6-28-2016

Rule-based Risk Monitoring Systems for Complex Datasets

Mona Haghighi
University of South Florida, monahaghighi@mail.usf.edu

Rule-based Risk Monitoring Systems for Complex Datasets

by

Mona Haghighi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Industrial and Management Systems Engineering
College of Engineering
University of South Florida

Co-Major Professor: Tapas Das, Ph.D.
Co-Major Professor: Shuai Huang, Ph.D.
Jose Zayas Castro, Ph.D.
Lawrence O. Hall, Ph.D.
Dave Morgan, Ph.D.

Date of Approval:
June 20, 2016

Keywords: Decision Tree, RuleFit, Machine Learning, Biomarker Identification, Item Response Theory

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this dissertation we present rule-based machine learning methods for solving problems with high-dimensional or complex datasets. We are applying decision tree methods on blood-based biomarkers and neuropsychological tests to predict Alzheimer's disease in its early stages. We are also using tree-based methods to identify disparity in dementia related biomarkers among three female ethnic groups. In another part of this research, we tried to use rule-based methods to identify homogeneous subgroups of subjects who share the same risk patterns out of a heterogeneous population. Finally, we applied a network-based method to reduce the dimensionality of a clinical dataset, while capturing the interaction among variables. The results show that the proposed methods are efficient and easy to use in comparison to the current machine learning methods.

CHAPTER 1: INTRODUCTION

Although an abundance of biomarkers can be measured in large-scale studies such as TEDDY/ADNI, it is unrealistic to measure all these biomarkers in the general population, due to the high cost, rigorous preprocessing and standardization protocols, and limited feasibility. Therefore, to proliferate outside the research arena, a risk monitoring system must be cost-effective, easy to implement and efficient for repeated use.

In this dissertation we target problems by identifying risk-predictive rules and also selecting and assembling the rules. So, the first problem is to generate rules from data and then optimally select and assemble a compact set of rules for accurately monitoring disease progression. The key for successful adoption of existing rule-discovery algorithms, when analyzing datasets with a large number of variables and interactions is to reduce the number of variables beforehand. While most existing high dimensional feature selection methods aim to identify the significant features, rule discovery concerns a more difficult problem that involves not only the features/variables (we will use the terms interchangeably) themselves, but also their interactions and ranges, which exponentially increases the dimensionality of the problem. Thus, the essential computational challenge is to identify highly synergistic groups of features such that high-quality rules are more likely to be discovered. To address this challenge, we propose to investigate a novel screening paradigm via a series of ensemble rule methods.

In the second chapter of this dissertation, we discuss trying to predict the elevation of brain amyloid burden using neuropsychological tests and blood-based biomarkers. Elevated brain

amyloid burden is the first symptom of Alzheimer's disease onset and we are applying decision tree models to predict that. To the best of our knowledge, there has not been any research, investigating Alzheimer's disease onset using a combination of neuropsychological tests and blood-based markers.

In the third chapter, we developed decision tree models to investigate plasma biomarkers of mild Alzheimer's Disease in females from three different ethnic groups. We tried to identify rules characterizing progression patterns in Hispanic, African American and Caucasian females. We used pathologic, inflammatory and cardiovascular markers measured in the plasma samples. The study tries to compare disease progression pattern in three different ethnic groups of females.

In Chapter four, we used the RuleFit algorithm to characterize homogeneous subgroups of subjects who participated in a pediatric study. We tried to find rules that identify subgroups with different risk patterns of parents who withdrew from The Environmental Determinants of Diabetes in Young (TEDDY) study in the first year. We analyzed the dataset from the survey which was filled out by parents at the inception of the study and identified eight predicting rules to predict withdrawal from the study.

In Chapter Five we tried to predict the progression pattern towards Alzheimer's disease, while doing dimensionality reduction using a network based system. We used the RuleFit algorithm to generate rules and the Latent Trait model for weighting the rules. Finally, we used the network based algorithm to select the most synergistic set of rules. In this study we tried to find fewer neuropsychological tests to administer, instead of a battery of 16 tests which take around four hours.

CHAPTER 2: IDENTIFYING COST-EFFECTIVE PREDICTIVE RULES FOR AMYLOID-β LEVEL BY INTEGRATING NEUROPSYCHOLOGICAL TESTS AND PLASMA-BASED MARKERS[1]

2.1 Introduction

Alzheimer's disease (AD) is a progressive, fatal neurodegenerative disorder, characterized by memory loss and other cognitive impairments. There is now a scientific consensus that the pathological events in AD initiate decades before clinical symptoms become apparent, and disease-modifying therapies will be most effective at the earliest stages of the disease. A major disease-modifying therapy which holds great promise in preventing AD is anti-amyloid preventative treatment [2], since abnormal amyloid-β deposition has been widely regarded as the initial event in a cascade of pathological processes, leading to synaptic dysfunction and neuronal death, and followed by the development of cognitive impairment and eventually dementia [3]. Despite the promises held by the developing anti-amyloid preventative treatments, the success of their clinical trials requires appropriately selected participants who are positive for Aβ pathology.

The identification of suitable individuals with elevated brain amyloid burden poses a great challenge in terms of feasibility and cost. To date, the advancement of molecular imaging tracers that bind to amyloid, such as the Pittsburgh Compound B (PiB), offers a non-invasive *in vivo* method to detect and quantify brain amyloid deposition [4,5]. However, this approach for pre-

[1] Portions of this Chapter were previously published in [1].

symptomatic detection is economically challenging for routine use given the current cost [6]. Similarly, the clinical use of other useful biomarkers such as $A\beta_{1-42}$ and phosphorylated tau in cerebral spinal fluid (CSF) is also limited, since lumbar puncture carries risks and is met with resistance in elderly subjects. Further it is unlikely to be used in primary health care centers to routinely screen large number of participants. Given the cost and limited availability of these brain amyloid measurement techniques, they are not reasonable first-line approaches for screening participants at risk of having elevated brain amyloid burden.

Recent studies have revealed the possibility of predicting elevated brain amyloid burden using more cost-effective measurements, such as neuropsychological tests and blood-based biomarkers. Some concurrent relationships between Aβ and cognition [7-9], metabolism decline [10] and brain atrophy [11], have been identified. A few studies have developed models to predict elevated Aβ level or Alzheimer's Disease, using either neuropsychological measures [12,13] or blood-based markers [14-17]. On the other hand, although neuropsychological measures and blood-based biomarkers have more practical applicability for routine use and are more cost effective, their predictive capabilities for detection of pre-symptomatic Alzheimer's disease are still limited [12-17]. For instance, by relying on neuropsychological measures alone, individuals with very high premorbid intellectual abilities experiencing incipient cognitive decline may go undetected, and false positives are possible in individuals with a low level of intellectual ability. It is also a well-known fact that the ceiling and floor effects limit the measurement capacity of many neuropsychological instruments [18-20]. Also, the set of blood-based biomarkers that have been reported as associated with AD are largely inconsistent in the literature [14-17], probably due to the inherent measurement uncertainty since these markers fluctuate over time [21,22]. Another possible reason is that uni-variate statistical methods were used for identifying these blood-based

biomarkers, falling short of recognizing the multivariate patterns that may be more robustly and reliably associated with AD pathology [14-17].

To date, we are aware of no prior work that has explicitly sought to identify these multivariate patterns which integrates both neuropsychological tests and blood-based markers, as existing research works focus on either neuropsychological tests or blood-based markers alone. As it is becoming increasingly apparent that uni-variate biomarkers are not sufficiently sensitive or specific for the diagnosis of complex, multifactorial disorders such as AD [23], it is more promising to consider applying multivariate data mining approaches to combine the neuropsychological measures and blood-based biomarkers and allow them to complement with each other, in order to identify biomarker signatures which are consistent with pre-clinical AD and specifically associated with amyloid pathology. Such an approach will be more practical for clinical use and be germane in designing large-scale prevention trials by enriching for those cases that are more likely to be amyloid positive by PET imaging. This would then require smaller numbers of individuals to be screened to populate anti-amyloid secondary prevention trials.

Therefore, our aim is to investigate the feasibility of extracting cost-effective, simple predictive rules of brain amyloid-β positivity for enriching the study population for clinical trials of anti-amyloid treatments, by integrating the neuropsychological tests and blood-based markers. We explore different strategies for building our prediction models, and compare their predictive performances. Moreover, rather than focusing on predictive regression models as in most of the relevant existing studies [12-17], we use the decision tree model since it can lead to simple decision rules that can be naturally translated into clinical settings for detecting amyloid positive cases. Furthermore, these rules will permit some individuals to be classified on the basis of only one, or

at most a few, measurements, whereas scores derived from regression-based prediction models, such as logistic regression or support vector machine, require that all covariates are available.

2.2 Materials and Methods

2.2.1 Subjects

All data used in the preparation of this article were obtained from the ADNI database (adni.loni.ucla.edu). ADNI is a naturalistic, longitudinal study of AD onset and progression being conducted at 57 sites in the United States and Canada. The ADNI was launched in 2003 by the National Institute of Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations. Further information about ADNI can be obtained from www.adni-info.org.

2.2.2 Analysis Dataset

Data used for the analyses presented here were accessed on May 11 2013 and comprise data from 50 normal adults aged 65 and older and 168 age-matched MCI (mild cognitive impairment) subjects for which blood proteomics data and Aβ status were available. Normal individuals were free of memory complaints or depression and had a Mini-Mental State Examination (MMSE) score of 28 to 30 and a Clinical Dementia Rating (CDR) score of 0. MCI individuals met Petersen criteria for single-domain or multi-domain amnestic MCI with MMSE scores of 24 to 27, CDR of 0.5, and an informant-verified memory complaint substantiated by abnormal education-adjusted scores on the Wechsler Memory Scale Revised—Logical Memory II. Other cognitive domains and everyday functioning were intact.

The variables included in this study are as follows. For neuropsychological measurements, we used the standard 11-item version of the ADAS - cog (including: word recall, commands, construction, naming, ideational praxis, orientation, word recognition, recall instructions, spoken language, word finding, comprehension) and 2 additional items (delayed word recall and number cancellation). We also included the individual total scores from both the 11-item and 13-item versions. For blood-based markers, we used the proteomics data set that was produced by the Biomarkers Consortium Project "Use of Targeted Multiplex Proteomic Strategies to Identify Plasma-Based Biomarkers in Alzheimer's Disease" [24]. We used 146 blood-based markers from the proteomic data downloaded from the ADNI web site. For measurements of amyloid burden, we used both PiB-PET imaging and CSF beta amyloid 1–42 ($A\beta_{1-42}$) level. The subjects were then dichotomized into either PiB positive (PiB retention summary measure > 1.5) or PiB negative (PiB retention summary measure ≤ 1.5), based on a threshold used in [25]. The CSF samples were acquired from these subjects by the ADNI Biomarker Core laboratory at the University of Pennsylvania Medical Center. The subjects were then dichotomized into either CSF $A\beta_{1-42}$ positive (CSF $A\beta_{1-42}$ level of ≤ 192 pg/mL) or CSF $A\beta_{1-42}$ negative (CSF $A\beta_{1-42}$ level of >192 pg/mL), based on a threshold used in [26]. Finally, a subject is classified as amyloid positive if the subject is positive either by PiB-PET or CSF $A\beta_{1-42}$ or both.

2.2.3 Statistics

Data for the 50 normal and 168 MCI were used for building the prediction models. As mentioned in the introduction, we explored different strategies for integrating the neuropsychological tests and blood-based markers, and compared the predictive performance of these integrated prediction models with the single-modality prediction models that are built on either neuropsychological tests or blood-based markers. Specifically, we: 1) evaluated the

predictive performance of the neuropsychological measurements; 2) evaluated the predictive performance of the blood-based biomarkers; 3) evaluated the predictive performance when the neuropsychological measurements and blood-based biomarkers were combined.

Therefore, we generated three decision tree models: model 1 (M1) built a decision tree that only uses the ADAS-cog; model 2 (M2) built a decision tree that only uses blood-based markers; model 3 (M3) built a decision tree that uses both ADAS-cog and blood-based markers. For creating each of the decision tree models, the conditional recursive partitioning technique [27] was used. This technique is a nonparametric methodology that creates a decision tree with respect to risk factors and their interactions that are most important in determining the outcome. Basically, it consists of three steps. The first step is tree building. A group of subjects (represented as a node on the tree) would split into child nodes if the testing statistic that measures the group differences between the two child nodes was significant for any feature beyond the 0.05 probability level. The significance level was adjusted for the number of multiple comparisons by the Bonferroni method. The cut-off point that determined the splitting of the node for a continuous variable was the point that maximized the test statistic with the P value less than 0.05. Each split resulted in the definition of two subgroups. The second step is termination of the tree building. There are multiple criteria that have been demonstrated to be effective in the termination of tree building. One approach that was adopted in our study was to terminate the tree building when there were a minimum pre-specified number of observations in each of the leaf nodes, i.e., 30 observations was used in our study. The third step was tree pruning that revised and reduced the size of the obtained tree after step 2. The main purpose of tree pruning was to achieve the optimal balance between the tree complexity (i.e., a tree with too many levels and leaf nodes will be cumbersome to use) and maintenance of prediction accuracy, (by deleting the internal nodes that do not substantially

improve accuracy). A simple but effective strategy recommended in the literature is to select the smallest tree whose model error falls within the one standard error rule [27], which was adopted in our study also. Decision tree analysis has been found valuable in many biomedical studies. For example, it was used for a cancer study to divide patients into homogenous groups based on the length of survival [28]. It has an advantage over the regression models in identifying prognostic factors because it relies on fewer modeling assumptions and has an established procedure that adapts to missing data through the use of surrogate measures. Also, because the method is designed to divide subjects into groups based on the heterogeneity of clinical outcome of interest, it defines groupings for outcome classification whereas regression models do not. Moreover, there is no need to explicitly include covariate interactions or transformations because of the recursive splitting structure of tree model construction. Analyses were performed using R, version 2.12 (http://www.r-project.org/), and the contributed libraries for our analyses, were "party", and "pROC".

2.2.4 Evaluation of Predictive Performances of Different Models

In attempt to evaluate the predictive performances of these models, we use both internal and pseudo-external validation. First, we randomly split the whole dataset into 2 subsets of two-thirds and one-third size. The first two-thirds dataset will be used for internal validation, and the one-third dataset will be used for pseudo-external validation.

For the internal validation, a 5-fold cross-validation procedure was used to randomly divide the two-third dataset into 5 equal sized parts. Then each part was left out as the internal testing dataset one at a time, and the remaining 4 parts were used as a training set to train the decision tree model, which was used to predict on the internal testing dataset. After all 5 iterations were finished, the predictions from all 5 test sets were pooled together to estimate the classification accuracy,

sensitivity, and specificity. This entire procedure on the two-thirds dataset was repeated 10 times, with different initial seeds, to yield robust estimates of the classification accuracy, sensitivity and specificity. Sensitivity refers to the ability to correctly classify the subjects who are amyloid positive; it measures the proportion of amyloid positives who are correctly identified as such. Specificity refers to the ability to correctly classify the subjects who are amyloid negative and measures the number of amyloid negatives who are correctly identified as such.

The external validation was conducted upon the one-third dataset once a decision tree model was built from the two-thirds dataset. This division resulted in similar numbers of control/ MCI and amyloid positive/negative cases in the train and test sets. The decision tree model was applied to the one-third dataset, and thereby, the classification accuracy, sensitivity, and specificity of the three models was estimated.

