effective to varying degrees in the datasets that were evaluated, creating parsimonious genetic representations for patient-specific groups, with the potential to discover novel variants.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am grateful for the contributions of many people towards the accomplishment of this work.

Foremost, I would like to express my deep gratitude to my thesis advisor, Shyam Visweswaran, who makes brilliance look so effortless. I was privileged to have his patient and insightful guidance as a student. The initial version of the algorithm at the heart of my dissertation was developed by Dr. Visweswaran along with Gregory F. Cooper, whom I also thank for his contributions to not only this work, but as a teacher and researcher at the Department of Biomedical Informatics (DBMI) at the University of Pittsburgh. The development of this algorithm benefitted from work done in early stages by Jose Posada Aguilar, Adriana Johnson, and Louisa Zhang. While a doctoral student, I received funding from the National Library of Medicine (NLM), and I am grateful for the experience this enabled from the entire community of students, staff and faculty at DBMI.

My development as a graduate student was enriched by Michael J. Becich, whose mentorship opened my eyes to a vast world of opportunities in healthcare and helped me find my niche. I would also like to thank my committee members, David C. Whitcomb and Xinghua Lu, for their expertise. May I also pause a moment in memory of M. Michael Barmada, whose untimely passing was a great loss to our academic community.

The journey undertaken towards this dissertation included contributions from many in my personal life as well. I would like to thank my parents, Biplab and Susmita Dutta, whose intellectual curiosity, integrity and unconditional love have made me who I am today. I would also like to thank my parents-in-law, Sam and Cathie Moscato, whose unwavering support and sense of humor has sustained me over the years. I am grateful for the support and friendship of Sudeshna Dutta, Monika Woszczyna, Katrina Romagnoli and Rubaiya Amin. My son, Aritro A. Moscato, through whose eyes I get to see anew the wonder and beauty of the world, who reminds me constantly of the joy in discovery and creation. My husband, Michael J. Moscato, for his generosity and optimism; his partnership is the rock that gives me wings.

1.0 INTRODUCTION

Fundamental activities in medicine include assessing disease risk, making diagnoses, estimating prognoses, and selecting appropriate therapy for an individual patient. The emerging area of precision medicine takes the approach that accounting for differences within clinically defined populations allow doctors and researchers to more precisely define groups of patients for specific treatment and prevention strategies[1]. This approach is particularly encouraging for patients with complex diseases, such as chronic pancreatitis, where groups of patients with the same disease often exhibit a variety of both causal factors and emergent clinical features, implying diversity in their disease mechanism. A patient-based approach, which works within the context of the population-based knowledge of the disease, has the potential to enable 1) more accurate prediction of individual risk, 2) discovery of disease subtypes among subpopulations, and 3) identification of individual genomic, environmental and clinical factors that explain risk, diagnosis, or prognosis in the individual. The precision medicine approach is in contrast to a "one-size-fits-all" approach, in which disease prevention and treatment strategies are developed for an average person, with less consideration for the differences between individuals.

In this dissertation, I present the development and evaluation of an algorithm for finding genetically defined subgroups within patient populations, the Personalized Decision Tree Algorithm (PDTA).

1.1 OVERVIEW OF THE PDTA

Predictive modeling consists of two steps: learning, which trains a model from a database of individuals, and inference, which predicts an outcome for a given individual. If a model performs well at prediction, it can be examined to identify features that are relevant to the prediction. The typical paradigm in predictive modeling consists of learning a single model from a database of individuals, which is then applied to predict outcomes for any future individual. Such a model is called a *population-wide* model because it is intended to be applied to an entire population of future individuals. In contrast, personalized modeling focuses on learning models that are tailored to the characteristics of the individual at hand. Personalized models that are optimized to perform well for a specific individual are likely to have better predictive performance than the typical population-wide models, which are optimized to have good predictive performance, on average, on all future individuals[2]. Moreover, personalized models have the potential to identify genomic, environmental and clinical factors that are specific to the individual for explaining risk, diagnosis or prognosis. Thus, personalized predictive models are likely to be useful in precision medicine.

The goal of precision medicine is to map treatment and prevention strategies to specific groups of people in an evidence-based manner. I propose that the PDTA will enable prediction of patient subgroups characterized by a set of distinct predictive variables. PDTA uses a Bayesian score, which prioritizes combinations of variables that maximize the posterior probability of observed phenotype or outcome, and focuses the search for predictors around the known features of an individual.

1.2 HYPOTHESIS AND AIMS OF THE DISSERTATION

I hypothesize that the incorporation of an individual's genomic and clinical factors in building predictive models will lead to better predictions, and identification of factors relevant to clinical outcomes in that individual. Additionally, this approach will identify a subpopulation of individuals in the data who are most similar to the individual and identify an interpretable set of characteristics that defines the subpopulation.

I also hypothesize that the personalized modeling approach will provide several advantages. It will enable the discovery of a parsimonious set of factors that are predictive for the individual for whom the model was developed. By building a Bayesian model that views each patient within a subpopulation of similar patients, we can discover rare genomic variants that are otherwise obscured in diverse patient populations.

To test my hypothesis, I propose the following aims:

Aim 1. Implement and extend a Personalized Decision Tree Algorithm (PDTA) to produce individualized predictions, uncover a subpopulation whose members are most similar to the individual at hand and identify relevant factors for the individual of interest, that include genomic factors.

Aim 2. Evaluate PDTA and its extensions on synthetic, semi-synthetic and real genomic data

1.3 INNOVATION

I have developed an algorithm that constructs personalized decision tree models based on the genomic factors of a specific individual. The algorithm uses a Bayesian score to evaluate models, and a personalized search strategy to identify high-scoring models. Unlike generic decision trees, which myopically choose splits based on best information gain at the local node, the PDTA considers the best score obtained over the entire model structure and parameter priors. The algorithm is also implemented to consider value merging to accommodate allelic combinations as seen in genotypic models. Models built with this algorithm present an interpretable and statistically sound way to analyze complex disease phenotypes to (1) discover rare variants in individuals, (2) discover subpopulations of individuals, (3) discover genetic signatures of subpopulations of individuals. In doing so, the algorithm improves our ability to develop personalized treatment and prevention, characterizing verifiable subpopulations within a heterogeneous disease population with a parsimonious set of predictive factors.

1.4 OVERVIEW OF THE DISSERTATION

In this section, I provide a brief overview of this dissertation.

Chapter 2 provides background relevant to the role of a personalized algorithm for genomics and precision medicine. Chapter 3, describes the algorithm and related theory. Chapter 4 describes the datasets, performance metrics, information sources and conventions used in this dissertation. Chapter 5 describes the evaluation of PDTA on several different datasets. Chapter 6 summarizes the findings and conclusions drawn from this work.

2.0 BACKGROUND

This chapter provides an overview of genomics and precision medicine relevant to the application and utility of the PDTA, followed by machine learning background relevant to the development of the PDTA. Section 2.1 gives an overview of genomics and gene variants in disease. Section 2.2 covers the importance and goals of precision medicine. Section 2.3 gives an overview of chronic pancreatitis as a case study for the application of PDTA. Section 2.4 places the development of PDTA within the context of existing machine learning applications.

2.1 GENOMIC BASIS OF DISEASE

Over the past century, the classification of diseases based on causative genes has settled broadly into two types: Mendelian or monogenic diseases, and complex or polygenic diseases.

Mendelian diseases, such as cystic fibrosis or sickle cell disease, are caused primarily by defect of a single gene, and tend to be rare in the general population. In this category of diseases, genomic sequencing of a small number of affected individuals may be sufficient to identify the causal variant and the associated gene. For example, Ng et al. sequenced the exome of four unrelated individuals with Freeman Sheldon syndrome (a rare inherited disorder) and eight healthy individuals, and were able to correctly identify the gene previously known to cause the syndrome[3]. In a following study, Ng et al. sequenced the exomes of four individuals with Miller syndrome (a rare malformation disorder) and identified a new casual variant[4].

Complex diseases, involving disruption in several genes, occur more frequently in the population, and exhibit complex features in their etiology as well as expression. They are likely to be heterogeneous, with variants in a number of different genes, or have variants in multiple loci of the same gene, all manifesting in the same or similar phenotypes. Even though they may cluster in families, they do not necessarily have a clear pattern of inheritance. Many complex diseases – such as heart disease, type 2 diabetes, and cancer – are influenced by lifestyle and environmental factors. If multiple genomic causes manifest small and varied effects on disease expression, large samples of affected and healthy individuals are needed to identify the causal variants. For example, a meta-analysis of 74,046 individuals identified 19 variants associated with Alzheimer's disease[5].

The simplest sequence variations in the DNA are the single nucleotide variants (SNVs). Classically, SNV referred to any variation in the DNA that may or may not have been well characterized, while an SNV with a minor allele frequency (MAF) exceeding 5% in the population was called a single nucleotide polymorphism (SNP). In recent years, however, the recognition of the value of rarer variants has led to the broadening of the term "SNP" to often be used interchangeably with the term "SNV". For example, the National Center for Biotechnology Information (NCBI) database of short genetic variations, dbSNP, while initially built as a database of SNPs, has expanded to include single nucleotide substitutions even if they have not been found to occur frequently enough in a population to be termed polymorphic[6]. According to the rare variant hypothesis, a significant proportion of inherited susceptibility to chronic diseases may be due to the summation of the effects of a series of low frequency variants of a variety of different genes, each conferring a moderate but readily detectable increase in relative risk[7]. Even if the frequency of a single variant is less than 5%, the collective power of numerous rare variants can raise disease association to a significant level. While disease-causing rare variants usually have MAF ranging from 0.1% to 3%, depending on the context, some have been found to have frequency lower than 0.01%[8].

Many pathogenic variants, as commonly sought in clinical genetic testing, have been discovered from genome wide association studies (GWASs)[9]. The paradigm of GWAS is based on the common disease-common variant hypothesis, which posits that disease-causing alleles are likely to be common gene variants that can be detected as statistically significant when comparing affected individuals with controls[10]. These variants alone often fail to provide a complete functional mechanism for the disease phenotype, and it has become increasingly clear that rare variants likely also play an important role in complex diseases, independently or in conjunction with other variants, common or rare[11,12]. Next generation sequencing methods such as whole exome sequencing provide a more complete set of rare and common variants from protein-coding regions (exons) of the DNA. Given this landscape, genomic causal mechanism can be broadly subdivided into three categories[13]: a large number of small-effect common variants across the entire allele frequency spectrum (the infinitesimal model)[14,15], a large number of large-effect rare variants (the rare allele model)[16] or some combination of genotypic, environmental and epigenetic interactions (the broad sense heritability model)[17,18]. Therefore, the development of statistical methods to detect rare variants with relevant effect sizes is crucial[12].

2.2 PRECISION MEDICINE FOR COMPLEX DISEASES

According to the National Institutes of Health (NIH), precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person"[1]. With the rapid increase in data collection and the ability to analyze larger datasets in tractable time, it is feasible to prioritize a patient's care to their individual needs. To unravel this potential, personalization of medicine requires two fundamental characterizations: a well-grounded understanding of who the patient is, and an equally robust understanding of the subpopulation that most resembles that patient, in the context of the decisions at hand. When these characterizations are represented probabilistically and can be used to drive decision-making in a rational manner, we can maximize the positive outcomes for the patient[19]. Next generation sequencing technologies provide detailed genomic information on each patient, but the clinical implications can only be realized when the results of genetic testing are actionable, thus informing prognosis or treatment[20]. With rapidly growing clinical databases and interoperability of electronic health records (EHRs), we get closer to being able to use genomic knowledge and clinical information to further enrich an actionable, personalized patient profile.

With the existing use of evidence-based recommendations in clinical practice, personalization can be implemented by capturing subgroups whose members share the most salient features with a specific patient[21,22]. This is the idea behind endeavors such as future point-of-service decision support scanning EHRs for patients with similar profiles[23]. The first reported use of querying medical records of past patients in near real-time to aid in treatment decision took place in 2011, when investigators at Stanford University searched for patients similar to a 13 year old girl with lupus nephritis to decide on anticoagulation therapy[24]. Other patient similarity analytics tools can incorporate predictive models of a patient's outcome that is trained from a subpopulation similar to the index patient[25].

The high dimensional data generated by next generation sequencing provide a comprehensive view of common and rare SNVs in a patient’s genomic profile, but the number of individuals comprising a cohort for study is usually small relative to the large number of SNVs to be assessed. In addition, if we are to uncover causal variants that are relevant only to a subset of patients, genetic signatures can appear to be highly variable across the tested populations, while showing patterns of similarity within subpopulations. The appropriate analysis methods should have the ability to make sound statistical judgments within the smaller sample sizes of subpopulations, and also be able to make use of the larger overarching phenotypic population, which is essentially the context that the smaller subpopulations are conditioned upon. Algorithms capable of discovering patient-specific subgroups based on their genomics dovetail with these goals of precision medicine, for both variant discovery as well as understanding the patient-specific pathogenicity in an actionable framework[26].

2.3 CHRONIC PANCREATITIS AS A CASE STUDY

Pancreatitis is an inflammatory condition of the pancreas. Repeated inflammatory damage, as recurrent acute pancreatitis (RAP) or chronic pancreatitis (CP), can lead to irreversible scarring and fibrosis of the pancreas, causing pain, incremental loss of function (leading to syndromes such as maldigestion and diabetes mellitus), and an increased risk of pancreatic cancer. The annual incidence of CP in industrialized countries is 3 to 10 per 100,000 population[27]. Although traditionally defined by a common clinical appearance and pathology, CP is a complex set of disorders of heterogeneous etiology, involving a variety of signaling pathways, metabolic factors, and environmental susceptibilities[28]. Patients with CP often differ in their clinical course, symptomology and environmental risk factors.

Recent insights into the genetics of CP present a strong case for the role of precision medicine in guiding treatment of patients with CP[29,30]. The major CP causing genes are involved in trypsin control in the pancreas. A heterozygous pathogenic variant of the trypsinogen gene *PRSS1*, which results in autosomal dominant inheritance of pancreatitis, is found in 60%-100% of families with hereditary pancreatitis; non-*PRSS1*-related hereditary pancreatitis shows a different disease course, with a later age of onset[31,32]. Not all pathogenic mutations in PRSS1 cause hereditary pancreatitis through the same mechanism; some, such as A16V exhibit variable penetrance, and can be seen with additional risk factors such as SPINK N34S to cause CP[33]. While the pathogenicity of PRSS1 A16V has been known for a while, another PRSS1 mutation, P17T was only discovered recently to mimic the effect of A16V[34]. Rare variants whose pathogenicity only occurs in the presence of other variants are difficult to discover through traditional association studies.

Multiple other genetic risk factors are common, such as loss-of-function variants in *CFTR* and *SPINK1*, and various combinations of these genetic risk factors may be seen in different generations of one family[35]. A growing list of variants in other genes, such as *CTRC*, *CASR*, *CLDN2*, *CPA1* have been emerging as well, with varying contributions to disease risk, along with variants that affect not only the coding sequences of proteins, but also their regulation[29]. The susceptibility to develop CP is also known to be increased by lifestyle factors such as smoking and alcohol use, with certain subgroups of patients showing a higher susceptibility to these factors than others[36]. Correspondingly, studies have shown evidence of SNVs in *PRSS1-PRSS2* and *CLDN2* loci with high association to alcohol-related CP[37].

Genes such as *PRSS1*, *SPINK1*, *CFTR* and *CTRC* are commonly used in multigene panels to diagnose CP, although it is suspected that other common and rare variants are also risk factors in CP[38,39]. In the treatment and prevention of CP, it is important to uncover not only independent risk factors for the disease, but also factors that act synergistically and as modifiers of other risk factors. The genetic and clinical/behavioral risk factors characterizing a group of patients offer clues to the underlying disease mechanism malfunctioning in that group of patients and can steer treatment needs by anticipating the dysfunction specific to that patient's biology. For example, damage to islet cells during the course of CP can give rise to diabetes in some CP patients[40]. Identifying genetic markers in the subgroup of patients prone to develop diabetes would open up the possibility of treatments specifically important for that subgroup of patients. A growing understanding of genetics and gene–environment interactions have led to clinical centers, such as at the University of Pittsburgh, classifying CP patients into general disease mechanism groups in their evaluation for pancreatic disease[30]. Improvements in personalization algorithms will support such endeavors.

2.4 MACHINE LEARNING METHODS

In building predictive models, a machine learning algorithm constructs a hypothesis space based on a finite set of observations, or instances. Generally speaking, the model is induced from these instances, known as the training dataset, and then used to infer outcomes on a separate set of instances, known as the test dataset. The best model – a combination of relational structure and parameters to ‘explain’ an underlying mechanism that can generate this data – is built based on a scoring criterion to assess the likelihood of the training dataset from the model’s predictions. The value of prediction is then evaluated on a number of criteria, such as accuracy, generalizability and reliability, to make a claim on the suitability of the learning algorithm.

The PDTA is most closely related to work categorized as instance-based learning. Section 2.4.1 covers the general area of instance-based learning. Section 2.4.2 focuses on a decision tree learning. Section 2.4.3 focuses on patient-specific machine learning.

2.4.1 Instance-based Machine Learning

Instance-based learning is a type of inductive learning method: simply put, it is learning by example. Unlike traditional classification methods, which learn from instances in the training data and then apply the model to test instances presented to it, instance-based learning methods incorporate test instances to create a model from training instances local to it[41]. It is often referred to as ‘lazy’ learning, since most of the computational work is not done upfront on the training dataset, but is done when the test instance becomes available. This constraint can cause a high computational load at runtime. However, these methods are useful for their ability to optimize locally around specific instances.

There are three primary components to instance-based algorithms: (1) A similarity or distance function, which computes the similarity between instances used to identify instances in the training dataset that are near the test instance; (2) A classification function, which decides the locality specified for the test instance; (3) A concept description updater, which tracks the performance of the classification, and makes decisions on the choice of instances to include in the concept description. The *k*-nearest neighbor classification algorithm, in which a chosen number, *k*, of the nearest neighbors of a classifier are used in order to create a local model for the test instance, is a canonical instance-based algorithm. Most instance-based algorithms derive from this concept.

Other instance-based methods take advantage of the values of predictors to infer a class distribution within the locality of the test instance. This has been implemented by incorporating the instance-based approach in general classification methods such as decision trees and naïve Bayes. The LazyDT algorithm, described by Friedman et. al. (see Section 2.4.2.2) conceptually constructs the decision tree considered best for each test instance, and was shown to have better accuracy in predictions compared to population-wide decision tree methods such as ID3 and C4.5[42]. The Lazy Bayesian Rules (LBR) learner, described by Zheng et. al., similarly modified a naïve Bayes model to search for the best model using the values of the predictors specific to the test instance, and was also shown to have superior performance when compared to a variety of population-wide methods[43,44]. This method was further extended by Visweswaran et. al. to perform Bayesian model averaging (BMA) over LBR models, and was shown to have higher accuracy on a range of datasets[45]. Other instance based algorithms that use Bayesian network models have also shown promise (see Section 2.4.3)[2,46].

2.4.2 Decision trees

A decision tree, also known as a classification tree, is a predictive model that contains a branching structure of a tree. The tree consists of interior nodes, branches and leaf nodes. Interior nodes denote predictor variables, branches denoted values that a predictor can take, and all leaf nodes denote the target or outcome variable. An interior node denotes a test that is performed with a predictor, and each value that the predictor can take is denoted by a branch emerging from the node. The root node at the top of the tree is an interior node that denotes the first predictor that is tested. Leaf nodes are terminal nodes that are found at the bottom of the tree; a leaf node denotes a decision or outcome and is represented by one value of the target variable. A path from the root to a leaf denotes a series of tests on predictors leading to a specific outcome at the leaf node.

The decision tree is a partitioning classifier: the tree partitions the input space into a set of non-overlapping regions, where each region corresponds to a distinct path in the tree. To construct a tree, decision tree methods usually employ a top-down, recursive, greedy divide-and-conquer strategy. At every stage, the attribute most efficacious to classification of the observations is selected as predictor for the next step – a strategy termed “greedy” because it only considers the best option at the current step, without considering options that could lead to a better tree at some later stage. The structure and parameters for the tree are typically derived by recursive splitting to optimize a criterion such as error or entropy. Commonly used criteria for predictor selection include the Gini index, information gain, and misclassification error[47]. The goal is to construct a tree with the shortest paths, i.e. the shallowest possible tree, in a top-down manner.

Each path in the tree defines a subpopulation of observed instances, and each member of the training dataset is assigned exactly one path in the tree. The test performed at each node of the decision tree is essentially an “if-then” rule, and a path in the tree is simply a conjunction of such rules terminating with a target at the leaf node. If-then rules and decision trees are expressive and human interpretable, and can capture both main effects and interactions among predictors. These qualities make decision tree models a popular choice for clinical decision making.

2.4.2.1 Lazy decision trees

Friedman et.al. implemented an instance-specific ('lazy') version of decision trees which they called LazyDT[42]. The LazyDT considers the given test instance and derives a decision path guided by its features. Similar to generic decision trees, it uses a greedy approach to successively select features to add to the path, and uses information gain as its splitting criterion. At each step the LazyDT chooses a feature optimizing information gain at that node. After a feature is selected, only those individuals in the training dataset who match the features in the path derived so far are retained to derive the remaining part of the path. When compared to several decision tree methods including ID3, C4.5 without pruning, and C4.5 with pruning, LazyDT had higher predictive accuracy overall and performed substantially better than ID3 and C4.5 without pruning. Compared to C4.5 and CART, LazyDT handled missing values more naturally, but could be limited by the fact that it did not perform pruning, and hence could be susceptible to overfitting[48].

One of the key motivations for LazyDT was to avoid the undue influence of features that may dominate when training data is considered as a whole, but are irrelevant to the test instance at hand. A single tree built in advance can lead to many irrelevant splits for a given test instance, thus fragmenting the data unnecessarily. Such fragmentation reduces the significance of tests at lower levels since they are based on fewer instances. A decision tree built for the given instance can avoid splits on features that are irrelevant for the specific instance.

2.4.2.2 Probabilistic decision trees

In a probabilistic decision tree, each leaf node represents a probability distribution over the target variable. The tree can be converted into a set of if-then rules, wherein a path in the tree is translated into a probabilistic rule. Each rule has a conjunction of features in the antecedent, and a probability distribution for the target in the consequent. When presented with a test case, the tree performs inference to identify a path such that the features in the path match the features in the test instance. The target probability distribution is estimated from the known outcomes of observed instances whose features matched the features in the identified path. Decision trees are very well suited to implementing Bayesian models, where we seek to reduce uncertainty as 'beliefs' that can be updated as more data becomes available. Bayesian scores tend towards simpler structure, but as the model sees more data, it is willing to recognize that a more complex structure is necessary; it is good at trading off fit-to-data for lower model complexity, thereby reducing the extent of overfitting[49]. Learning decision trees as a Bayesian network was well developed by researchers near the turn of the century, and has gone on to form the foundation for Microsoft's proprietary Decision Trees Algorithm to learn causal interaction models by obtaining approximate posterior distributions[50-52].

Buntine first described a Bayesian approach to learn probabilistic decisions trees from data[53]. He developed an algorithm to search and identify the tree with the maximum posterior probability, using the Bayesian Dirichlet (likelihood) equivalent uniform (BDeu) score to maximize the posterior probability of the tree structure. While BDeu has been a popular scoring choice for model selection criterion, especially in cases with no prior knowledge, it requires the selection of an equivalent sample size (ESS) hyperparameter, which reflects the degree of prior belief one would have needed to be confident about the given choice of Dirichlet prior parameters[50]. There is no generally accepted rule for determining the correct assignment of ESS. The value of ESS controls how much smoothing occurs in estimating probability parameters. Empirical studies have shown that the network structure is very sensitive to the choice of ESS; when the chosen value of ESS was reduced, for increasing sample size, simpler network structures tended to be selected; when ESS became large, the number of arcs in the structure usually increased[54,55]. An asymptotic analysis of the BDeu score showed a monotonic increase in arcs added to the network structure as the ESS increases[56]. When using the BDeu score, it is important to test for its influence on the resultant model.

2.4.3 Patient-specific models

Patient-specific modeling can refer to a broad and diverse set of computational methods, all of which aim to improve personalization of healthcare. The idea of treatment schemes applied across subgroups of patients – also called stratified medicine[57] – is not a new one. Some of the significant barriers to realizing the full potential of personalized medicine are issues of data collection and harmonization. I will not discuss these issues. My focus is on the analysis of this data to uncover rare variants with significant association with patient subgroups.

One way to personalize for patient-based discovery is by building local models for a subset of the problem space, to be applied to patient subgroups. An example is patient-specific prediction for cancer survival[58]. Commonly used survival plots (using the Kaplan-Meier estimator) attempt prediction by aggregating individual patient characteristics. By implementing a local regression method for learning patient-specific survival time distribution based on all available patient attributes, Lin et. al. were able to improve the accuracy of survival predictions on a cohort of cancer patients.

In contrast to such local modeling methods, personalized models have an element of computation "on the fly" for every new input vector (as characterized by instance-based approaches, Section 2.4.1), to generate an individual model based on the closest data instances to the new instance taken from a dataset[59]. These can employ the use of similarity metrics, as done in one study to predict patient-specific risk profiles[60]. Ng et. al. developed a trainable similarity metric to find clinically similar patients, then created personalized risk factor profiles by analyzing the parameters of the trained personalized models. Another way to use a patient-specific heuristic was demonstrated by Visweswaran et al, averaging over a set of Markov blanket models[48]. Both studies found that the personalized models gave better predictions than population-trained models.

Using decision trees to build personalized models is essentially the task of creating a decision path customized to the features of a patient. Since decision trees are comprised of non-overlapping series of nodes, a path identified as best match for a patient defines a subpopulation in the data, and each member of a dataset assigned to the tree is assigned to exactly one path. Previous work in creating personalized decision paths has shown promising results[48,61]. While those models were mostly focused on developing good predictive models, the PDTA is focused on uncovering variants that may be rare in the general population, but have relevance to predicting disease state within a subpopulation of similar patients.

3.0 ALGORITHMIC METHODS

This chapter describes the PDTA that derives a personalized decision tree model. Briefly, the first phase builds a tree from the full population of patients, implementing what I call the standard decision tree algorithm (SDTA). The next phase implements the PDTA, personalized to a single patient, to custom-build a tree (which is just a path) for the patient. For ease of exposition, I will first describe the SDTA in detail and then I will explain how the PDTA builds on the SDTA and differs from it. After a summary of the main ideas underpinning the two algorithms, a detailed description follows.

The goal of both the SDTA and PDTA is to optimally predict a discrete target (or class) variable, such as a clinical outcome, from a moderately large number of predictors such as genomic variants and other clinical factors. The SDTA derives a probabilistic decision tree model that I term the standard decision tree (SDT) model, and the PDTA learns a personalized probabilistic decision tree model that I call the personalized decision tree (PDT) model. Since these are decision tree models, each path in the model implicitly defines a disease-specific group of factors that characterize a subpopulation of individuals. If we use genomic variants as predictors for the model, the group of variants comprising a path serves as the genetic signature for a specific subpopulation. Given a person *P*, PDTA identifies a personalized path that represents a subpopulation of individuals from the training dataset who are genetically similar to *P* at a small number of loci.

The SDT is a decision tree that has class probabilities (i.e., a probability distribution for the prediction) at its leaf nodes, and paths in the tree that can be interpreted as probabilistic rules. The SDT is an example of a population model that is derived from a group of patients (each patient represented as an instance in a training dataset), and is applied to predict the outcome for any future patient. The PDT model consists of a combination of a decision tree and a patient-specific (PS) path, where the path is tailored to the characteristics of the patient (an instance chosen from the test dataset) whose outcome is being predicted, and the decision tree is an alternate model for others who are not similar to the chosen patient. Personalized models that are optimized to perform well for a specific individual can identify features such as genomic factors that are specific for an individual, even if the factors are not common enough to reach statistical significance in a population-wide search, thus enabling precision medicine and the search for rare variants.

3.1 NOTATION AND DEFINITIONS

In the following exposition, random variables are denoted by italicized uppercase letters, such as X and Y. Instantiations of random variables are denoted by corresponding italicized lowercase letters, x and y. Vectors of variables are denoted by bold uppercase letters, such as $\boldsymbol{X}$, $\boldsymbol{Y}$, and their instantiations denoted by bold lowercase letters $\boldsymbol{x}$ and $\boldsymbol{y}$. A feature is a specification of a variable, denoted as a variable-value pair. Thus, $X = x$ is a feature that specifies that variable X is assigned the value x; $\boldsymbol{X} = \boldsymbol{x}$ is a feature vector that specifies that the variables in $\boldsymbol{X} = (X_1, X_2, \dots, X_i, \dots, X_m)$ have the values given by $\boldsymbol{x} = x_1, x_2, \dots, x_i, \dots, x_m$.