2.3 Results

2.3.1 Demographics

Characteristics of the 218 participants that are used in our study are summarized in Table 1 (so are the characteristics of the training and testing dataset that are generated by randomly splitting the 218 participants). Participants are well matched for age (P=0.4866, Kruskal-Wallis test) and education. There are more men than women (60.0%, 71%, for Normal and MCI, respectively), and the proportion of men is greater in the MCI group.

2.3.2 Estimation of the Three Models Based on the Training Dataset

The three models were built using two-thirds of the data for training. The estimated model of M1, using only the cognitive function data, is shown in Figure 1, together with the classification results of all 218 subjects (including both the training dataset and testing dataset). Here, M1

identified 3 variables from the ADAS-cog tests, which are Delayed Word Recall, Orientation, and the TOTALMOD (The modified ADAS-Cog 13-item scale includes all original ADAS-Cog items with the addition of a number cancellation task and a delayed free recall task, for a total of 85 points. [29]). M1 identified two highly enriched subgroups, Node 1 (majority is amyloid negative) and Node 5 (majority is amyloid positive). These two subgroups are characterized by two rules, M1_Rule1: TOTALMOD ≤ 7.33 -> amyloid negative, M1_Rule2: (TOTALMOD > 13.67 AND Delayed Word Recall >6 AND Orientation >0) -> amyloid positive. Note that, Node 4 also has a relatively homogenous subgroup where the majority is amyloid positive.

The model M2 (as shown in Figure 2) automatically identified 5 blood-based markers that were predictive of the amyloid-positivity out of the 146 blood-based markers. These 5 markers are APOE (Apolipoprotein E), PAP (Prostatic Acid Phosphatase), TTR (Transthyretin), MMP10 (Matrix Metalloproteinase-10) and MYOGLOBN (Myoglobin). M2 also identified two highly enriched subgroups, Node 1 (majority is amyloid negative) and a merger of Node 5 and Node 6 (majority is amyloid positive). These two subgroups are characterized by two rules, M2_Rule1: (APOE > 1.785 AND TTR > 2.569) -> amyloid negative, M2_Rule2: (APOE ≤1.785 AND PAP ≤ - 0.638 AND MMP10 > - 1.481) -> amyloid positive.

Model M3 (as shown in Figure 3) used all the ADAS-cog variables and the blood-based markers as potential predictors. It identified one ADAS-cog variable, the TOTALMOD, and three blood-based markers, the APOE (Apolipoprotein E), FSH (Follicle-Stimulating Hormone) and IGM (Immunoglobulin M), which were predictive of cases that were the amyloid pathology. M3 identified three homogenous subgroups, Node 1 (majority is amyloid negative), Node 4 (majority is amyloid positive) and Node 5 (majority is amyloid positive). These three subgroups are characterized by three rules, M3_Rule1: (TOTALMOD≤ 13.67 AND IGM>0.176) -> amyloid

negative, M3_Rule2: (TOTALMOD>13.67 AND FSH≤ 1.079 AND APOE≤ 1.69) -> amyloid positive, M3_Rule3: (TOTALMOD > 13.67 AND FSH > 1.079) -> Amyloid Positive.

2.3.3 Performance Evaluation of the Three Models Using the Internal Validation

The sensitivities, specificities and the AUC values of all the three models that are obtained by both the internal cross-validation and pseudo-external validation are shown in Table 2. From the internal validation results, it is evident that M3 has the maximum predictive capability since its AUC, sensitivity and specificity are all larger than other models. This indicates that an integration strategy tends to produce better prediction model than the simple strategy that is used by M1 and M2. Comparing M1 with M2, it can be seen that M1 has larger prediction capability than M2, in terms of the mean values of the AUC, sensitivity and specificity. Note that M1 is a decision tree model based on ADAS-cog, while M2 is a decision tree based on blood-based markers. However, the difference between M1 and M2 is not statistically significant considering the standard derivations of the performance values. All these observations hold on the pseudo-external validation results.

2.3.4 Application of the Three Models to the Pseudo-External Validation Dataset

All the three models were estimated using the two-third training dataset. The remaining one-third testing data was used to evaluate their predictive performances. The results are shown in Figure 4. It is evident that all the models were predictive of the testing data, showing that over-fitting is thereby not likely. It can also be seen that the prediction performance of M3 is superior to the other models, i.e., the 95% CI of the AUC of M3 doesn't overlap with the 95% CI of the AUC of M1 and M2, which demonstrated that the integration of both the neuropsychological tests with blood-based markers is effective. The sensitivity and specificity can also be extracted by analyzing Figure 4. For example, with the specificity being fixed at 0.75, the sensitivities of the

three models are approximately 0.61, 0.56, 0.77, respectively. Note that our cross-validation randomly split the whole dataset into two subsets without intentionally balancing the subsets for amyloid positivity, yet maintained a similar distribution of positive and negative cases (see the characteristics in Table 1).

2.4 Discussion

Our study identified effective prediction models for detecting subjects with elevated amyloid burden. All three models identified simple rules that are predictive of brain amyloid level. (These rules use cost-effective measurements, and also permit some individuals to be classified on the basis of only one, or at most a few, measurements. For example, in M2 (Figure 2), 43% of the 216 subjects have the ApoE plasma value > 1.785, and only 33% of this group are amyloid positive. In contrast, 57% of the 216 subjects have the ApoE plasma value ≤ 1.785, and 67% of this group are amyloid positive. This implies that by implementing the decision rule, ApoE ≤ 1.785 -> amyloid positive, it will enrich the amyloid positive population two fold. Therefore, as long as these rules can be clinically validated, we believe that these simple decision rules can be naturally translated into clinical settings, such as enrichment screening for Alzheimer's prevention trials of anti-amyloid treatments. The black area in each box plot at the end of terminal nodes shows the percentage of amyloid positives; the number of subjects in each terminal node is included at the top of the nodes. The trees were shown to be unchanging through cross-validations, since the algorithm used for tree building in our study is based on CTREE (Conditional Inference Tree) which results in unchanging trees in comparison to traditional decision trees. Since the CTREE conducts multiple test procedures to fulfill the stopping criteria, while the traditional decision trees apply information measures such as the Gini index.

Some comments on the M1 decision tree model that only uses ADAS-cog variables. It is evident from the tree (shown in Figure 1) that, the risk of being amyloid positive increases from the left nodes to the right nodes, while at the same time, the scores of the ADAS-cog items that are used by M1 also increase. This trend is consistent with the nature of ADAS-cog as higher scores of the ADAS-cog variables imply greater cognitive impairment. Also, the item, delayed word recall, has been found to be associated with amyloid pathology in recent studies that used cohorts of cognitively normal subjects which were different from ours [8,30]. On the other hand, the correlation between the ADAS-cog items, the "orientation", with amyloid pathology, requires further investigation.

The interpretation of the result in Figure 4, i.e., that M1 slightly outperforms M2, needs to be interpreted cautiously. First of all, the difference between the prediction performance of M1 and M2 is not statistically significant. Secondly all, M1 has lower specificity than M2 (Table 2). Thirdly, the prediction performance of M1 may be biased towards higher accuracy, since ADAS-cog has been used as a major criterion in ADNI for defining the diagnostic groups, and there is a significant correlation between the clinical diagnosis (e.g., defines MCI or NC) with amyloid positivity in our cohort (p-value < 0.05). So, it is biased to use the neuropsychological measurements to predict the outcome, because for some subjects the outcomes were defined by these neuropsychological measurements.

An integration of ADAS-cog with blood-based markers improved the prediction accuracy. From Figure 4, it is clear that the integrative model, M3, outperforms M1 (ADAS-cog only) and M2 (blood-based markers only). This implies that the ADAS-cog and blood-based markers provide supplementary predictive information.

The blood-based markers that are found predictive of the amyloid deposition are ApoE (Apolipoprotein E), PAP (Prostatic Acid Phosphatase), TTR (Transthyretin), MMP10 (Matrix Metalloproteinase-10), MYOGLOBN (Myoglobin), IGM (Immunoglobulin M) and FSH (Follicle-Stimulating Hormone). Most of these blood-based markers have been found to be associated with amyloid pathology or Alzheimer's in previous studies. For example, the association between the APOE level in plasma with brain amyloid burden has been identified in [17,31,32], where Thambisetty et al. [32] used the BLSA cohort that is a different cohort from ours. Our result is consistent with the studies in [17,31,32] that showed that the level of APOE in plasma, independent of genotype, is also a marker of risk. The PAP, as an amyloidogenic protein, has been found to form amyloid fibrils independent of those formed by Aβ [33]. The IGM has been reported in [34] to be protective in Amyloid formation since they may serve as a "buffering system" to keep free potential toxic endogenous peptides. This protective effect is also consistent with the results in M2 (shown in Figure 2), i.e., the subgroup of node 2, who have a lower level of TTR, is more likely to be amyloid positive than the subgroup with higher level of TTR (node 1) since TTR binds with amyloid-β (Aβ) and has been suggested to protect against Aβ deposition [35]. So, node 1, which has higher level of TTR, has the expected lower number of amyloid positive subjects. Also, the evidence shows that matrix metalloproteinases (including the MMP10) play an important role in the pathogenesis of AD and may be involved in the processing pathway of amyloid beta [36]. It has been shown that the chromogranin peptides are markers for human hippocampal pathways, and have a potential as neuronal markers for synaptic degeneration in Alzheimer's disease [37]. Evidence that supports the associations between FSH and MYOGLOBN with the amyloid pathology can be found [38,39].

We also compared our results with the 16 blood-based markers found to associated with brain amyloid burden in [17], and found only the association of APOE with amyloid is mutual. As that study employed a uni-variate linear regression model for identifying the blood-based markers on the ADNI cohort, the associations between the 16 blood-based markers with amyloid burden were reported to be quite weak [17]. These associations were not significant after adjusting for other covariates such as age. On the other hand, our method identified a different set of blood-based markers that were highly predictive of amyloid burden when used in combination as rules, indicating that our method has the advantage of identifying the blood-based markers that are correlated with amyloid pathology in a nonlinear and multivariate way.

Our study has limitations. First, we only used the ADAS-Cog as the representative neuropsychological measurement. Although ADAS-Cog is a standard tool in pivotal clinical trials, those intended to provide evidence for a drug marketing approval, to detect therapeutic efficacy in cognition, it is not considered sensitive enough to measure disease progression in early disease stages. As our ultimate goal is to identify an enrichment decision model for detecting amyloid positive cases from cognitively normal subjects, a better alternative may be the Neuropsychological Test Battery [40]. Also, since our study relied on one single cohort for estimating and validating the decision tree model, whether the enrichment decision model can be generally applied to other research studies remains to be confirmed. Moreover, we used both normal aging and MCI subjects for analysis, whether the decision tree models can extrapolate to general normally aging subjects needs to be further validated.

Our future work includes a large-scale study that will use all the potential clinical variables rather than ADAS-cog only. We will include a number of AD-related neuropsychological measurements, such as MMSE, Boston Naming Test, Verbal Learning Test, Clinical Dementia

Rating scale, to name a few. Three recent studies have revealed that some of these neuropsychological measurements are predictive of amyloid pathology [12,13,41]. Existing research has also revealed that some variables measuring the activities of daily living are also associated with the Alzheimer's [42]. We need to validate our enrichment decision model on other cohorts. Moreover, although the integration model, M3, has demonstrated its effectiveness, it is possible that a better integration strategy may exist, which can further boost the prediction accuracy. Overall, the results indicate that the neuropsychological measurements with blood-based markers can lead to an effective and accurate prediction model for detecting subjects with elevated amyloid burden. This prediction model has led to several simple rules, which have a great potential for use in clinical settings, such as enrichment screening for Alzheimer's prevention trials of anti-amyloid treatments.

Table 1 Demographics of the 218 participants

	NC (Total)	MCI (Total)	Training NC	Training MCI	Validation NC	Validation MCI
Number	50	168	37	108	13	60
AGE	83.82(6.3)	82.06(6.9)	85.32(5.2)	81.5(6.9)	79.5(7.6)	83.1(7.0)
Education	15.9(3.03)	16.2(2.56)	15.8(3.0)	16.2(2.6)	16.3(3.1)	16.2(2.6)
Gender (% male)	0.6	0.71	0.59	0.71	0.62	0.72
MMSE	29.28(0.93)	26.98(1.99)	29.16(0.93)	26.98(1.99)	29.62(0.87)	27(2.01)

Table 2 The prediction performance statistics of the three models

Model	Internal Validation			Pseudo-External Validation		
	Sens % (s.d. %)	Specs % (s.d. %)	AUC % (s.d. %)	Sens %	Specs %	AUC %
M1	0.71 (0.07)	0.65 (0.05)	0.79 (0.04)	0.71	0.67	0.77
M2	0.63 (0.04)	0.67 (0.06)	0.75 (0.02)	0.65	0.68	0.74
M3	0.71 (0.05)	0.68 (0.06)	0.82 (0.02)	0.69	0.72	0.85

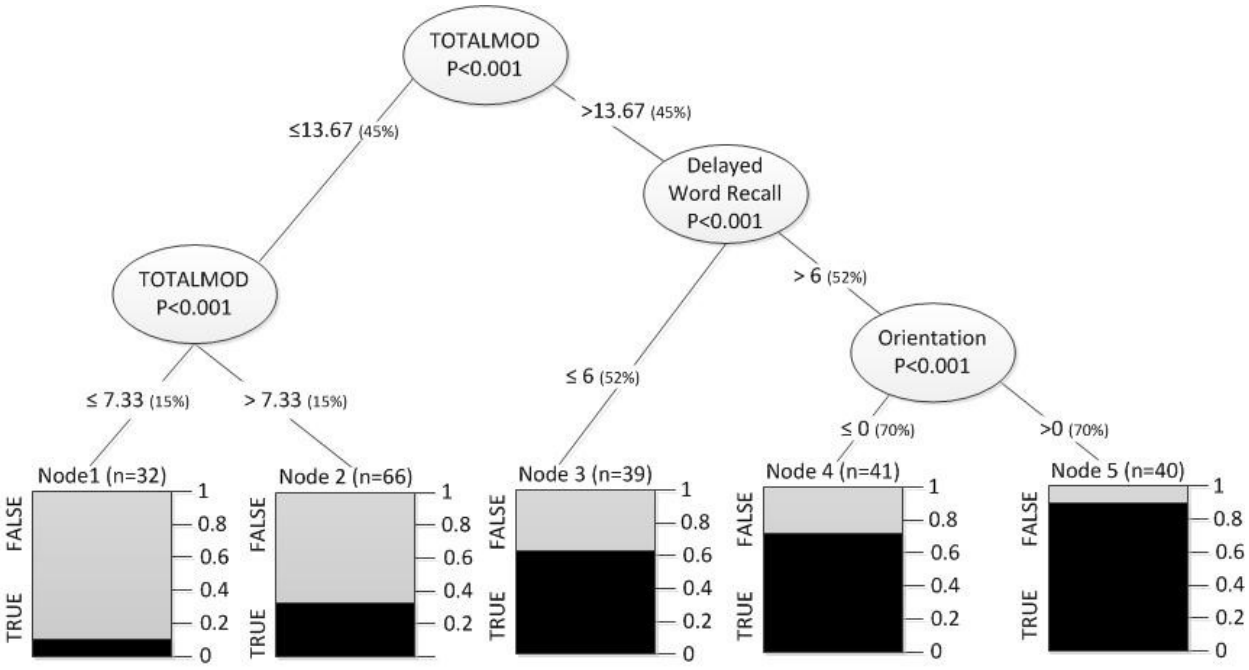


Figure 1 The decision tree model of ADAS-cog (M1)