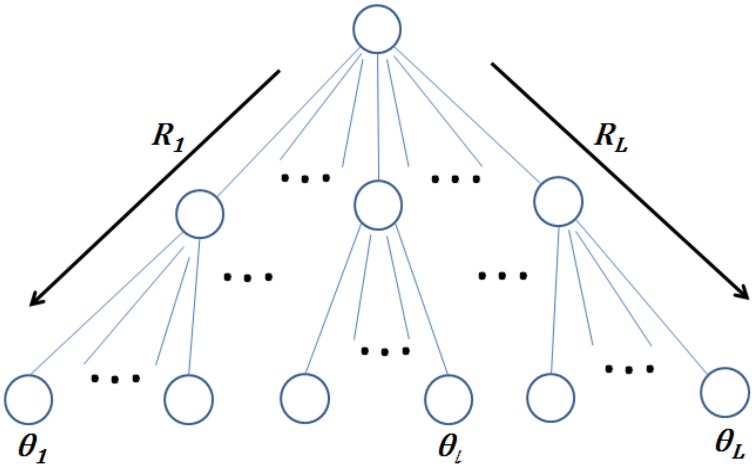
The dataset D = ($\boldsymbol{X}$, $\boldsymbol{Y}$) contains observations on m predictor variables $\boldsymbol{X}$ and one target variable Y, for n individual patients, and is comprised of a set of patients $\boldsymbol{d}_1$, $\boldsymbol{d}_2$, … , $\boldsymbol{d}_i$, … , $\boldsymbol{d}_n$. The ith patient, $\boldsymbol{d}_i$ = ($\boldsymbol{x}_i$, y_i) = (x_{i1}, x_{i2}, …, x_{im}, y_{i}) is instantiated by the observations on the predictors and the target for that patient (Figure 1 (a)). A model is derived from a training dataset D that contains values for $\boldsymbol{X}$ and $\boldsymbol{Y}$. A model is typically evaluated on a test dataset D_{test} that is distinct from the training dataset and consists of a set of test instances. In the case of the SDTA, only D is used to derive the model, which is then applied to the values of $\boldsymbol{X}$ in the test instance to predict a value of Y. In the case of PDTA, D is used along with the values of $\boldsymbol{X}$ in the test instance to derive the model, which is then applied to the values of $\boldsymbol{X}$ in the test instance to predict a value of Y.

Both the SDT and PDT models use decision trees, modeled as M = (T, $\boldsymbol{\theta}$), where T denotes the model structure and $\boldsymbol{\theta}$ is a parameter vector for the probability distributions over target $\boldsymbol{Y}$. If L is the number of leaf nodes in T, then T consists of L paths $\boldsymbol{R} = \boldsymbol{R_1}, \boldsymbol{R_2}, \ldots, \boldsymbol{R_l}_{\ldots}, \boldsymbol{R_L}$ with parameter vector $\boldsymbol{\theta}_l = (\theta_{l1}, \theta_{l1}, \ldots, \theta_{lk}, \ldots, \theta_{lK})$, where K is the number of values Y can take (Figure 1(b)). Path $\boldsymbol{R_l}$ consists of a conjunction of J features $X_1 = x_1$ ^ $X_2 = x_2$ ^ $X_j = x_j$. . . ^$X_J = x_J$ where the predictor list $\boldsymbol{X_R} = (X_1, X_2, . . ., X_j, . . ., X_J)$ is a subset of variables $\boldsymbol{X}$ in the domain and the list $\boldsymbol{x_R} = (x_1, x_2,. . ., x_j,. . ., x_J)$ is the list of corresponding values.

(a)

$$\begin{array}{llll}
 & \boldsymbol{X} = & X_1,\ X_2,\ X_3,\ \dots,\ X_m & Y \\
\boldsymbol{d_1}: & \boldsymbol{X_1} = & x_{11}, x_{12}, x_{13}, \dots, x_{1m} & y_1 \\
\boldsymbol{d_2}: & \boldsymbol{X_2} = & x_{21}, x_{22}, x_{23}, \dots, x_{2m} & y_2 \\
\vdots & & & \\
\boldsymbol{d_n}: & \boldsymbol{X_n} = & x_{n1}, x_{n2}, x_{n3}, \dots, x_{nm} & y_n
\end{array}$$

(b)



$$\boldsymbol{\Theta_1} = (\Theta_{11}, \Theta_{12}, \dots, \Theta_{1K})$$
$$\boldsymbol{\Theta_2} = (\Theta_{21}, \Theta_{22}, \dots, \Theta_{2K})$$
$$\vdots$$
$$\boldsymbol{\Theta_L} = (\Theta_{L1}, \Theta_{L2}, \dots, \Theta_{LK})$$

Figure 1. Schematic representation of terminology used in this dissertation

3.2 STANDARD DECISION TREE (SDT)

This section provides a detailed description of the SDT model, and then describes the SDTA that uses a Bayesian approach to learn the optimal SDT from a dataset *D*. This is essentially a search for a decision tree structure with a high posterior probability, given the observations in *D*. The SDTA employs a score-and-search approach where a scoring metric (or score, for short) is adopted to evaluate candidate tree structures, while a heuristic search strategy is used to find a tree structure with the best score. For computational efficiency, a heuristic search is employed instead of an exhaustive search of the model space of the tree[62].

3.2.1 SDT model

I will first describe the components of the SDT model in terms of structure (the graphical relationships), parameters (computation of probabilities) and inference. I will then discuss the Bayesian score to compute the parameters for best model that fits *D*. This is followed by a section describing how the model is used to perform inference for a test instance.

Model structure and parameters. An SDT model $M = (T, \boldsymbol{\theta})$ has the structure *T* and the vector of parameters $\boldsymbol{\theta}$. The structure *T* has *L* leaf nodes, and a set of paths $\boldsymbol{R}$. The size of $\boldsymbol{\theta}$ is determined by the number of parameters *K* of the multinomial distribution $P(Y|\boldsymbol{X}_{\boldsymbol{R}_l} = \boldsymbol{x}_{\boldsymbol{R}_l})$ over the target variable *Y* conditioned on $\boldsymbol{X}_{\boldsymbol{R}_l} = \boldsymbol{x}_{\boldsymbol{R}_l}$. (Each conditional probability distribution is a multinomial; K is the number of values Y can take.) The values for $\boldsymbol{\theta}_{lk} = (\theta_{l1}, \theta_{l2},. . ., \theta_{lk},. . ., \theta_{lK})$ at each leaf are estimated from individuals in the training dataset D that satisfy $\boldsymbol{X}_{\boldsymbol{R}_l} = \boldsymbol{x}_{\boldsymbol{R}_l}$, using the BDeu score.

Parameter estimation. The parameters are estimated in a Bayesian manner. The estimate for parameter θ_{lk} is given by the expectation (mean) of the posterior probability distribution of the parameter given data:

$$\theta_{lk} \equiv E\left[P\left(Y = k \middle| X_{R_l} = x_{R_l}\right)\right] = \frac{\alpha_{lk} + N_{lk}}{\alpha_l + N_l} \,, \qquad (1)$$

where N_l is the number of individuals in the training dataset D that satisfy $\boldsymbol{X}_{R_l} = \boldsymbol{x}_{R_l}$; N_{lk} is the number of individuals that satisfy $\boldsymbol{X}_{R_l} = \boldsymbol{x}_{R_l}$ and $Y = k$; α_{lk} is a parameter prior that can be interpreted as belief equivalent to having previously (prior to obtaining D) seen that α_{lk} individuals satisfy $\boldsymbol{X}_{\boldsymbol{R}_l} = \boldsymbol{x}_{\boldsymbol{R}_l}$ and $Y = k$; $N_l = \sum_k N_{lk}$ and $\alpha_l = \sum_k \alpha_{lk}$.

The parameter α is the priori equivalent sample size, or ESS, as discussed earlier for BDeu score. This parameter controls how much smoothing occurs in estimating probability parameters: the higher the value of α, the greater the smoothing that occurs.

Inference. Given a test instance $\boldsymbol{X}_{test} = \boldsymbol{x}_{test}$, the parameterized model M is used to compute the probability distribution over the target variable Y_{test} of the test instance. Inference is performed by identifying the relevant path $\boldsymbol{R}_{test}$ that fits the test instance's features, and then the parameters in the leaf node of $\boldsymbol{R_{test}}$ are returned. Identification of the relevant path is done by sequentially identifying the outbound branch at each interior node, starting at the root node, by the value of the predictor associated with the node in the test instance.

3.2.2 Description of SDTA

The SDTA employs a heuristic search-and-score approach: In each iteration, every leaf node in the tree is examined for the possibility of replacing it with an interior node corresponding to a predictor variable that has not already been used. The tree with this replacement is evaluated by a score, as a measure of how well the tree represents the distribution the data was drawn from.

Given the data *D*, the score of tree *T* is a measure of how probable the structure *T* is, over all possible parameterizations of *T*. It can therefore be calculated by the posterior probability:

$$P(T|D) = \frac{P(D|T)P(T)}{P(D)}. \tag{2}$$

For a fixed dataset *D*, *P(D)* is constant, hence we can consider the score as:

$$score(T) = P(D|T)P(T). \tag{3}$$

In this score, the term *P(T)* refers to the prior distribution over the structure T, hence known as the structure prior, described in Section 3.2.2.1. The term *P(D|T)* captures the posterior probability of the data given a particular parameterization of T. I use the Bayesian scoring function BDeu to optimize for this posterior term, described in Section 3.2.2.2. The numbers involved in Equation 2 can become very small, so computation is done in log space to avoid underflow. Following the derivations of the structure prior (Equation 5) and BDeu score (Equation 9) we get:

$$logscore(T) = \left[\sum_{i \in S} \log(p_i) + \sum_{i \notin S} log(1-p_i)\right]$$

$$+ \left[\sum_{l=1}^{L} \left(\log(\Gamma(\alpha_l)) - \log(\Gamma(\alpha_l + N_l))\right)\right.$$

$$\left. + \left[\sum_{k=1}^{K} \left(\log(\Gamma(\alpha_{lk} + N_{lk})) - \log(\Gamma(\alpha_{lk}))\right)\right]\right] \quad (4)$$

3.2.2.1 Structure Prior

The term *P*(*T*) the denotes the prior distribution over the structure *T*. I assume a binomial prior with uniform probability:

$$P(T) = \prod_{i \in S} p_i \prod_{i \notin S} (1-p_i)\,, \quad (5)$$

where *S* is the subset of variables in *D* that are in *T*. For computational ease, I calculate this in log space:

$$\log P\,(T) = \sum_{i \in S} \log(p_i) + \sum_{i \notin S} log(1-p_i) \quad (6)$$

We can incorporate intuition from a domain expert through the structure prior. Let *q* be the expected number of predictors (ENP) in *D* that a domain expert expects to be predictive of *Y*. Since the total number of variables in *D* is *m*, a simple binomial prior is given by $p_i = q / m$, when $i \in S$, and otherwise $p_i = 1 - (q / m)$. Further information on the effect of variables can be incorporated by replacing the uniform *p* with informative scores.

3.2.2.2 Bayesian score: BDeu

The best model is one which not only maximizes the probability of generating the observed data, but also fits the unobserved data in the full distribution from which the observations are sampled. A good scoring function, therefore, should be asymptotically correct: the learned distribution with maximum score should converge (with high probability) to the underlying distribution, as the size of the data increases.

Considering $\boldsymbol{\theta}$ as the full set of parameterizations for *T*, the joint probability can be marginalized over all choices of $\boldsymbol{\theta}$:

$$P(T,D) = \int_{\theta} P(T,D,\theta)\,d\theta.$$

Applying the chain rule:

$$P(T,D) = \int_{\theta} P(D|T,\theta)P(\theta|T)P(T)\,d\theta. \tag{7}$$

Since P(T) is not dependent on θ, this can be moved outside of the integral:

$$P(T,D)$$

$$= P(T)\int_{\theta} P(D|T,\theta)P(\theta|T)\,d\theta.$$

Since *P(T, D)* = *P(D|T)·P(T)*, dividing both sides by *P(T)* gives the marginal likelihood:

$$P(D|T) = \int_{\theta} P(D|T,\theta)\,P(\theta|T)\,d\theta. \tag{8}$$

The first term in the integral, $P(D \mid T, \boldsymbol{\theta})$, is the likelihood of the data *D* given structure *T*. For the second term, $P(\boldsymbol{\theta} \mid T)$, if I assume a Dirichlet probability density function that specifies a parameter prior distribution over $\boldsymbol{\theta}$, the marginal likelihood has a closed form solution (derived in Heckerman et al 1995 [50]:

$$P(D|T) = \prod_{l=1}^{L} \frac{\Gamma(\alpha_l)}{\Gamma(\alpha_l + N_l)} \left[\prod_{K=1}^{K} \frac{\Gamma(\alpha_{lk} + N_{lk})}{\Gamma(\alpha_{lk})} \right], \tag{9}$$

where Γ is the gamma function; K is the number of values of Y; L is the number of leaf nodes in T; N_{lk} is the number of individuals in D that satisfy the tests in path R_l and $Y = k$; α_{lk} is the parameter prior ESS where α_l is a set $\{\alpha_{l1}, \alpha_{l2}, \ldots, \alpha_{lk}\}$; $N_l = \sum_k N_{lk}$ and $\alpha_l = \sum_k \alpha_{lk}$.

For computational ease, I calculate this in log space:

$$\log P(D|T) = \left[\sum_{l=1}^{L} \left(\log(\Gamma(\alpha_l)) - \log(\Gamma(\alpha_l + N_l)) \right) \right.$$

$$\left. + \left[\sum_{k=1}^{K} \left(\log(\Gamma(\alpha_{lk} + N_{lk})) - \log(\Gamma(\alpha_{lk})) \right) \right] \right] \tag{10}$$

3.2.2.3 Search Strategy

The space of tree structures is very large; an algorithm that examines every tree structure to identify the highest-scoring tree is computationally intractable except for very small values of *m*. Thus, heuristic search methods are appropriate for searching this model space. The SDTA uses greedy search to identify a high scoring model. Because greedy search is not guaranteed to be optimal, SDTA is not guaranteed to find the best scoring model in the model space. However, in practice, greedy search performs well and identifies an excellent scoring tree structure.

The pseudocode for growing the SDT is given in Figure 2. It follows a general process laid out in Buntine (1992)[62]. In each iteration, every leaf node is examined for the possibility of replacing it with an interior node corresponding to a predictor variable that has not already been used. The replacement that best improves the score (Equation 3) over the current tree is chosen; the node is added, extending the tree by one level. If no replacement can be found that improves the score of the current tree, the search terminates and the current tree is returned.

```
SDTA (X, D)
    Input:            X is a set of predictor variables
                      D is a training dataset of individuals described using X, a set of predictor
                      variables, and target variable Y
    Output:           Tree

Tree = null
free_X = list of all predictor variables
score = Bayesian score of Tree computed using Equation 4
LOOP
      best_Score = score
      found_Best_X = False
      temp_Tree = Tree
      FOR each leaf_Node in temp_Tree
                FOR each X in free_X
                          Grow temp_Tree by replacing leaf_Node with X
                          temp_Score = Bayesian score of temp_Tree computed using Equation 4
                IF temp_Score > best_Score
                                    found_Best_X = True
                                    best_Score = temp_Score
                                    best_X = X
                          END IF
                END FOR
      END FOR
      IF found_Best_X = True
                Grow Tree by replacing leaf_Node with best_X
                Remove best_X from free_X
      ELSE
                EXIT from LOOP
      END IF
END LOOP
Return Tree
```

Figure 2: Pseudocode for the SDTA

3.2.3 Example of SDT model

In order to illustrate how the SDTA works, consider a training dataset $D = (\boldsymbol{X}, \boldsymbol{Y})$, consisting of 200 individuals with two SNVs, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, and a target variable $\boldsymbol{Y}$ (either diseased or healthy). Each SNV has three genotypes coded as 0, 1, and 2 (for major homozygote, heterozygote, and minor homozygote, respectively). Target Y can take two values 1 and 0 denoting *diseased* and *healthy* respectively. Of the 200 persons, 20 have $Y = 1$, and 180 have $Y = 0$.

Figure 3 shows an SDT model built using SDTA and the Bayesian score described in Section 3.2.2. The counts of healthy and diseased individuals assigned to each node are shown as [*number of healthy*, *number of diseased*]. This model contains only variables X_1 and X_2 as being informative for predicting Y in D. For example, the probability of being healthy or diseased in the bottom right leaf node, with leaf counts [2, 1], can be computed from the Bayesian estimator in Equation 1. Setting the prior α_{lk} to 1, which represents a uniform prior probability distribution,

$N_{lk =}$ number of healthy = 2

N_l = number of healthy + number of diseased = 3

$\alpha_l = 1 + 1 = 2.$

Therefore the probability of being healthy at this leaf node is (1+2)/(2+3) = 0.6.

Inference. Let A be a test individual (not present in D) for whom we want to predict Y, and for whom the SNV features are $X_1 = 2$, $X_2 = 1$, $X_3 = 0$, $X_4 = 2$, and $Y = 1$. The SDTA will find the path with features matching those of A. In this case, they are $X_1 = 2$, $X_2 = 1$, which leads to the leaf node with counts of [4, 3]. The probability of disease for A is (1+3)/(2+3+4) = 0.44

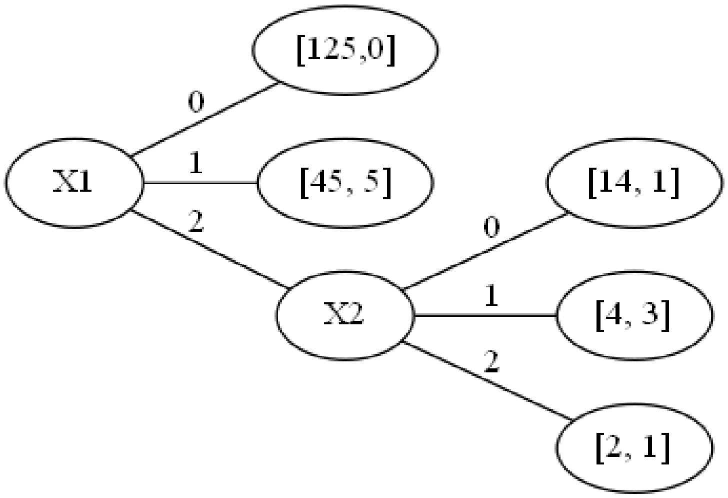


Figure 3: Example SDT model

3.3 PERSONALIZED DECISION TREE (PDT)

This section provides a detailed description of the PDT model, and then describes the PDTA that uses a Bayesian approach to create patient-specific (PS) path from the SDTA. Again, the model searches for a tree with highest posterior probability, this time using both the dataset *D* and the single test instance from D_{test}.

3.3.1 PDT model

The model constructed by the PDTA takes as input the features of the current person and outputs a probability distribution over the target variable. First, an SDT is built from *D* and fixed. This tree is used by the PDT to generate the PS path and its corresponding residual decision tree (RDT). The extension of the PS path is terminated if the addition of the best candidate predictor does not improve the score while building the model. Consequently, if there is no PS path that leads to improvement of PDT over the SDT, no PS path + RDT is created, and the PDT is identical to the SDT.

Model structure and parameters. Structurally, the PDT is comprised of two concurrent decision networks: One is the PS path that is personalized to the subject from D_{test}, the other is the RDT (similar to the SDT, but with certain predictors removed when they are chosen to be added to the PS). All training cases are assigned to a single path in either the PS or RDT – the model is scored considering both as a unit model M_{PDT}, built from D and the single instance from D_{test}. The PDT model differs from the SDT model by having L+1 paths, with the additional PS path set as $\boldsymbol{R}_{L+1}$.

Parameter estimation. The parameterization θ proceeds similar to that in SDT; the parameter is given by the same Equation. 1:

$$\theta_{lk} \equiv E[P(Y = k|X_{R_l} = x_{R_l})] = \frac{\alpha_{lk} + N_{lk}}{\alpha_l + N_l}, \qquad (11)$$

- with the notable difference: in this case, l ranges from 1 to L+1, to account for the additional PS path.

Inference. Given a test subject $\boldsymbol{X}_{test} = \boldsymbol{x}_{test}$, the parameterized model M is used to compute the probability distribution over the target variable Y_{test} of the test subject. If there is a PS path, then parameters in the leaf node of $\boldsymbol{R}_{L+1}$ are returned. If there is no PS path, then inference is performed on the RDT by sequentially identifying the outbound branch at each interior node, starting at the root node, by the value of the predictor associated with the node in the test subject the identify the relevant path ***R**_{test}*

the identify the relevant path $\boldsymbol{R_{test}}$ that fits the test subject features. The parameters in the leaf node of $\boldsymbol{R_{test}}$ are then returned.

3.3.2 Description of PDTA

PDTA uses a Bayesian approach similar to STDA. Below I detail the ways in which PDTA differs from STDA.

3.3.2.1 Model score

The derivation of the model score for PDTA is identical to that for SDTA. Since there are *L*+1 paths to consider in a PDT, I modify Equation 4 (Section 3.2.2) to set the score of model *M* as:

$$
\begin{aligned}
&logscore(M_{PDT}) \\
&\quad = \left[\sum_{i\in S}\log(p_i) + \sum_{i\notin S} log(1-p_i)\right] \\
&\quad + \Bigg[\sum_{l=1}^{L+1}\big(\log(\Gamma(\alpha_l)) - \log(\Gamma(\alpha_l + N_l))\big) \\
&\quad + \Bigg[\sum_{k=1}^{K}\big(\log(\Gamma(\alpha_{lk} + N_{lk})) - \log(\Gamma(\alpha_{lk}))\big)\Bigg]\Bigg] \qquad (12)
\end{aligned}
$$

3.3.2.2 Search strategy

The PDTA fixes the tree derived by SDTA, then uses a heuristic greedy search to derive $\boldsymbol{R_{L+1}}$. All of the predictors, including the ones in the tree, are allowed to be candidates for inclusion in the PS path.

The pseudocode for growing the PDT is given in Figure 4. The search begins by selecting a predictor X_{chosen} for the PS path, and assigning to it the set of individuals from D who all have the same value for the X_{chosen} as the current patient from D_{test}; that set is denoted $D_{(L+1)}$. The individuals in $D_{(L+1)}$ are removed from the paths in full tree to which they had been originally assigned, leaving the remainder RDT. This process of selecting a predictor and extending the PS path continues. Similar to the SDT search method, the PS path is extended by adding one best predictor at a time, and the goodness of a candidate predictor is evaluated by temporarily adding it to the PS path and computing the score of M_{PDT}. The extension of the PS path is terminated if the addition of the best candidate predictor does not improve the score of the current model.

The PDTA may not find a distinct person-specific path for every person. When PDTA is unable to improve upon any of the paths in the SDT, it uses one of the SDT paths to predict the current person.

```
PDTA (X, D, $D_{test}$)
   Input:              X is a set of predictor variables
                       D is a training dataset of individuals described using X, a set of predictor
                       variables, and target variable Y
                       $D_{test}$ is data for individual of interest described using X
    Output:    Residual_Tree + PS where PS is patient-specific path personalized with $D_{test}$


       Residual_Tree = SDTA (X, D)
       PS = null
       free_X = list of all predictor variables
       score = Bayesian score of Residual_Tree + PS computed using Equation 12
       LOOP
       best_Score = score
       found_Best_X = False
       temp_Residual_Tree = Residual_Tree
       temp_PS = null
       FOR each X in free_X
              Grow temp_PS by replacing leaf_Node with X = value in $D_{test}$
              Remove individuals from temp_Residual_Tree who are assigned to temp_PS
              temp_Score = Bayesian score of temp_Residual_Tree + temp_PS computed
                                                   using Equation 12
       IF temp_Score > best_Score
                     found_Best_X = True
                     best_Score = temp_Score
                     best_X = X
              END IF
       END FOR
       IF found_Best_X = True
              Grow PS by replacing leaf_Node with best_X
              Remove best_X from free_X
              Remove individuals from Residual_Tree who are assigned to PS
       ELSE
              EXIT from LOOP
       END IF
   END LOOP
   Return Residual_Tree + PS
```

Figure 4: Pseudocode for PDTA

3.3.3 Example PDT model

In order to illustrate how the PDT works, let us consider the same training dataset D from Section 3.2.3. Figure 5 shows the PDT model built using PDTA and the score as described in Section 3.3.2.1, comprised of a PS path and an RDT. As in the SDT example, the parameters in the leaf nodes are estimated from *D* using the Bayesian estimator in Equation 1.

The PS path, which contains the SNV $X_2 = 2$, is applicable only to the subgroup in D whose members have X_2 with the value 2, as does the current person. The PS path does not contain the predictor X_1 because it did not improve the prediction for this subgroup. The residual tree serves as the default predictive model for those individuals in *D* who do not satisfy the PS path (i.e., for individuals whose features are not similar to the current person). The shaded leaf nodes in the RDT indicate leaf nodes from which individuals were removed to populate the leaf node of the PS path.

Inference. Given test individual *A*, and the PS path which was built to match the features in *A*, and $X_1 = 0$, $X_2 = 2$, the probability of disease is (5+1) / (5+1+15+1) = 0.273, which is higher than the 0.008 predicted by the model in Figure 5.

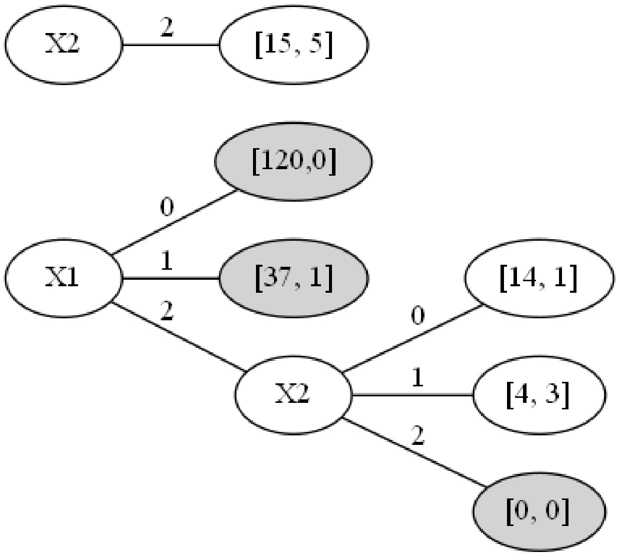


Figure 5: Example PDT model

4.0 EXPERIMENTAL METHODS

This chapter describes the datasets, the performance metrics, the gene and variant information and comparison algorithms. Section 4.1 describes the datasets used in the experiments. Section 4.2 describes the performance metrics used for model evaluation. Section 4.3 describes details of gene and variant information used in this dissertation. Section 4.4 gives details of the comparison algorithms.

4.1 DATASETS

This section describes the datasets used to develop and evaluate the algorithm. Each of them has a single binary target variable per sample to indicate its case/control status. The datasets vary in size, complexity, variable type and biological realism, offering a multifaceted view of algorithm performance. Section 4.1.1 describes the Synthetic dataset (SD), which was created to emulate a SNV dataset, with preset causality and frequency in the generation of variables. Section 4.1.2 describes the small biological dataset (SBD), comprised of a curated selection of SNVs and wildtype/mutation pairs which are biologically realistic, with lower complexity than a full exomic dataset. Section 4.1.3 describes the semi-synthetic dataset (SSD), which has the full complexity of a real genomic dataset, but the true causal variables are known. Section 4.1.4 describes a whole exome SNV dataset containing samples marked as cases of chronic pancreatitis or recurrent acute pancreatitis (WE-CP), along with creation of a reduced variable version (WE-CP-R) of the dataset. Section 4.1.5 describes a subset comprised of the pancreatitis cases only, from WE-CP, matched for presence or absence of diabetes as phenotype (WE-DB). This section also contains details of the creation of the reduced variable version (WE-DB-R). Table 1 provides an overview of all the datasets.

Table 1: Summary of datasets

	Dataset	Number of samples	Number of variables	Any variable values missing?	Biological?	Train set size	Test set size
1	SD	10,000	1,000	No	No	9000	1000
2	SBD	2,201	155	Yes	Yes	1761	440
3	SSD	6,970	24,487	No	Mixed	5577	1393
4	WE-CP	2,135	246,012	Yes	Yes	1708	427
	WE-CP-R	2,135	516	Yes	Yes	1708	427
5	WE-DB	1420	245,573	Yes	Yes	1137	283
	WE-DB-R	1420	638	Yes	Yes	1137	283

4.1.1 Synthetic Dataset (SD)

The SD consists of 1,000 SNVs and a binary disease variable that was modelled as a function of 35 “causal” SNVs. Of the 35 SNVs, 25 of them were modeled as rare variants, with minor allele frequencies (MAFs) that were sampled uniformly from the range 0.0001 to 0.01 and odds ratios in the range 2 to 10. The remaining 10 of the 35 SNVs were modeled as common variants, with MAFs in the range 0.05 to 0.50 and odds ratios in the range 1.05 to 1.50. This choice was motivated by current thinking that common variants have smaller effects than rare variants. The remaining 965 SNVs ("noise" variants) ranged from common to rare, but they do not have an effect on the disease. The dataset consists of 10,000 samples of which 13.3% had disease and the remaining were healthy.

4.1.2 Small Biological Dataset (SBD)

The SBD consists of 2,201 individuals, of whom 980 are diseased (diagnosed with CP or RAP) and 1,221 are healthy. I pre-processed the dataset to remove some of the individuals and the non-genetic variables. The final dataset contained 2,201 individuals and 155 genetic variables, of which 142 were SNVs and 13 were binary variables (classified as either wildtype or mutant). This dataset was curated to create a small biological dataset containing a mix of clinical variables that may or may not be associated with CP or RAP.

4.1.3 Semi-Synthetic Dataset (SSD)

The SSD is a mini-exome semi-synthetic dataset that was constructed for the Genetic Analysis Workshop 17 (GAW17) held in 2010 at Boston, Massachusetts[63,64]. The data was obtained from 697 unrelated individuals whose exomes were sequenced in the 1000 Genomes Project and the genomic data consists of 24,487 autosomal SNVs that map to 3,205 genes. This is a mini-exome dataset since the 3,205 genes comprise a subset of all human genes. The synthetic portion of the dataset consisted of four quantitative risk factors that were simulated as normally distributed phenotypes. The genes associated with each of the risk factors were chosen from cardiovascular risk and inflammation pathways.