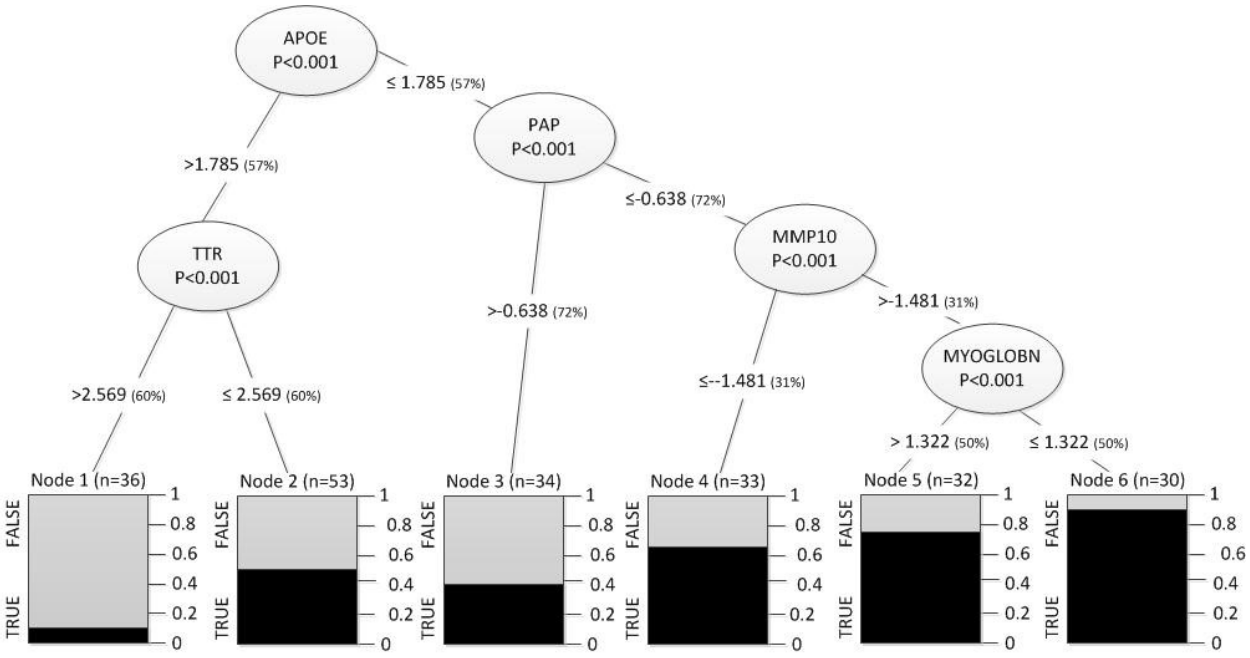


Figure 2 The decision tree model of blood-based markers (M2)

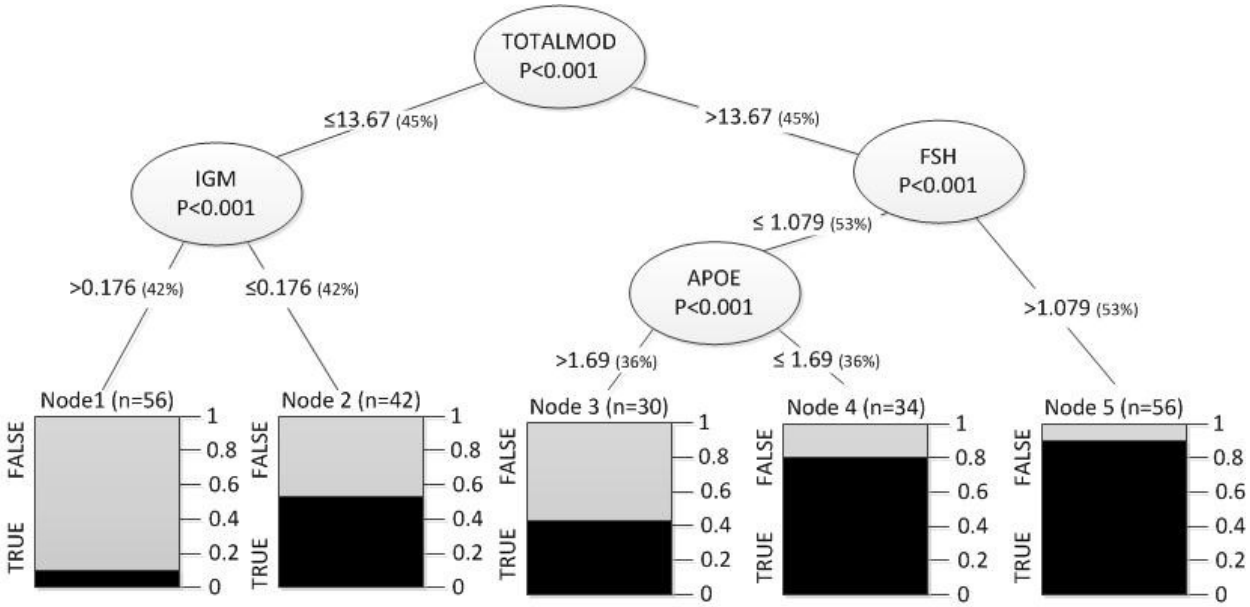


Figure 3 The decision tree model of both ADAS-cog and blood-based markers (M3)

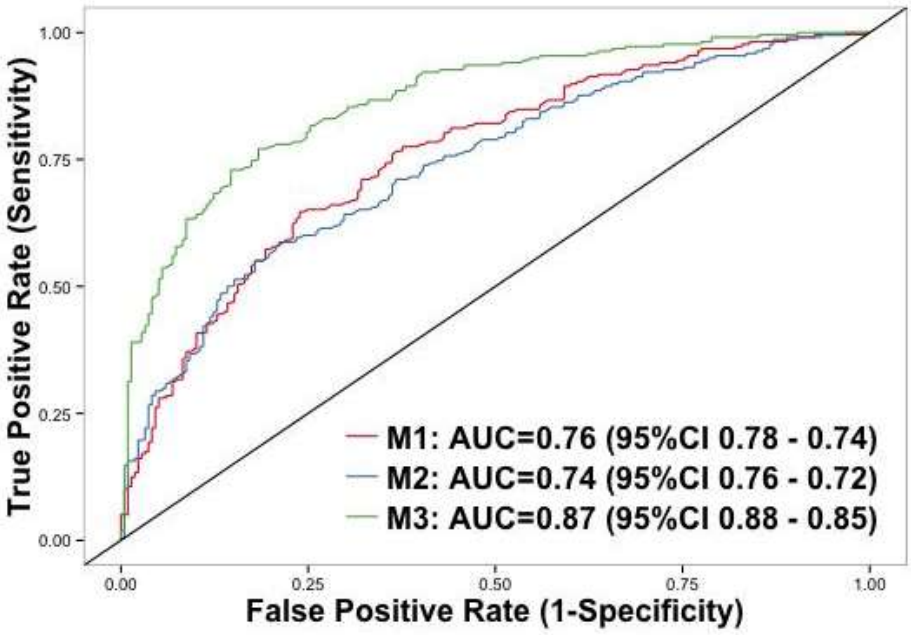


Figure 4 The receiver operator curves of the three models on the testing dataset

CHAPTER 3: THE DISPARITY OF DEMENTIA RELATED PLASMA BIOMARKERS AMONG ETHNIC MILD AD FEMALE PATIENTS

3.1 Introduction

The new criteria and guidelines for diagnosis of Alzheimer's Disease (AD) highlight the importance of identifying the need for biomarkers that are specific and sensitive to identify the disease in its early stages. In recent publications, patients' cerebral spinal fluid (CSF) has been used to identify specific biomarkers that predict AD progression [43]. For example, studies performed on CSF samples identified increased expression of both total tau and phosphorylated-tau specific to AD patients compared to the normal controls [44]. In addition, studies have linked low levels of amyloid beta 42 ($A\beta_{42}$) in CSF with AD progression, highlighting the specificity of $A\beta_{42}$ as a CSF biomarker [44]. Although studying relevant AD biomarkers provide promising diagnostic and therapeutic potential, CSF collection is an invasive and expensive method [44]. The utilization of biological samples such as plasma, which are routinely collected during patients' clinical visits, represents a more feasible option and would help facilitate early discovery of AD biomarkers [43,45].

Although the pathological biomarkers (amyloid beta and tau levels) are important early indicators of AD disease progression, supplemental cardiovascular and inflammatory biomarkers would allow for a more effective early diagnosis. Despite the association of cardiovascular risk factors with the increased risk of Alzheimer's Disease in the African American population [46], few studies have focused on the neuro-pathological association between cardiovascular risk factors

and AD in this population [46-48]. In addition, few studies have been performed entailing inflammatory markers as potential biomarkers in AD disease progression, despite laboratory research and genome wide association studies (GWAS) recognizing the impact of inflammation on AD pathology [49-51]. According to Alzheimer's Association reports, African Americans are twice as likely and Hispanics are about one and one-half times more likely to have Alzheimer's disease (AD) compared to Caucasians (Alzheimer Organization Association, 2012). Also, risk factors such as high blood pressure, diabetes [52] and gender [53,54] are associated with Alzheimer's disease and other dementias [55]. There are no known genetic factors explaining the increased prevalence in African American and Hispanic communities; however, the aforementioned conditions are more prevalent in these groups [56]. Most longitudinal and cross sectional studies lack analyses of the role minority patients play in Alzheimer's disease studies.

Our study aims to identify specific vascular, inflammatory and pathological plasma biomarkers that are significantly associated with the incidence of Alzheimer's disease in female patients. Moreover, we aimed to identify biomarkers that specifically link the incidence of mild AD to African Americans compared to Hispanic and Caucasian ethnic female patients. The study consisted of a 13-biomarker panel, which were measured in the plasma samples of African American, Hispanic and Caucasian female patients diagnosed as MCI/mild AD at the Byrd Alzheimer's Institute, USF, between the years 2006-2011. Further, we utilized computational analysis and applied correlation models to identify plasma biomarker levels that correlate with the early AD status present in each of ethnic groups. These findings provide the necessary data for designing longitudinal studies where large ethnic groups can be included and the biomarker selection can differentiate among ethnic groups and disease status. This may open the door to an individualized medication selection process that will ensure positive drug responses.

3.2 Materials and Methods

Plasma samples were attained from the USF, Alzheimer Disease Research Center (ADRC) data base, belonging to African American, Hispanic or Caucasian female patients that have visited the Byrd Alzheimer's Institute between years 2006-2011. For a diagnosis of MCI and mild dementia, both International Statistical Classification of Diseases, 10th Revision, and Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition criteria were met [57]. The study included 50 female minority patients of African American or Hispanic descent and 25 of Caucasian descent. Of the 75 female patients, 30 samples belong to age-matched female controls (n=10/group) and 45 samples belong to the female patient population diagnosed with Mild Alzheimer's disease (n=15/group). As a measure of neuropsychological performance, participants were analyzed based on the Florida Cognitive Activities Scale (FCAS). Given the results of the neurological examinations, diagnostic classification of the patients was determined by the Mini-Mental State Examination (MMSE) scores (21-30) and the Clinical Dementia Rating (CDR) values (0-2). A complete clinical history and neuropsychological testing were maintained from the institute's ADRC database. All patients have signed a waiver consent form prior to the clinical visit. Plasma from all control and AD subjects was prepared according to the clinical laboratory improvement amendments (CLIA) standards, which will make these samples amenable for clinical trial work. The research protocol for this study was approved by the University of South Florida IRB committee.

3.2.1 Decision Tree Analysis

We developed decision tree models on the data gathered for each ethnic group in order to identify homogeneous subgroups embedded in each ethnicity whose members have similar biomarker levels according to the following models: model #1 (M1) a decision tree on data of

African American females; model #2 (M2) a decision tree on Caucasian females and model #3 (M3) a tree on Hispanic females. The decision tree method is described in Chapter 2.

3.3 Results

We applied a decision tree learning algorithm, a *computational* modeling algorithm, to our data set in order to identify homogeneous subgroups with similar biomarker levels that links them to the risk of developing AD in each ethnic group. We present three models (Figure 5) for African-American, Caucasian and Hispanic women.

The decision tree model learned from the African American female cohort (M1) identified three biomarkers: $A\beta_{40}$, Plasminogen activator/inhibitor (PAI.1) and Eotaxin, which can characterize one homogeneous subgroup, Node 7 and one almost homogeneous subgroup Node 3. Node 3 with the majority of age-controls is based on the M1_Rule1: ($A\beta_{40} \leq 126.37$ pg/ml and PAI.1 > 19229 pg/ml) -> Age Control. Further, in this model, Nodes 6,7 identify two mostly to completely homogenous mild AD groups, which can be characterized following the M1_Rule2: ($A\beta_{40} > 126.27$ and Eotaxin ≤ 103.14) -> Mild AD, M1_Rule3: M1_Rule2: ($A\beta_{40} > 126.27$ and Eotaxin > 103.14) -> Mild AD (Fig. 5a).

Model #2 (M2) was built based on the data set from the Caucasian female cohort and identified three biomarkers: Cystatin C, Fibrinogen and Plasminogen activator/inhibitor (PAI.1), which were linked to three subgroups based on the following rules as: M2_Rule1 identifies Node 3 as a homogenous age-matched control group characterized as: (Cystatin C $\leq$ 1312.95 ng/ml and Fibrinogen > 2352.15 μg/ml) -> Age Control. M2_Rule2 identifies Node 6 as an almost homogeneous population of mild AD patient group (n=7/8 subjects) characterized as: (Cystatin C $\leq$ 1312.95 ng/ml and Fibrinogen $\leq$ 2352.15 μg/ml and PAI.1 $\leq$ 22105 pg/ml) -> Mild AD.

Meanwhile, M2_Rule3 identifies Node 7 as a homogenous mild AD group characterized as: Cystatin C > 1312.95 ng/ml -> Mild AD (Figure 5 b).

Model #3 (M3) is built on the Hispanic cohort dataset and identified $A\beta_{40}$ and Cystatin C as biomarkers linked to one homogeneous subgroup of mild ADs (Figure 5 c). This subgroup (Node 5) is characterized by M3_Rule1: ($A\beta_{40}$ > 88.5 pg/ml and Cystatin C > 1076.79 ng/ml) -> Mild AD. Interestingly, when the Cystatin C rule in this decision tree was applied to the other two ethnicities, the majority of Caucasians and African American individuals who satisfy this rule were also Alzheimer's disease patients. This suggests that Cystatin C is associated with mild AD and should be studied further as a potential early detector of Alzheimer's disease.

For each model, the sensitivity and specificity were estimated. Both sensitivity and specificity of M1 (93% and 70%, respectively) and M2 (87% and 90%, respectively) are satisfactory; however, M3 has a sensitivity of only 66.7% and was unable to predict a homogeneous mild AD subgroup albeit this model performed with 100% specificity. Table 3 provides a comparison between predictive powers of the three models after 5 fold cross validation. There's a big standard deviation in the predictive performance of the models, which could be due to the small number in the population the models were built on.

We also developed a decision tree on the whole 75 example dataset, and tried to compare the variables which showed up in each analysis. $A\beta_{40}$, Total.Tau and CystatinC were the biomarkers that appeared in the decision tree on the whole dataset. However, CystatinC appeared only in M2 and M3; $A\beta_{40}$ in M1 and M3 and PAI.2 in M1 and M2.

3.4 Discussion

In this study we analyzed pathological, vascular and inflammatory biomarkers in the plasma of African American, Caucasian and Hispanic female Mild AD patients and age-matched control individuals. The goal of this study was to determine effective potential biomarkers that are specific for each ethnicity and can be used for early detection of Alzheimer's disease (AD).

Cystatin C is highly relevant to the progression of Alzheimer's disease [58]. In fact, early research links cystatin C with amyloid beta found in the vascular walls and senile plaque cores in the brains of patients with Alzheimer's disease [59]. The evidence suggests that cystatin C could even protect the brain against amyloid-induced toxicity by binding to amyloid beta protein and inhibiting Aβ42 oligomer and fibril formation [59-61]. A study examining an AD patient cohort consisting of elderly men at the age 77 years found that the reduction of serum cystatin C levels was significantly associated with the increased risk of Alzheimer's disease [62].

In our study cystatin C levels were significantly increased in the mild AD female population compared to the age-matched controls. Further, we report that cystatin C was significantly higher in Hispanic mild AD female patients compared to the Hispanic age-matched control group. The computational analysis on the Hispanic cohort dataset identified cystatin C as a biomarker linked to a homogeneous mild AD subgroup suggesting that cystatin C can play an important role in the early detection of Alzheimer's disease.

Interestingly, when this rule (cystatin C>1076.79 ng/ml -> Mild AD) was applied to the Caucasian and African American cohorts, the majority of subjects that satisfied the rule were mild AD patients. These results are consistent with studies that associate cystatin C with Alzheimer's disease, and suggest cystatin C as a therapeutic biomarker for the early detection of Alzheimer's disease [63].

The decision tree model identified Aβ40 and Eotaxin levels that were significantly associated with Alzheimer's disease in the African American female cohort. Hereby, our findings suggest that plasma Eotaxin levels in the African American female population could be a risk marker for this ethnic and gender patient cohort, a relationship that has yet to be studied. Such efforts have been made and in fact, a study measuring chemokine and cytokine levels in a cohort consisting of 13 controls and 11 Alzheimer's disease patients reported significant elevation of serum Eotaxin levels in AD patients compared to the control group; however the study did not account for differences between ethnic groups in their patient cohort [64].

Despite the non-significant changes measured in the cardiovascular biomarkers levels in our patient cohort, analysis identified the combination of cystatin C ≤ 1312.95 ng/ml, Fibrinogen ≤ 2352.15 μg/ml and PAI.1 ≤ 22105 pg/ml associated with 87% Caucasian female mild AD patients. One limitation in applying these analyses is the low number of patients enrolled in this study; therefore, the models applied here are only suggestive of relevant biomarkers. Cardiovascular markers have been shown to be modified during Alzheimer's disease progression [47,48], for example, Jaeho Oh et al. [58] reported increased levels of plasma PAI-1 levels in MCI and AD subjects as compared to normal controls. The authors also reported that the PAI-1 levels were gradually increased as the dementia progressed [59]. In addition, Genome Wide Association Studies (GWAS) confirmed the APOE e4 allele as a risk factor and identified ABC7, a membrane transporter protein, to be a strong genetic risk factor of AD in African American cohort [65]. Both genes are involved in cholesterol transport, and given that cholesterol metabolism involvement in vascular conditions and its implication in Alzheimer's disease [66] it presents a potential marker for future studies.

When comparing the decision tree on the total population of the 75 subjects (Figure 6) and trees on each ethnic group, we can see that Cystatin C is an identifying biomarker for Caucasians and Hispanics, but not for African Americans. Fibrinogen was a biomarker for Caucasians, and not for the other ethnicities. And Aβ40 which was the main identifying biomarker for the whole dataset, does not seem to be an important one for Caucasian women.

One limitation in applying these analyses is the small sample size enrolled in this study; therefore, the models applied here are only suggestive of relevant biomarkers. To our knowledge, very few studies have analyzed the relation of ethnicity to AD in such an inclusive panel of biomarkers. Considering that clinical manifestations are often the result of complex extrinsic social factors, there is growing awareness that these factors may influence some of the disparities in clinical presentation and treatment [61,62]. We believe that finding biomarkers that link biological risk factors to cognitive function and determining biomarker disparity among ethnic groups can significantly advance research for the development of effective preventive therapeutic interventions in the treatment of Alzheimer's disease.

Table 3 Prediction power of the three models after 5-fold cross validation

Model	Sens (s.d.)	Specs (s.d.)	Acc (s.d.)
African Americans	0.61 (0.16)	0.7 (0.14)	0.6 (0.14)
Caucasians	0.66 (0.23)	0.57 (0.23)	0.6 (0.25)
Hispanics	0.67 (0.23)	0.6 (0.14)	0.64 (0.17)
Whole data	0.96 (0.06)	0.54 (0.22)	0.8 (0.1)

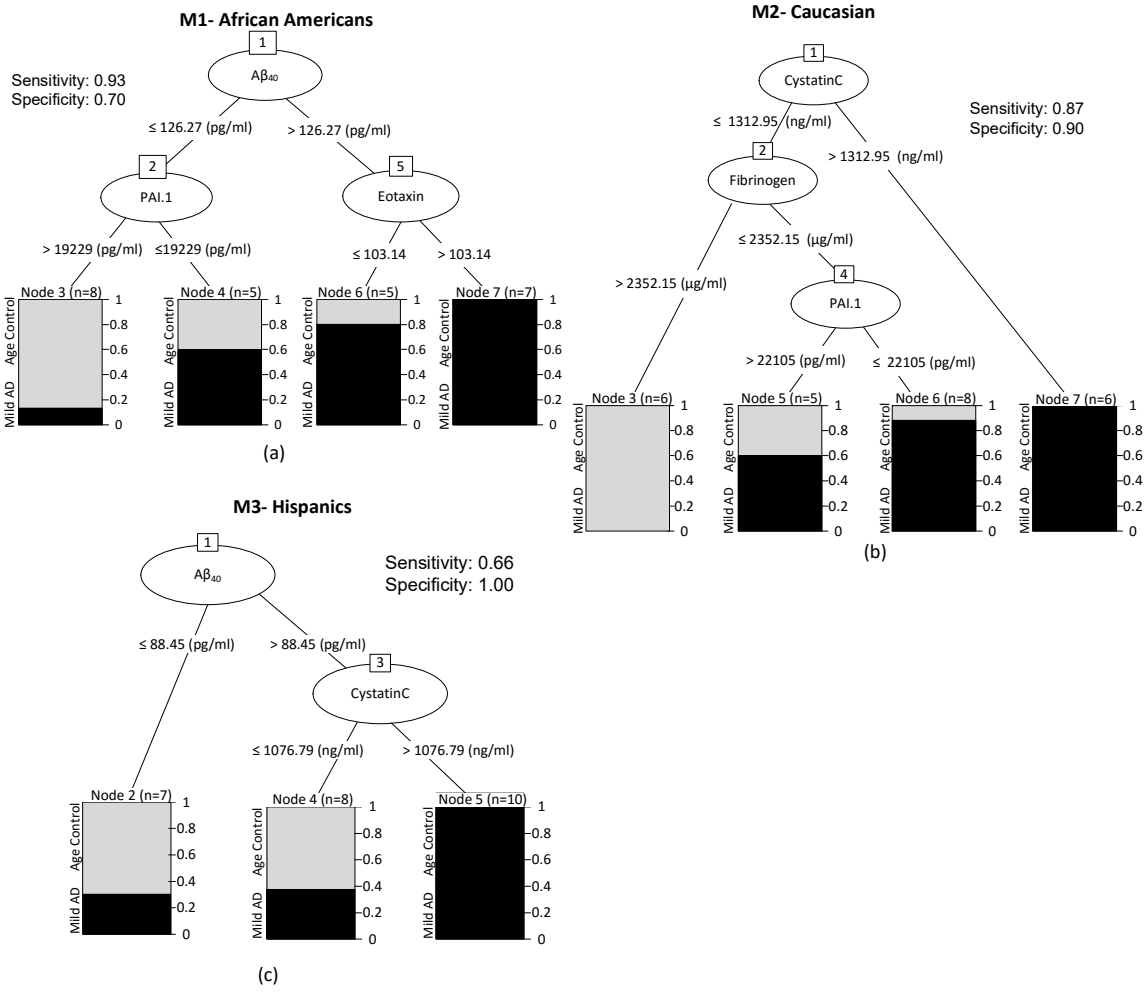


Figure 5 Decision tree models for each ethnic group.

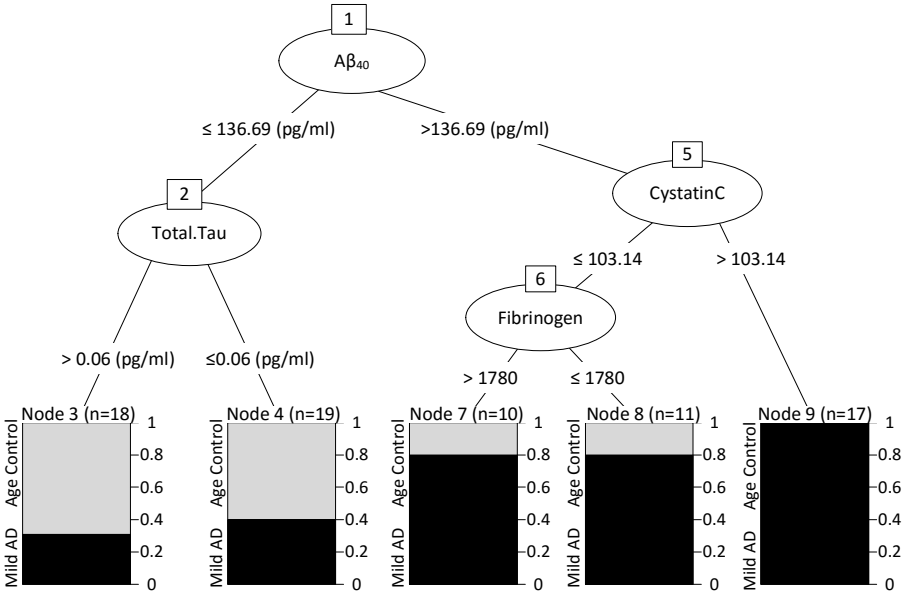


Figure 6 Decision tree on the whole 75 subjects of the study.

CHAPTER 4: A COMPARISON OF RULE-BASED ANALYSIS AND REGRESSION METHODS IN UNDERSTANDING THE RISK FACTORS FOR STUDY WITHDRAWAL IN A PEDIATRIC STUDY

4.1 Introduction

Understanding the factors associated with the risk of individuals withdrawing from a study is an important first step towards identifying the eventual health needs of different individuals within a population [67]. This lays the foundation to develop and deliver appropriate resources to the right targets, called "tailored health interventions". Evidence suggests that individuals respond better to tailored care than standardized care that is designated for the average population [68-71]. Therefore, health professionals need to identify the subgroups of individuals characterized by different patterns of risk factors. However, rather than identifying subgroups, traditional intervention studies often focus on the outcome of interest for the population as a whole [67,72,73]. One commonly adopted approach is to use logistic regression to identify factors associated with study withdrawal [74-76]. However, this approach only models the average effects of the risk factors. Consequently, it is likely that the interventions developed from regression models will be geared toward the average member of the population, with less consideration of the special needs of different subgroups [77].

The aim of our study was to illustrate the use of the rule-based analysis [78-80] as an exploratory technique in an epidemiologic context. The rule-based analysis [78-80] is particularly useful for identifying the subgroups embedded in a dataset - whose members share similar risk

patterns - that influence the outcome of interest. A rule describes the range of values for one or more risk factors that are associated with either an increase or decrease in risk for withdrawal in a subset of individuals. Thus, rules provide a way to define the risk pattern of subsets of individuals where each rule may indicate a specific unmet health need or warning signal for study withdrawal. By identifying the rules from observational studies, a comprehensive set of risk-predictive rules can be considered as a set of sensors, providing us personalized risk estimation by looking into the risk patterns for each individual.

Specifically, we used a recently developed rule-discovery algorithm for the rule-based analysis, the RuleFit method [80], which is one example from a huge array of rule-based methods that are promising for epidemiologic research. The RuleFit method has an advantage over logistic regression because it relies on a nonparametric model with fewer modeling assumptions, random forests [79]. RuleFit creates rules from a random forest which is capable of identifying the risk predictive rules. Also, the rule-based analysis permits an individual's risk to be predicted on the basis of only one, or at most a few, risk factors, whereas scores derived from regression models require that all covariates be available.

We demonstrate rule-based analysis using data from a large multinational epidemiological natural history study of type 1 diabetes mellitus (T1DM), the Environmental Determinants of Diabetes in the Young (TEDDY) study [81]. Specifically, we used the rule-based analysis for predicting study withdrawal during the first year of the TEDDY study, by effectively integrating the psychosocial, demographic, and behavioral risk factors collected at study inception. We compare the rule-based analysis with a previous analysis that was conducted on the same data [76]. The previous analysis used traditional logistic regression methods to identify features, called factors, collected at study inception that were strongly associated with study withdrawal during

the first year of TEDDY [76]. However, the way these factors interact with each other and the way these interactions might define subgroups in the study population with different risk levels remains unknown. Therefore, we tested the hypothesis that the rule-based analysis can identify risk-predictive rules useful for stratifying the study population into different subgroups with different risk levels for study withdrawal in the first year of TEDDY. The previous analysis [76] provided us an opportunity to critically evaluate the potential added value of a rule-based analysis over and above that provided by traditional logistic regression methods. Also, we considered how the rule-based method could lead to more informed intervention strategies or prioritization of the intervention allocation to the study participants. An example of intervention allocation could be assigning a nurse to the participants who are at higher risk of withdrawal from the study to follow up for each appointment. By conducting this comparison, we also hoped to identify some practical guidelines for when should we use rule-based methods and when a regression model would be more preferable, enriching the analytic toolbox of today's epidemiologists to address emerging data challenges.

4.2 Materials and Methods

4.2.1 The TEDDY Study

TEDDY is a natural history study that seeks to identify the environmental triggers of autoimmunity and T1DM onset in genetically at-risk children identified at three centers in the United States (Colorado, Washington, and Georgia/Florida) and three centers in Europe (Finland, Germany, and Sweden). Infants from the general population with no immediate family history of T1DM, as well as infants who have a first degree relative with T1DM, are screened for genetic risk at birth using human leukocyte antigen genotyping. Parents with infants at increased genetic risk for T1DM are invited to participate in TEDDY. Parents are fully informed of the child's

increased genetic risk and the protocol requirements of the TEDDY study, including the requirement that eligible infants must join TEDDY before the infant is 4.5 months of age. The TEDDY protocol is demanding with study visits for blood draws and other data and sample collection scheduled every three months during the first four years of the child's life and biannually thereafter. Parents are also asked to keep detailed records of the child's diet, illnesses, life stresses and other environmental exposures. TEDDY obtains written consent from the parents shortly after child's birth for obtaining genetic and other samples from the infant and also parents. Details on study design and methods have been previously published [81]. The study methods were carried out in accordance with the approved guidelines by local Institutional Review or Ethics Boards and monitored by an External Evaluation Committee formed by the National Institutes of Health.

4.2.2 Study Sample

This analysis focused on two groups of families from the general population used in the previous logistic regression study [76] : 2,994 families who had been active in TEDDY for ≥ 1 year and 763 families who withdrew from TEDDY during the first year. Both the prior and current analyses were limited to general population families because study withdrawal among the first degree relatives population was rare.