For evaluation of PDTA, I used one of the 4 phenotypes as target, which was influenced by 72 SNVs in 13 genes and no other external factors. Of the causative SNVs for the phenotype, 38.4% are private variants – with only one individual out of 697 having the variant – and 12.8% SNVs are common with MAF > 0.05. The data in the first ten GAW17 replicates were pooled to create the SSD with 24,487 SNVs and 6,970 samples.

4.1.4 Whole Exome Chronic Pancreatitis (WE-CP)

The primary data set was obtained using an IBD Exome chip with 246,212 markers, collected as part of the NAPS2 project[65]. The dataset had 2135 samples, of which 1,420 patients were diagnosed with either CP or RAP, and 715 were controls. Since missingness is treated as a feature (i.e., it can be assigned to a variable as a value) in PDTA, all variants that were missing in the diseased cases at a statistically significant difference from controls were removed, leaving 246,012 SNVs in the dataset used as WE-CP. 1708 samples were used to train the model and427 samples were used to test the model. Notably, no commonly known SNPs for PRSS1 were present in WE-CP.

In order to reduce the number of unpredictive SNVs in the dataset, all SNVs were tested for association with the phenotype using a chi-square test, and only those with $p < 10^{-4}$ were kept in a new dataset, WE-CP-R. WE-CP-R contained 516 SNVs, out of which 272 had MAF < 0.05.

4.1.5 Whole Exome Pancreatitis with Diabetes (WE-DB)

All control subjects were removed from WE-CP to create WE-DB, with 1,420 CP-positive patients. This dataset was divided into cases and controls based on presence of endocrine insufficiency. 374 patients had diabetes and 1046 patients were controls. PDTA was applied to WE-DB using 1137 samples to train the model, and 283 to test predictions.

In order to reduce the number of unpredictive SNVs in WE-DB, all SNVs were tested for association with the phenotype using a chi-square test, and only those with $p < 10^{-4}$ (642 SNVs) were saved in a new dataset, WE-CP-R. After adjusting for missingness, WE-DB-R was left with 638 SNVs, out of which 395 had MAF < 0.05.

4.2 PERFORMANCE METRICS

4.2.1 Area under the ROC curve (AUC)

Receiver Operating Characteristic (ROC) curves are commonly used to evaluate the predictive performance of machine learning algorithms, with the area under ROC (AUC) used as a summary statistic of discrimination[66]. In brief, the ROC curve illustrates the tradeoff between true positive rate (also known as sensitivity) and false positive rate (incorrectly identified controls). An algorithm with perfect discriminability would have an AUC of 1; one that performs no better than chance would have an AUC of 0.5. To evaluate the predictive performance of an algorithm I computed the AUC from the predictions for each individual in the test set. Confidence intervals for the AUCs were computed using method of DeLong, using the pROC package in R[67,68].

4.2.2 Minimal Split Ratio (MSR)

In this work, I created a metric called Minimal Split Ratio (MSR), as a measurement of how susceptible the model may be to noise. The MSR of a tree model is based on the proportion of terminal leaf nodes classifying on a low count of samples versus a robust number of samples, with a slight smoothing function.

4.2.3 Causal variable detection

For the datasets in which the casual variables are known, I report positive predictive value (PPV), true positive rate (TPR), and the F1 score. When the algorithm predicts an instance as positive for a condition, a 'true positive' (TP) is when the instance was correctly identified as positive, a 'false positive' (FP) is when the instance was incorrectly identified as positive. Likewise, the prediction of an instance as negative for the condition can be either 'true negative' (TN), when the absence of the condition is correctly identified, or a 'false negative' (FN), when the presence of the condition was missed. In evaluating a predictor, the acceptable tradeoff between TP, FP, TN and FN depends on the priorities of the application. The PPV, also known as precision, gives the ratio of TP to the total number of instances predicted as positive. The TPR, also known as sensitivity or recall, gives the ratio of TP to the total number of instances that were truly positive for the condition. The F1 score is a weighted average of PPV and TPR, ranging in value between 0 and 1, with values closer to 1 better than values closer to 0.

4.3 GENE AND VARIANT INFORMATION

4.3.1 PDTA conventions

The PDTA is designed to consider all possible values of a variable in determining the edge, or branch, to be placed emerging from a node. Since I am interested in representing SNVs at the nodes, I categorize the possible values of the variable as:

0: Major homozygous, or wildtype
1: Heterozygous
2: Minor homozygous, or biallelic mutant
3: Missing

Therefore, listing the paths in a tree (Figure 6) where the root node, rs123, branches to node rs456 when it is heterozygous (SNV=1), and to node rs789 when it is minor homozygous (SNV=2), would be represented by the two paths:

```
rs123--[1]--rs456
rs123--[2]—rs789
```

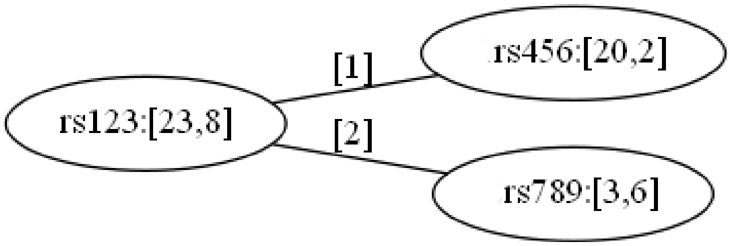


Figure 6: Example tree with two paths, node counts and edges. Brackets inside the node show counts of instances at that node for the binary outcome (such as [*healthy*, *disease*])

In all analyses, SNV with MAF < 0.05 are referred to as rare SNVs.

4.3.2 Tools and Databases

PLINK was used for association analysis, MAF calculation and manipulation of WE-CP and WE-DB[69]. NCBI dbSNP was referenced for information on all model SNVs, retrieved before June 2018. The global MAF I use from dbSNP are the values reported by NHLBI Trans-Omics for Precision Medicine (TOPMed)[70]. A full list of databases used for referencing genes and SNVs in this work is shown in Table 2.

Table 2: Resources for gene and variant information

	Web location	Description
dbSNP [71]	www.ncbi.nlm.nih.gov/snp/	Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants
ClinVar [72]	www.ncbi.nlm.nih.gov/clinvar/	Public archive of relationships among genomic variation and human phenotype
Pancreas Genetics [73]	www.pancreasgenetics.org	Database of gene variants assembled from published reports and personal submissions from investigators
Pancreatic Cancer Database [74]	www.pancreaticcancerdatabase.org	Database of manually curated molecular alterations associated with pancreatic cancer from research articles
SNPedia [75]	www.snpedia.com	Bioinformatics wiki supporting genome annotation, interpretation and analysis

4.4 COMPARISON ALGORITHMS

4.4.1 Regular decision tree

To compare the performance of PDTA to a regular decision tree, I used the decision tree classifier in Scikit-learn with default parameters[76]. This package uses an optimized version of the

Classification and Regression Trees (CART) algorithm, a non-parametric learning method which grows a tree based on information gain at each node[77].

4.4.2 K2 Bayesian score

To assess the choice of BDeu score in PDTA, I compare its performance to implementations of PDTA which instead used another Bayesian score, K2[78]. The BDeu and K2 scores both derive from the Bayesian Dirichlet (BD) score; the K2 score is a specialization of the BD score which assumes an uninformed prior assignment. Unlike BDeu, the K2 score does not require specification of parameter priors. Variations of the K2 score have been used in other studies as uniform prior score metric (UPSM) and Dirichlet prior score metric (DPSM)[79,80].

4.4.3 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) approaches scoring Bayesian networks as an information theoretic task: it estimates the quality of each model according to relative information loss when that model is used to represent the process that generated the data[81].

4.4.4 Minimum Description Length (MDL) score

The Minimum Description Length (MDL) score also approaches scoring Bayesian networks as an information theoretic task[82]. MDL is often used interchangeably with the BIC, as they are equivalent under certain conditions[80]. In many applications, MDL has been shown to outperform other scores such as BDeu and AIC, but it may not work well on smaller datasets[83,84].

5.0 EXPERIMENTAL RESULTS

This chapter describes the results of applying the algorithms described in Chapter 3 to the datasets described in Chapter 4. Section 5.1 describes the iterative development of PDTA on SD. Section 5.2 describe the evaluation of the final developed PDTA on the rest of the datasets.

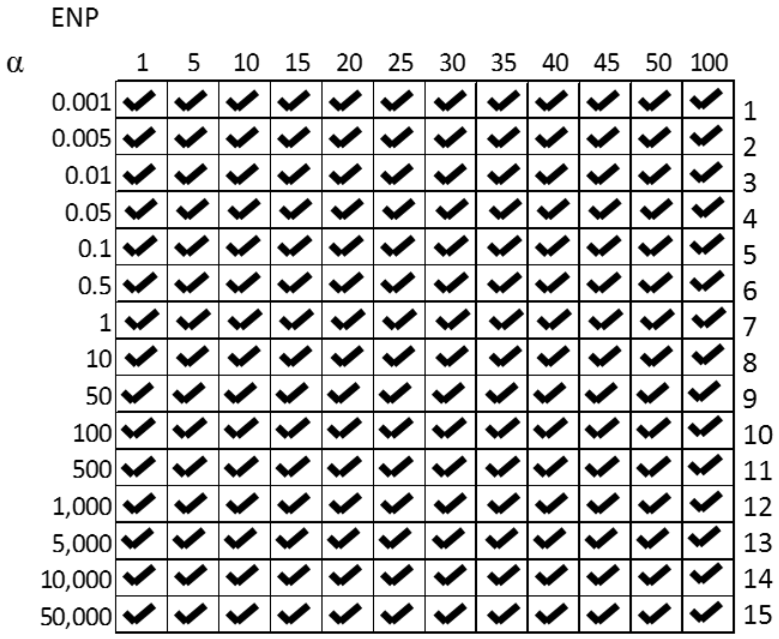
5.1 EVALUATION AND ITERATIVE REFINEMENT OF PDTA ON SD

In this section, I describe the algorithmic performance on a synthetic dataset, and test improvements to PDTA, first optimizing the initial tree building part of the algorithm (SDTA), followed by the instance-based personalization part of the algorithm (PS paths). The score used by PDTA for choosing nodes requires the setting of values for two hyperparameters: α and ENP. In Section 5.1.1, I present an assessment of the effect that these hyperparameter values have on the model. This is followed by further experiments in refining the algorithm for PDTA model selection (5.1.2, 5.1.3). In Section 5.1.4, I summarize my observations on SD.

5.1.1 Empirical Optimization of PDT on SD

The SD is a dataset where the true causal predictors are known. Therefore, we can assess the performance of PDTA on SD not only for accuracy in classification (measured by AUC), but also for accuracy in the selection of causal variables (measured by F1, PPV and TPR). The PDTA score requires two hyperparameters to be set: ENP, which influences the structure prior, and *α*, which influences the BDeu score. First, SDTA was applied to the SD with the combination of *α* and ENP values as shown in Table 3.

Table 3: Hyperparameter settings used in SDTA on SD



α \ ENP	1	5	10	15	20	25	30	35	40	45	50	100	
0.001	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1
0.005	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2
0.01	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	3
0.05	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	4
0.1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	5
0.5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	6
1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	7
10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	8
50	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
100	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
500	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	11
1,000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	12
5,000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	13
10,000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	14
50,000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	15

The trend of SDTA performance with these settings is shown in terms of AUC (Figure 7a) and F1 score (Figure 7b). In general, a higher F1 score indicates the model is better at retrieving the true causal variables as predictive variables (nodes in the tree), while minimizing false negatives. It is important to note that there are only 35 causal variables in the SD, so the F1 score for large trees which have more than 35 nodes will automatically suffer from a lower PPV. An additional metric, which I call MSR (Figure 7c), is a measurement of how susceptible the model may be to noise, based on the proportion of leaves classifying on a low count of samples versus a robust number of samples. Lower values of ENP produce SDTs with better AUC, F1 and MSR across all values of α.

Table 4: 10 best SDTs on SD, ranked by lowest MSR

Model	MSR	AUC	95% CI	α	ENP	# of leaves	# of nodes	# of causal SNVs	PPV	TPR	F1
SDT1	0.431	0.987	0.982-0.992	50	1	49	24	24	1.000	0.686	0.814
SDT2	0.449	0.987	0.982-0.992	10	1	47	23	23	1.000	0.657	0.79
SDT3	0.449	0.987	0.982-0.992	1	1	47	23	23	1.000	0.657	0.79
SDT4	0.479	0.986	0.980-0.992	50	5	71	35	28	0.800	0.800	0.80
SDT5	0.489	0.983	0.974-0.993	0.001	1	43	21	20	0.952	0.571	0.71
SDT6	0.490	0.983	0.974-0.993	0.005	1	47	23	22	0.957	0.629	0.76
SDT7	0.490	0.983	0.974-0.993	0.01	1	47	23	22	0.957	0.629	0.76
SDT8	0.490	0.986	0.980-0.992	0.05	1	49	24	23	0.958	0.657	0.78
SDT9	0.490	0.986	0.980-0.992	0.1	1	49	24	23	0.958	0.657	0.78
SDT10	0.490	0.986	0.980-0.992	0.5	1	49	24	23	0.958	0.657	0.78

Higher values of ENP and α tend to grow larger trees (Figure 8a). The number of nodes in a tree increases monotonically with increase in the value of ENP (Figure 8b), but not with increase in the value of α (Figure 8c). Lower MSR values are seen with smaller tree sizes (Figure 8d), and many of these also have the highest AUC values (Figure 8e). The tree with the lowest

MSR value, SDT1 in Table 4 (AUC = 0.987, 95% C.I.: 0.982-0.992, F1 = 0.814) performed almost as well as the tree with overall highest AUC value (AUC = 0.989, 95% C.I.: 0.985-0.994, F1 = 0.889), which was produced with parameters $\alpha = 1000$, ENP = 1.

As can be seen in Figures 1 and 2, trees with lower MSR values tend to have better performance as measured by both AUC and F1, and lower ENP values are sufficient to generate high scoring SDTs across all measures of performance. From these observations, I hypothesize that larger SDTs, with higher ENP values – which take much longer to build – may be a poor use of resources, and we can select the best SDT by choosing the model with the lowest MSR value. Therefore, the α and ENP values for lowest MSR SDT, SDT1, are applied to test PDTA on the SD. The best ENP value observed for SDT was retained as tree ENP, and a range of path ENPs were used to test for PS path creation (Table 5).

Table 5: Results of PDTA on SD with best SDT parameters
(SDT1 in Table 4)

Model	α	**Tree ENP**	**Path ENP**	**# of times PS chosen (out of 1000)**	**SDT AUC**	**95% CI**	**PDT AUC**	**95% CI**
PDT1	50	1	1	0	0.987	0.982-0.992	0.987	0.982-0.992
PDT2	50	1	20	12	0.987	0.982-0.992	0.986	0.981-0.992
PDT3	50	1	50	12	0.987	0.982-0.992	0.986	0.981-0.992
PDT4	50	1	100	12	0.987	0.982-0.992	0.986	0.981-0.992

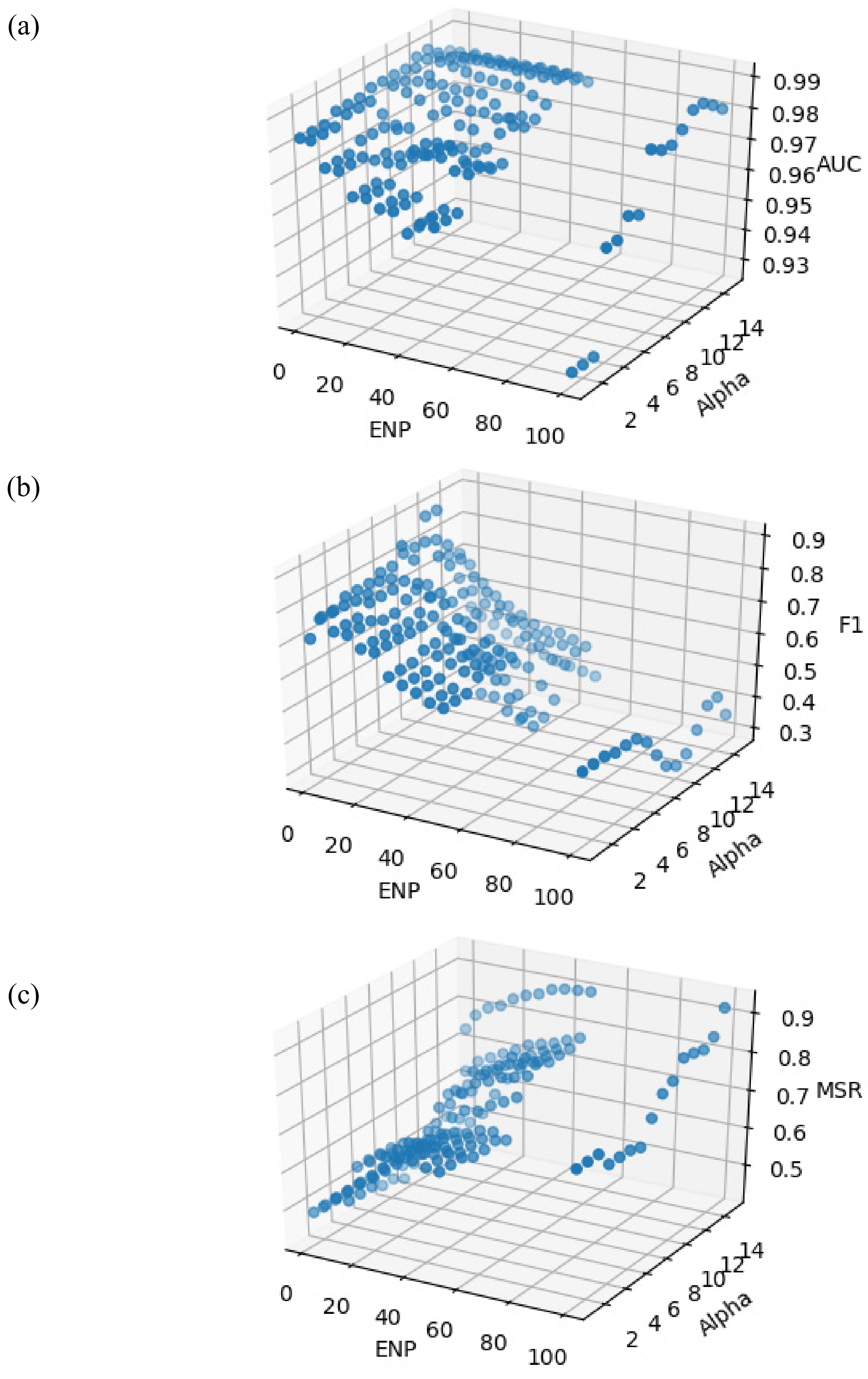


Figure 7: Trend of SDTA on SD
(a) AUC, (b) F1, (c) MSR

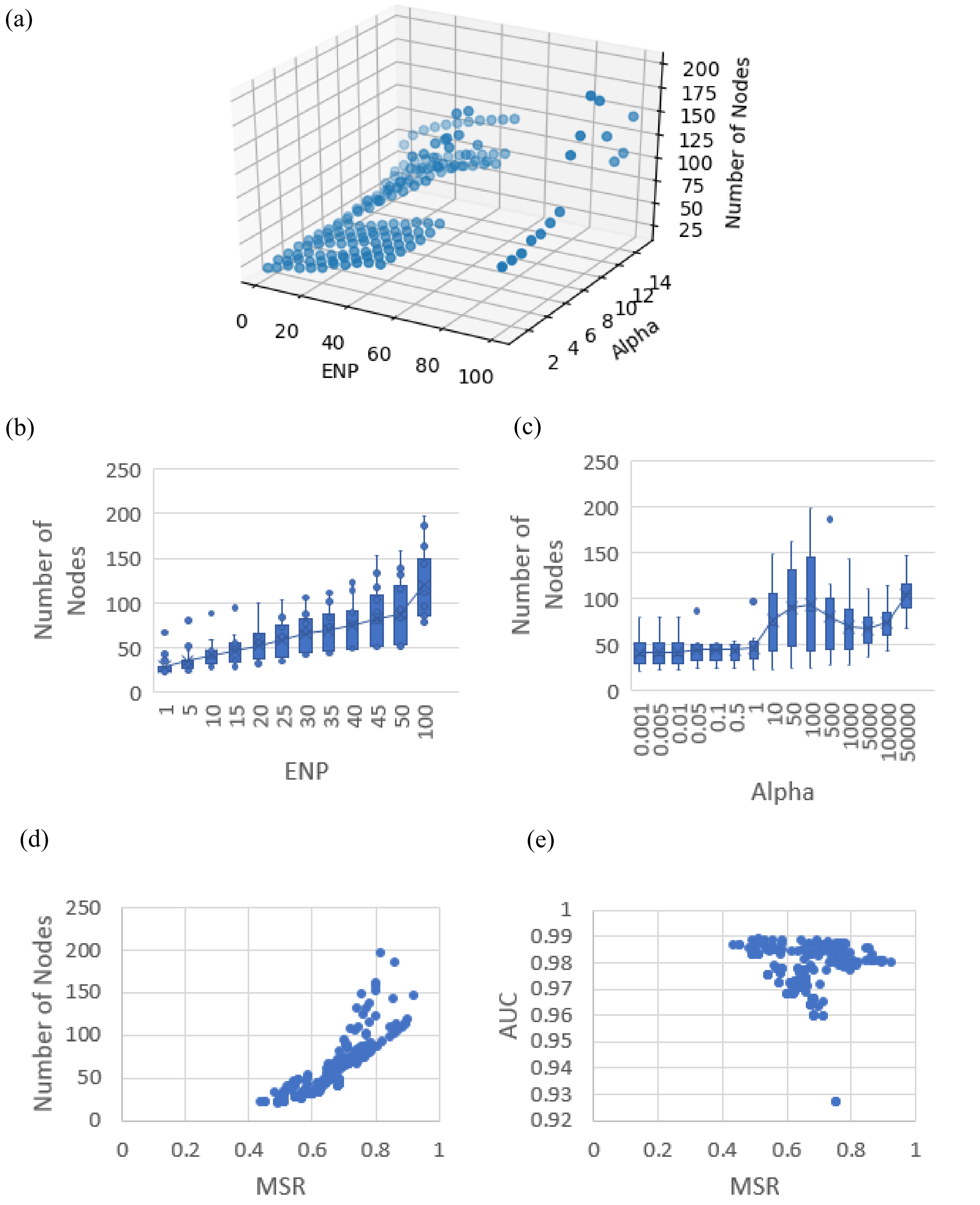


Figure 8: Trend of SDTA on SD, continued

(a) Size of trees trend by both hyperparameters; (b) Size of trees grow monotonically with ENP, (c) Size of trees grow with Alpha, but not monotonically, (d) & (e) Smaller trees have lower MSR but still high AUCs.

As seen in Table 5, in PDT1 (the lowest attempted path ENP), no PS paths were selected. That means that all patients in the test dataset matched best to a path in the SDT. In PDT2, however, 12 out of the 1000 test instances were assigned to PS paths, but there was no improvement in AUC. Increasing the path ENP (PDT3, PDT4) neither recruited additional PS paths to the model, nor improved AUC. Therefore, the best model generated for SD by PDTA is SDT1 from Table 4. The frequency of paths in SDT1 chosen by test instances is shown in Table 6. All SNVs chosen as internal nodes in the SDT are causal variables, and most of the nodes chosen per path are rare causal variables (Table 6).

When PS paths were selected by test instances, there were 3 unique paths, 2 of which included rare causal variables. PDT2 (Figure 9a) had the lowest path ENP value where PS paths were created. As path ENP increased, PS paths were created again in PDT3and PDT4, but they differed from the paths in PDT2 only in the addition of a single SNV, v372 (Figure 9b). This SNV, v372, was one of the noise variants in SD, not a causal variant. Recall the earlier observation that higher tree ENPs tend to generate larger SDTs without a corresponding improvement in performance (Figure 8a); it appears, similarly, that higher path ENPs will add more nodes to the PDT, even without adding value to the model.

Table 6: Paths selected from SDT1 by SD test instances, ranked by frequency

Path		# of times selected	Disease Prediction	Target	Path length	Terminal Leaf	# of causal SNVs	# of rare causal SNVs
SDT_Path1	v24--[0]--…--v30--[0]	514	[1.0,0.0]	[514,0]	19	[4781,5]	19	18
SDT_Path2	v24--[0]--…--v27--[0]	204	[1.0,0.0]	[204,0]	22	[1773,7]	22	20
SDT_Path3	v24--[0]--…--v27--[1]	58	[0.96,0.04]	[58,0]	22	[357,13]	22	20
SDT_Path4	v24--[0]--…--v30--[2]	32	[0.98,0.02]	[32,0]	19	[258,4]	19	18
SDT_Path5	v24--[0]--…--v2--[1]	27	[0.1,0.9]	[1,26]	2	[16,148]	2	2
SDT_Path6	v24--[0]--…--v17--[1]	14	[0.03,0.97]	[0,14]	3	[3,111]	3	3
SDT_Path7	v24--[0]--…--v22--[1]	12	[0.22,0.78]	[4,8]	5	[24,84]	5	5
SDT_Path8	v24--[0]--…--v4--[1]	12	[0.48,0.52]	[4,8]	9	[61,66]	9	9
SDT_Path9	v24--[0]--…--v14--[1]	11	[0.47,0.53]	[6,5]	10	[46,53]	10	10
SDT_Path10	v24--[0]--…--v8--[1]	11	[0.26,0.74]	[3,8]	6	[29,85]	6	6
SDT_Path11	v24--[0]--…--v12--[1]	11	[0.49,0.51]	[5,6]	8	[75,77]	8	8
SDT_Path12	v24--[0]--…--v9--[1]	10	[0.71,0.29]	[5,5]	15	[63,26]	15	15
SDT_Path13	v24--[0]--…--v5--[1]	10	[0.54,0.46]	[3,7]	11	[63,54]	11	11
SDT_Path14	v24--[0]--…--v13--[1]	8	[0.79,0.21]	[6,2]	16	[79,21]	16	16
SDT_Path15	v24--[0]--…--v10--[1]	8	[0.04,0.96]	[0,8]	4	[4,96]	4	4
SDT_Path16	v24--[0]--…--v7--[1]	8	[0.79,0.21]	[5,3]	17	[51,13]	17	17
SDT_Path17	v24--[1]--…--v26--[0]	7	[0.17,0.83]	[0,7]	2	[13,64]	2	1
SDT_Path18	v24--[0]--…--v19--[1]	6	[0.05,0.95]	[1,5]	12	[1,27]	12	12
SDT_Path19	v24--[0]--…--v23--[1]	5	[0.81,0.19]	[4,1]	21	[31,7]	21	20
SDT_Path20	v24--[0]--…--v27--[2]	5	[0.94,0.06]	[5,0]	22	[23,1]	22	20
SDT_Path21	v24--[0]--…--v11--[1]	5	[0.18,0.82]	[2,3]	13	[5,24]	13	13
SDT_Path22	v24--[0]--…--v1--[1]	5	[0.22,0.78]	[4,1]	14	[6,22]	14	14
SDT_Path23	v24--[1]--…--v26--[1]	4	[0.01,0.99]	[0,4]	2	[0,74]	2	1
SDT_Path24	v24--[0]--…--v25--[1]	3	[0.02,0.98]	[0,3]	8	[0,32]	8	7
SDT_Path25	v24--[0]--…--v25--[2]	2	[0.1,0.9]	[0,2]	8	[0,4]	8	7
SDT_Path26	v24--[1]--…--v26--[2]	2	[0.02,0.98]	[0,2]	2	[0,26]	2	1
SDT_Path27	v24--[0]--…--v6--[1]	2	[0.59,0.41]	[2,0]	20	[9,6]	20	19
SDT_Path28	v24--[0]--…--v0--[1]	2	[0.69,0.31]	[2,0]	18	[21,9]	18	18
SDT_Path29	v24--[0]--…--v14--[2]	1	[0.5,0.5]	[0,1]	10	[0,0]	10	10
SDT_Path30	v24--[0]--…--v25--[0]	1	[0.42,0.58]	[0,1]	8	[19,26]	8	7

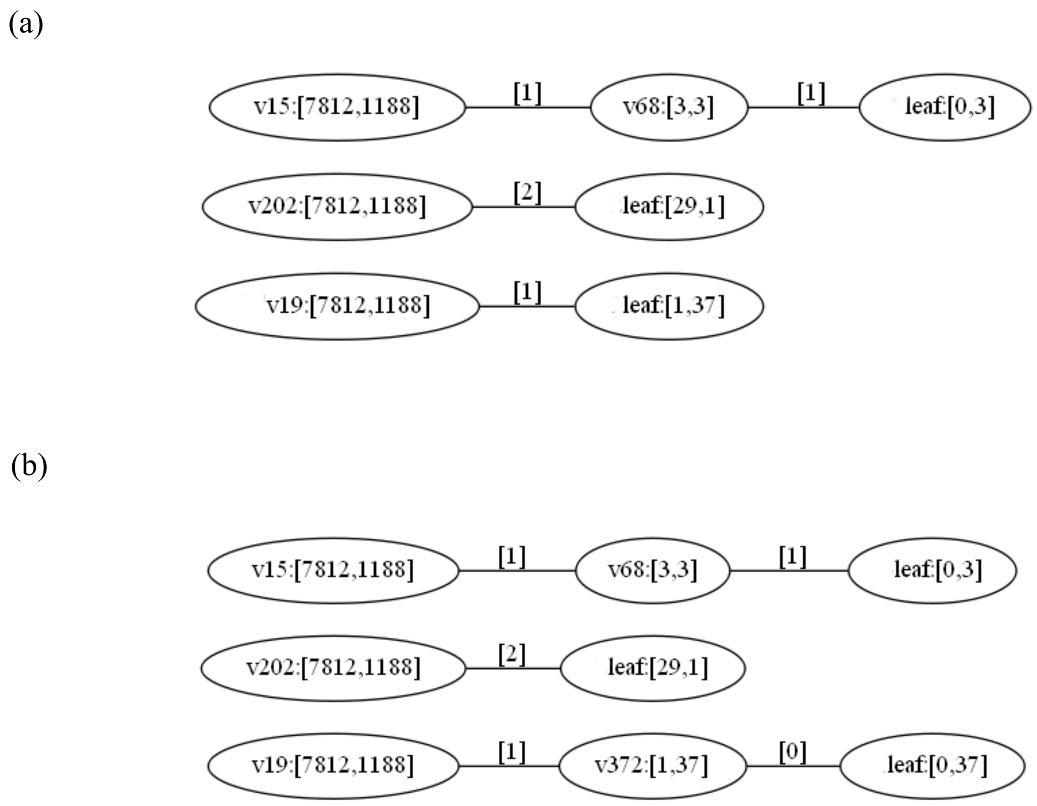


Figure 9: PS paths generated by PDTA
(a) in PDT2, Table 5, (b) in PDT3 and PDT4, Table 5

5.1.2 Refinements to improve generalization

The PDTA only builds trees by growing them, as the BDeu score is expected to control well for overfitting[49]. However, it could be worthwhile to check if the tree can be improved by taking measures to reduce overfitting at the leaves. There are two common ways to accomplish this: one is by incorporating a termination rule in the tree growth phase, such that the tree stops adding nodes when it has reached a minimum number of sample counts; another is by adding a tree pruning phase after the tree growth phase. Both of these methods were tried on SDTA with the following results.

When a termination rule was added to the SDTA, the effect of parameter settings (Figure 10, 11) are qualitatively different from the original SDTA (Figure 7). There are a few similarities in trends: trees with lower MSR values (Figure 10c) show better performance by both AUC (Figure 10a) and F1 (Figure 10b), higher values of ENP and α tend to grow larger trees (Figure 11a), and the growth in number of nodes in a tree increases monotonically with value of ENP (Figure 11b), but not with value of α (Figure 11c). Compared to the original SDTA, however, the resulting trees were much smaller (Figure 11a), and performed significantly worse (Figure 10a, 11a). Smaller trees tend to have better MSR values (Figure 11d), and SDTs with lower MSR values have higher AUCs (Figure 11e), so the choice of best model by lowest MSR value is still sound. Table 7 shows details on the 10 best trees, ranked by the lowest MSR values. The tree with lowest MSR value, SDT1 in Table 7 (AUC = 0.677, 95% CI: 0.623-0.732; F1 = 0.2) performed only slightly less well than the tree with overall highest AUC value (AUC = 0.689, 95% CI: 0.637-0.742; F1 = 0.34), which resulted from parameters α = 50, ENP = 15. These results show that adding a termination rule to the tree-building phase diminishes the performance of SDTA.

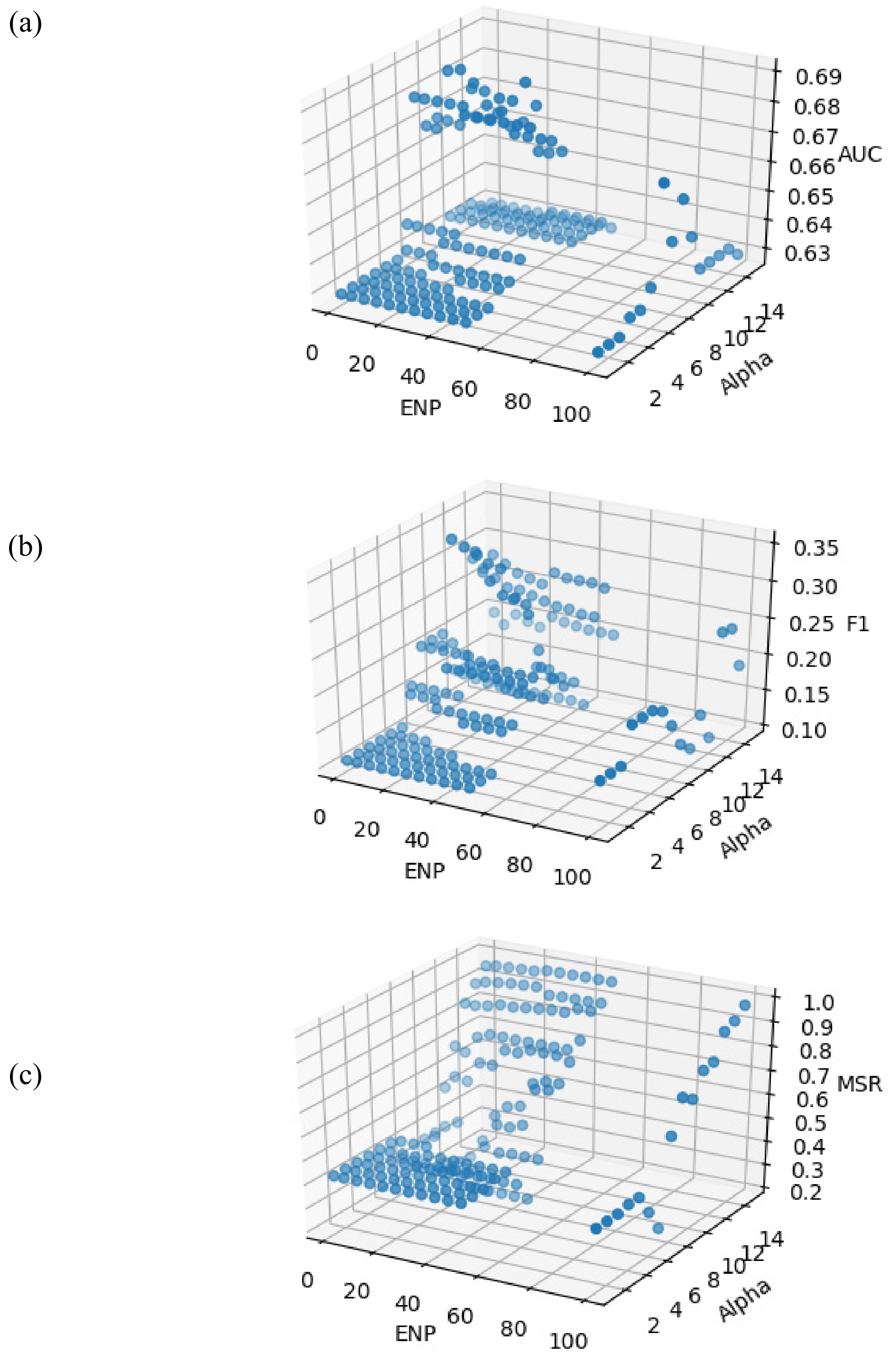


Figure 10: Trend of SDTA on SD with termination rule
(a) AUC, (b) F1, (c) MSR

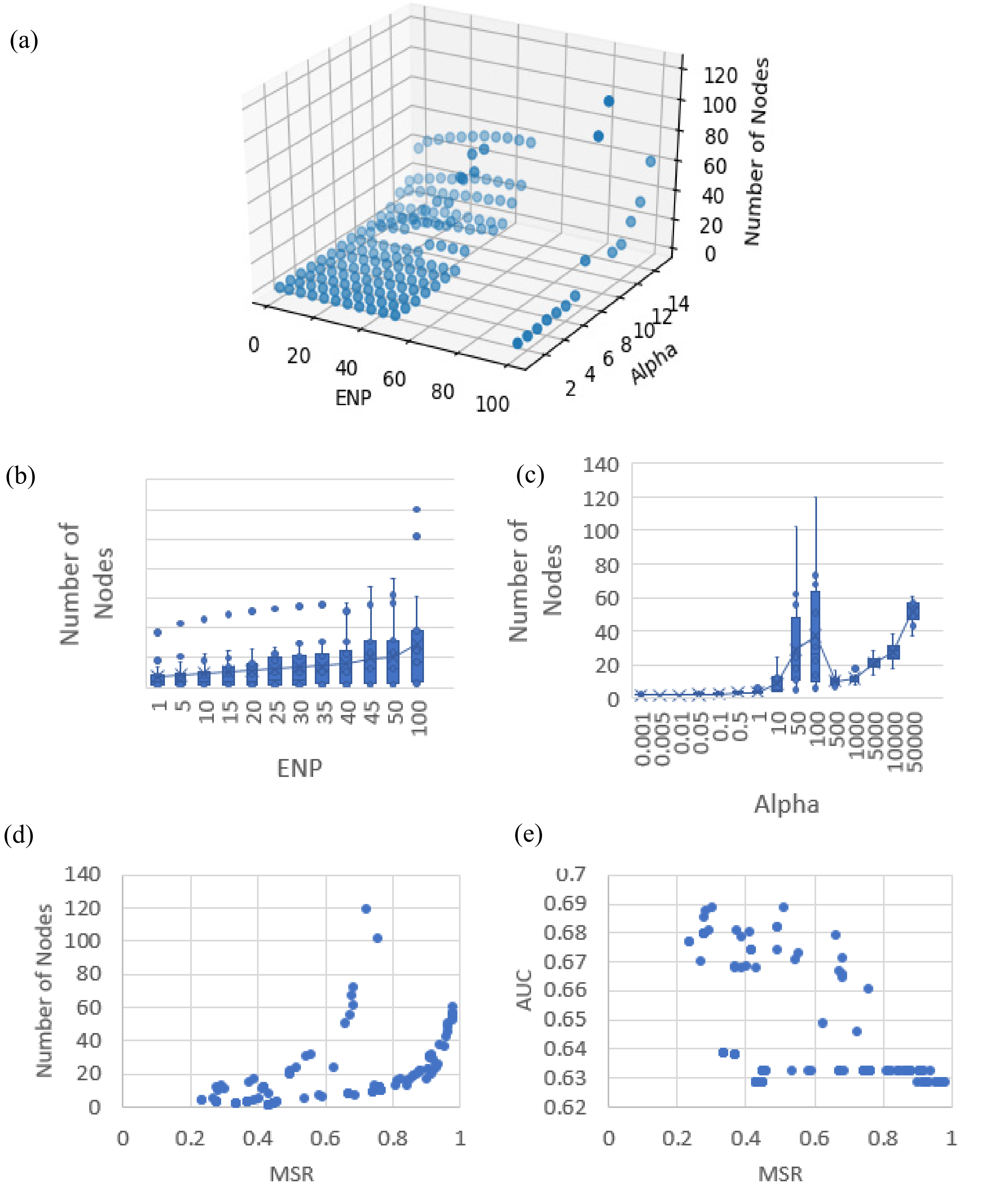


Figure 11: Trend of SDTA on SD with termination rule, continued
(a) Size of trees trend by both hyperparameters, (b) Size of trees grow monotonically with ENP, (c) Size of trees grow with Alpha, but not monotonically, (d) & (e) Smaller trees have lower MSR but still high AUCs

Instead of adding a termination rule, the other method of reducing overfit was implemented by allowing SDTA to grow a full tree based solely on the best score, with a tree pruning phase added after the growth phase. As seen in Figures 12 and 13, the trends of parameter settings on performance metrics are almost identical to the trends seen for the original SDTA (Figures 7 and 8). The best performing models (Table 8) also look similar to the results obtained before pruning was applied (Table 4), and the best performing tree SDT1 in Table 8 (AUC = 0.987, 95% CI: 0.982-0.992); F1 = 0.814) performed just as well as SDT1 in Table 4.

Based on these observations, I concluded that:

(1) The addition of a termination rule to the tree building phase of the algorithm is detrimental to PDTA.

(2) The addition of pruning following the tree building phase of the algorithm is not detrimental to PDTA.

Table 7: 10 best SDTs on SD with termination rule, ranked by lowest MSR

Model	MSR	AUC	95% CI	*α*	ENP	# of leaves	# of nodes	# of causal SNVs	PPV	TPR	F1
SDT1	0.231	0.677	0.623-0.732	10	25	11	5	4	0.800	0.114	0.200
SDT2	0.231	0.677	0.623-0.732	10	30	11	5	4	0.800	0.114	0.200
SDT3	0.267	0.671	0.616-0.726	1	100	13	6	4	0.667	0.114	0.195
SDT4	0.273	0.680	0.624-0.736	10	10	9	4	4	1.000	0.114	0.205
SDT5	0.273	0.680	0.624-0.736	10	15	9	4	4	1.000	0.114	0.205
SDT6	0.273	0.680	0.624-0.736	10	1	9	4	4	1.000	0.114	0.205
SDT7	0.273	0.680	0.624-0.736	10	20	9	4	4	1.000	0.114	0.205
SDT8	0.273	0.680	0.624-0.736	10	5	9	4	4	1.000	0.114	0.205
SDT9	0.273	0.680	0.624-0.736	1	25	9	4	4	1.000	0.114	0.205
SDT10	0.273	0.680	0.624-0.736	1	30	9	4	4	1.000	0.114	0.205

Table 8: 10 best SDTs on SD with pruning, ranked by lowest MSR

Model	MSR	AUC	95% CI	α	ENP	# of leaves	# of nodes	# of causal SNVs	PPV	TPR	F1
SDT1	0.431	0.987	0.982-0.992	50	1	49	24	24	1.000	0.686	0.814
SDT2	0.449	0.987	0.982-0.992	10	1	47	23	23	1.000	0.657	0.793
SDT3	0.468	0.987	0.982-0.992	1	1	45	22	22	1.000	0.629	0.772
SDT4	0.469	0.986	0.980-0.992	0.5	1	47	23	23	1.000	0.657	0.793
SDT5	0.475	0.986	0.981-0.992	10	5	59	29	28	0.966	0.800	0.875
SDT6	0.479	0.986	0.980-0.992	50	5	69	34	28	0.824	0.800	0.812
SDT7	0.492	0.989	0.984-0.994	500	1	57	28	28	1.000	0.800	0.889
SDT8	0.507	0.989	0.984-0.994	1000	5	73	36	28	0.778	0.800	0.789
SDT9	0.507	0.989	0.984-0.994	500	5	65	32	28	0.875	0.800	0.836
SDT10	0.508	0.988	0.983-0.993	100	5	63	31	28	0.903	0.800	0.848

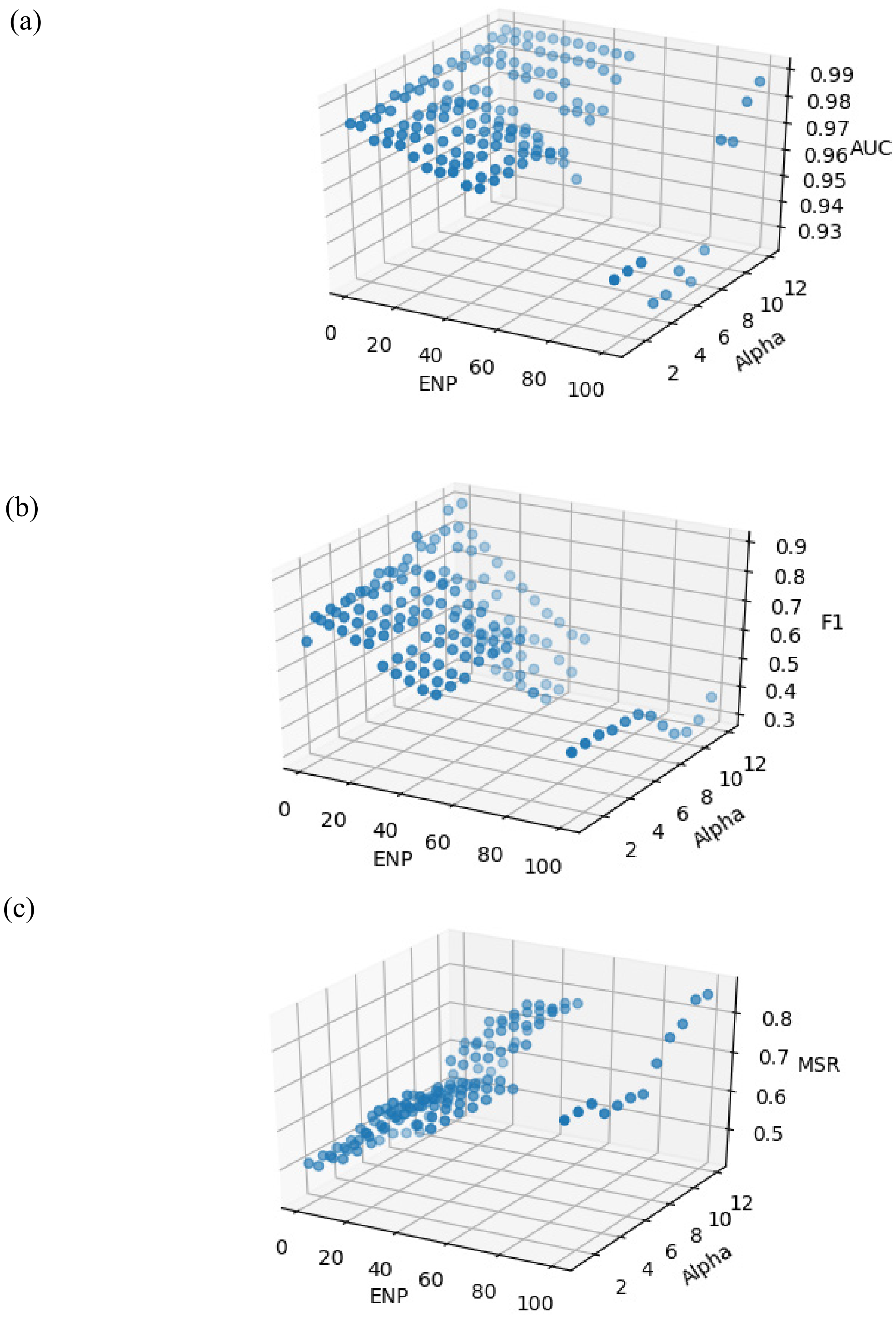


Figure 12: Trend of SDTA on SD with pruning
(a) AUC, (b) F1, (c) MSR

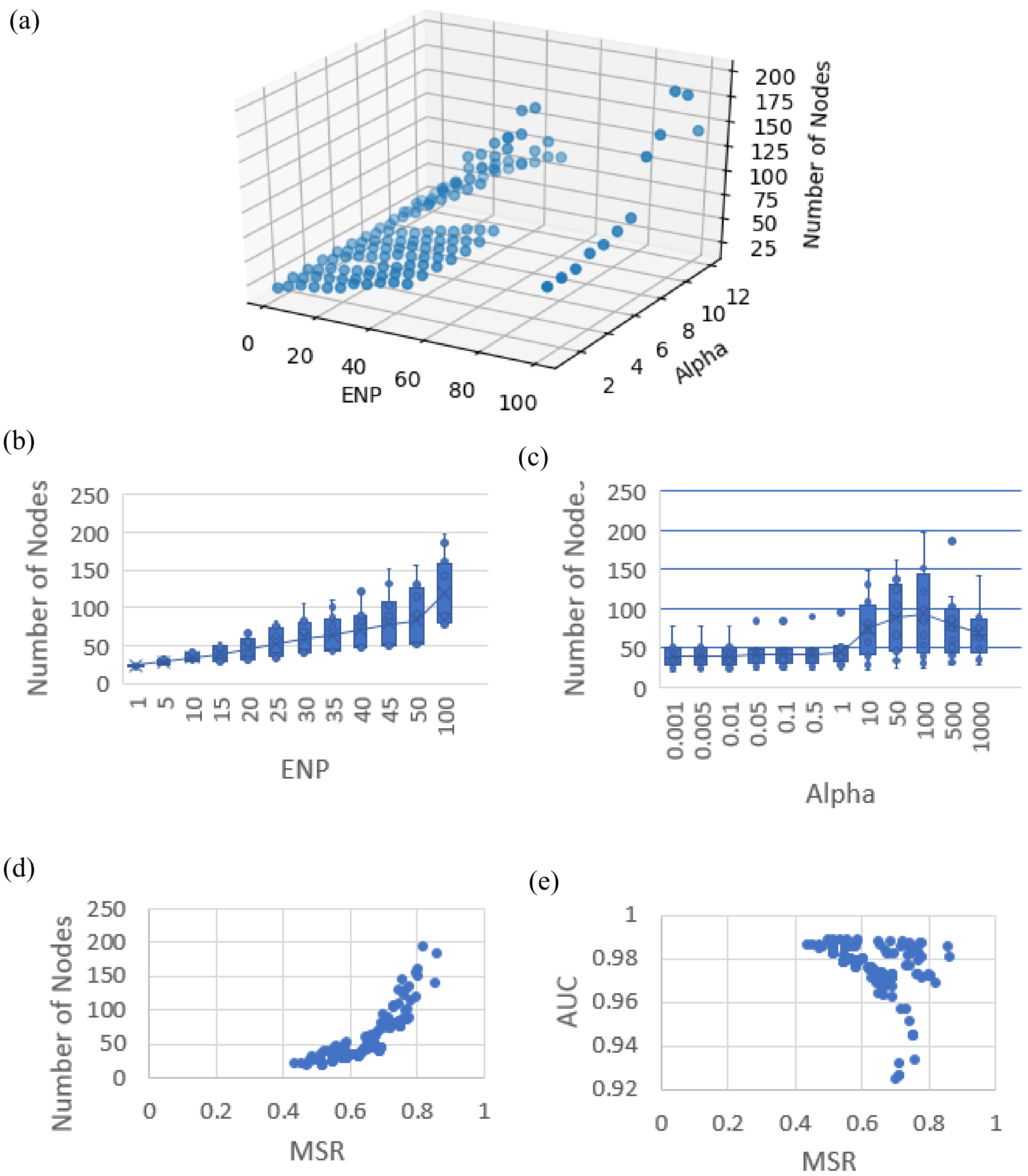


Figure 13: Trend of SDTA on SD with pruning, continued
(a) Size of trees trend by both hyperparameters; (b) Size of trees grow monotonically with ENP; (c) Size of trees grow with Alpha, but not monotonically; (d) & (e) Smaller trees have lower MSR but still high AUCs

5.1.3 Refinements for genetic models

The current implementation of PDTA assumes a genotype model where each of three genotypes at an SNV is treated as an independent value. However, in genetic studies, it is common to test using genotype classes, such as dominant or recessive models [85]. Consider the typical case where the possible alleles at a locus are *A* (wildtype) and *a* (variant). A dominant genetic model assumes that having one or more copies of the *a* allele will result in disease. A recessive genetic model assumes that two copies of the *a* allele must be present to result in disease. In order to allow for such effects, PDTA was implemented to calculate the score using not only independent values for each variable, but also combinations of variable values, i.e., the branch emerging from a node can take one or more values.

When this feature was added to the SDTA, the effect of parameter settings exhibited trends similar to the original SDTA in Figure 7: trees with lower MSR values (Figure 14c) still had better performance by both AUC (Figure 14a) and F1 (Figure 14b). The one notable difference was an increase in size of many of the trees; more nodes were being recruited to the trees for the same hyperparameters, compared to results seen in Figure 7 (Table 9).

I conclude that adding the genotype adaptation to PDTA does not reduce the performance of PDTA.

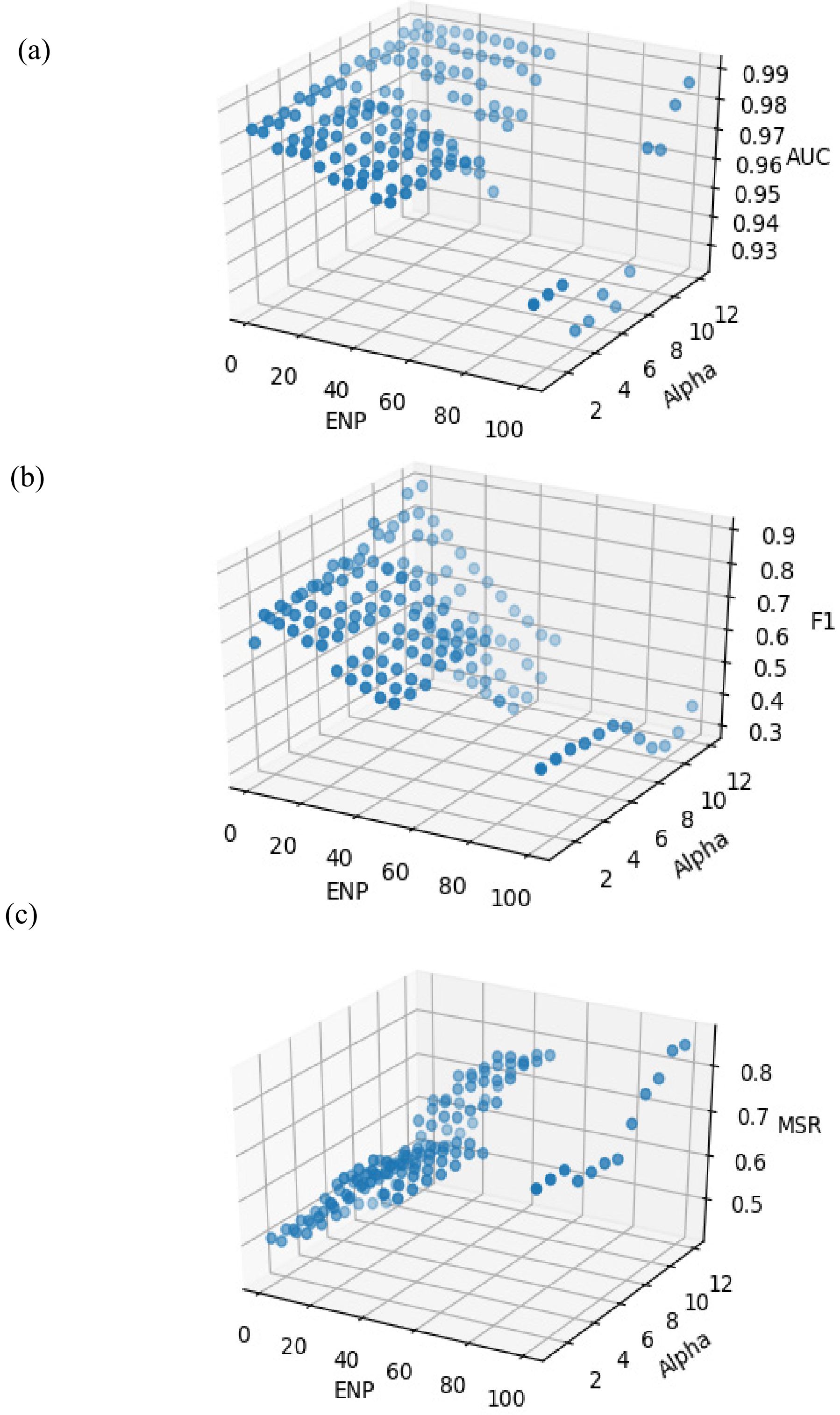


Figure 14: Trend of SDTA on SD with pruning and genotype adaptation
(a) AUC, (b) F1, (c) MSR

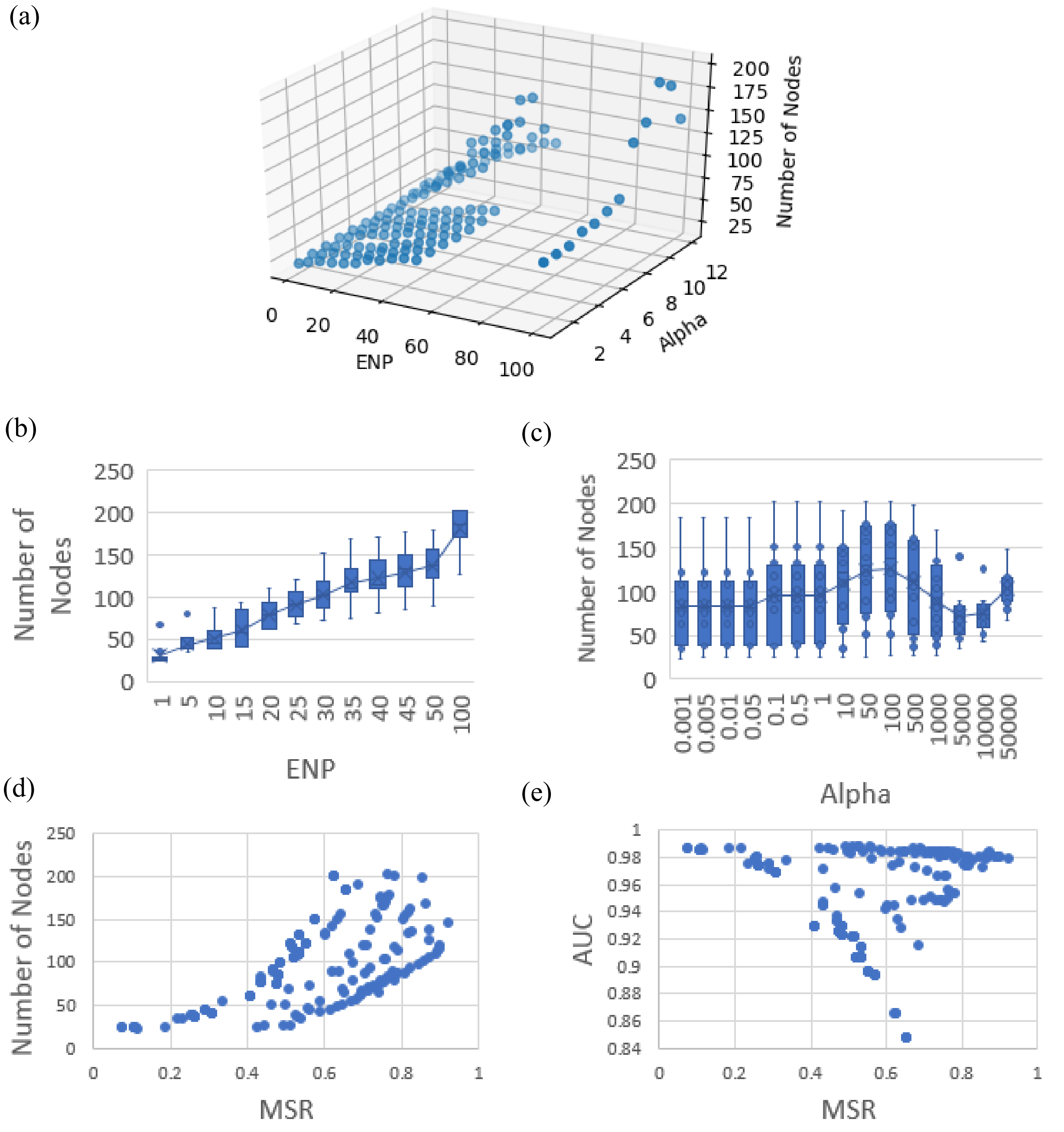


Figure 15: Trend of SDTA on SD with pruning and genotype adaptation, continued
(a) Size of trees trend by both hyperparameters, (b) Size of trees grow monotonically with ENP, (c) Size of trees grow with Alpha, but not monotonically, (d) & (e) Smaller trees have lower MSR but still high AUCs

Table 9: 10 best SDTs on SD with pruning and genotype adaptation, ranked by lowest MSR

Model	MSR	AUC	95% CI (Delong)	α	ENP	# of leaves	# of nodes	# of causal SNVs	PPV	TPR	F1
SDT1	0.071	0.987	0.982-0.992	0.5	1	26	25	23	0.920	0.657	0.767
SDT2	0.071	0.987	0.982-0.992	1	1	26	25	23	0.920	0.657	0.767
SDT3	0.103	0.987	0.982-0.992	0.01	1	27	26	23	0.885	0.657	0.754
SDT4	0.103	0.987	0.982-0.992	0.05	1	27	26	23	0.885	0.657	0.754
SDT5	0.103	0.987	0.982-0.992	0.1	1	27	26	23	0.885	0.657	0.754
SDT6	0.107	0.987	0.982-0.993	0.005	1	26	25	22	0.880	0.629	0.733
SDT7	0.111	0.986	0.980-0.992	0.001	1	25	24	21	0.875	0.600	0.712
SDT8	0.182	0.987	0.982-0.993	10	1	31	26	24	0.923	0.686	0.787
SDT9	0.214	0.987	0.982-0.993	10	5	40	35	27	0.771	0.771	0.771
SDT10	0.231	0.976	0.961-0.991	0.001	5	37	35	22	0.629	0.629	0.629

Table 10: Results of PDTA on SD with pruning and genotype
with best SDT building parameters (SDT1 in Table 9)

α	Tree ENP	Path ENP	# times PS chosen (out of 1000)	SDT AUC	95% CI	PDT AUC	95% CI
0.5	1	1	0	0.987	0.982-0.992	0.987	0.982-0.992
0.5	1	20	0	0.987	0.982-0.992	0.987	0.982-0.992
0.5	1	50	0	0.987	0.982-0.992	0.987	0.982-0.992
0.5	1	100	6	0.987	0.982-0.992	0.987	0.982-0.992

5.1.4 Conclusions

The PDTA performed very well, with AUC = 0.98, on the SD dataset, and uncovers rare variables as predictors in patient subgroups. Based on further tests, the PDTA was updated with two features: the genotype adaptation during the tree building phase, and the pruning step after the tree building phase. The AUC performance of the original PDTA on SD was already close to 0.99 so there was not much room for improvement. These additions have pragmatic value to the interpretability of the model, which is important for my purpose of patient-specific subgroupings; so since they did not reduce the performance on SD when added to PDTA, they were added as the final version of PDTA for all future tests of this algorithm.