4.2.3 Study Variables

Study variables were selected from data collected on the screening form at the time of the child's birth and from interview and questionnaire data collected at the baby's first TEDDY visit. These variables included: demographic characteristics---TEDDY country (Finland, Germany, Sweden, United States); mother's age (in years); child's gender; maternal health during pregnancy---number of illnesses, gestational diabetes or type 2 diabetes (yes/no); mother's lifestyle behaviors during pregnancy---smoked at any time during pregnancy (yes/no), alcohol consumption (no

alcohol, 1-2 times per month, ≥ 3 times per month during each trimester), employment status (worked during all 3 trimesters/did not work at all or reduced work hours); baby's health status---birth complications (yes/no), health problems since birth (yes/no), hospitalizations after birth (yes/no); number of stressful life events during and after pregnancy; mother's emotional status including worry and sadness during pregnancy (rated on 5 point scales), anxiety about the child's risk of developing diabetes measured by a six-item scale adapted from the State component of the State-Trait Anxiety Inventory [68-70]; the accuracy of the mother's perception of the child's risk for developing diabetes (accurate: indicating the child's T1DM risk was higher or much higher than other children's T1DM risk; inaccurate: indicating the child's T1DM risk was the same, somewhat lower or much lower than other children's T1DM risk); and whether the child's father completed the initial study questionnaire (yes/no). The study variables are listed in Table 4.

4.2.4 Previous Logistic Regression Results

Multiple logistic regression was used to identify significant predictors of early withdrawal from TEDDY. Variables were entered in blocks in the following order: demographic variables (country of residence, child's gender, mother's age); pregnancy/birth variables (maternal diabetes, illness in mother or child, birth complications, maternal smoking; maternal drinking; maternal employment outside the home, maternal worry or sadness during pregnancy, number of stressful life events occurring during pregnancy or after the child's birth); father's participation in TEDDY defined by father's completion of a brief questionnaire; and mother's reactions to the baby's increased T1DM risk (anxiety and accuracy of mother's perception of the child's T1DM risk). Nine percent of the study sample (N=326) had missing data on one or more variables. The withdrawal rate for participants with complete data (19%) was substantially lower than the withdrawal rate among those with some missing data (35%). Consequently, the analysis was first

completed for those with no missing data and then rerun for the full sample using multiple imputation to generate appropriate parameter estimates for missing data using the Proc MI and Proc MIANALYZE procedures available from SAS 9.1 [71]. Table 5 provides the results of the final logistic regression model for the sample of 3,431 TEDDY participants with no missing data. The model was highly significant (Chi-Square = 264.87 (12), p <.0001) and accurately placed 81.6% of the sample into their respective group (Actives versus Withdrawals). The data in Table 5 also provides the final logistic regression model for the total sample, with multiple imputation methods used to replace missing data. Because the early withdrawal rate was higher among participants with missing data, we added a variable to the imputed model, >1 missing data point (yes/no). The presence of >1 missing data points predicted early drop-out over and above all other variables in the model. The descriptive information for each of the significant predictors is provided in Table 6.

4.2.5 Statistical Methods

Basic Idea of the RuleFit Method. We used RuleFit [80] to discover the hidden rules that may be predictive of the risk of early withdrawal in subsets of TEDDY individuals. A rule consists of several interacting risk factors and their ranges. We are interested in the rules by which the subjects can be stratified by distinct risk levels. For example, a rule consisting of State Anxiety Inventory Score > 45 and Dad Participation = NO would be useful if the subjects who can be characterized by this rule have a higher risk of early withdrawal. RuleFit is a computational algorithm that can scale up for high-dimensional applications (e.g., with a large number of variables) for rule discovery, and is capable of exhaustively searching for potential rules on a large number of candidate risk factors. It has two phases, the “rule generation phase” and “rule pruning phase.”

4.2.5.1 Rule Generation

At this stage, random forests [79] is used to exhaustively search for candidate rules over the potential risk factors. Random forests is a high-dimensional rule discovery approach that extends traditional decision tree models [12]. Specifically, a random forest generates a number of trees, with each tree being generated from a subpopulation generated by bootstrapping the original dataset. Since we can extract rules from trees in a forest and each rule could characterize a subpopulation, the random forest is actually a collection of rules that are trying to characterize the whole dataset.

4.2.5.2 Rule Pruning

As a heuristic search approach, random forests may produce a large number of rules that can be redundant or irrelevant to predicting early withdrawal due to overfitting. To address this, the sparse regression model [82,83] can be applied to select a minimum set of risk-predictive rules, by using all the potential rules as predictors and the withdrawal status as the outcome. The sparse regression model is a high-dimensional variable selection model that can be applied on a large number of variables, and has been widely used in bioinformatics and systems biology [60,84]. Unlike other rule pruning methods, which remove antecedents with little predictive power, LASSO ejects the whole redundant rule out of the tree. That is the path from the root to the leaf, which makes up the rule.

In what follows, we illustrate the details of how the RuleFit method uses the three models, a decision tree, random forests, and sparse linear regression models, in the rule generation stage and the rule pruning stage:

4.2.5.3 Stage 1 of RuleFit - Rule Generation

Rule generation is computationally challenging, since the number of potential rules grows exponentially in relationship to the number of risk factors, e.g., even for 100 variables, let the maximum number of variables in a rule be 3, and 4 possible cutoff values for each variable, then the potential number of rules is $\sim 10^7$. Given such a large number of potential rules, an intelligent rule generator is needed to narrow down the search by effectively detecting high-quality risk-predictive rules. Decision trees with rule generation from them provide such an intelligent rule generator. We can extract rules from a decision tree by which we can segment the population into different subgroups. For example, we used a decision tree model for analyzing the TEDDY dataset to divide the population into subgroups based on the percentage of study withdrawals in each subgroup. The decision tree model is a nonparametric method that automatically explores the given risk markers for a tree that has high accuracy in predicting study withdrawal. In our analysis, we built a decision tree on all the data as shown in Figure 7, three subgroups with distinct risk levels were identified and could be characterized by rules defined by maternal age, smoking status, number of missing data, and a geographical indicator for Finland. For example, the leftmost node characterizes a subgroup of subjects, in which all of them have Maternal age < 27.5 and Finland = NO. The risk of study withdrawal in this subgroup is 0.38. This analysis demonstrated that the decision tree model is a powerful tool for detecting the subgroups that can be characterized by rules. Note that, the cut-off value of each marker used in Figure 7 was automatically determined by the Recursive Partitioning Algorithm (RPA) [12].

One limitation of the decision tree is that only exclusive rules can be identified. For instance, the decision tree in Fig.1 implies that each participant can only be characterized by one single rule, which doesn't consider the possibility that a participant may have multiple risk patterns

characterized by different factors or different interactions between factors. As a remedy, random forests [79] generates a number of trees: in each iteration, we build a decision tree on a bootstrapped sample of the training set randomly choosing a test feature from a specified number of top ranked features or "risk factors", and this process iterates until the pre-specified number of trees are created [79].

To understand a random forest, it is worth mentioning that the essence of this iterative procedure is to generate a large number of substantially different trees, since the more similar the trees are, the less advantage building multiple trees has. In order to achieve this goal, randomization methods are used, which is the reason for the name "random forests". Each tree is built on a bootstrapped training set using a subset of risk factors (features), the heterogeneity of the subjects is well addressed in the random forest model, increasing the likelihood of detecting meaningful risk-predictive rules for different subgroups [80]. As each tree can be decomposed into a number of rules, e.g., in Figure 7, we could extract at least five rules where each rule corresponds to a path to a leaf node in the tree, with random forests we can collect many rules.

4.2.5.4 Stage 2 of RuleFit - Rule Pruning

While in most machine learning methods, rule pruning is about deleting antecedents from rules, in RuleFit the whole rule is ejected from the tree. Here, rule pruning is essentially a procedure of selecting a subset of rules out of a pool of q candidate rules, denoted as $R = [R_1, R_2, \dots, R_q]$, which are predictive of the output variable Y. This problem is particularly challenging in high-dimensional settings where we have a large number of generated rules so that q is large. One solution to select the most critical rules is to adopt the *Least Absolute Shrinkage Selection Operator* (LASSO) [82], which is a sparse linear regression model that is capable of identifying a subset of

relevant variables out of a huge list of candidate variables. Specifically, the formulation of LASSO is

$$\min_{\beta} \|\mathrm{Y} - \mathrm{R}\beta\|_2^2 + \lambda\|\beta\|_1. \qquad (4\text{-}1)$$

Here, R is a binary variable standing for each rule, 1 if the rule is fulfilled and 0 otherwise and the square error term, $\|\mathrm{Y} - \mathrm{R}\beta\|_2^2$ is used to measure the model fit. The L1-norm penalty term $\|\boldsymbol{\beta}\|_1$, is defined as the sum of the absolute values of all elements. The user-specified penalty parameter, λ, aims to achieve an optimal balance between the model fitness and model complexity – a larger λ will result in a sparser estimate for β. It has been shown that LASSO is efficient on variable selection both from theoretical research [82] and empirical studies [60,83-85]. Efficient algorithms have been developed to solve the optimization problem, such as the shooting algorithm [82], proximal gradient algorithms [83], etc. Through LASSO, we expect that the rules with critical risk factor patterns will be identified with controlled redundancy. In our study, since the output variable Y, i.e., the withdrawal status, is a binary variable, sparse logistic regression [83] is a better choice than linear regression, which can be readily applied in the R package of RuleFit [80].

In summary, RuleFit is computationally efficient since efficient algorithms have been developed for both Random Forests and sparse linear regression models. RuleFit has an automated cross-validation procedure for tuning its parameters, such as the number of trees, the size of the trees and the penalty parameter λ in LASSO, which can be used to obtain a set of high-quality rules. More details about RuleFit can be found in [80]. Figure 8 also provides a schematic description of the Rulefit algorithm.

4.3 Results

4.3.1 Identified Risk-Predictive Rules

Table 7 provides the risk-predictive rules identified by the RuleFit algorithm for the Active and Withdrawn families used in the previous logistic regression analysis [76]. We generated 2000 rules with average number of 4 terminal nodes, which resulted in generating 300~350 trees and picked the 8 top rules selected by the algorithm. Changing the parameter max-rules (the total number of rules generated) and also different seeds resulted in a different rule set; however, the top rules stayed the same. The risk factors identified in the risk-predictive rules are the same as those identified in the previous logistic regression analysis: demographic factors including maternal age and country, maternal lifestyle factors during pregnancy including as smoking, drinking, and working outside the home, psychosocial factors including the mother's perception of the child's risk and her anxiety about the child's risk, dad participation, and the number of missing data points. In addition, the interaction between the state anxiety inventory score with the risk perception accuracy found in the previous study, which was further validated in the rule-based analysis (see Table 5 and Table 6). However, the rule-based analysis was more powerful at detecting the interactions between the risk factors, by including two or more variables and their ranges in a single rule. In addition, the rule-based approach identified the number of negative life events as a risk factor, a variable that was not significant in the prior logistic regression analysis. And the rule-based approach found no significant role for child gender, which had a weak effect in the prior analysis (see Table 5). Note that the rules shown in Table 7 were identified by LASSO from 2000 candidate rules generated by random forests.

4.3.2 Investigation of the Risk Levels of Withdrawal When Matching Risk Patterns

We next investigated the risk level of matching each of the rules by computing the study withdrawal rate for each subgroup that matched a rule. Figure 9 illustrates the withdrawal rates of each of the eight identified rules as well as the overall withdrawal rate of the whole study population. The number of subjects in each subgroup is also shown in Figure 9. It is clear that matching any of the first four rules will boost the risk of early withdrawal dramatically, while endorsing any of the latter four rules will help decrease the risk significantly. Approximately 10 percent of the study population matched no rules and their withdrawal rates were relatively high, suggesting that there may be other important subgroups that were not detectable with the available measures.

4.3.3 Investigation of the Redundancy of the Rules

One important technical issue in rule-based analysis is the control of redundancy of rules. Two rules are redundant if a participant matches one rule and this participant will match the other rule. Obviously, it is less desirable to have two rules that largely overlap with each other. We investigated the redundancy of the 8 rules and present the results in Figure 10. Figure 10 can be read in this way: the pie graph on row i (corresponds to rule i) and column j (corresponds to rule j) records the proportion of the participants matching rule i who also match rule j. It can be seen that, there is some overlap between some rules, such as rule 1 and rule 4, rule 5 and rule 7. The reason for a correlation between two rules may be that both rules share some common risk factors, e.g., both rule 1 and rule 4 involve maternal age < 27.5 in their definitions.

4.4 Discussion

In this dissertation, rule-based analysis [80] has been proposed to enrich the toolbox of epidemiological intervention studies that have been relying on regression models. We used data from the TEDDY study and demonstrated that the rule-based analysis can effectively identify risk-predictive rules from the psychosocial, demographic, and behavioral risk factors. The 8 identified rules by RuleFit are found predictive of early withdrawal during the first year of the TEDDY study. The 8 rules involve different sets of risk factors, highlighting the different nature of the withdrawal risk for each of these subgroups. Note that these 8 rules are not exclusive, giving the flexibility that an individual can show multiple risk patterns simultaneously.

We also compared the rule-based analysis with the previous analysis that was conducted on the same data [76]. We found that both methods detected almost the same set of risk factors, providing validation of our rule-based analysis. Note that the previous analysis only identified the average effects of these risk factors across the whole population, without considering how these risk factors interact with each other in determining the risk of early withdrawal. The rule-based analysis was superior at detecting interactions between the risk factors in each rule.

As each rule characterizes a distinct risk pattern that consists of different risk factors, a further investigation of the particular characteristics of each rule may help identify the special health needs of the subgroup whose members match this rule, leading to tailored interventions. For example, as revealed in Rule 3, for mothers who are highly anxious about their child's T1D risk with a state anxiety inventory score > 45, the lack of participation of the father increases the risk of study withdrawal. In an effort to tailor an intervention to this specific subgroup, a study nurse might be assigned to the family having this risk pattern to enhance the psychological support for the mother and encourage the participation of the father. On the other hand, the rules are also

helpful for developing general-purpose interventions. For instance, as smoking during pregnancy was selected in multiple rules, investigations may be conducted to understand why this behavior is related to the risk of study withdrawal. If smoking during pregnancy was found to be an indicator of less health-conscious attitudes, a tailored intervention might be developed for mothers who smoked during pregnancy to increase their health consciousness in an effort to reduce their risk of study withdrawal. As tailored interventions are developed and deployed, it is also important to evaluate the efficacy of these interventions for the subgroups separately, in order to identify the best intervention strategy for each subgroup.

The rule-based analysis also identified negative life events as a risk factor for early withdrawal, which was not detected by the logistic regression model used in the previous study [76]. Previous studies have linked negative life events with immune system functioning [86,87] and the onset of T1DM [88,89]. While the mechanism underlying the linkage between the negative life events and study withdrawal remains unknown, it is reasonable to expect that mothers experiencing numerous negative life stresses may not have the personal resources to remain in the study. Certainly tailoring an intervention to this subgroup of individuals seems warranted.

The rule-based method has a number of advantages when handling complex datasets. It can be used with a mix of nominal, ordinal, integer or continuous variables and it can combine a mixture of variables –-demographic, biological, psychological---without interpretation difficulty. Also, as rules are scale independent, data does not need to be standardized. Finally, the rules will permit some individuals to be classified on the basis of only one, or at most a few, risk factors, whereas risk scores derived from regression models require that all the risk factors are available.

There are limitations of the rule-based approach for epidemiologic studies. First, it is not suitable for studying the overall impact of a single independent variable on the outcome variable.