Unlike some of the observations in the literature, increasing values of α did not lead to a monotonic increase in edges recruited to the tree[56]. This underscores the importance of empirical testing of the algorithm for score behavior.

5.2 EVALUATION OF PDTA

In section 5.1, I presented the evaluation and iterative refinement of the algorithm on a synthetic dataset. Although SD was designed as a realistic dataset, real genomic datasets are likely to have noise and mixtures of distributions that are more difficult to discover. Therefore, it is important to evaluate the performance of PDTA on real genomic datasets. However, we do not know all of the true causal variables in real datasets, so we cannot assess the algorithm for true positives in the straightforward way we could with SD.

In this section, I present results from the refined version of PDTA on several different datasets, each of which has different properties, and is informative in different ways on PDTA performance. In section 5.2.1, I use SBD, which is curated from a real genomic dataset, while having lower dimensionality than genomic datasets. In section 5.2.2 I use SSD, which represents the full complexity of a real genomic dataset, with synthetic causal variables; nevertheless, the true causal variables are available for evaluating algorithmic performance. In section 5.2.3 I use

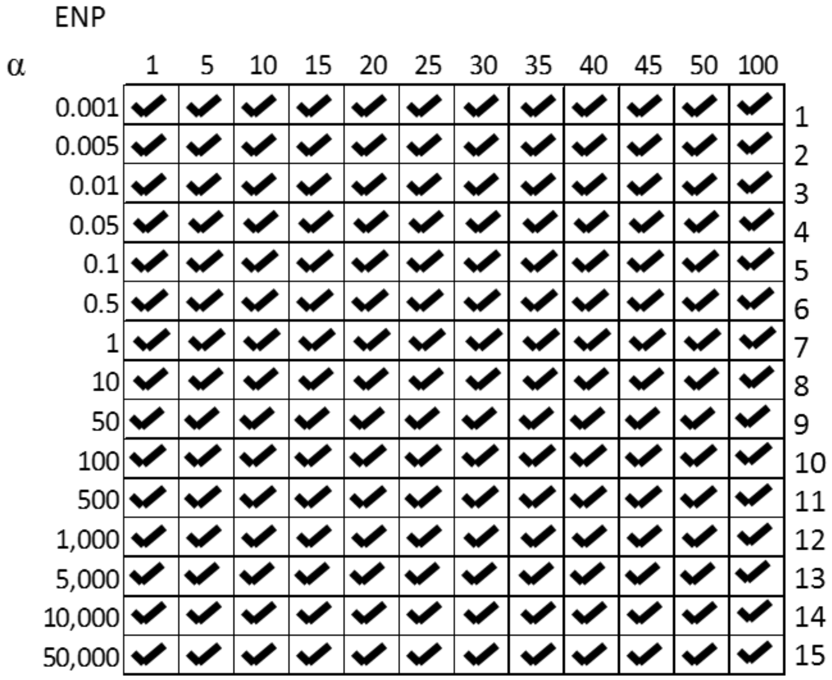
CP-WE, which is a whole exome dataset. In section 5.2.4 I use CP-DB, which has a smaller number of samples than CP-WE, but has the same dimensionality as CP-WE.

5.2.1 Evaluation of refined PDTA on SBD

5.2.1.1 Empirical Optimization of PDT on SBD

To evaluate the performance of PDTA on SBD, first a search for the best SDT is performed over the same settings as explored in Section 5.1.1 (Table 11).

Table 11: Hyperparameter settings used in SDTA on SBD



	ENP												
α	1	5	10	15	20	25	30	35	40	45	50	100	
0.001	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	1
0.005	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2
0.01	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	3
0.05	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	4
0.1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	5
0.5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	6
1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	7
10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	8
50	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
100	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
500	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	11
1,000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	12
5,000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	13
10,000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	14
50,000	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	15

There are some similarities in the trends seen on SBD, when compared to the previous observations on SD (Section 5.1), and some differences. The highest AUC values are achieved with low ENP and high α (Figure 16a), while the best MSR value is achieved with low ENP and low α values (Figure 16b). It is therefore more difficult, than in the case of SD, to assess a clear overlap between the zone of high AUC and low MSR (Figure 16d). MSR values rise with increasing tree size (Figure 16c), and the lowest MSR value on SBD does not achieve the highest AUC (Figure 16d, Table 12), although the confidence intervals for the SDT with the highest AUC and the SDT with best MSR overlap. A perusal of the top ten SDTs ranked by lowest MSR (Table 12) and the top ten SDTs ranked by highest AUC (Table 13) shows that SDTs with lower MSR were much smaller (Table 12: mean number of nodes = 7.2 ± 0.42) than SDTs with higher AUCs (Table 13: mean number of nodes = 60.3 ± 48.94).

Higher values of ENP and α tend to produce larger SDTs (Figure 17 a, b, c), and, consistent with SD observations (Section 5.1), SDTA on SBD also shows a monotonic increase in tree size with increasing value of ENP (Figure 17b), but not α (Figure 17 c). In fact, at ENP = 100, the tree included all 155 variables in the SBD as nodes (Figure 17b), regardless of α value. Holding ENP steady at 100, the α values do not affect AUC or MSR monotonically (Figure 17 d, e).

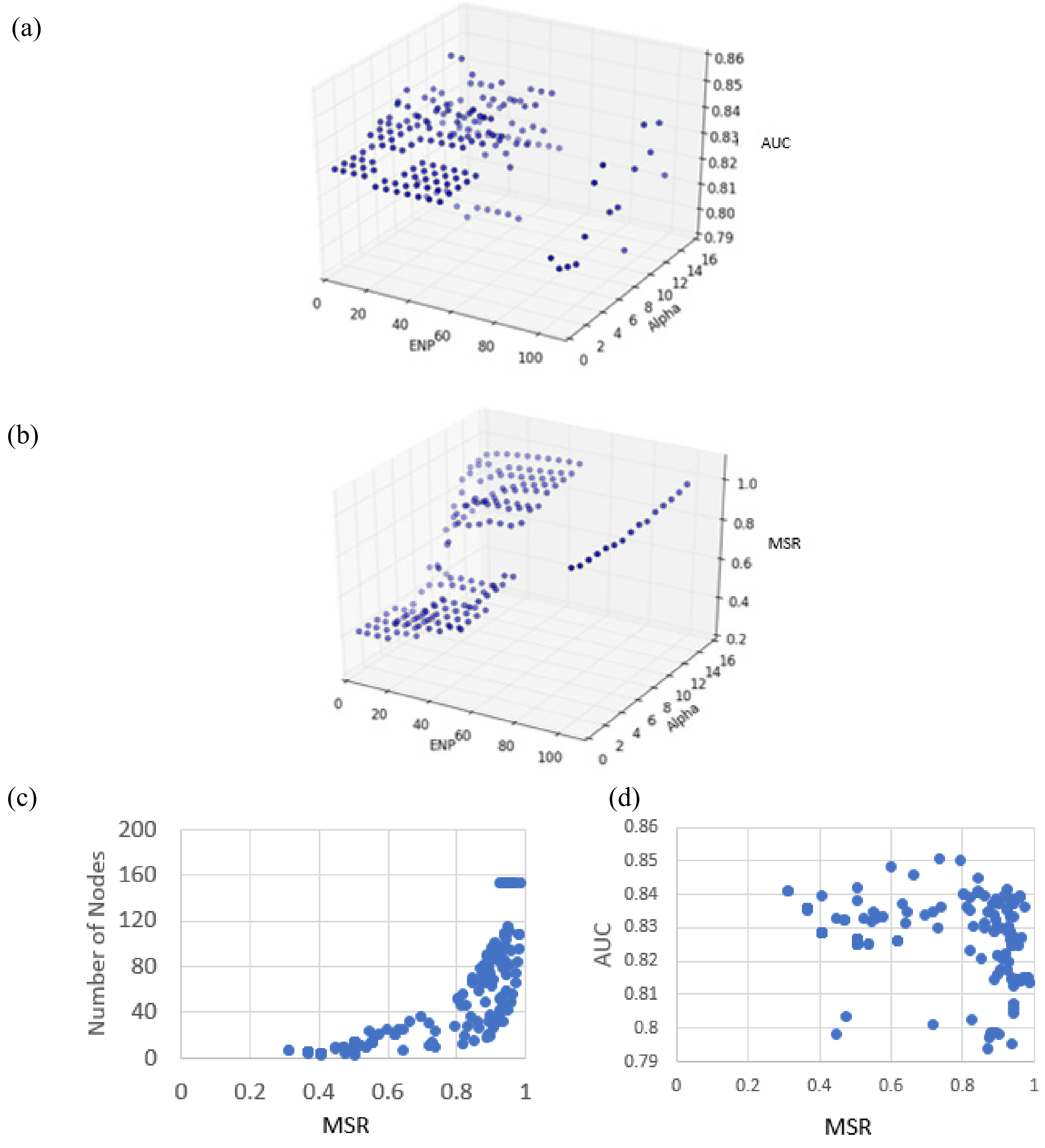


Figure 16: Trend of SDTA performance on SBD
(a) AUC, (b) MSR, (c) MSR increases with size of trees, (d) In a departure from trend seen with SD, lower MSRs are not among the highest AUCs

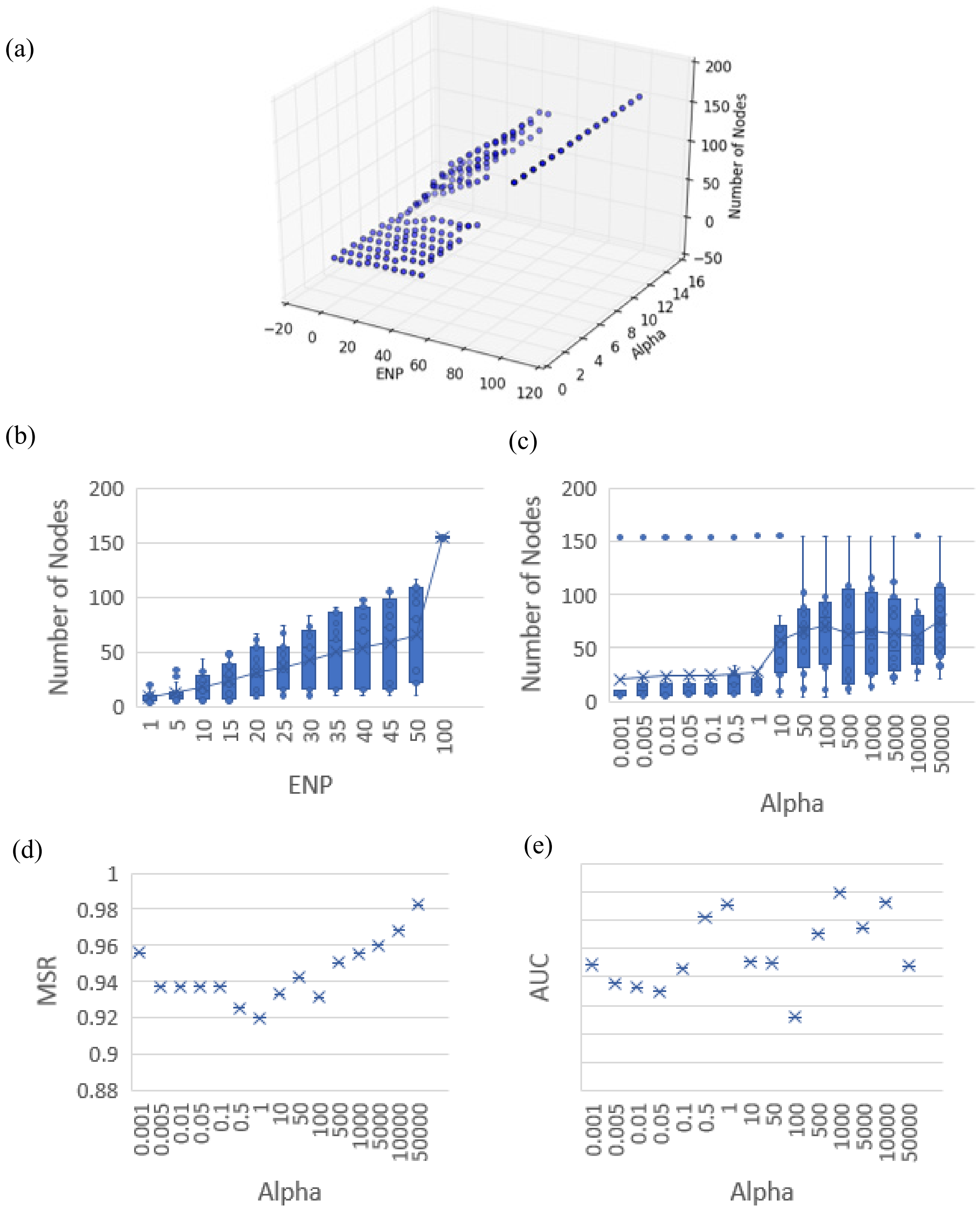


Figure 17: Trend of SDTA performance on SBD, continued
(a) Size of trees trend by both hyperparameters, (b) Size of trees grow monotonically with ENP, (c) Size of trees grow with α, but not monotonically, (d) & (e) Influence of α on MSR and AUC when ENP = 100

Table 12: 10 best SDTs on SBD, when ranked by lowest MSR

Model	MSR	AUC	95% CI	α	ENP	# of leaves	# of nodes
SDT1	0.308	0.842	0.810-0.874	1	10	11	8
SDT2	0.308	0.842	0.810-0.874	1	5	11	8
SDT3	0.364	0.836	0.804-0.869	0.5	10	9	7
SDT4	0.364	0.836	0.804-0.869	0.5	15	9	7
SDT5	0.364	0.836	0.804-0.869	0.5	1	9	7
SDT6	0.364	0.836	0.804-0.869	0.5	5	9	7
SDT7	0.364	0.836	0.804-0.868	0.01	15	9	7
SDT8	0.364	0.836	0.804-0.868	0.01	20	9	7
SDT9	0.364	0.836	0.804-0.868	0.05	10	9	7
SDT10	0.364	0.836	0.804-0.868	0.05	15	9	7

Table 13: 10 best SDTs on SBD, when ranked by highest AUC

Model	AUC	95% CI	MSR	α	ENP	# of leaves	# of nodes
SDT1	0.851	0.8155-0.8868	0.731	1000	10	65	25
SDT2	0.851	0.8153-0.8862	0.788	1000	15	78	30
SDT3	0.849	0.8137-0.8834	0.594	0.5	45	30	27
SDT4	0.847	0.8114-0.8817	0.658	0.5	50	36	33
SDT5	0.845	0.8095-0.8815	0.838	1000	20	97	38
SDT6	0.842	0.8069-0.8779	0.500	50	1	10	4
SDT7	0.842	0.8066-0.8776	0.921	5000	30	149	54
SDT8	0.842	0.8099-0.8735	0.308	1	10	11	8
SDT9	0.842	0.8099-0.8735	0.308	1	5	11	8
SDT10	0.841	0.806-0.8768	0.839	10	35	116	68

5.2.1.2 Model Assessment on SBD

The SDTs generated for different hyperparameter values do not show a clear overlap between the zone of high AUC and low MSR. This calls into question the assumption that the best performing models can be selected by low MSR. Therefore, I examine the results from running a full PDTA with the parameters which led to the SDT with lowest MSR (SDT1 and SDT2 from Table 12), followed by results from a full PDTA with parameters which led to the SDT with the highest AUC (SDT1 from Table 13).

SDT1 and SDT2 from Table 12 are the best tree models as measured by MSR. These two models are identical, achieved by using parameter values α = 1 and tree ENP = 5, 10. The parameter values of SDT2 are applied to test PDTA on SBD (Table 14). None of the paths in the SDT contained rare SNVs. As the value of path ENP increases, the number of PS paths selected increases, without much corresponding improvement in PDT AUC. In PDT3 in Table 14, a single PS path was selected 3 times, but it did not contain any rare variables. In PDT4, two PS paths were selected 4 times, and one of them contained 1 rare variable. For the highest path ENP values, in PDT5 and PDT6, each of the PS paths were comprised of all 155 variables in the SBD. Despite the inclusion of all variables for patient specific signatures, PDTA did not significantly improve performance over the model that was obtained from SDTA.

SDT1 from Table 13 is the best SDT measured by AUC. Its parameter values, α = 1000 and tree ENP = 10, are applied to test PDTA on SBD (Table 15). In keeping with earlier observations of ENP behavior, as the path ENP is increased, more variables are included in PS paths. Several of the SDT paths had rare SNVs as nodes. At the highest path ENP values, PDT5 and PDT6, each of the PS paths contained all 155 variables in the SBD. In PDT4, 7 unique PS paths, comprised of 2-3 nodes each, were selected 12 times. None of the PS paths contained any rare variables.

Table 14: Results of PDTA on SBD with best SDT parameters by MSR (SDT2 of Table 12)

Model	α	Tree ENP	Path ENP	# times PS chosen (out of 440)	SDT AUC	95% CI	PDT AUC	95% CI
PDT1	1	5	1	0	0.842	0.810-0.874	0.842	0.810-0.874
PDT2	1	5	20	0	0.842	0.810-0.874	0.842	0.810-0.874
PDT3	1	5	50	3	0.842	0.810-0.874	0.842	0.810-0.873
PDT4	1	5	70	4	0.842	0.810-0.874	0.843	0.812-0.875
PDT5	1	5	80	7	0.842	0.810-0.874	0.843	0.812-0.875
PDT6	1	5	100	13	0.842	0.810-0.874	0.843	0.812-0.875

Table 15: Results of PDTA on SBD with best SDT parameters by AUC (SDT1, Table 13)

Model	α	Tree ENP	Path ENP	# times PS chosen (out of 440)	SDT AUC	95% CI	PDT AUC	95% CI
PDT1	1000	10	1	0	0.851	0.816-0.887	0.851	0.816-0.887
PDT2	1000	10	20	0	0.851	0.816-0.887	0.851	0.816-0.887
PDT3	1000	10	50	3	0.851	0.816-0.887	0.849	0.813-0.885
PDT4	1000	10	70	12	0.851	0.816-0.887	0.849	0.813-0.885
PDT5	1000	10	80	14	0.851	0.816-0.887	0.847	0.811-0.883
PDT6	1000	10	100	29	0.851	0.816-0.887	0.840	0.804-0.877

5.2.1.3 Summary of PDTA performance on SBD

A tree comprised of all variables could be useful as a detailed representation of individual patient signatures, but to gauge the ability of the algorithm to predict patient subgroups with a tractable set of distinct predictive variables, I prefer models that can attain high AUC with lower ENP, uncovering smaller sets of variables predictive of a patient subset. A model with high AUC that includes all available predictors in the paths would be the least useful. Given two models with AUCs that are similar, the smaller one (if it has a robust number of paths) would be preferable to the larger one.

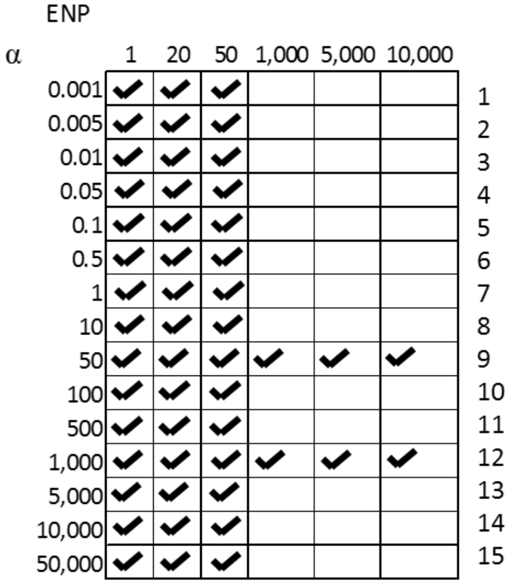
The observations of PDTA performance on SBD cast doubt upon my earlier assumption (in Section 5.1) that MSR is a good criterion for selecting the best model. When the best PDTA was parameterized according to best MSR, there were no rare variants in the SDT, and only a single rare variant found in one of the PS paths. When the best PDTA was parameterized according to best AUC, there were many rare variants among the paths in the SDT. Prioritizing MSR may be biasing the algorithm towards population-based prediction, away from individual-based prediction. This implies that it may be better not to prioritize MSR in selection of the best tree, and with incorporation of pruning after the tree growth phase, concerns regarding overfitting are reduced.

5.2.2 Evaluation of refined algorithm on SSD

5.2.2.1 Empirical Optimization of PDT on SSD

To evaluate the performance of PDTA on SSD, first a search for the best SDT is performed. The causal variables in the SSD dataset are known; therefore, the predictive performance of the models generated can be assessed not only for accuracy in classification, but also for accuracy in the selection of causal variables (measured by F1, PPV and TPR). Due to the large number of variables in the SSD, the computational runtime of PDTA is substantially longer. In sections 5.1.1 and 5.2.1, the results on two datasets indicated that the number of nodes in the SDT rose monotonically with ENP, and that the use of higher ENP may not be necessary to obtain the best performing tree models. Therefore, a smaller selection of ENP values was chosen to evaluate settings for SDTA to apply to SSD (Table 16).

Table 16: Hyperparameter settings used in SDTA on SSD



	ENP						
α	1	20	50	1,000	5,000	10,000	
0.001	✓	✓	✓				1
0.005	✓	✓	✓				2
0.01	✓	✓	✓				3
0.05	✓	✓	✓				4
0.1	✓	✓	✓				5
0.5	✓	✓	✓				6
1	✓	✓	✓				7
10	✓	✓	✓				8
50	✓	✓	✓	✓	✓	✓	9
100	✓	✓	✓				10
500	✓	✓	✓				11
1,000	✓	✓	✓	✓	✓	✓	12
5,000	✓	✓	✓				13
10,000	✓	✓	✓				14
50,000	✓	✓	✓				15

If MSR is a good criterion to use, we would need to select models with low MSR and high AUC. However, PDTA on SSD showed very little overlap between favorable values of AUC and MSR (Figure 18 a & c). Higher α and higher ENP values generated trees with better AUCs and worse MSR, and favoring either one over the other misses the zone of best F1 (Figure 18 b). None of the higher ENP values generated trees with F1 greater than 0.1 (Figure 18b), and higher ENP values lead to larger trees (Figure 19a & b) without corresponding improvement in the number of causal SNVs included in the model (Table 17).

Table 17: 10 best SDTs on SSD, when ranked by highest AUC

Model	SDT AUC	95% CI (Delong)	MSR	α	ENP	# of leaves	# of nodes	# of causal SNVs	PPV	TPR	F1
SDT1	0.602	0.573-0.632	0.331	1000	5000	119	73	2	0.027	0.028	0.028
SDT2	0.602	0.573-0.631	0.293	1000	1000	73	43	2	0.047	0.028	0.035
SDT3	0.598	0.569-0.628	0.338	1000	10000	158	102	2	0.020	0.028	0.023
SDT4	0.595	0.566-0.624	0.241	1000	50	27	15	2	0.133	0.028	0.046
SDT5	0.590	0.561-0.620	0.065	50	1000	43	37	3	0.081	0.042	0.055
SDT6	0.589	0.560-0.619	0.065	50	10000	75	68	4	0.059	0.055	0.057
SDT7	0.587	0.558-0.616	0.208	1000	20	22	12	2	0.167	0.028	0.048
SDT8	0.585	0.556-0.615	0.077	50	5000	63	56	4	0.071	0.056	0.063
SDT9	0.579	0.551-0.608	0.130	50	50	21	16	3	0.188	0.042	0.068
SDT10	0.576	0.548-0.604	0.167	100	1	10	6	2	0.333	0.028	0.051

In section 5.2.1, I made the case for choosing the best model by prioritizing high AUC with smaller SDT size. Looking at the SDTs that achieve best 10 AUC values (Table 17), the smallest is SDT10, which has 6 nodes and parameter settings α = 100, ENP = 1. This model successfully found 2 of the causal variables in the SSD in its SDT paths. In order to find one more causal variable, the minimum tree size in Table 17 increased to 16 (SDT9) and 37 (SDT5). SDT5 has a higher AUC, so it was selected as the more desirable model, and its parameter values applied to run PDTA on SSD. However, no PS paths were selected, so SDT5 remained the best model.

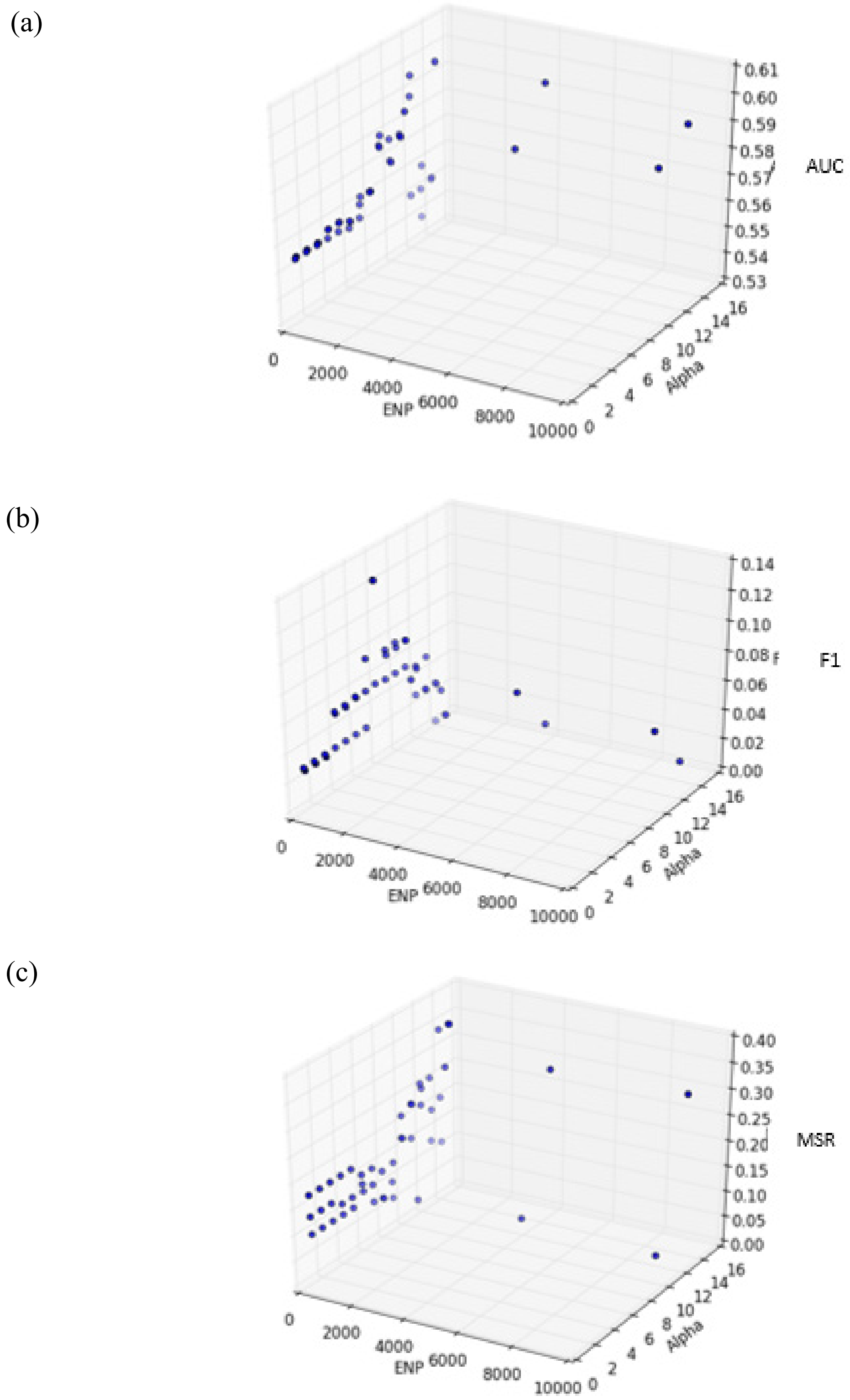


Figure 18: Trend of SDTA on SSD
(a) AUC, (b) F1, (c) MSR

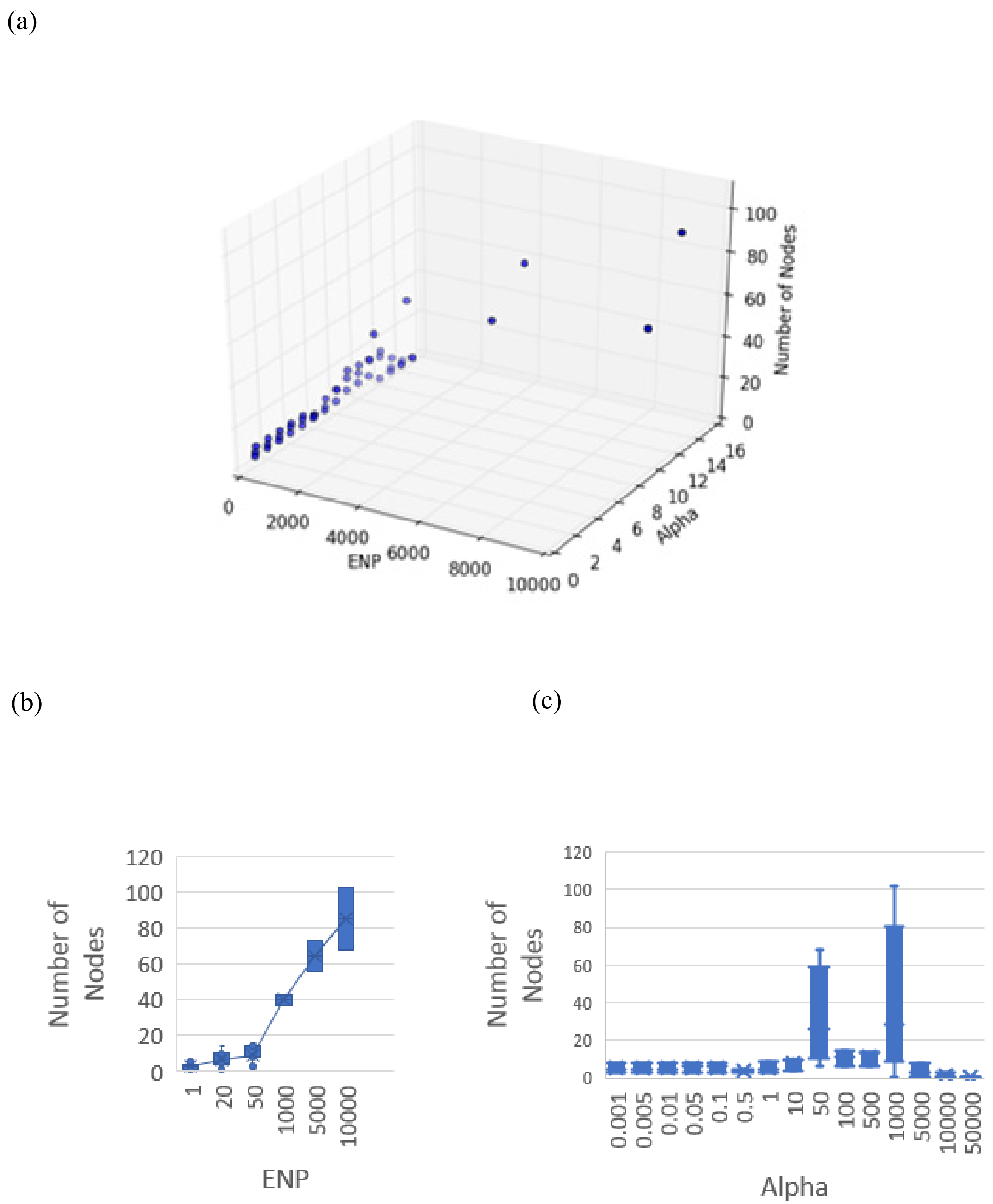


Figure 19: Trend of SDTA on SSD, continued
(a) Size of trees trend by both hyperparameters, (b) Size of trees grow monotonically with ENP,
(c) Size of trees grow with α, but not monotonically

Table 18: SDT paths selected by test samples of SSD, ranked by frequency (SDT5 of Table 17)

Path		# of times selected	Disease Prediction	Path length	Terminal Leaf	# of causal SNVs	# of rare SNVs
SDT_Path1	C6S5380—[0]--…—C6S458—[0, 1]	483	[0.58,0.42]	19	[1163,854]	3	14
SDT_Path2	C6S5380—[1]--…—C17S4431—[0]	264	[0.52,0.48]	7	[541,505]	1	3
SDT_Path3	C6S5380—[0]--…—C19S4493—[0, 2]	83	[0.67,0.33]	18	[225,112]	3	13
SDT_Path4	C6S5380—[0]--…—C20S2249—[0]	79	[0.37,0.63]	10	[123,208]	2	4
SDT_Path5	C6S5380—[0]--…—C1S5270—[2]	59	[0.61,0.39]	6	[154,97]	2	3
SDT_Path6	C6S5380—[2]--…—C12S689—[1, 2]	34	[0.4,0.6]	3	[42,64]	1	0
SDT_Path7	C6S5380—[0]--…--C9S3563—[1]	27	[0.47,0.53]	2	[48,55]	1	1
SDT_Path8	C6S5380—[1]--…—C12S741—[2]	23	[0.29,0.71]	5	[28,69]	1	2
SDT_Path9	C6S5380—[0]--…—C20S1784—[2]	22	[0.63,0.37]	8	[49,29]	2	3
SDT_Path10	C6S5380—[0]--…—C13S1994—[0]	21	[0.59,0.41]	7	[53,36]	2	3
SDT_Path11	C6S5380—[2]--…—C12S689—[0]	21	[0.16,0.84]	3	[15,84]	1	0
SDT_Path12	C6S5380—[0]--…—C13S1994—[2]	20	[0.21,0.79]	7	[10,40]	2	3
SDT_Path13	C6S5380—[0]--…—C6S764—[0]	20	[0.43,0.57]	4	[47,63]	2	2
SDT_Path14	C6S5380—[1]--…—C17S2836—[0]	20	[0.13,0.87]	3	[10,70]	1	1
SDT_Path15	C6S5380—[1]--…—C19S3805—[2]	17	[0.17,0.83]	3	[7,36]	1	1
SDT_Path16	C6S5380—[1]--…—C17S2836—[1, 2]	15	[0.43,0.57]	3	[32,43]	1	1
SDT_Path17	C6S5380—[0]--…—C6S458—[2]	14	[0.82,0.18]	19	[30,6]	3	14
SDT_Path18	C6S5380—[0]--…—C6S764—[1, 2]	14	[0.16,0.84]	4	[7,39]	2	2
SDT_Path19	C6S5380—[2]--…--C22S1809—[2]	13	[0.67,0.33]	2	[25,12]	1	0
SDT_Path20	C6S5380—[0]--…—C11S5926—[1]	12	[0.06,0.94]	8	[0,8]	3	6
SDT_Path21	C6S5380—[0]--…—C9S3563—[2]	12	[0.15,0.85]	2	[10,58]	1	1
SDT_Path22	C6S5380—[0]--…—C1S2069—[2]	10	[0.69,0.31]	9	[21,9]	2	3
SDT_Path23	C6S5380—[1]--…—C8S5023—[2]	9	[0.16,0.84]	6	[3,18]	1	2
SDT_Path24	C6S5380—[0]--…—C20S2249—[1, 2]	9	[0.11,0.89]	10	[3,28]	2	4
SDT_Path25	C6S5380—[0]--…—C19S5128—[2]	9	[0.86,0.14]	15	[27,4]	3	12
SDT_Path26	C6S5380—[0]--…—C14S1880—[1]	9	[0.88,0.12]	18	[19,2]	3	14
SDT_Path27	C6S5380—[0]--…—C7S397—[2]	8	[0.2,0.8]	5	[4,18]	2	3
SDT_Path28	C6S5380—[0]--…—C1S2631—[1, 2]	8	[0.89,0.11]	16	[20,2]	3	13
SDT_Path29	C6S5380—[0]--…—C11S5292—[1, 2]	7	[0.19,0.81]	6	[6,27]	3	4
SDT_Path30	C6S5380—[0]--…—C12S3093—[1]	7	[0.89,0.11]	14	[21,2]	3	12
SDT_Path31	C6S5380—[0]--…—C7S397—[1]	7	[0.48,0.52]	5	[11,12]	2	3
SDT_Path32	C6S5380—[1]--…—C15S1347—[1, 2]	7	[0.04,0.96]	4	[0,13]	1	2
SDT_Path33	C6S5380—[0]--…—C19S56—[2]	5	[0.04,0.96]	7	[0,15]	3	5

Table 18 (continued)

Path		# of times selected	Disease Prediction	Path length	Terminal Leaf	# of causal SNVs	# of rare SNVs
SDT_Path34	C6S5380--[0]--...--C10S2157--[1, 2]	4	[0.15,0.85]	10	[2,14]	3	8
SDT_Path35	C6S5380--[0]--...--C9S1219--[1, 2]	4	[0.21,0.79]	12	[3,13]	3	10
SDT_Path36	C6S5380--[0]--...--C1S716--[1, 2]	3	[0.13,0.87]	10	[3,24]	3	7
SDT_Path37	C6S5380--[0]--...--C1S716--[0]	3	[0.58,0.42]	10	[10,7]	3	7
SDT_Path38	C6S5380--[0]--...--C9S3331--[1, 2]	3	[0.2,0.8]	11	[3,14]	3	9
SDT_Path39	C6S5380--[0]--...--C19S4493--[1]	2	[0.97,0.03]	18	[18,0]	3	13
SDT_Path40	C6S5380--[0]--...--C5S3215--[1]	2	[0.24,0.76]	13	[4,14]	3	11
SDT_Path41	C6S5380--[0]--...--C11S5926--[2]	2	[0.94,0.06]	8	[8,0]	3	6
SDT_Path42	C6S5380--[1]--...--C17S4431--[1, 2]	2	[0.23,0.77]	7	[6,22]	1	3

5.2.2.2 Model assessment on SSD

All of the paths in SDT5 contained known causal variables in them, and all but one also had rare variables (Table 18).

The largest patient subgroup found (SDT_Path1 in Table 18) had ambiguous disease prediction. The lack of good predictors for this large group is reflected in the modest AUC of SDT5. However, there were smaller groups – for example: SDT_Path11, SDT_Path14, SDT_Path17, SDT_Path11, SDT_Path25 – which had much stronger predictive value for patient-based subgroups, with a mixture of causal and rare variants.

5.2.2.3 Summary of PDTA performance on SSD

The best model obtained by applying PDTA on SSD did not have very good scores overall: AUC = 0.59 (95% CI: 0.561-0.620), F1 = 0.6, PPV = 0.1. A closer look at the paths showed poor performance for subgroups characterizing large sections of the population, but several pockets of good prediction for smaller groups. Even though the SDT appears to be a population-wide tree, as compared to the PS paths created in a final PDT, it is actually prioritizing individualized groups fairly well.

The observations on SBD made me doubt the value of MSR in selecting the best model; the observations on SSD further validated my suspicions. For all following experiments on genomic datasets, MSR was no longer used in model selection.

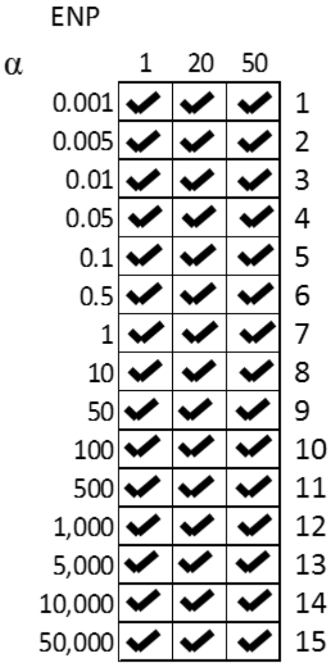
5.2.3 Evaluation of refined algorithm on WE-CP

5.2.3.1 Empirical Optimization of PDT on WE-CP and WE-CPR

In all datasets examined so far, the number of nodes in the SDT rose monotonically with ENP, which leads to longer runtimes for the algorithm. Since WE-CP is a very large dataset, the search for the best SDT was applied to WE-CP with a small set of low ENP values (Table 19). For these values, none of the SDTs formed trees with more than 3 or 4 nodes. Therefore, the experiments were repeated on a subset of WE-CP with reduced dimensionality (WE-CP-R) (see Section 4.1.4), along with additional higher values of ENP (Table 20).

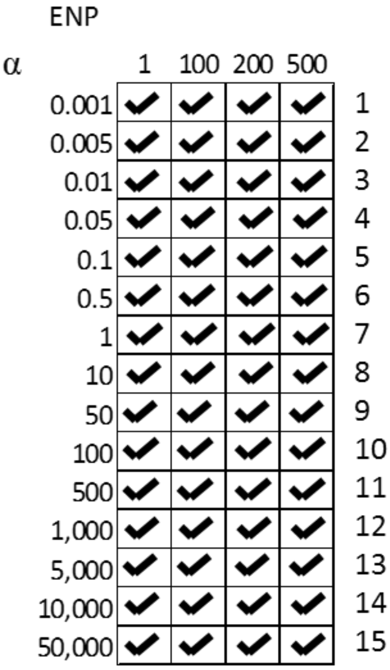
Despite the reduced dimensionality of WE-CP-R compared to WE-CP, the parameter value ENP = 1 failed to produce full trees with more than 3 or 4 nodes. However, at higher ENP values, better SDTs were formed, with most of the higher AUC models including all available SNVs as nodes (Table 21). The best AUC achieved was 0.57 (95% CI: 0.528-0.62). Choosing the best model by prioritizing high AUC with smaller SDT size, the parameters for SDT2 in Table 21 were applied to test PDTA on WE-CP-R (Table 22).

Table 19: Hyperparameter settings used in SDTA on WE-CP



	ENP			
α	1	20	50	
0.001	✓	✓	✓	1
0.005	✓	✓	✓	2
0.01	✓	✓	✓	3
0.05	✓	✓	✓	4
0.1	✓	✓	✓	5
0.5	✓	✓	✓	6
1	✓	✓	✓	7
10	✓	✓	✓	8
50	✓	✓	✓	9
100	✓	✓	✓	10
500	✓	✓	✓	11
1,000	✓	✓	✓	12
5,000	✓	✓	✓	13
10,000	✓	✓	✓	14
50,000	✓	✓	✓	15

Table 20: Hyperparameter settings used in SDTA on WE-CP-R



	ENP				
α	1	100	200	500	
0.001	✓	✓	✓	✓	1
0.005	✓	✓	✓	✓	2
0.01	✓	✓	✓	✓	3
0.05	✓	✓	✓	✓	4
0.1	✓	✓	✓	✓	5
0.5	✓	✓	✓	✓	6
1	✓	✓	✓	✓	7
10	✓	✓	✓	✓	8
50	✓	✓	✓	✓	9
100	✓	✓	✓	✓	10
500	✓	✓	✓	✓	11
1,000	✓	✓	✓	✓	12
5,000	✓	✓	✓	✓	13
10,000	✓	✓	✓	✓	14
50,000	✓	✓	✓	✓	15

Table 21: 10 best SDTs on WE-CP-R, ranked by highest AUC

Model	AUC	95% CI	MSR	α	ENP	# of leaves	# of nodes
SDT1	0.574	0.528-0.620	0.985	0.1	500	520	516
SDT2	0.570	0.513-0.627	0.872	50	100	319	153
SDT3	0.565	0.510-0.621	0.977	0.5	500	520	516
SDT4	0.564	0.518-0.610	0.987	0.001	500	519	516
SDT5	0.564	0.510-0.618	0.972	10	500	1104	516
SDT6	0.557	0.506-0.607	0.990	50000	500	1150	516
SDT7	0.555	0.501-0.610	0.872	10	200	241	123
SDT8	0.553	0.507-0.599	0.989	0.005	500	520	516
SDT9	0.553	0.500-0.606	0.780	1	100	57	54
SDT10	0.553	0.521-0.585	0.842	0.5	100	36	35

Table 22: Results of PDTA on WE-CP-R with best SDT parameters
(SDT2 in Table 21)

Model	α	Tree ENP	Path ENP	# times PS chosen (out of 427)	SDT AUC	95% CI (Delong)	PDT AUC	95% CI (Delong)
PDT1	50	100	1	0	0.570	0.513-0.627	0.570	0.513-0.627
PDT2	50	100	100	0	0.570	0.513-0.627	0.570	0.513-0.627
PDT3	50	100	200	0	0.570	0.513-0.627	0.570	0.513-0.627
PDT4	50	100	300	3	0.570	0.513-0.627	0.568	0.512-0.625
PDT5	50	100	400	19	0.570	0.513-0.627	0.574	0.517-0.630
PDT6	50	100	450	42	0.570	0.513-0.627	0.585	0.529-0.642
PDT7	50	100	500	42	0.570	0.513-0.627	0.585	0.529-0.642

5.2.3.2 Model assessment on WE-CP-R

The best SDT model contained rare variables in many of its paths (Table 23). At the PS path creation phase, the number of PS paths increased with the value of path ENP, with AUC rising slightly at the higher path ENPs (Table 22). The model with the highest AUC was PDT6, where, 42 out of 427 test instances selected one of 39 individual PS paths. However, all of the PS paths included all 516 SNVs as nodes, failing to generalize to more than 1 or 2 test instances each, and had near ambiguous (~50%) disease prediction. Therefore, I conclude that the PS paths are not useful for finding patient subgroups, and instead proceed with the subgroups found when all test instances chose a path in the SDT. This was the case in PDT1, PDT2 and PDT3: 110 unique paths in their SDT were selected by the 427 test instances. The largest subgroup had 127 patients; the smallest (39 subgroups) had 1. The top 8 most frequently selected paths are shown in Table 23. Since all of these subgroups are from the same tree, they share the same root node, rs7332962. The 5 largest patient subgroups are detailed in Tables 24-28. The interrelationship of these 5 paths is shown in Figure 20.

Table 23: Top 8 SDT paths selected by test samples of WE-CP-R (out of 110) ranked by frequency (PDT1 of Table 2)

Path		**# times selected (out of 427)**	**Disease Prediction**	**True Target**	**Path length**	**Terminal Leaf**	**# of rare SNVs**
SDT_Path1	rs7332962--[0, 2]--...--rs79636164--[0]	127	[0.3,0.7]	[38,89]	19	[132,313]	14
SDT_Path2	rs7332962--[0, 2]--...--rs17107315:[36,56]--[0, 2]	20	[0.36,0.64]	[7,13]	16	[32,56]	10
SDT_Path3	rs7332962--[0, 2]--...--rs12919410:[32,39]--[1, 2]	15	[0.35,0.65]	[5,10]	12	[19,36]	5
SDT_Path4	rs7332962--[0, 2]--...--rs11628525:[13,100]--[0]	10	[0.0,1.0]	[1,9]	10	[0,34]	5
SDT_Path5	rs7332962--[0, 2]--...--rs4764427:[7,45]--[1]	9	[0.0,1.0]	[1,8]	13	[0,21]	5
SDT_Path6	rs7332962--[0, 2]--...--rs17702641:[6,56]--[0]	7	[0.0,1.0]	[2,5]	16	[0,25]	5
SDT_Path7	rs7332962--[0, 2]--...--rs2854128:[8,31]--[2]	7	[0.0,1.0]	[3,4]	22	[0,18]	9
SDT_Path8	rs7332962--[0, 2]--...--rs291096:[10,43]--[0]	7	[0.01,0.99]	[1,6]	20	[0,12]	9

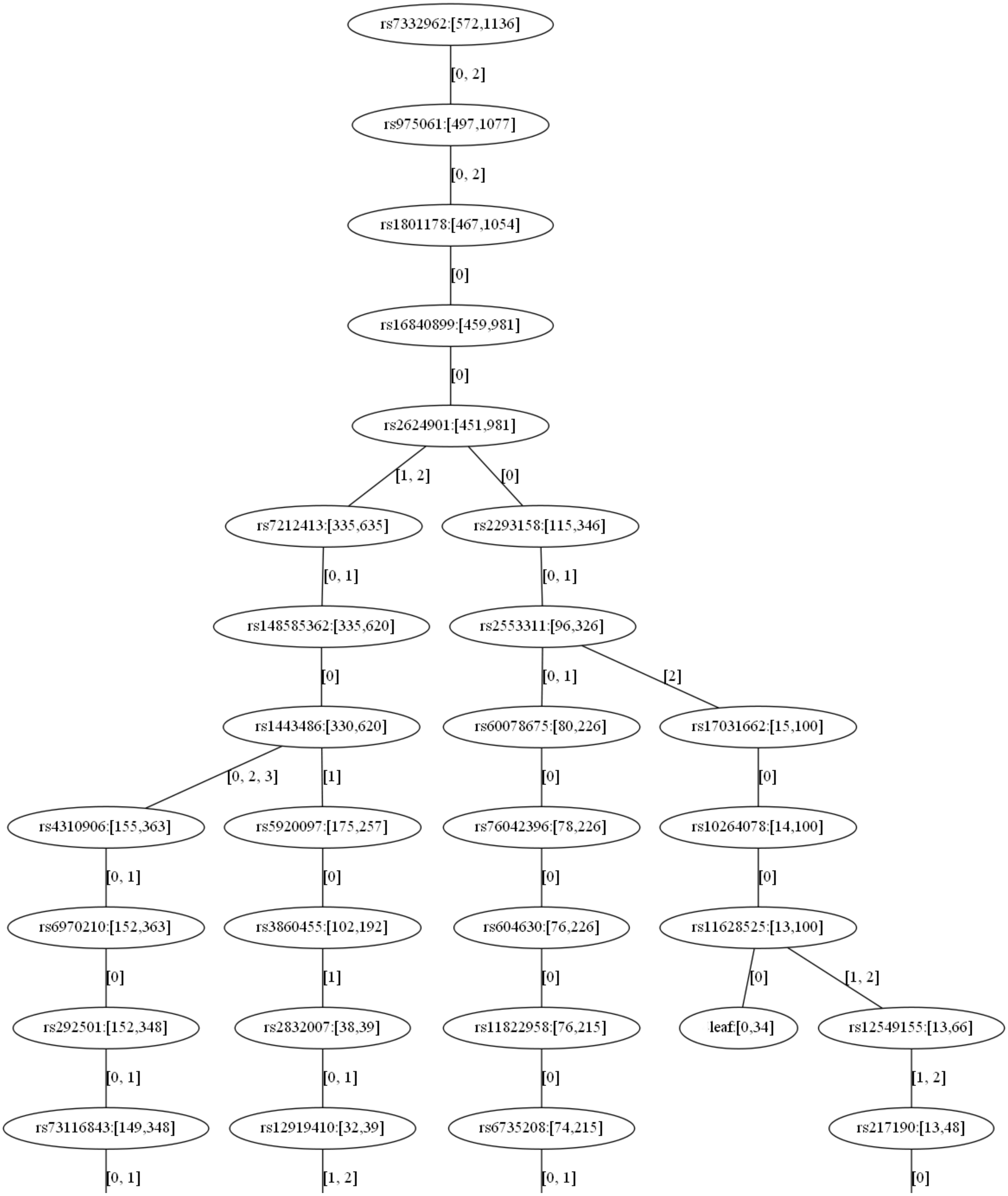


Figure 20: Interrelation of the 5 patient subgroups highlighted in Tables 24-28

The most commonly chosen path, SDT_Path1 in Table 23, correctly predicted presence of disease in 89 out of 127 patients. The path is comprised of 19 SNVs, out of which 14 are rare SNVs. Few of the SNVs have known functional consequence and clinical significance from dbSNP (Table 24). However, a review of the literature shows many of these SNVs or their associated genes have roles in pancreas related function. GABRP, MUC16, MAP1S, MEGF6, ARHGAP24, and OR1L3 have all been shown to be associated with pancreatic cancer, a common late stage consequence of CP[86-95]. Out of these, in this patient subgroup, only the SNV GABRP shows mutation in its emergent branch (Node #1, Table 24). The others having been selected as predictive nodes with major homozygous (no mutation) emergent edges, for this patient subgroup, indicates that there are other patient subgroups where they are predictive in the mutated state.

The CFTR gene has many variants that are pathogenic for CP[29] and the variant in this patient subgroup, rs1801178, is predicted to be deleterious by SIFT/Polyphen[96] although in this particular patient subgroup, it is not mutated. RHBDL3 plays a role in pancreatic development[97] and PIK3R6 has been implicated in Type II diabetes[98]. Both genes have mutated variants in this patient subgroup.

Table 24: SNVs in SDT_Path1 on WE-CP-R
with known functional consequence and clinical significance from dbSNP
(*) Indicates a gene with known or suspected role in pancreas related function or disorder

Node	SNP	Edge	Functional consequence	Clinical significance	Associated Gene	Node Count	Global MAF (TOPMED)
root	rs7332962	[0, 2]	intron variant	NA	CENPJ RNF17	[572,1136]	T=0.0718/9022
1	rs975061	[0, 2]	synonymous codon	NA	GABRP *	[497,1077]	T=0.0478/6001
2	rs1801178	[0]		Likely benign (for cystic fibrosis)	CFTR * CFTR-AS1	[467,1054]	G=0.00007/9
3	rs16840899	[0]	Missense	NA	SORCS2	[459,981]	T=0.0184/2305
4	rs2624901	[1, 2]	intron variant	NA	KIAA1257	[451,981]	T=0.4512/56652
5	rs7212413	[0, 1]		NA		[335,635]	A=0.1594/20014
6	rs148585362	[0]	Missense	NA	MUC16 *	[335,620]	C=0.0021/262
7	rs1443486	[0, 2, 3]	intron variant, upstream variant 2KB	NA	LOC105371731 RHBDL3 *	[330,620]	G=0.4507/56590
8	rs4310906	[0, 1]	intron variant	NA	PIK3R6 *	[155,363]	A=0.0825/10360
9	rs6970210	[0]	intron variant, missense	NA	INMT-MINDY4 INMT	[152,363]	G=0.0419/5259
10	rs292501	[0, 1]	intron variant, missense, syn codon, utr variant 3 prime	NA	C7orf49 TMEM140	[152,348]	T=0.0891/446
11	rs73116843	[0, 1]	intron variant	NA	ANKRD31	[149,348]	T=0.0514/6457
12	rs16970731	[0]	intron variant	NA	GPATCH8	[146,348]	A=0.0410/5152
13	rs77093026	[0]	missense, nc transcript variant	NA	LOC105372299 MAP1S	[145,340]	A=0.0284/3568
14	rs61746548	[0]	Missense	NA	MEGF6 *	[140,339]	T=0.0472/5933
15	rs10493753	[0, 1]	missense, nc transcript variant	NA	SPATA1	[140,326]	C=0.0998/12528
16	rs1482097	[0]	intron variant	NA	ARHGAP24 *	[137,326]	A=0.1100/13810
17	rs1799980	[0]	Missense	Benign	SCNN1B	[137,313]	T=0.0398/4993
18	rs79636164	[0]	Missense	NA	OR1L3 *	[134,313]	T=0.0631/7928

The second most commonly chosen path, SDT_Path2 in Table 23, correctly predicted disease in 13 out of 20 patients (Table 25). The path is comprised of 16 SNVs, out of which 10 are rare SNVs. The terminal node in this path is (wildtype and biallelic mutant) rs17107315, also known as the N34S variant in SPINK, a well-known risk factor for CP[99].