This is because a single independent variable may play a role in multiple rules, which results in difficulty investigating its overall effect on the whole population. Also, domain insight is very important in the identification of the rules using RuleFit. Due to the automatic nature of the rule-based approach, it is tempting to simply enter all possible candidate variables into the program without justification of which independent variables should be considered. It has been recommended in the literature [90] that prior knowledge regarding the relationship between the independent and dependent variables should be incorporated with the rule-based models. One of the reasons the rule-based approach yielded remarkably similar findings to the logistic regression approach in terms of identifying risk factors per se, is that considerable thought was put into variable selection and measurement by the TEDDY group. Rule-based models should not be used for blind exploration of large data sets and should benefit from careful *a priori* variable selection.

In general, through the study of the rule-based method in the TEDDY cohort and comparison with a previous study that used logistic regression methods, we could draw the following practical guidance for how to integrate rule-based analysis methods into the existing epidemiological toolbox. If there is a strong hypothesis that multiple subgroups may exist in the dataset, the rule-based method could be a very useful approach. On the other hand, subgroups may vary from dataset to dataset, and the rules (and the risk factors involved in these subgroups) identified by the rule-based method may vary from dataset to dataset as well. It is important to understand that the rule-based method is a customized method that is tailored for analyzing an individual dataset, so whether or not the results identified from one dataset could be generalized to another dataset depends on the subgroup structure of the new dataset. While flexibility of an analytic method usually comes with the risk of overfitting, a customized method also needs customized expertise or solid domain knowledge of the dataset. Finally, rule-based methods can

be considered as opportunistic methods that aim to discover sub-groups, but the results identified by rule-based methods are not necessary exclusive. For example, it is possible that there are more rules besides the eight rules identified from the TEDDY cohort by the RuleFit.

In summary, we believe that the rule-based approach will be useful in many epidemiologic studies, particularly with heterogeneous populations consisting of subgroups of individuals. The distinct risk factors that define each subgroup could also reflect a different mechanism of withdrawing from the study, leading to development of different intervention strategies. Besides the utility in designing tailored interventions, it can also help with the prioritization of the intervention targets, e.g., we could choose to eliminate a particularly high-risk subgroup at the beginning of a clinical study. Note that the RuleFit algorithm introduced here is one example from a huge array of the rule-based methods that are promising for epidemiologic research in general. How to properly adopt them for addressing the analytic challenges in epidemiologic studies will be an important future research topic.

Table 4 Study variables of the model

Country	United States
	Finland
	Germany
	Sweden
Child's Gender	
Maternal Age	
Maternal Health During Pregnancy	Number of illnesses
	Gestational Diabetes/ Type2 Diabetes
Maternal Lifestyle during pregnancy	Smoking (Yes/No)
	Alcohol Consumption
	Employment during pregnancy
	Baby's health status/ birth complications (Yes/No)
	Health problems since birth (Yes/No)
	Mother's Emotional status during pregnancy including sadness or worry (Yes/No)
Mother's Perception of the Child's risk of Developing Type I Diabetes	
State Anxiety Inventory Score	
Dad participation	
Missing Data points	

Table 5 Logistic regression results [76]

		Sample with No Missing Data (N=3431)						Sample with missing data imputed (N=3757)		
Predictor variable		Estimate	SE	P-value	OR	95% Confidence Interval		β	SE	P-value
Intercept		1.126	0.424	0.008				0.982	0.400	0.014
Country	United States	ref						ref		
	Finland	-0.420	0.130	0.001	0.657	0.509	0.848	-0.431	0.123	0.0004
	Germany	0.278	0.222	0.211	1.321	0.854	2.042	0.154	0.218	0.481
	Sweden	-0.342	0.110	0.002	0.711	0.572	0.882	-0.346	0.104	0.002
Child sex female	No	ref								
	Yes	0.160	0.092	0.081	2.316	1.840	2.915	0.217	0.086	0.012
Maternal age (years)		-0.058	0.009	<0.0001	0.944	0.927	0.961	-0.053	0.009	<0.0001
Maternal Lifestyle Behaviors during Pregnancy										
Smoked	No	ref						ref		
	Yes	0.841	0.117	<0.0001	2.318	1.841	2.918	0.803	0.117	<0.0001
Alcohol consumption in last trimester	None	ref								
	1-2 times/month	-0.343	0.148	0.020	0.709	0.531	0.948	-0.280	0.140	0.045
	>2 times/month	-0.424	0.319	0.183	0.654	0.350	1.222	-0.401	0.299	0.180
Worked all trimesters	No	ref						ref		
	Yes	-0.396	0.095	<0.0001	0.673	0.559	0.811	-0.364	0.090	<0.0001
Dad participation	No	ref						ref		
	Yes	-0.569	0.162	0.0005	0.566	0.412	0.778	-0.608	0.146	<0.0001
Risk perception	Underestimate	ref						ref		
	Accurate	-1.257	0.375	0.0008	0.284	0.137	0.593	-1.032	0.354	0.004
State Anxiety Inventory score		0.001	0.006	0.835	1.001	0.989	1.014	0.001	0.006	0.825
State Anxiety Inventory score x risk perception		0.023	0.009	0.011	1.023	1.005	1.041	0.018	0.009	0.039
>1 missing data points								1.321	0.464	0.007

Table 6 Characteristics of TEDDY actives and withdrawals [76]

Characteristic	Actives (n = 2994)	Withdrawals (n =763)	Total Sample (n = 3757)
Country	N (%)	N (%)	N
Finland	747(84%)	140(16%)	887
Germany	106(75%)	36(25%)	142
Sweden	1052(82%)	231(18%)	1283
United States	1089(75%)	356(25%)	1445
Child sex	N (%)	N (%)	N
Male	1538 (81%)	352 (19%)	1890
Female	1456 (78%)	411 (22%)	1867
Maternal age (years)	M (SD)	M (SD)	M (SD)
	30.8 (5.0)	28.5 (5.7)	30.4(5.2)
Maternal Lifestyle Behaviors During Pregnancy			
Smoking	N (%)	N (%)	N
Smoked	296(63%)	171(37%)	467
Did not smoke	2602(84%)	510(16%)	3112
Data missing	96(54%)	82(46%)	178
Alcohol consumption at 3rd trimester	N (%)	N (%)	N
Alcohol 1-2 times per month	474(87%)	72(13%)	546
Alcohol ≥ 3 time per month	105(89%)	13(11%)	118
No alcohol	2359(79%)	609(21%)	2968
Data missing	56(45%)	69(55%)	125
Employment status	N (%)	N (%)	N
Worked all 3 trimesters	1418(85%)	251(15%)	1669
Reduced work, quit, or did not work at all	1426(77%)	417(23%)	1843
Data missing	150(61%)	95(39%)	245
Dad Participation in TEDDY	N (%)	N (%)	N
Participated	2813(82%)	624(18%)	3437
Did Not Participate	181(57%)	139(43%)	320
Maternal Reactions to Child's Increased TIDM Risk			
Risk perception	N (%)	N (%)	N
Accurate	1809(84%)	355(16%)	2164
Underestimate	1132(77%)	343(23%)	1475
Data missing	53(45%)	65(55%)	118
State Anxiety Inventory score	M (SD)	M (SD)	M (SD)
Total Sample	38.7(9.7)	40.8(10.6)	39.1(9.9)
Risk Perception: Accurate	38.8(10.2)	41.7(10.4)	39.3(9.6)
Risk Perception: Underestimate	38.4(10.2)	39.9(10.8)	38.8(10.4)
	N (%)	N (%)	N
Data missing	46 (42%)	63 (58%)	109
Missing Data	N (%)	N (%)	N
≤1 missing data points	2944 (81%)	695 (19%)	3639
> 1 missing data points	50 (42%)	68 (58%)	118

Table 7 The 8 rules identified by the RuleFit method.

Rule 1 (risk increasing rule)	Rule 2 (risk increasing rule)
Maternal age < 27.5 Finland = NO	Smoker during pregnancy = YES Accurate risk perception = NO State anxiety inventory score > 45
Rule 3 (risk increasing rule)	**Rule 4 (risk increasing rule)**
State anxiety inventory score > 45 Dad participation = NO	Maternal age < 27.5 Accurate risk perception = NO Alcohol consumption in last trimester < 2 times per month
Rule 5 (risk decreasing rule)	**Rule 6 (risk decreasing rule)**
Worked all trimesters = YES Smoker during pregnancy= NO	Finland = NO Alcohol consumption in last trimester >0 Number of negative events < 2
Rule 7 (risk decreasing rule)	**Rule 8 (risk decreasing rule)**
Smoker during pregnancy = NO State anxiety inventory score < 45 Number of missing data points ≤ 1	Maternal age > 27.5 Smoker during pregnancy = NO Number of missing data points ≤ 1

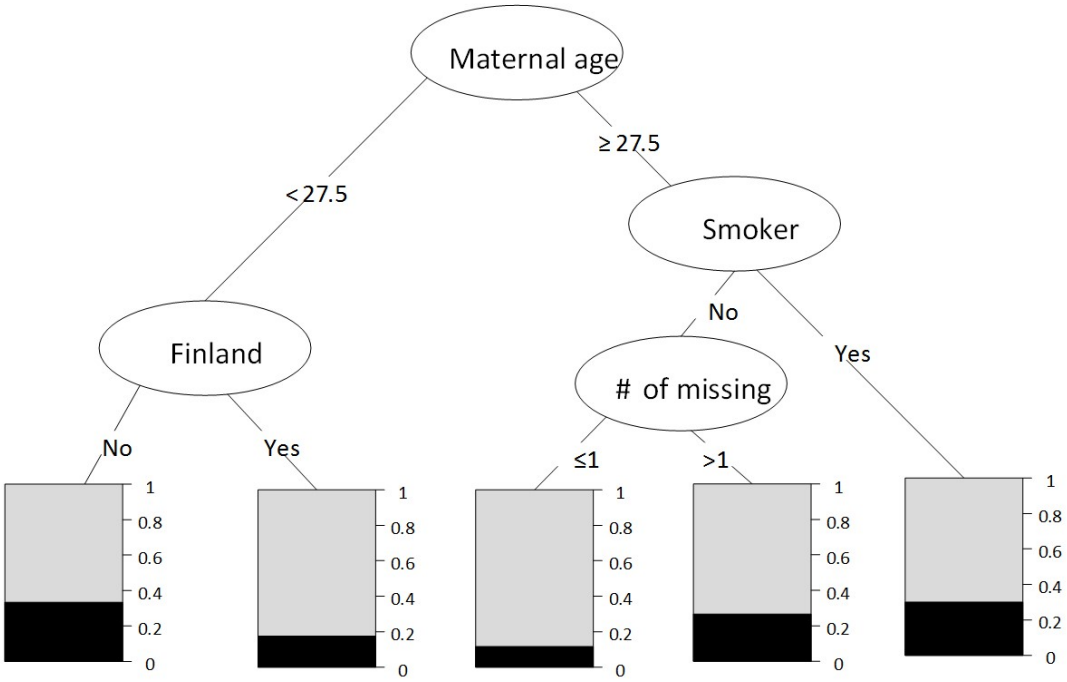


Figure 7 A decision tree learned from the TEDDY data.

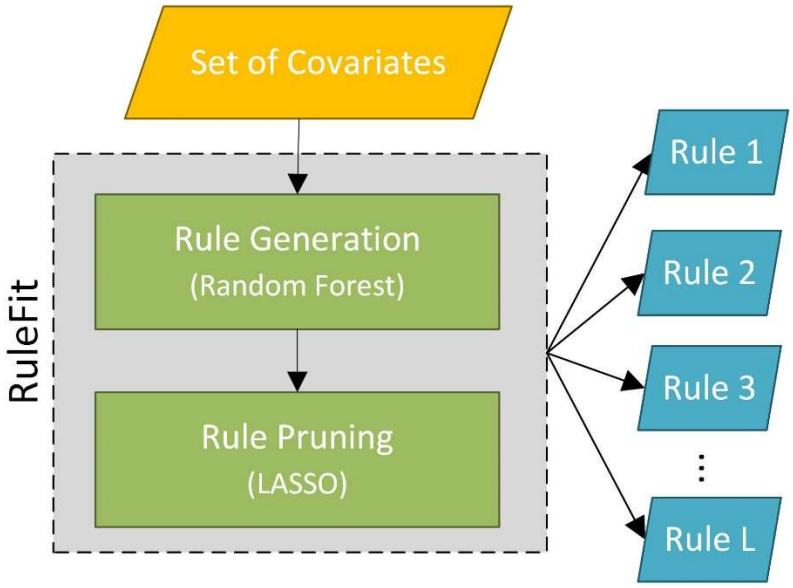


Figure 8 Flow diagram of the RuleFit algorithm.

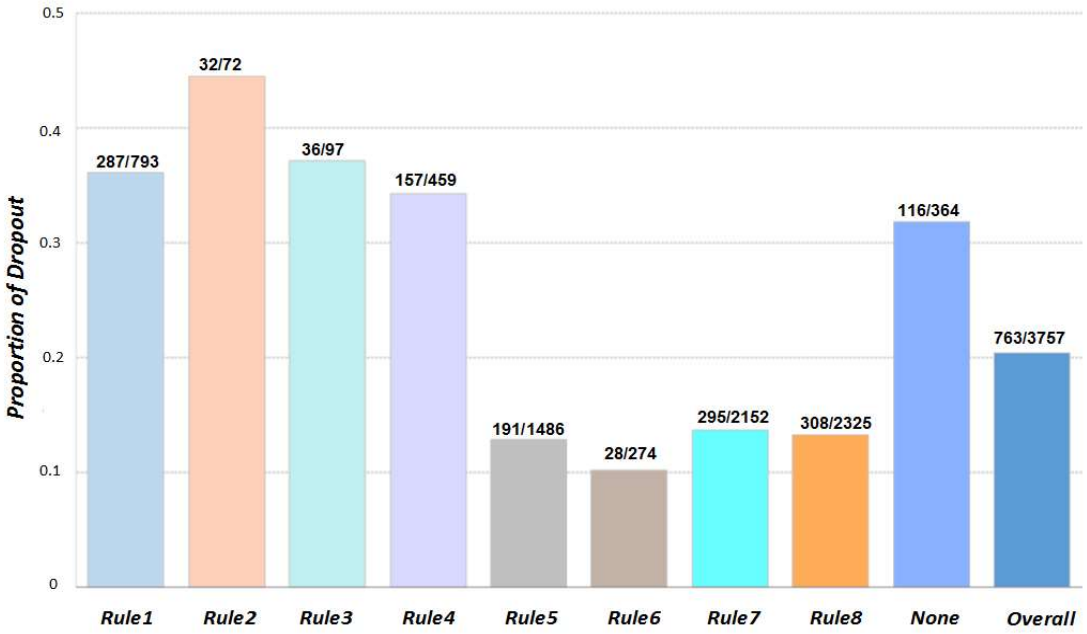


Figure 9 Proportion of early withdrawal of the eight rules and the overall population.

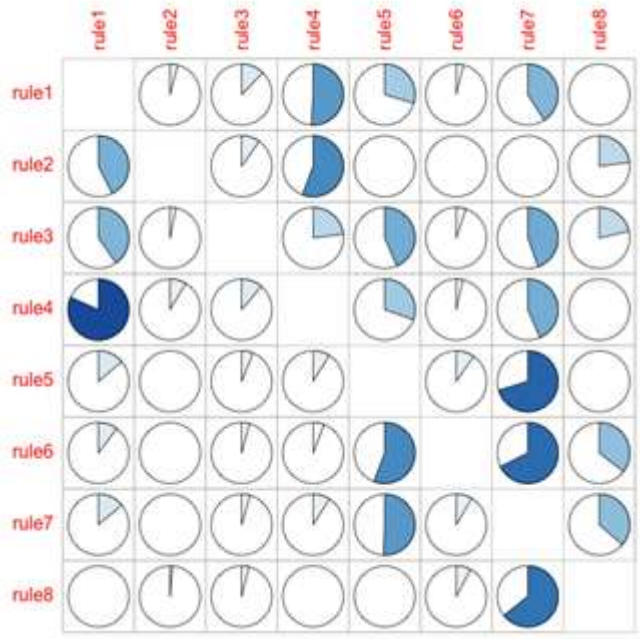


Figure 10 Investigation of the redundancy of the 8 rules.