For the first 4 nodes, SDT_Path2 is the same as SDT_Path1; they diverge at rs2624901, which branches with mutated alleles in SDT_Path1, and wildtype allele in SDT_Path2. The remaining SNPs in SDT_Path2 are distinct from those in SDT_Path1. Again, few of the SNVs have known functional consequence and clinical significance from dbSNP (Table 25), but a review of the literature shows some of them or their associated genes have roles in pancreas related function. In addition to GABRP and CFTR, which have variants in the portion of the path common to SDT_Path1, SDT_Path2 also includes SNVs associated with STAT5A, PCNT, and HYOU1 which have been shown to have roles in pancreatic endocrine function[100-103]. Out of these, the SNV associated with STAT5A is heterozygous for this patient subgroup. Another SNV in this patient group is associated with gene ECM1, which may have a role in pancreatic cancer[104].

Table 25: SNVs in SDT_Path2 on WE-CP-R
with known functional consequence and clinical significance from dbSNP
(*) Indicates a gene with known or suspected role in pancreas related function or disorder

	SNP	Edge	Functional consequence	Clinical significance	Associated Gene	Node Count	Global MAF (TOPMED)
root	rs7332962	[0, 2]	intron variant	NA	CENPJ, RNF17	[572,1136]	T=0.0718/9022
1	rs975061	[0, 2]	synonymous codon	NA	GABRP *	[497,1077]	T=0.0478/6001
2	rs1801178	[0]		Likely benign in cystic fibrosis	CFTR, CFTR-AS1 *	[467,1054]	G=0.00007/9
3	rs16840899	[0]	Missense	NA	SORCS2	[459,981]	T=0.0184/2305
4	rs2624901	[0]	intron variant	NA	KIAA1257	[451,981]	T=0.4512/56652
5	rs2293158	[0, 1]	intron variant	NA	STAT5A *	[115,346]	C=0.2950/37044
6	rs2553311	[0, 1]		NA		[96,326]	A=0.4150/52113
7	rs60078675	[0]	Missense	Likely benign	PCNT *	[80,226]	T=0.0224/2814
8	rs76042396	[0]	Missense	NA	DNASE2B	[78,226]	A=0.0360/4517
9	rs604630	[0]	Missense	NA	CTSW	[76,226]	A=0.0674/8464
10	rs11822958	[0]	nc transcript variant, synonymous codon	NA	HYOU1 *	[76,215]	T=0.0476/5982
11	rs6735208	[0, 1]	Missense	With other allele (in nemaline myopathy)	NEB	[74,215]	T=0.3268/41030
12	rs73721657	[0]	Missense	NA	KCP	[74,204]	T=0.0557/6998
13	rs435004	[0]	intron variant	NA	ST6GALNAC3	[72,204]	T=0.0932/11699
14	rs3737240	[0]	Missense	NA	ECM1 *	[72,195]	T=0.2823/35446
15	rs17107315	[0, 2]	Missense	With other allele (in chronic pancreatitis)	SPINK1 *	[36,56]	C=0.0070/885

The third most commonly chosen path, SDT_Path3 in Table 23, correctly predicted disease in 10 out of 15 patients (Table 26). The path is comprised of 12 SNVs, out of which 5 are rare SNVs. This path shares its first 4 nodes with the previous two patient subgroups, and then the next 3 nodes with SDT_Path1. This common portion of the path contains variants associated with MUC16, CFTR and GABRP. It diverges from SDT_Path1 at rs1443486, which branches with the heterozygous variant to continue on SDT_Path3. This SNV is associated with gene RHBDL3, which plays a role in pancreatic development [97]. The full list of nodes in SDT_Path3 are shown in Table 26, along with information from dbSNP.

Table 26: SNVs in SDT_Path3 on WE-CP-R
with known functional consequence and clinical significance from dbSNP
(*) Indicates a gene with known or suspected role in pancreas related function or disorder

	SNP	Edge	Functional consequence	Clinical significance	Associated Gene	Node Count	Global MAF (TOPMED)
root	rs7332962	[0, 2]	intron variant	NA	CENPJ RNF17	[572,1136]	T=0.0718/9022
1	rs975061	[0, 2]	synonymous codon	NA	GABRP *	[497,1077]	T=0.0478/6001
2	rs1801178	[0]		Likely benign (cystic fibrosis)	CFTR * CFTR-AS1	[467,1054]	G=0.00007/9
3	rs16840899	[0]	missense	NA	SORCS2	[459,981]	T=0.0184/2305
4	rs2624901	[1, 2]	intron variant	NA	KIAA1257	[451,981]	T=0.4512/56652
5	rs7212413	[0, 1]		NA		[335,635]	A=0.1594/20014
6	rs148585362	[0]	missense	NA	MUC16 *	[335,620]	C=0.0021/262
7	rs1443486	[1]	intron variant, upstream variant 2KB	NA	LOC105371731 RHBDL3 *	[330,620]	G=0.4507/56590
8	rs5920097	[0]		NA		[175,257]	T=0.2594/32574
9	rs3860455	[1]	intron variant	NA	NPAS2	[102,192]	T=0.3128/39273
10	rs2832007	[0, 1]	intron variant	NA	N6AMT1	[38,39]	A=0.3952/49626
11	rs12919410	[1, 2]		NA		[32,39]	T=0.4272/53641

The fourth most commonly chosen path, SDT_Path4 in Table 23, correctly predicted disease in 9 out of 10 patients (Table 27). The path is comprised of 10 SNVs, out of which 5 are rare SNVs. This path shares its first 4 nodes with the previous three patient subgroups, branches off at rs2624901 with SDT_Path2, then branches off from SDT_Path2 at rs2553311 with minor homozygous allele. The full list of nodes and known functional consequence and clinical significance from dbSNP (Table 27) do not show any relationship to CP, but a review of the literature shows HYAL4 and SIPA1L1 may be involved in cancer progression [105,106].

Table 27: SNVs in SDT_Path4 on WE-CP-R
with known functional consequence and clinical significance from dbSNP
(*) Indicates a gene with known or suspected role in pancreas related function or disorder

	SNP	Edge	Functional consequence	Clinical significance	Associated Gene	Node Count	Global MAF (TOPMED)
root	rs7332962	[0, 2]	intron variant	NA	CENPJ RNF17	[572,1136]	T=0.0718/9022
1	rs975061	[0, 2]	synonymous codon	NA	GABRP *	[497,1077]	T=0.0478/6001
2	rs1801178	[0]	intron variant missense	Likely benign (cystic fibrosis)	CFTR * CFTR-AS1	[467,1054]	G=0.00007/9
3	rs16840899	[0]	missense	NA	SORCS2	[459,981]	T=0.0184/2305
4	rs2624901	[0]	intron variant	NA	KIAA1257	[451,981]	T=0.4512/56652
5	rs2293158	[0, 1]	intron variant	NA	STAT5A *	[115,346]	C=0.2950/37044
6	rs2553311	[2]		NA		[96,326]	A=0.4150/52113
7	rs17031662	[0]	intron variant	NA	LOC105369938 MYBPC1	[15,100]	T=0.1107/13900
8	rs10264078	[0]	intron variant	NA	HYAL4 *	[14,100]	G=0.0584/7327
9	rs11628525	[0]	intron variant	NA	SIPA1L1 *	[13,100]	G=0.3293/41355

The fifth most commonly chosen path, SDT_Path5 in Table 23, correctly predicted disease in 8 out of 9 patients. The path is comprised of 13 SNVs, out of which 5 are rare SNVs. This patient subgroup shares the entire path of SDT_Path4 until its final node rs11628525, where SDT_Path5 branches off with the mutant variant, then continues for three more nodes (Table 28). The SNV rs1162525 is associated with SIPAIL1, which may be implicated in pancreatic cancer[106].

Table 28: SNVs in SDT_Path5 on WE-CP-R
with known functional consequence and clinical significance from dbSNP
(*) Indicates a gene with known or suspected role in pancreas related function or disorder

	SNP	Edge	Functional consequence	Clinical significance	Associated Gene	Node Count	Global MAF (TOPMED)
root	rs7332962	[0,2]	intron variant	NA	CENPJ RNF17	[572,1136]	T=0.0718/9022
1	rs975061	[0,2]	synonymous codon	NA	GABRP *	[497,1077]	T=0.0478/6001
2	rs1801178	[0]	intron variant missense	Likely benign (cystic fibrosis)	CFTR * CFTR-AS1	[467,1054]	G=0.00007/9
3	rs16840899	[0]	missense	NA	SORCS2	[459,981]	T=0.0184/2305
4	rs2624901	[0]	intron variant	NA	KIAA1257	[451,981]	T=0.4512/56652
5	rs2293158	[0,1]	intron variant	NA	STAT5A *	[115,346]	C=0.2950/37044
6	rs2553311	[2]		NA		[96,326]	A=0.4150/52113
7	rs17031662	[0]	intron variant	NA	LOC105369938 MYBPC1	[15,100]	T=0.1107/13900
8	rs10264078	[0]	intron variant	NA	HYAL4 *	[14,100]	G=0.0584/7327
9	rs11628525	[1,2]	intron variant	NA	SIPA1L1 *	[13,100]	G=0.3293/41355
10	rs12549155	[1,2]	intron variant	NA	LOC105375706	[13,66]	A=0.4049/50845
11	rs217190	[0]	intron variant	NA	ABLIM1	[13,48]	G=0.2279/28612
12	rs4764427	[1]	intron variant	NA	LOC102724227	[7,45]	A=0.4589/57623

5.2.3.3 Summary of PDTA performance on WE-CP-R

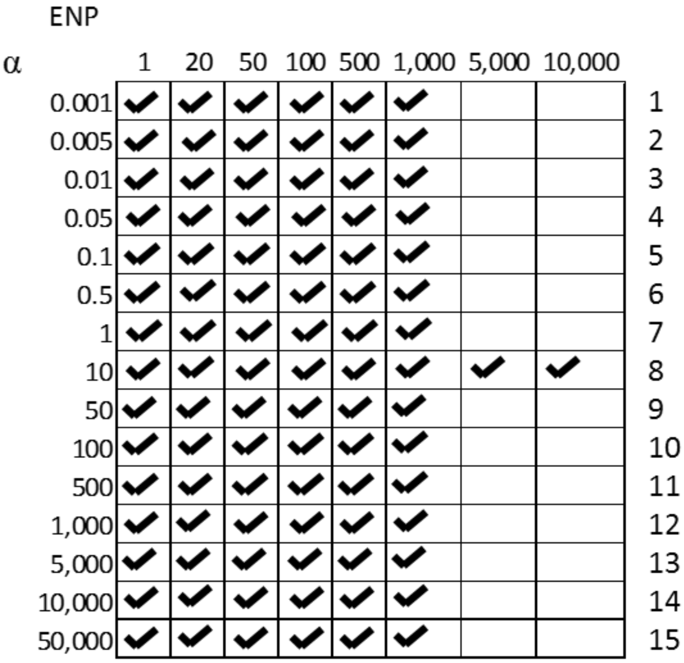
When PDTA was applied to WE-CP, only lower ENP values were used. This was due to the great computational burden of large trees that would be created at high ENP for a dataset of this magnitude. Upon reducing the number of variables, PDTA was able to build models with AUC = 0.57 (95% CI: 0.513-0.627), within which several patient-specific subgroups were discovered. These subgroups included many rare variants. Although few of the predictors in these subgroups were found in databases of known pathogenic variants, many were found to have associations with genes that are under investigation for pancreas related issues. Therefore, some of the rare variants discovered by PDTA may be novel variants.

5.2.4 Evaluation of refined algorithm on WE-DB

5.2.4.1 Empirical Optimization of PDT on WE-DB

The final dataset used to evaluate PDTA was the WE-DB. This is a subset of patient samples from the WE-CP dataset, containing only the cases which were positive for CP/RAP (see Section 4.1.5). Consequently, the number of instances is much lower, while the number of SNVs is comparable to that in WE-DB. I expected, as seen in WE-CP, that low ENP values would not be good for tree building. However, high ENP values lead to very long runtimes for a dataset of this size. Therefore, SDTA was applied to WE-DB with the hyperparameter settings in Table 29. As expected, the lower ENP values (ENP = 1, 20, 50) failed to form trees with more than 3 or 4 nodes. For higher ENP values, larger trees are formed, but even the largest SDT was comprised of 79 nodes (SDT5 in Table 30; α = 10, ENP = 10,000) – a substantially tractable number of variables for WE-DB, which contains 245,573 variables.

Table 29: Hyperparameter settings used in SDTA on WE-DB



ENP

α	1	20	50	100	500	1,000	5,000	10,000	
0.001	✓	✓	✓	✓	✓	✓			1
0.005	✓	✓	✓	✓	✓	✓			2
0.01	✓	✓	✓	✓	✓	✓			3
0.05	✓	✓	✓	✓	✓	✓			4
0.1	✓	✓	✓	✓	✓	✓			5
0.5	✓	✓	✓	✓	✓	✓			6
1	✓	✓	✓	✓	✓	✓			7
10	✓	✓	✓	✓	✓	✓	✓	✓	8
50	✓	✓	✓	✓	✓	✓			9
100	✓	✓	✓	✓	✓	✓			10
500	✓	✓	✓	✓	✓	✓			11
1,000	✓	✓	✓	✓	✓	✓			12
5,000	✓	✓	✓	✓	✓	✓			13
10,000	✓	✓	✓	✓	✓	✓			14
50,000	✓	✓	✓	✓	✓	✓			15

Table 30: 10 best SDTs on WE-DB, ranked by highest AUC

Model	SDT AUC	95% CI	MSR	α	ENP	# of leaves	# of nodes
SDT1	0.570	0.507-0.633	0.560	5000	500	23	9
SDT2	0.570	0.501-0.640	0.551	10	1000	76	65
SDT3	0.566	0.496-0.636	0.375	50	100	14	7
SDT4	0.556	0.491-0.621	0.667	10	5000	103	76
SDT5	0.555	0.490-0.620	0.693	10	10000	112	79
SDT6	0.555	0.496-0.613	0.478	1	1000	21	18
SDT7	0.550	0.484-0.615	0.576	5000	1000	31	13
SDT8	0.547	0.487-0.608	0.481	0.1	1000	25	24
SDT9	0.547	0.487-0.608	0.481	0.05	1000	25	24
SDT10	0.547	0.487-0.608	0.481	0.01	1000	25	24

SDT1 in Table 30 achieved the highest AUC, building a tree with 9 nodes. Choosing this as the best model, its parameter values were applied to test PDTA on WE-DB (Table 31). Even with a fairly high path ENP, PDTA failed to produce any PS paths on this dataset. Therefore, I continue with SDT1 as the best model.

Table 31: Results of PDTA on WE-DB with best SDT parameters (SDT1 in Table 30)

Model	α	tree ENP	path ENP	# of times PS chosen (out of 283)	SDT AUC	95% CI	PDT AUC	95% CI
PDT1	5000	500	500	0	0.570	0.507-0.633	0.570	0.507-0.633
PDT2	5000	500	1000	0	0.570	0.507-0.633	0.570	0.507-0.633

5.2.4.2 Model assessment on WE-DB

The paths in the SDT of the best model, SDT1 in Table 30, were dominated by rare variants (Table 32). 15 unique paths in the SDT were selected by the 283 test samples, but only the most frequently selected path, SDT_Path1, had an unambiguous disease prediction. Therefore, I conclude that the SDT_Path1 is the only clearly defined patient subgroup we can get from this model.

Table 32: SDT paths selected by test samples of WE-DB, ranked by frequency (SDT1 of Table 30)

Path		# of times selected	Disease Prediction	Target	Path length	Terminal Leaf	# of rare SNVs
SDT_Path1	rs1500701--[0]--...--rs114883808--[0]	209	[0.75,0.25]	[162,47]	9	[731,167]	8
SDT_Path2	rs1500701--[1]—leaf	15	[0.49,0.51]	[11,4]	1	[21,25]	1
SDT_Path3	rs1500701--[0]--...--rs10069511--[1]	14	[0.5,0.5]	[10,4]	3	[18,19]	3
SDT_Path4	rs1500701--[0]--...--rs41266136--[1]	11	[0.5,0.5]	[9,2]	6	[23,21]	6
SDT_Path5	rs1500701--[0]--...--rs61741082--[1]	8	[0.5,0.5]	[3,5]	5	[10,8]	5
SDT_Path6	rs1500701--[0]--...--rs2234256--[1]	6	[0.5,0.5]	[2,4]	4	[11,12]	4
SDT_Path7	rs1500701--[0]--...--rs72634778--[1]	6	[0.48,0.52]	[3,3]	7	[4,11]	7
SDT_Path8	rs1500701--[0]--...--rs74452694--[2]	6	[0.49,0.51]	[3,3]	8	[4,8]	7
SDT_Path9	rs1500701--[0]--...--rs114883808--[1]	2	[0.5,0.5]	[1,1]	9	[6,8]	8
SDT_Path10	rs1500701--[3]—leaf	1	[0.5,0.5]	[0,1]	1	[0,0]	1
SDT_Path11	rs1500701--[0]--...--rs74452694--[3]	1	[0.5,0.5]	[1,0]	8	[0,0]	7
SDT_Path12	rs1500701--[0]--...--rs61741082--[2]	1	[0.5,0.5]	[1,0]	5	[0,1]	5
SDT_Path13	rs1500701--[0]--...--rs2234256--[2]	1	[0.5,0.5]	[1,0]	4	[0,2]	4
SDT_Path14	rs1500701--[0]--rs73277460--[1]	1	[0.49,0.51]	[1,0]	2	[7,12]	2
SDT_Path15	rs1500701--[0]--...--rs10069511--[3]	1	[0.5,0.5]	[1,0]	3	[0,1]	3

SDT_Path1 was chosen for 209 out of the 283 test samples, and it correctly predicted the absence of disease in 162 of them. 8 out of the 9 SNVs comprising nodes in SDT_Path1 are rare variables. The known information from dbSNP on functional consequence and clinical significance of the SNVs in SDT_Path1 do not show any connection to pancreatic function or disorder (Table 33). However, mutations in the CTAGE5 gene, associated with the second SNV in this path, rs73277460, is known to interfere with proinsulin trafficking [107]. This is consistent with the absence of mutation in rs73277460 in SDT_Path1, which predicts absence of disease. This subgroup also includes SNVs with are associated with HRNR and HNRNPH1 which may play a role in pancreatic cancer [108,109].

Table 33: SNVs in SDT_Path1 on WE-DB
with known functional consequence and clinical significance from dbSNP
(*) Indicates a gene with known or suspected role in pancreas related function or disorder

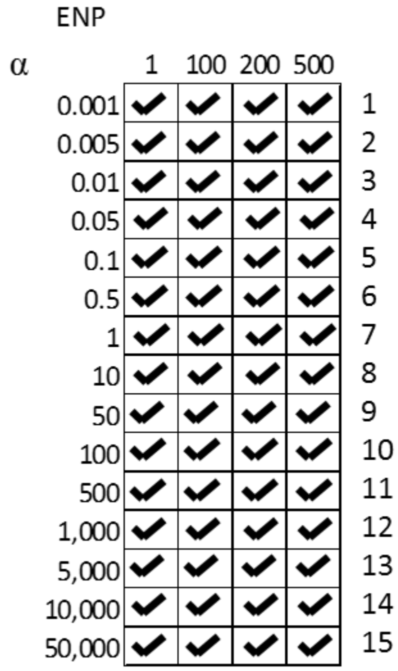
	SNP	Edge	Functional consequence	Clinical significance	Associated Gene	Node Count	Global MAF (TOPMED)
root	rs1500701	[0]	intron variant	NA	EPHA6	[837,300]	G=0.0649/8144
1	rs73277460	[0]	missense, nc transcript variant	NA	CTAGE5 *	[816,274]	G=0.0209/2622
2	rs10069511	[0]	intron variant	NA	LINC01933	[809,261]	A=0.0289/3625
3	rs2234256	[0]	intron variant, missense, upstream variant 2KB	Association: Alzheimer's disease	LOC105375056, TREM2, TREML1	[790,241]	G=0.0413/5186
4	rs61741082	[0]	missense	NA	CHTF18	[778,227]	C=0.0366/4599
5	rs41266136	[0]	missense	NA	HRNR *	[768,217]	T=0.0162/2033
6	rs72634778	[0]	missense, nc transcript variant, upstream variant 2KB, utr variant 5 prime	NA	LOC107984912, ZBTB48	[745,195]	A=0.0059/735
7	rs74452694	[0, 1]	intron variant, upstream variant 2KB, utr variant 5 prime	NA	HNRNPH1 *	[741,184]	T=0.1005/12616
8	rs114883808	[0]	intron variant, missense	NA	LOC101928583, SLC7A14	[737,176]	T=0.0072/904

5.2.4.3 Empirical Optimization of PDT on WE-DB-R

In the case of WE-CP, the use of a lower dimensional dataset resulted in better performance of PDTA. Therefore, I used the same method to create a lower dimensional dataset from WE-DB, WE-DB-R (see Section 4.1.5), to see if PDTA will give better results. First, SDTA was applied to WE-DB-R with the hyperparameter settings shown in Table 34.

As seen in previous cases, raising the value of ENP created larger trees, and in general higher ENPs and higher α resulted in better AUCs (Table 35). SDT1 achieved the best AUC at 0.62 (95% CI: 0.545-0.690), but it recruited all 638 available SNVs as nodes, thereby lowering its value as predictor of patient subgroups (Table 35). The next highest ranked model, SDT2, was built using 42 nodes, and had an AUC of 0.61 (95% CI: 0.553-0.672); as the smaller tree without much loss in predictive power, this was selected as the best model to proceed with the next phase of PDTA (Table 36).

Table 34: Hyperparameter settings used in SDTA on WE-DB-R



	ENP				
α	1	100	200	500	
0.001	✓	✓	✓	✓	1
0.005	✓	✓	✓	✓	2
0.01	✓	✓	✓	✓	3
0.05	✓	✓	✓	✓	4
0.1	✓	✓	✓	✓	5
0.5	✓	✓	✓	✓	6
1	✓	✓	✓	✓	7
10	✓	✓	✓	✓	8
50	✓	✓	✓	✓	9
100	✓	✓	✓	✓	10
500	✓	✓	✓	✓	11
1,000	✓	✓	✓	✓	12
5,000	✓	✓	✓	✓	13
10,000	✓	✓	✓	✓	14
50,000	✓	✓	✓	✓	15

Table 35: 10 best SDTs on WE-DB-R, ranked by highest AUC

Model	AUC	95% CI	MSR	α	ENP	# of leaves	# of nodes
SDT1	0.617	0.5449-0.6892	0.994	50000	400	1344	638
SDT2	0.613	0.5531-0.672	0.967	10000	200	89	42
SDT3	0.611	0.5515-0.6696	0.973	50000	200	148	61
SDT4	0.608	0.5513-0.6654	0.923	10000	100	50	22
SDT5	0.596	0.5243-0.6682	0.986	1000	400	1321	638
SDT6	0.596	0.5228-0.6683	0.957	1000	200	413	178
SDT7	0.591	0.5201-0.6628	0.904	100	100	279	132
SDT8	0.591	0.5199-0.6623	0.918	100	200	329	154
SDT9	0.590	0.5208-0.6601	0.673	1000	50	50	24
SDT10	0.590	0.5207-0.6602	0.980	1	400	646	638

Table 36: Results of PDTA on WE-DB-R with best SDT parameters (SDT2 in Table 35)

Model	α	Tree ENP	Path ENP	# of times PS (out of 283)	SDT AUC	95% CI	PDT AUC	95% CI
PDT1	10000	200	1	0	0.613	0.5531-0.672	0.613	0.553-0.672
PDT2	10000	200	100	0	0.613	0.5531-0.672	0.613	0.553-0.672
PDT3	10000	200	200	0	0.613	0.5531-0.672	0.613	0.553-0.672
PDT4	10000	200	400	71	0.613	0.5531-0.672	0.605	0.542-0.667

5.2.4.4 Model assessment on WE-DB-R

PS paths were produced only in PDT4 of Table 36, which had the highest path ENP setting, but all the PS paths selected all 638 SNVs as nodes, thus failing to produce patient subgroups. Therefore, I proceed with SDT2 of Table 35 as the best model. 31 unique paths were selected by the 283 test samples, and all of them contained rare SNVs. However, just as was the case in WE-DB, only the most frequently selected path, SDT_Path1 (Table 37) had an unambiguous disease prediction. Therefore, SDT_Path1 is the only clearly defined patient subgroup we can get from this model.

SDT_Path1 was chosen for 233 of the 283 test samples, and it correctly predicted the absence of disease in 184 of them. The information from dbSNP on functional consequence and clinical significance of the SNVs in SDT_Path1 do not show any relation to pancreatic function or disorder (Table 38). However, there were numerous genes associated with SNVs in this patient group which have been implicated in pancreas related diseases. TBC1D31 is involved in β-cell replication[110]; CENPF may be disrupted in islet cells in diabetes[111]; ESRRG is involved in regulation of β-cell maturation[112]. There were also several genes in this subgroup – MUC6, TPD52, MMP1, FBXW8, C5, PALB – that have been investigated for a possible role in pancreatic cancer[88,89,113-117] though the predictor for this task was for the presence or absence of diabetes, if genetic markers for pancreatic cancer in CP patients are in a subgroup that is not likely to develop diabetes, there may be a mechanistic connection that relates the co-occurrence of pathogenic/protective variants.

Table 37: Top 8 SDT paths selected by test samples of WE-DB-R (out of 31) ranked by frequency (SDT2 of Table 35)

Path		# of times selected	Disease Prediction	Target	Path length	Terminal Leaf	# of rare SNVs
SDT_Path1	rs199833698:[837,300]--[0]--…--rs17043116:[783,192]--[0, 1, 3]	233	[0.77,0.23]	[184,49]	27	[781,190]	22
SDT_Path2	rs199833698:[837,300]--[0]--…--rs55959319:[787,204]--[1]	4	[0.5,0.5]	[3,1]	18	[1,2]	16
SDT_Path3	rs199833698:[837,300]--[0]--…--rs2081807:[8,10]--[2]	3	[0.47,0.53]	[2,1]	4	[0,8]	3
SDT_Path4	rs199833698:[837,300]--[0]--…--rs7086208:[789,210]--[2]	3	[0.49,0.51]	[2,1]	14	[0,2]	12
SDT_Path5	rs199833698:[837,300]--[1]--…--rs73507220:[8,4]--[1]	3	[0.49,0.51]	[1,2]	3	[0,2]	2
SDT_Path6	rs199833698:[837,300]--[0]--…rs12614237:[13,13]--[0]	3	[0.51,0.49]	[3,0]	5	[2,0]	4
SDT_Path7	rs199833698:[837,300]--[0]--…--rs11762428:[6,2]--[1]	2	[0.49,0.51]	[1,1]	5	[0,2]	3
SDT_Path8	rs144825978:[796,241]--[0]--…--rs76280974:[783,196]--[1]	2	[0.5,0.5]	[1,1]	23	[0,1]	18

Table 38: SNVs in SDT_Path1 on WE-DB-R
with known functional consequence and clinical significance from dbSNP
(*) Indicates a gene with known or suspected role in pancreas related function or disorder

	SNP	**Edge**	**Functional consequence**	**Clinical significance**	**Associated Gene**	**Node Count**	**Global MAF (TOPMED)**
root	rs199833698	[0]	Missense	NA	MUC6 *	[837,300]	A=0.0314/3949
1	rs75916244	[0]	missense, nc transcript variant, synonymous codon	NA	APOO	[823,280]	C=0.0227/2851
2	rs80284803	[0]	Missense	NA	HYDIN	[818,269]	G=0.000008/1
3	rs2230795	[0]	missense, nc transcript variant	Likely benign	ELP1, IKBKAP	[810,257]	T=0.0359/4502
4	rs146538633	[0]	intron variant,missense, upstream variant 2KB, utr variant 5 prime	NA	CHTF18, RPUSD1	[797,244]	T=0.0046/576
5	rs144825978	[0]	Missense	NA	RGAG1, RTL9	[796,241]	T=0.0087/1090
6	rs117461662	[0]	intron variant, missense, nc transcript variant	NA	LOC105375919, TPD52 *	[794,239]	G=0.0032/404
7	rs111534710	[0]	Missense	NA	FAM47A	[791,230]	G=0.0129/1623
8	rs200410849	[0]	missense, upstream variant 2KB	NA	UMODL1	[791,227]	C=0.0006/76
9	rs379999	[0, 1]		NA		[791,222]	C=0.0856/10744
10	rs148980271	[0]	intron variant, missense	NA	MMP1, * WTAPP1	[791,219]	T=0.0008/101
11	rs114540180	[0]	missense, splice acceptor variant	NA	GIMAP6	[791,216]	C=0.0188/2362
12	rs144993453	[0]	intron variant, missense	NA	RBPMS	[789,212]	A=0.0029/366
13	rs7086208	[0, 1]	intron variant	NA	FRMD4A	[789,210]	A=0.1011/12692
14	rs147449897	[0]	missense, nc transcript variant	NA	FBXW8 *	[789,208]	G=0.0041/515
15	rs41311881	[0]	Missense	NA	C5 *	[789,207]	G=0.0052/653
16	rs35988863	[0]	Missense	NA	KRT71	[789,206]	A=0.0131/1650
17	rs55959319	[0]	missense, nc transcript variant	NA	TBC1D31 *	[787,204]	A=0.0169/2121
18	rs3795518	[0, 1]	Missense	NA	CENPF *	[786,202]	A=0.1179/14799
19	rs5904862	[0, 1, 2]	missense, nc transcript variant, syn codon	NA	CDKL4	[783,199]	G=0.2442/30668
20	rs11892364	[0, 1]		NA		[783,198]	A=0.0174/2188
21	rs4743820	[0, 1, 2]	intron variant, upstream variant 2KB	NA	LINC00484, LOC100507103	[783,197]	C=0.3941/49484
22	rs76280974	[0]	intron variant, missense	NA	WDR78	[783,196]	C=0.0080/1005
23	rs138789658	[0]	missense, utr variant 5 prime	Uncertain significance	PALB2 *	[783,195]	C=0.0054/679
24	rs139102003	[0]	Missense	NA	CTSO	[783,194]	A=0.0070/880
25	rs75987633	[0, 1]		NA		[783,193]	T=0.0300/3770
26	rs17043116	[0, 1, 3]	intron variant		ESRRG *	[783,192]	A=0.0665/8352

5.2.4.5 Summary of PDTA performance on WE-DB and WE-DB-R

The WE-DB and WE-DB-R datasets did not have a large sample size. Also, the samples were heavily unbalanced in favor of controls – there were almost 3 times as many controls than diabetic patients. This may have driven PDTA to predict the main subgroup defined by absence of disease rather than presence of disease. Nevertheless, PDTA found a predictive patient subgroup in both WE-DB and WE-DB-R, and the subgroups contained many rare SNVs. Reducing the variable space from WE-DB to WE-DB-R did not give us more patient subgroups, but gave a model with better AUC and more predictive variants.