CHAPTER 5: HIGH-THROUGHPUT SCREENING FOR RULE DISCOVERY FROM ADRC (ALZHEIMER'S DISEASE RESEARCH CENTERS) DATASETS

5.1 Introduction

Alzheimer's disease (AD) is the most common form of dementia for which there has not been any disease modifying therapy so far. It is expected that by 2050 there could be 100 million people worldwide suffering from AD. However, if the disease is detected at its initial stages, there would be a chance to delay the progression of the disease. Considering the cognitive decline in the patient and its consequences on the quality of life of the patients and her/his family, selecting biomarkers that could help in detecting the disease at its early stages is of high importance.

Complex diseases such as Alzheimer's are influenced by a combination of genetic and environmental risk factors and their interactions. Neuropsychological evaluations are the very initial stage of differentiating normal subjects from the ones who are prone to progression to AD. Several studies have shown a correlation between Neuropsychological tests and biomarkers that can detect the progression process to dementia. [4,91,92]

In this study we use a dataset from the Alzheimer's disease Research Center (ADRC) in Tampa, Florida. The ADRCs provide information and referral services to adults with mental illnesses and assign a degree of urgency to clients who require assistance. They also manage the availability of financial resources for certain key long-term care programs targeted for elders to ensure financial viability and stability. The dataset consists of the results of neurological examinations, medical information, family history and a battery of neuropsychological tests.

The battery of neuropsychological tests of ADRC includes 16 different Neuropsychological tests, which contain 290 different variables. Since administrating so many tests would take significant time and some of these tests would measure the same information, we are trying to find redundant tests in this battery of tests and see how we could come to the same conclusion about the pathological group of a subject by using a subset of the neuropsychological tests instead of administering all of them.

Previous methods available for dimensionality reduction and high-dimensional data analysis mainly focus on personalized feature selection and rarely pay attention to interaction between the variables. Our challenge is to reduce the number of variables, and identify highly synergistic groups of variables, then discover rules from a high-dimensional dataset considering not only the variables themselves, but also their interactions and ranges. We use statistical models to select rules and minimize uncertainty.

5.2 Data and Methods

5.2.1 Data

The data set of this study comes from ADRC at Byrd Alzheimer's institute. The original dataset consisted of 353 variables including demographic information, Neurological examinations, medical history and a battery of psychological tests. A set of 25 summarizing variables were selected by the Byrd Alzheimer's institute and we limited our analysis to them. These variables included age, gender, number of years of schooling in U.S., Mini Mental State Examination (MMSE), Digit Span (Forward and Backward score), Hopkins Verbal Learning test (total trial 1-3, delayed recall, Recognition true positives), Trail Making (A, B), Digit Symbol ss, Block Design ss, Category Fluency Total, Verbal Fluency total, Sit Retrieval (Total trials 1–3, The

neuropsychological test Bag with a 20 min delay, Recognition recall total), Judgment of Line Orientation, Mohs Time, Visual Reproduction (immediate total score, delayed total score), Boston Naming test, spontaneous (total), similarities (WAIS) ss, Wide Range Achievement Test – 3rd edition.

Based on the results of all the neuropsychological tests the subjects are categorized into three groups of Normal Controls (NC), Mild Cognitive Impairments (MCI) and Alzheimer Disease (AD). Our dataset included 548 NC, 163 MCI and 549 AD. Since we are using binary classification methods, we apply the data analysis process for each pair of pathological groups separately.

5.2.2 Methods

The data analysis process is done in four different steps. In the first step, we use the RuleFit Algorithm [80] to generate predictive rules of progression to Alzheimer's Disease. In the second step, we use the latent trait model [93] in order to weight the detected rules found in Step 1. In Step 3 we use the Maximum Weighted Multiple Clique Algorithm [94] and select the most synergistic clique of rules. Finally, in Step 4 we sum up the number of risk increasing rules each individual matches and subtract the risk decreasing rules that they match from that and calculate a risk score for each individual. Figure 11 provides a schematic illustration of the data analysis process. We will describe each step of the process in more detail in future sections.

5.2.2.1 Step 1 – Generating Rules Using RuleFit Algorithm

We are using the RuleFit algorithm [80] to discover rules that are predicting the outcome of interest, which in our case is progression to Alzheimer's Disease. Each rule consists of one or several variables and their ranges. Rulefit is an algorithm that is capable of extracting rules from a dataset with a large number of variables as discussed in 4.2.5. We defined our rules using single

variables; in Step 4 of our analysis we try to capture the interaction between variables using the Maximum Weighted Multiple Clique Problem (MWMCP).

5.2.2.2 Step 2 - Latent Trait Model

Item response theory has been used in psychometrics for measuring various kinds of latent traits such as ability, intelligence, knowledge, etc. In our problem disease, risk is our latent trait which is not directly measurable; however, the rules are essentially measurable evidence associated with the underlying disease risk. There are several functions for modeling latent traits; we use the logistic function, which has been widely used in the biological sciences to model the growth of animals and plants. We preferred this model due to its simplicity and since we are modeling a progression process. The function for the logistic model is shown in Equation 2, in which θ is the disease risk, a is the discrimination parameter and b is the difficulty parameter.

$$P(\theta) = \frac{1}{1+e^{-a(\theta-b)}} \qquad (5\text{-}1)$$

Based on the function, we can compute the Item Information curve, which is a measure of precision and how much information each rule provides. The function for information is shown in Equation 3

$$I(\theta) = a^2 P(\theta)(1 - P(\theta)) \qquad (5\text{-}2)$$

We can get the area under each information curve as $\int_{-\infty}^{+\infty} I(\theta)$ as the information of each single rule and the non-overlapping area under each pair of curves as the interactive information from a pair of rules $\int_{-\infty}^{+\infty}(max(I_1(\theta), I_2(\theta)) - min(I_1(\theta), I_2(\theta)))$. We use the information that we calculate at this stage as the weight of each rule in Step 3.

5.2.2.3 Step 3 - Maximum Weighted Multiple Clique Algorithm

in this step, we use a network-based formulation to select a subset of rules that have the most synergistic power for differentiating each pair of pathological groups. In this network, each node represents a potential biomarker and could be weighted by the predictive power of that biomarker, and the edges between each pair of nodes can be weighted with the synergistic power of the two biomarkers. We use each node to represent a rule and the edges as the i of each pair of rules; it should be mentioned that both nodes and edges are binary variables; each node is weighted by the total area under its information curve, and the edges are weighted by the total of their non-overlapping area. We use an optimization model developed by [94] which tries to find a clique of nodes with the maximum total weights of both nodes and edges; they have called this problem a maximum weighted multiple clique problem (MWMCP). They derived an algorithm that provides an optimal or near-optimal solution using the column generation method [95,96].

$$
\begin{aligned}
&\max \sum_{i}\sum_{k} p(v_i)\, X_{ik} + \sum_{i}\sum_{j>i}\sum_{k} w(e_{ij}) Z_{ijk} \\
&s.t. \quad X_{ik} + X_{jk} \le 1 \qquad \forall\, i,j,k : j > i\,,\, e_{ij} \notin E \\
&Z_{ijk} \le \frac{1}{2}\left(X_{ik} + X_{jk}\right) \quad \forall\, i,j,k : j > i \\
&Z_{ijk} \ge X_{ik} + X_{jk} - 1 \quad \forall\, i,j,k : j > i \\
&X_{ik}, Z_{ijk} \in \{0,1\} \qquad \forall\, i,j,k \\
&i = 1, \dots, n \\
&j = 1, \dots, m \\
&k = 1, \dots, k
\end{aligned}
\tag{5-3}
$$

In Equation 4 X_{ik} would be 1 if node i is selected in clique k and 0 otherwise and Z_{ijk} would be 1 if edge i-j is selected in clique k and 0 otherwise. $p(v_i)$ is the weight of node i and $w(e_{ij})$ is the weight for edge i-j. The constraints in this guarantee that each node belongs to only

one clique and the variables are binary. The goal in this problem is to find multiple cliques of nodes with the maximum weight of nodes and edges.

5.2.2.4 Step 4- Finding the Risk Score for Each Individual

In this step rules are labeled as risk-increasing or risk-decreasing based on the number of individuals in each pathological group of the ones who match each rule. We should mention that in the RuleFit method, each individual can match more than one rule. The risk score is determined for each person, by counting the number of risk increasing rules and subtracting the risk-decreasing rules that they match.

5.3 Results

In the first step we used the RuleFit algorithm to identify rules for each pair of pathological groups. Table 8, Table 9 and Table 10Table 10 provide the risk-predictive rules identified by the RuleFit algorithm for each pair of pathological groups. It can be seen that Sit Retrieval test, Trail timing and Visual Reproduction appeared in all three sets of comparisons. Demographic variables such as age and number of years of schooling in the US only appeared in comparison of Normal Controls and MCIs; however, the fact that these rules would lead to an increase or decrease in the risk of developing AD depends on the cut off values and also the direction of the inequality.

In the second step of our analysis, we coded each rule as a binary variable; if an individual matched the rule, the rule would be coded as 1 and 0 otherwise. Then we applied item response theory with a logistic function on the rules. Each rule is used as a risk factor to predict the latent trait which in our case is the risk of progression to Alzheimer's disease. The item information curves for rules are shown in Figure 12 and as described in the methods section, we used the area under the item information curve as the weight of each rule and the non-overlapping area under

the item information curve of each pair of rules as the weight of the interaction between each pair of rules. Plots of item information can be used to see how much information an item contributes. The adjacency matrix, which includes the weights is the result of Step 2.

In step 3 the adjacency matrix from step 2 is plugged into the MWMCP algorithm and a set of variables with the maximum synergistic weights are selected. It should be mentioned that at each iteration of the algorithm, we tuned the threshold for removing insignificant edges in the model so that the desired number of rules (in our case four) are selected. Since we tried to select as few rules as possible we changed the threshold so that four rules would be selected at each run. The top four rules that were selected by the MWMCP to separate the NC from AD were Sit Retrieval Total Trials 1 - 3 ≤ 20.5 -> AD, Verbal Fluency Total ≥ 9.5 -> NC, Visual Reproduction Immediate Total Score ≤ 15.5 -> AD, Sit Retrieval Recognition Recall Total ≥ 23.5 -> NC. From the rules that identify NC and MCI, the top selected rules were Visual Reproduction Delayed Total Score ≤ 10.5 -> MCI, Sit Retrieval Total Trials 1 - 3 ≤ 19 -> MCI, Number of years of schooling in U.S. ≥ 17.5 -> NC, Hopkins Verbal Total Trial 1 - 3 ≤ 23.5 -> MCI. And the top four rules that identify MCI and AD are Sit Retrieval Total Trials 1 - 3 ≥ 18.5 -> MCI, Category Fluency Total ≥ 24.5 -> MCI, Hopkins Verbal Delayed Recall ≤ 2.5 -> AD, Mini Mental State Examination≥ 27.5 -> MCI.

Finally, in step 4 we summed up the number of risk increasing rules that each individual matches and subtracted the number of risk-decreasing rules from that and ended up with a risk score. We further investigated the risk level of progressing to Alzheimer's disease for the positive risk scores versus negative ones. If we categorize individuals with a positive score at higher risk of AD and the ones with zero or negative score at lower risk of developing AD, we could get a reasonable level of prediction performance.

5.4 Conclusion

We tried to summarize a battery of neuropsychological tests with as few tests as possible and see if instead of 16 different tests, we could use one or two. We found Mini-Mental State Exam, Hopkins verbal learning, verbal fluency test, Sit retrieval and visual reproduction as the most important tests in categorizing the subjects. A comparison of the network-based methods with logistic regression and support vector machines with linear and Gaussian Kernels is provided in Table 11. The results compare the predictive performance of the network-based model with the other methods based on the variables selected by our algorithm and also all the 26 variables in the study. Our network based method, provides acceptable performance in comparison with other methods and also the models using all variables, except for the case for comparing NC and MCI. Although in many cases, we can find other methods with better performance, we should consider the fact that the main goal of our analysis is dimension reduction and selecting a subset of tests that could lead to an acceptable result.

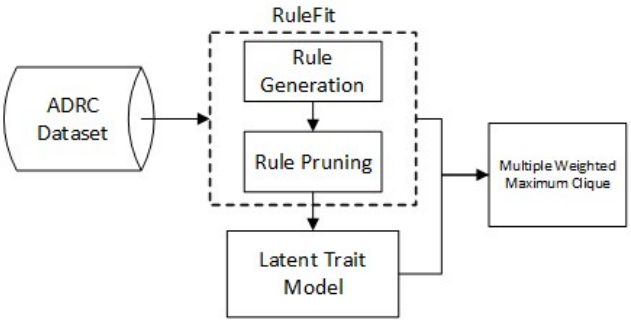
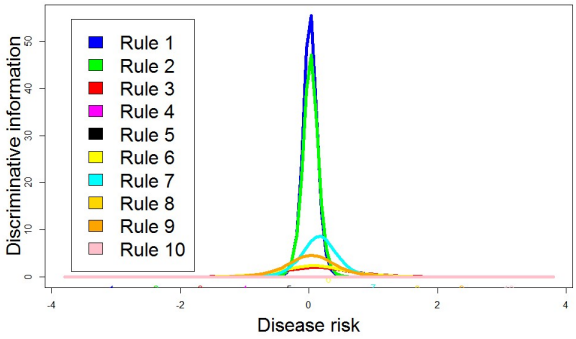
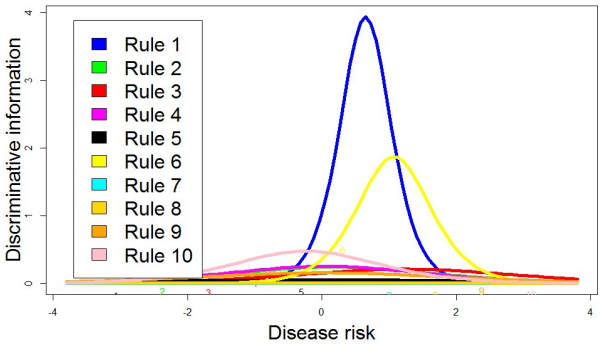


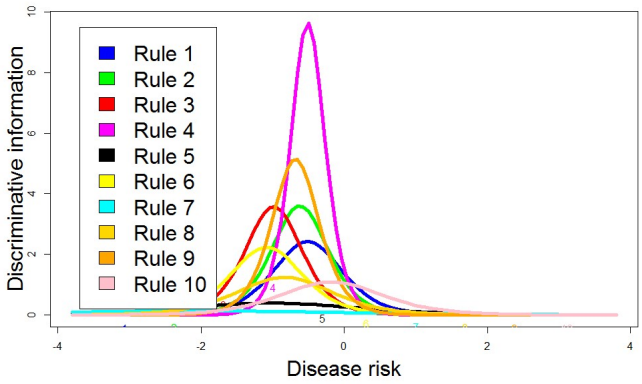
Figure 11 Data analysis process



(a) NC-AD



(b) NC-MCI



(c) MCI-AD

Figure 12 Item information curves

Table 8 Identified rules by RuleFit algorithm for NC-AD pathological groups

Rule 1 (risk increasing rule)	Rule 2 (risk increasing rule)
Sit Retrieval Total Trials 1 - 3 ⩽ 20.5	Visual Reproduction Delayed Total Score ⩽ 7.5
Rule 3 (risk increasing rule)	Rule 4 (risk increasing rule)
Category Fluency Total ≤ 33.5	Boston Naming Spontaneous (total) ≤ 39.5
Rule 5 (risk increasing rule)	Rule 6 (risk decreasing rule)
Visual Reproduction Immediate Total Score ≤ 15.5	Sit Retrieval bag a 20 min delay ≥ 9.5
Rule 7 (risk decreasing rule)	Rule 8 (risk decreasing rule)
Sit Retrieval Recognition Recall Total ≥ 23.5	Wide Range Achievement Test - raw score ≥ 40.5
Rule 9 (risk decreasing rule)	Rule 10 (risk decreasing rule)
Trails B Time ≥ 217.5	Verbal Fluency Total ≥ 9.5