In the case of WE-CP, I did not test PDTA with larger ENPs. Upon testing WE-CP-R with larger ENPs, PDTA performed better and uncovered rare variants. Large ENPs on large datasets take a large amount of time to compute. The comparison of WE-DB with WE-DB-R works in favor of the instinct that it is not worth wasting time on pushing large and time-consuming models, when reduced variable versions give better predictive performance while still finding rare variants as predictors, without the extra computational investment.

A further consideration about WE-DB and WE-DB-R: the patients in this dataset were not aligned for disease severity. It is possible that many of the control cases were actually patients who had not yet developed diabetes, but had the genetic profile that would eventually lead them to develop diabetes. Incorrect labeling of targets is a known source of poor predictor performance. It would be worthwhile to redo this analysis with a larger sample size, with a higher proportion of case to control, and diabetic vs non-diabetic CP patients who are at comparable stages of disease. This may be better suited to find variants for subpopulations of CP patients who are prone to develop diabetes.

5.2.5 Comparison to other methods

The PDTA is computationally much more intensive than regular decision tree implementations. Therefore, it is important to know if PDTA performs better than a regular decision tree package such as available in Scikit-learn[76]. Also, because of different theoretical underpinnings, different scoring functions can result in different optimal networks. To see how much PDTA is influenced by the choice of score, I compare performance of PDTA implementation with other implementations which use as score another Bayesian score, K2, and two information theoretic scores, AIC and MDL, across all 5 datasets (Table 39). PDTA is seen to outperform the regular decision tree on all 5 datasets. The choice of BDeu as the score for PDTA gives superior or equivalent performance compared to K2, AIC and MDL.

Table 39: PDTA performance compared to other methods

AUC	PDTA AUC (95% CI)	Scikit-learn decision tree	PDTA Using K2	PDTA Using AIC	PDTA Using MDL
SD	0.989 (0.985-0.994)	0.851	0.987	0.982	0.987
SBD	0.851 (0.816-0.887)	0.720	0.840	0.838	0.844
SSD	0.602 (0.573-0.632)	0.563	0.597	0.590	0.583
WE-CP-R	0.574 (0.528-0.620)	0.524	0.562	0.563	0.556
WE-DB-R	0.617 (0.545-0.689)	0.532	0.621	0.565	0.570

6.0 CONCLUSIONS

The PDTA was developed to enable discovery of patient-specific groups in complex diseases, where the discovery of rare variants is of great value. The results of evaluation on 5 different datasets, ranging from synthetic to real, shows that it is good at discovering rare variants as predictors. In the case of the CP-related datasets, several validated SNPs as well as unknown SNVs were identified that could be novel undiscovered disease variants that modulate or cause CP in conjunction with other variants.

Our initial assumption was that the PDTA would have a separation in predictions from its two phases: the SDT was expected to be useful for population-wide characterization, while the PS path would give patient-specific groups. It turned out that the PS paths rarely improved over the personalization achieved by SDT. Patient-specific subgroups with rare variables as predictors surfaced within the SDT as a result of the Bayesian tree building method. This is nice, because it truly exemplifies a hierarchical view of a population that can be divided broadly (for presence or absence of a disease) but also stratified for how groups of patients with characteristic groups of variants can have individual disease risks, particularly valuable for uncovering rare variants that normally are not caught in large statistical studies. For example, in the dataset where chronic pancreatitis was used as the target phenotype, two subgroups emerged which included variants in diabetes-related genes. With the knowledge of a patient's variant subgroup, precision medicine can be applied for prognosis and treatment decisions targeting the pathways relevant to that individual's biological sources of dysfunction.

Although the PDTA does a good job of finding patient subgroups, it exhibits some signs of tension between generalizing to the population, and finding robust subgroups. The SSD was created specifically as a computational challenge to researchers for the discovery of rare variants. When PDTA was applied to SSD, the overall performance measures were not so good, with poor predictability for the majority of the population, but there were many subgroups within the tree that had high predictive power. When PDTA was applied to WE-DB, the presence of a much larger number of controls relative to diseased samples resulted in the largest predictive group to characterize the control population. The sample size and composition of the dataset can have a crucial effect on how well the PDTA balances the search for populations of interest.

There were several rare variants implicated by PDTA with causal function in association with other rare and common variants. These can be novel discoveries to be tested further using methods such as candidate gene studies. When PDTA was applied to the CP cohort with diabetes as target phenotype, a patient subgroup was found to aggregate variants for insulin dysfunction with variants involved in pancreatic cancer. Diabetes and pancreatic cancer are both known to be eventual side effects of chronic pancreatitis, but there is no clear timeline of the development of these effects in relation to each other in the broader CP population. Moreover, there is evidence that suggests co-occurrence of risk factors for the development of diabetes and cancer in subpopulations of CP patients[118,119]. The variants discovered by PDTA could point to the mechanistic basis for such associations.

6.1 FINDINGS AND CONTRIBUTIONS

PDTA provides (1) a prediction for the disease state, and (2) a set of variables in a patient-specific subgroup that represent informative predictors for the individual patient of interest within the larger context of the diseased population, and may provide clues to the disease mechanisms active in that type of person. The set of probabilistic rules obtained from all individuals in the test set provides (1) a set of predictions for the population as a whole that is represented by the set, and (2) the union of features in all the rules provides clues to the disease mechanisms that are active in the entire population.

The implementation of PDTA for genomic data, specifically, provides parsimonious genetic signatures containing rare variants that have the potential to be novel causal variants in subpopulations of patients with chronic pancreatitis.

6.2 FUTURE WORK

Observing the behavior of PDTA on several different datasets suggests many possible avenues of further improvement.

6.2.1 Dimensionality Reduction

In the case of large genomic datasets, their lower dimensional counterparts appeared to give better prediction results, while still selecting rare variants as predictors. It would be worth considering some form of dimensionality reduction to genomic datasets before application of the

PDTA. Dimensionality reduction techniques are common in the biomedical domain. Many dimensionality reduction methods construct new features as a combination of variables, to create a more parsimonious set of variables. This would be unsuitable for the PDTA, which is trying to find rare variants. Any choice of dimensionality reduction should preserve interpretability in terms of the individual contributions of variants in the dataset, while providing enrichment of the variable space.

6.2.2 Genotypic specificity

In the development of PDTA, I described the genotypic adaptation to assignment of variable values, to mimic the idea of different kinds of genetic models (Section 5.1.3). This allows for variables in the model to assume more than one allelic value in calculating the branches emerging from that node. In some cases this may make sense: if a heterozygous variant has the same functional impact as a biallelic mutant variant, for example, they would have the same predictive effect for the disease phenotype. However, in some cases, the combination of functionally distinct conditions are not realistic. Consultation with a domain expert would help in refining this variable selection step such that only functionally meaningful combinations would contribute to tree growth.

BIBLIOGRAPHY

1 National Library of medicine (NLM). *What is Precision Medicine?* <https://ghr.nlm.nih.gov/primer/precisionmedicine/definition> (2014-2018).

2 Visweswaran, S. *et al.* Learning patient-specific predictive models from clinical data. *Journal of biomedical informatics* 43, 669-685, doi:10.1016/j.jbi.2010.04.009 (2010).

3 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272-276, doi:10.1038/nature08250 (2009).

4 Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42, 30-35, doi:10.1038/ng.499 (2010).

5 Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 45, 1452-1458, doi:10.1038/ng.2802 (2013).

6 Kitts A, P. L., Ward M, et al. *The Database of Short Genetic Variation (dbSNP)*, <https://www.ncbi.nlm.nih.gov/books/NBK174586/> (2013).

7 Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 40, 695-701, doi:10.1038/ng.f.136 (2008).

8 Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annual review of genetics* 44, 293-308, doi:10.1146/annurev-genet-102209-163421 (2010).

9 Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine* 363, 166-176, doi:10.1056/NEJMra0905980 (2010).

10 Shields, R. Common disease: are causative alleles common or rare? *PLoS biology* 9, e1001009, doi:10.1371/journal.pbio.1001009 (2011).

11 Gorlov, I. P., Gorlova, O. Y., Frazier, M. L., Spitz, M. R. & Amos, C. I. Evolutionary evidence of the effect of rare variants on disease etiology. *Clinical genetics* 79, 199-206, doi:10.1111/j.1399-0004.2010.01535.x (2011).

12 Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11, 773-785, doi:10.1038/nrg2867 (2010).

13 Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* 13, 135-145, doi:10.1038/nrg3118 (2012).

14 Fisher, R. A. *The genetical theory of natural selection: a complete variorum edition*. (Oxford University Press, 1930).

15 Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* 9, 255-266, doi:10.1038/nrg2322 (2008).

16 Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 11, 415-425, doi:10.1038/nrg2779 (2010).

17 Feldman, M. & Lewontin, R. The heritability hang-up. *Science* 190, 1163-1168, doi:10.1126/science.1198102 (1975).

18 Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11, 446-450, doi:10.1038/nrg2809 (2010).

19 Kohane, I. S. The twin questions of personalized medicine: who are you and whom do you most resemble? *Genome medicine* 1, 4, doi:10.1186/gm4 (2009).

20 Jameson, J. L. & Longo, D. L. Precision medicine--personalized, problematic, and promising. *The New England journal of medicine* 372, 2229-2234, doi:10.1056/NEJMsb1503104 (2015).

21 Krishna, R. Model-based benefit-risk assessment: can Archimedes help? *Clinical pharmacology and therapeutics* 85, 239-240, doi:10.1038/clpt.2008.240 (2009).

22 Bellows, J., Patel, S. & Young, S. S. Use of IndiGO individualized clinical guidelines in primary care. *Journal of the American Medical Informatics Association* 21, 432-437 (2014).

23 Gallego, B. *et al.* Bringing cohort studies to the bedside: framework for a 'green button' to support clinical decision-making. *Journal of Comparative Effectiveness Research* 4, 191-197, doi:10.2217/cer.15.12 (2015).

24 Frankovich, J., Longhurst, C. A. & Sutherland, S. M. Evidence-based medicine in the EMR era. *The New England journal of medicine* 365, 1758-1759, doi:10.1056/NEJMp1108726 (2011).

25 Brown, S. A. Patient Similarity: Emerging Concepts in Systems and Precision Medicine. *Frontiers in physiology* 7, 561, doi:10.3389/fphys.2016.00561 (2016).

26 Manrai, A. K., Ioannidis, J. A. & Kohane, I. S. Clinical genomics: From pathogenicity claims to quantitative risk estimates. *JAMA* 315, 1233-1234, doi:10.1001/jama.2016.1519 (2016).

27 Witt, H., Apte, M. V., Keim, V. & Wilson, J. S. Chronic pancreatitis: challenges and advances in pathogenesis, genetics, diagnosis, and therapy. *Gastroenterology* 132, 1557-1573, doi:10.1053/j.gastro.2007.03.001 (2007).

28 Whitcomb, D. C. *et al.* Chronic pancreatitis: An international draft consensus proposal for a new mechanistic definition. *Pancreatology* 16, 218-224, doi:10.1016/j.pan.2016.02.001 (2016).

29 Whitcomb, D. C. Genetic Risk Factors for Pancreatic Disorders. *Gastroenterology* 144, 1292-1302, doi:10.1053/j.gastro.2013.01.069 (2013).

30 Whitcomb, D. C. What is personalized medicine and what should it replace? *Nature Reviews Gastroenterology and Hepatology* 9, 418-424 (2012).

31 Whitcomb, D. C. *et al.* Hereditary pancreatitis is caused by a mutation in the cationic trypsinogen gene. *Nat Genet* 14, 141-145, doi:10.1038/ng1096-141 (1996).

32 LaRusch, J. & Whitcomb, D. C. Genetics of pancreatitis. *Current opinion in gastroenterology* 27, 467-474, doi:10.1097/MOG.0b013e328349e2f8 (2011).

33 Moran, R. A., Quesada-Vazquez, N., Sinha, A., de-Madaria, E. & Singh, V. K. High Penetrance of the PRSS1 A16V Mutation in a Kindred With SPINK1 N34S and CFTR TG11-5T Co-mutations. *Pancreas* 45, e2-4, doi:10.1097/MPA.0000000000000445 (2016).

34 Nemeth, B. C., Szucs, A., Hegyi, P. & Sahin-Toth, M. Novel PRSS1 Mutation p.P17T Validates Pathogenic Relevance of CTRC-Mediated Processing of the Trypsinogen Activation Peptide in Chronic Pancreatitis. *The American journal of gastroenterology* 112, 1896-1898, doi:10.1038/ajg.2017.393 (2017).

35 LaRusch, J., Barmada, M. M., Solomon, S. & Whitcomb, D. C. Whole exome sequencing identifies multiple, complex etiologies in an idiopathic hereditary pancreatitis kindred. *JOP: Journal of the pancreas* 13, 258 (2012).

36 Yadav, D., Eigenbrodt, M. L., Briggs, M. J., Williams, D. K. & Wiseman, E. J. Pancreatitis: prevalence and risk factors among male veterans in a detoxification program. *Pancreas* 34, 390-398 (2007).

37 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* 44, 1349-1354, doi:10.1038/ng.2466 (2012).

38 LaRusch, J., Solomon, S. & Whitcomb, D. C. *GeneReviews® [Internet]* (ed Ardinger HH Adam MP, Pagon RA, et al.) (University of Washington, Seattle (WA), 2014).

39 Masson, E., Chen, J. M., Scotet, V., Le Marechal, C. & Ferec, C. Association of rare chymotrypsinogen C (CTRC) gene variations in patients with idiopathic chronic pancreatitis. *Hum Genet* 123, 83-91, doi:10.1007/s00439-007-0459-3 (2008).

40 Ewald, N. & Hardt, P. D. Diagnosis and treatment of diabetes mellitus in chronic pancreatitis. *World journal of gastroenterology* 19, 7276-7281, doi:10.3748/wjg.v19.i42.7276 (2013).

41 Aggarwal, C. C. *Data classification: algorithms and applications*. (CRC Press, 2014).

42 Friedman, J. H., Kohavi, R. & Yun, Y. Lazy decision trees. *AAAI/IAAI, Vol. 1.* 717-724.

43 Zheng, Z. & Webb, G. I. Lazy learning of Bayesian rules. *Machine Learning* 41, 53-84 (2000).

44 Ting, K. M., Zheng, Z. & Webb, G. Learning lazy rules to improve the performance of classifiers. *Research and Development in Intelligent Systems XVI* 122-131 (Springer, 2000).

45 Visweswaran, S. & Cooper, G. F. Instance-specific Bayesian model averaging for classification. *Advances in Neural Information Processing Systems.* 1449-1456.

46 Visweswaran, S. & Cooper, G. F. Learning instance-specific predictive models. *Journal of Machine Learning Research* 11, 3333-3369 (2010).

47 Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M. & Edwards, D. D. *Artificial intelligence: a modern approach*. Vol. 2 (Prentice hall Upper Saddle River, 2003).

48 Visweswaran, S., Ferreira, A., Ribeiro, G. A., Oliveira, A. C. & Cooper, G. F. Personalized Modeling for Prediction with Decision-Path Models. *PloS one* 10, e0131022, doi:10.1371/journal.pone.0131022 (2015).

49 Koller, D. F., N. *Probabilistic Graphical Models: Principles and Techniques*. (2009).

50 Heckerman, D., Geiger, D. & Chickering, D. M. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20, 197-243, doi:10.1007/bf00994016 (1995).

51 Meek, C. & Heckerman, D. Structure and parameter learning for causal independence and causal interaction models. *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence* 366-375 (Morgan Kaufmann Publishers Inc., Providence, Rhode Island, 1997).

52 Microsoft. *Microsoft Decision Trees Algorithm Technical Reference*. <https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-decision-trees-algorithm-technical-reference?view=sql-server-2017>

53 Buntine, W. Theory refinement on Bayesian networks. *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence.* 52-60 (Morgan Kaufmann Publishers Inc., Providence, Rhode Island, 1991).

54 Silander, T., Kontkanen, P. & Myllymäki, P. On sensitivity of the MAP Bayesian netwok structure to the equivalent sample size parameter. *Proceedings of the Twenty-third Conference on Uncertainty in Artficial Intelligence.* 360-367 (2007).

55 Steck, H. & Jaakkola, T. S. On the Dirichlet prior and Bayesian regularization. *Proceedings of the 15th International Conference on Neural Information Processing Systems.* 713-720 (MIT Press, 2002).

56 Ueno, M. Learning networks determined by the ratio of prior and data. *Proceedings of the Twenty-sixth Conference on Uncertainty in Artficial Intelligence.* 598-605 (2010).

57 Trusheim, M. R., Berndt, E. R. & Douglas, F. L. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nat Rev Drug Discov* 6, 287-293 (2007).

58 Lin, H.-c., Baracos, V., Greiner, R. & Chun-nam, J. Y. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems.* 1845-1853.

59 Kasabov, N. Global, local and personalised modeling and pattern discovery in bioinformatics: An integrated approach. *Pattern Recognition Letters* 28, 673-685 (2007).

60 Ng, K., Sun, J., Hu, J. & Wang, F. Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity. *AMIA Summits on Translational Science Proceedings* 2015, 132-136 (2015).

61 Ferreira, A., Cooper, G. F. & Visweswaran, S. Decision Path Models for Patient-Specific Modeling of Patient Outcomes. *AMIA Annual Symposium Proceedings* 2013, 413-421 (2013).

62 Buntine, W. Learning classification trees. *Statistics and computing* 2, 63-73 (1992).

63 Almasy, L. *et al.* Genetic Analysis Workshop 17 mini-exome simulation. *BMC proceedings* 5 Suppl 9, S2, doi: 10.1186/1753-6561-5-S9-S2 (2011).

64 Ghosh, S. *et al.* Identifying rare variants from exome scans: the GAW17 experience. *BMC proceedings* 5 Suppl 9, S1, doi:10.1186/1753-6561-5-S9-S1 (2011).

65 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* 8, 520-531, doi:10.1159/000152001 (2008).

66 Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29-36 (1982).

67 DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837-845 (1988).

68 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* 12, 77, doi:10.1186/1471-2105-12-77 (2011).

69 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81, 559-575, doi:10.1086/519795 (2007).

70 National Heart, Lung, and Blood Institute (NHLBI). *Trans-Omics for Precision Medicine*. <https://www.nhlbiwgs.org/> (2014-2018).

71 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 29, 308-311 (2001).

72 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* 46, D1062-D1067, doi:10.1093/nar/gkx1153 (2018).

73 Nemeth, B. C. & Sahin-Toth, M. Human cationic trypsinogen (PRSS1) variants and chronic pancreatitis. *American journal of physiology. Gastrointestinal and liver physiology* 306, G466-473, doi:10.1152/ajpgi.00419.2013 (2014).

74 Thomas, J. K. *et al.* Pancreatic Cancer Database: an integrative resource for pancreatic cancer. *Cancer biology & therapy* 15, 963-967, doi:10.4161/cbt.29188 (2014).

75 Cariaso, M. & Lennon, G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic acids research* 40, D1308-1312, doi:10.1093/nar/gkr798 (2012).

76 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825-2830 (2011).

77 Breiman, L., Friedman, J., Olshen R. A. & Stone, C. J. *Classification and Regression Trees*. (Wadsworth, Inc., 1984).

78 Cooper, G. F. & Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9, 309-347, doi:10.1007/bf00994110 (1992).

79 Yang, S. & Chang, K. *Comparison of score metrics for Bayesian network learning*. Vol. 32 (2002).

80 Broom, B. M., Do, K. A. & Subramanian, D. Model averaging strategies for structure learning in Bayesian networks with limited data. *BMC bioinformatics* 13 Suppl 13, S10, doi:10.1186/1471-2105-13-S13-S10 (2012).

81 Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. *Proc 2nd Intl Sym Inform Theory*, 267-281, doi:10.1007/978-1-4612-1694-0_15 (1973).

82 Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 461-464 (1978).

83 Liu, Z., Malone, B. & Yuan, C. Empirical evaluation of scoring functions for Bayesian network model selection. *BMC bioinformatics* 13 Suppl 15, S14, doi:10.1186/1471-2105-13-S15-S14 (2012).

84 Allen, T. V. & Greiner, R. Model selection criteria for learning belief nets: An empirical comparison. *Proceedings of the Seventeenth International Conference on Machine Learning* 1047-1054 (Morgan Kaufmann Publishers Inc., 2000).

85 Lewis, C. M. Genetic association studies: design, analysis and interpretation. *Briefings in bioinformatics* 3, 146-153 (2002).

86 Takehara, A. *et al.* Gamma-aminobutyric acid (GABA) stimulates pancreatic cancer growth through overexpressing GABAA receptor pi subunit. *Cancer research* 67, 9704-9712, doi:10.1158/0008-5472.CAN-07-2099 (2007).

87 Johnson, S. K. & Haun, R. S. The gamma-aminobutyric acid A receptor pi subunit is overexpressed in pancreatic adenocarcinomas. *JOP : Journal of the pancreas* 6, 136-142 (2005).

88 Yue, T. *et al.* The prevalence and nature of glycan alterations on specific proteins in pancreatic cancer patients revealed using antibody-lectin sandwich arrays. *Molecular & cellular proteomics : MCP* 8, 1697-1707, doi:10.1074/mcp.M900135-MCP200 (2009).

89 Streppel, M. M. *et al.* Mucin 16 (cancer antigen 125) expression in human tissues and cell lines and correlation with clinical outcome in adenocarcinomas of the pancreas, esophagus, stomach, and colon. *Human pathology* 43, 1755-1763, doi:10.1016/j.humpath.2012.01.005 (2012).

90 Song, K. *et al.* Transforming Growth Factor TGFbeta Increases Levels of Microtubule-Associated Protein MAP1S and Autophagy Flux in Pancreatic Ductal Adenocarcinomas. *PloS one* 10, e0143150, doi:10.1371/journal.pone.0143150 (2015).

91 Murphy, S. J. *et al.* Integrated Genomic Analysis of Pancreatic Ductal Adenocarcinomas Reveals Genomic Rearrangement Events as Significant Drivers of Disease. *Cancer research* 76, 749-761, doi:10.1158/0008-5472.CAN-15-2198 (2016).

92 Cao, D. *et al.* Identification of novel highly expressed genes in pancreatic ductal adenocarcinomas through a bioinformatics analysis of expressed sequence tags. *Cancer biology & therapy* 3, 1081-1089; discussion 1090-1081 (2004).

93 Vila-Casadesus, M. *et al.* Deciphering microRNA targets in pancreatic cancer using miRComb R package. *Oncotarget* 9, 6499-6517, doi:10.18632/oncotarget.24034 (2018).

94 Kwon, M. S. *et al.* Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. *BMC genomics* 16 Suppl 9, S4, doi:10.1186/1471-2164-16-S9-S4 (2015).

95 Wei, P., Tang, H. & Li, D. Insights into pancreatic cancer etiology from pathway analysis of genome-wide association study data. *PloS one* 7, e46887, doi:10.1371/journal.pone.0046887 (2012).

96 George Priya Doss, C. *et al.* A novel computational and structural analysis of nsSNPs in CFTR gene. *Genomic medicine* 2, 23-32, doi:10.1007/s11568-008-9019-8 (2008).

97 Petri, A. *et al.* The effect of neurogenin3 deficiency on pancreatic gene expression in embryonic mice. *Journal of molecular endocrinology* 37, 301-316, doi:10.1677/jme.1.02096 (2006).

98 Zhang, Y., Han, D., Yu, P., Huang, Q. & Ge, P. Genome-scale transcriptional analysis reveals key genes associated with the development of type II diabetes in mice. *Experimental and therapeutic medicine* 13, 1044-1150, doi:10.3892/etm.2017.4042 (2017).

99 Witt, H. *et al.* Mutations in the gene encoding the serine protease inhibitor, Kazal type 1 are associated with chronic pancreatitis. *Nat Genet* 25, 213-216, doi:10.1038/76088 (2000).

100 Jackerott, M. *et al.* STAT5 activity in pancreatic beta-cells influences the severity of diabetes in animal models of type 1 and 2 diabetes. *Diabetes* 55, 2705-2712, doi:10.2337/db06-0244 (2006).

101 Zu, Y. *et al.* Pericentrin Is Related to Abnormal beta-Cell Insulin Secretion through F-Actin Regulation in Mice. *PloS one* 10, e0130458, doi:10.1371/journal.pone.0130458 (2015).

102 Deng, W. H. *et al.* Effects of ORP150 on appearance and function of pancreatic beta cells following acute necrotizing pancreatitis. *Pathol Res Pract* 207, 370-376, doi:10.1016/j.prp.2011.03.006 (2011).

103 Yu, K. H. *et al.* Stable isotope dilution multidimensional liquid chromatography-tandem mass spectrometry for pancreatic cancer serum biomarker discovery. *Journal of proteome research* 8, 1565-1576, doi:10.1021/pr800904z (2009).

104 Chen, R. *et al.* Pancreatic cancer proteome: the proteins that underlie invasion, metastasis, and immunologic escape. *Gastroenterology* 129, 1187-1197, doi:10.1053/j.gastro.2005.08.001 (2005).

105 McAtee, C. O., Barycki, J. J. & Simpson, M. A. Emerging roles for hyaluronidase in cancer metastasis and therapy. *Advances in cancer research* 123, 1-34, doi:10.1016/B978-0-12-800092-2.00001-0 (2014).

106 Zhang, L. *et al.* Identification of a putative tumor suppressor gene Rap1GAP in pancreatic cancer. *Cancer research* 66, 898-906, doi:10.1158/0008-5472.CAN-05-3025 (2006).

107 Fan, J. *et al.* cTAGE5 deletion in pancreatic beta cells impairs proinsulin trafficking and insulin biogenesis in mice. *The Journal of cell biology* 216, 4153-4164, doi:10.1083/jcb.201705027 (2017).

108 Gutknecht, M. F. *et al.* Identification of the S100 fused-type protein hornerin as a regulator of tumor vascularity. *Nature communications* 8, 552, doi:10.1038/s41467-017-00488-6 (2017).

109 Honore, B., Baandrup, U. & Vorum, H. Heterogeneous nuclear ribonucleoproteins F and H/H' show differential expression in normal and selected cancer tissues. *Experimental cell research* 294, 199-209, doi:10.1016/j.yexcr.2003.11.011 (2004).

110 Klochendler, A. *et al.* The Genetic Program of Pancreatic beta-Cell Replication In Vivo. *Diabetes* 65, 2081-2093, doi:10.2337/db16-0003 (2016).

111 Segerstolpe, A. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell metabolism* 24, 593-607, doi:10.1016/j.cmet.2016.08.020 (2016).

112 Yoshihara, E. *et al.* ERRgamma Is Required for the Metabolic Maturation of Therapeutically Functional Glucose-Responsive beta Cells. *Cell metabolism* 23, 622-634, doi:10.1016/j.cmet.2016.03.005 (2016).

113 Pan, S. *et al.* Proteomics portrait of archival lesions of chronic pancreatitis. *PloS one* 6, e27574, doi:10.1371/journal.pone.0027574 (2011).

114 Ito, T. *et al.* Expression of the MMP-1 in human pancreatic carcinoma: relationship with prognostic factor. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 12, 669-674 (1999).

115 Wang, H. *et al.* The CUL7/F-box and WD repeat domain containing 8 (CUL7/Fbxw8) ubiquitin ligase promotes degradation of hematopoietic progenitor kinase 1. *The Journal of biological chemistry* 289, 4009-4017, doi:10.1074/jbc.M113.520106 (2014).

116 Sendler, M. *et al.* Complement Component 5 Mediates Development of Fibrosis, via Activation of Stellate Cells, in 2 Mouse Models of Chronic Pancreatitis. *Gastroenterology* 149, 765-776 e710, doi:10.1053/j.gastro.2015.05.012 (2015).

117 Tischkowitz, M. D. *et al.* Analysis of the gene coding for the BRCA2-interacting protein PALB2 in familial and sporadic pancreatic cancer. *Gastroenterology* 137, 1183-1186, doi:10.1053/j.gastro.2009.06.055 (2009).

118 Li, D., Yeung, S. C., Hassan, M. M., Konopleva, M. & Abbruzzese, J. L. Antidiabetic therapies affect risk of pancreatic cancer. *Gastroenterology* 137, 482-488, doi:10.1053/j.gastro.2009.04.013 (2009).

119 Andersen, D. K. *et al.* Pancreatitis-diabetes-pancreatic cancer: summary of an NIDDK-NCI workshop. *Pancreas* 42, 1227-1237, doi:10.1097/MPA.0b013e3182a9ad9d (2013).