Table 9 Identified rules by RuleFit algorithm for NC-MCI pathological groups

Rule 1 (risk increasing rule)	Rule 2 (risk increasing rule)
Visual Reproduction Delayed Total Score ≤ 10.5	Sit Retrieval bag a 20 min Delay ≤ 6.5
Rule 3 (risk increasing rule)	Rule 4 (risk increasing rule)
Sit Retrieval Total Trials 1 - 3 ≤ 19.5	Wide Range Achievement Test - Raw Score ≤ 49.5
Rule 5 (risk increasing rule)	Rule 6 (risk increasing rule)
Trails B Time ≤ 93.5	Hopkins Verbal Total Trial 1 - 3 ≤ 23.5
Rule 7 (risk decreasing rule)	Rule 8 (risk decreasing rule)
Boston Naming Spontaneous (total) ≥ 47.5	Trails A Time ≤ 36.5
Rule 9 (risk decreasing rule)	Rule 10 (risk decreasing rule)
Age ≤ 72.5	Number of years of schooling in U.S. ≥ 17.5

Table 10 Identified rules by RuleFit algorithm for MCI-AD pathological groups

Rule 1 (risk increasing rule)	Rule 2 (risk increasing rule)
Visual Reproduction Delayed Total Score ≤ 7.5	Sit Retrieval Bag a 20 min Delay ≥ 2.5
Rule 3 (risk increasing rule)	Rule 4 (risk increasing rule)
Digit Symbol ss ≤ 10.5	Hopkins Verbal Delayed Recall ≤ 2.5
Rule 5 (risk increasing rule)	Rule 6 (risk increasing rule)
Verbal Fluency Total ≤ 20.5	Category Fluency Total ≤ 24.5
Rule 7 (risk decreasing rule)	Rule 8 (risk decreasing rule)
Sit Retrieval Total Trials 1 - 3 ≥ 18.5	Trails B Time ≤ 285.5
Rule 9 (risk decreasing rule)	Rule 10 (risk decreasing rule)
Boston Naming Spontaneous (total) ≥ 38.5	Mini Mental State Examination≥ 27.5

Table 11 Predictive performance of the network based model compared to available methods

		NC - AD			NC - MCI			MCI-AD		
		ACC	Sen	Spec	ACC	Sen	Spec	ACC	Sen	Spec
Selected Variables by Network-based	Network-Based	0.93	0.91	0.96	0.84	0.52	0.97	0.78	0.76	0.83
	SVM (linear)	0.915	0.87	0.97	0.78	0.27	0.92	0.72	0.82	0.48
	SVM (Gaussian)	0.95	0.92	0.97	0.82	0.36	0.96	0.76	0.84	0.57
	Logistic Regression	0.92	0.88	0.94	0.85	0.59	0.93	0.76	0.85	0.51
		NC - AD			NC - MCI			MCI-AD		
		ACC	Sen	Spec	ACC	Sen	Spec	ACC	Sen	Spec
All 26 Variables	SVM (linear)	0.9	0.97	0.66	0.76	0.95	0.85	0.87	0.90	0.78
	SVM (Gaussian)	0.98	0.97	0.96	0.83	0.92	0.64	0.98	0.85	0.77
	Logistic Regression	0.96	0.97	0.94	0.79	0.96	0.87	0.91	0.70	0.86

CHAPTER 6: SUMMARY

In this dissertation we applied rule-based analysis on medical data and tried to extract predictive rules out of large and inter-correlated data. Our methods were easy to implement, and the comparison of our results with current machine learning methods showed reasonable predictive performance.

First, we developed decision tree models on ADNI dataset and extracted rules out of it to predict elevated brain amyloid level, which is the first and most important symptom of Alzheimer's Disease onset. The current diagnostic methods for Alzheimer's disease are expensive and not available in all clinics. We predicted the amyloid level using blood proteomics and neuropsychological tests, which are inexpensive and available in all clinics. We have published a paper on this study in collaboration with Byrd Alzheimer's Institute in the Journal of Alzheimer's Disease [1].

Second, we tried to compare the interaction of plasma biomarkers of Alzheimer's in three different female ethnic groups. We applied decision tree models and extracted predictive rules of Alzheimer's disease for each ethnicity. Few studies have targeted pathological, cardiovascular and inflammatory biomarkers in plasma at the same time. We have submitted a manuscript on this study in the Journal of Alzheimer's Disease in collaboration with the Byrd Alzheimer's institute.

In our third study, we used the RuleFit algorithm and tried to identify homogeneous subgroups out of a population who participated in a pediatric study. We detected eight rules which characterized eight different subpopulations in the participants. We compared our results with a

previous regression model applied on the same dataset. We have submitted a manuscript on this study to the Scientific Reports Journal.

Finally, we did a redundancy analysis on a battery of neuropsychological tests from Alzheimer's Disease Research Center (ADRC). We used the RuleFit algorithm for extracting rules out of the dataset and applied Maximum Weighted Multiple Clique Problem (MWMCP) to select a clique of the rules with the maximum weight. The Latent trait model was used to weight the rules. As a result, we found five tests which could result in the same conclusion about the pathological group of the patients instead of using the full suite of tests.

REFERENCES

[1] M. Haghighi, A. Smith, D. Morgan, B. Small, S. Huang, Identifying cost-effective predictive rules of amyloid-β level by integrating neuropsychological tests and plasma-based markers, J. Alzheimer's Dis. 43 (2015) 1261-1270.

[2] J. Hardy, D.J. Selkoe, The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics, Science. 297 (2002) 353-356.

[3] V.L. Villemagne, R. Cappai, K.J. Barnham, R.A. Cherny, C. Opazo, K.E. Novakovic, C.C. Rowe, C.L. Masters, The Aβcentric Pathway of Alzheimer's Disease, in: Anonymous Abeta Peptide and Alzheimer's Disease, Springer, 2007, pp. 5-36.

[4] W. Klunk, A. Cohen, B. Snitz, H. Aizenstein, E. Halligan, L. Weissfeld, C. Mathis, J. Price, R. Nebes, Conversion to MCI and amyloid-positivity in normal controls, Alzheimer's & Dementia: The Journal of the Alzheimer's Association. 9 (2013) P150.

[5] G.W. Small, V. Kepe, L.M. Ercoli, P. Siddarth, S.Y. Bookheimer, K.J. Miller, H. Lavretsky, A.C. Burggren, G.M. Cole, H.V. Vinters, PET of brain amyloid and tau in mild cognitive impairment, N. Engl. J. Med. 355 (2006) 2652-2663.

[6] K.A. Johnson, S. Minoshima, N.I. Bohnen, K.J. Donohoe, N.L. Foster, P. Herscovitch, J.H. Karlawish, C.C. Rowe, M.C. Carrillo, D.M. Hartley, S. Hedrick, V. Pappas, W.H. Thies, Appropriate use criteria for amyloid PET: a report of the Amyloid Imaging Task Force, the Society of Nuclear Medicine and Molecular Imaging, and the Alzheimer's Association, J. Nucl. Med. 54 (2013) 476-490.

[7] K.E. Pike, G. Savage, V.L. Villemagne, S. Ng, S.A. Moss, P. Maruff, C.A. Mathis, W.E. Klunk, C.L. Masters, C.C. Rowe, Beta-amyloid imaging and memory in non-demented individuals: evidence for preclinical Alzheimer's disease, Brain. 130 (2007) 2837-2844.

[8] H.J. Aizenstein, R.D. Nebes, J.A. Saxton, J.C. Price, C.A. Mathis, N.D. Tsopelas, S.K. Ziolko, J.A. James, B.E. Snitz, P.R. Houck, Frequent amyloid deposition without significant cognitive impairment among the elderly, Arch. Neurol. 65 (2008) 1509-1517.

[9] K.E. Pike, K.A. Ellis, V.L. Villemagne, N. Good, G. Chételat, D. Ames, C. Szoeke, S.M. Laws, G. Verdile, R.N. Martins, Cognition and beta-amyloid in preclinical Alzheimer's disease: data from the AIBL study, Neuropsychologia. 49 (2011) 2384-2390.

[10] V. Leinonen, I. Alafuzoff, S. Aalto, T. Suotunen, S. Savolainen, K. Någren, T. Tapiola, T. Pirttilä, J. Rinne, J.E. Jääskeläinen, Assessment of β-amyloid in a frontal cortical brain biopsy specimen and by positron emission tomography with carbon 11–labeled Pittsburgh Compound B, Arch. Neurol. 65 (2008) 1304-1309.

[11] P. Bourgeat, G. Chetelat, V.L. Villemagne, J. Fripp, P. Raniga, K. Pike, O. Acosta, C. Szoeke, S. Ourselin, D. Ames, K.A. Ellis, R.N. Martins, C.L. Masters, C.C. Rowe, O. Salvado, AIBL Research Group, Beta-amyloid burden in the temporal neocortex is related to hippocampal atrophy in elderly subjects without dementia, Neurology. 74 (2010) 121-127.

[12] A. Bahar-Fuchsa, V. Villemagnea, K. Onga, G. Chetélatd, F. Lamba, C.B. Reiningere, M. Woodwardf, C.C. Rowea, Prediction of Amyloid-ß Pathology in Amnestic Mild Cognitive Impairment with Neuropsychological Tests, Journal of Alzheimer's Disease. 33 (2013) 451-462.

[13] C.A. Luis, L. Abdullah, G. Ait-Ghezala, B. Mouzon, A.P. Keegan, F. Crawford, M. Mullan, Feasibility of Predicting MCI/AD Using Neuropsychological Tests and Serum beta-Amyloid, Int. J. Alzheimers Dis. 2011 (2011) 786264.

[14] D.A. Llano, V. Devanarayan, A.J. Simon, Alzheimer's Disease Neuroimaging Initiative (ADNI), Evaluation of plasma proteomic data for Alzheimer disease state classification and for the prediction of progression from mild cognitive impairment to Alzheimer disease, Alzheimer Dis. Assoc. Disord. 27 (2013) 233-243.

[15] S. Burnham, N. Faux, W. Wilson, S. Laws, D. Ames, J. Bedo, A. Bush, J. Doecke, K. Ellis, R. Head, A blood-based predictor for neocortical Aβ burden in Alzheimer's disease: results from the AIBL study, Mol. Psychiatry. 19 (2014) 519-526.

[16] J.D. Doecke, S.M. Laws, N.G. Faux, W. Wilson, S.C. Burnham, C. Lam, A. Mondal, J. Bedo, A.I. Bush, B. Brown, Blood-based protein biomarkers for diagnosis of Alzheimer disease, Arch. Neurol. 69 (2012) 1318-1325.

[17] S.J. Kiddle, M. Thambisetty, A. Simmons, J. Riddoch-Contreras, A. Hye, E. Westman, I. Pike, M. Ward, C. Johnston, M.K. Lupton, Plasma based markers of [11C] PiB-PET brain amyloid burden, PloS one. 7 (2012) e44260.

[18] D. Cramer, D.L. Howitt, The Sage Dictionary of Statistics: A Practical Resource for Students in the Social Sciences, Sage, 2004.

[19] A.S. Kaufman, IQ Testing 101. Springer Publishing Co, 2009.

[20] W.P. Vogt, R.B. Johnson, Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences, Sage, 2011.

[21] M. Thambisetty, S. Lovestone, Blood-based biomarkers of Alzheimer's disease: challenging but feasible, Biomarkers in medicine. 4 (2010) 65-79.

[22] N. Rifai, M.A. Gillette, S.A. Carr, Protein biomarker discovery and validation: the long and uncertain path to clinical utility, Nat. Biotechnol. 24 (2006) 971-983.

[23] K. Gustaw-Rothenberg, A. Lerner, D.J. Bonda, H. Lee, X. Zhu, G. Perry, M.A. Smith, Biomarkers in Alzheimer's disease: past, present and future, Biomarkers in medicine. 4 (2010) 15-26.

[24] BC Plasma Proteomics Data Primer,.

[25] W.J. Jagust, D. Bandy, K. Chen, N.L. Foster, S.M. Landau, C.A. Mathis, J.C. Price, E.M. Reiman, D. Skovronsky, R.A. Koeppe, The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core, Alzheimer's & Dementia. 6 (2010) 221-229.

[26] L.M. Shaw, H. Vanderstichele, M. Knapik-Czajka, C.M. Clark, P.S. Aisen, R.C. Petersen, K. Blennow, H. Soares, A. Simon, P. Lewczuk, Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects, Ann. Neurol. 65 (2009) 403-413.

[27] A. Zeileis, T. Hothorn, K. Hornik, Model-based recursive partitioning, Journal of Computational and Graphical Statistics. 17 (2008) 492-514.

[28] L. Gordon, R.A. Olshen, Tree-structured survival analysis, Cancer Treat. Rep. 69 (1985) 1065-1069.

[29] R.C. Mohs, D. Knopman, R.C. Petersen, S.H. Ferris, C. Ernesto, M. Grundman, M. Sano, L. Bieliauskas, D. Geldmacher, C. Clark, Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. Alzheimer Disease & Associated Disorders. 11 (1997) 13-21.

[30] E. Stomrud, O. Hansson, H. Zetterberg, K. Blennow, L. Minthon, E. Londos, Correlation of longitudinal cerebrospinal fluid biomarkers with cognitive decline in healthy older adults, Arch. Neurol. 67 (2010) 217-223.

[31] K. Rasmussen, A. Tybjærg-Hansen, B. Nordestgaard, R. Frikke-Schmidt, Apolipoprotein E plasma level and genotype—/INS; Risk of dementia in 76,000 individuals from the general population, J. Neurol. Sci. 333 (2013) e342-e342.

[32] M. Thambisetty, R. Tripaldi, J. Riddoch-Contreras, A. Hye, Y. An, J. Campbell, J. Sojkova, A. Kinsey, S. Lynham, Y. Zhou, Proteome-based plasma markers of brain amyloid-β deposition in non-demented older individuals, J. Alzheimer's Dis. 22 (2010) 1099-1109.

[33] A. ChenI Heinrichs, S. Mason, Inhibiting amyloid formation, Nature Structural & Molecular Biology. 18 (2011) 747.

[34] A. Marcello, O. Wirths, T. Schneider-Axmann, M. Degerman-Gunnarsson, L. Lannfelt, T.A. Bayer, Circulating immune complexes of Aβ and IgM in plasma of patients with Alzheimer's disease, J. Neural Transm. 116 (2009) 913-920.

[35] L. Velayudhan, R. Killick, A. Hye, A. Kinsey, A. Güntert, S. Lynham, M. Ward, R. Leung, A. Lourdusamy, A.W. To, Plasma transthyretin as a candidate marker for Alzheimer's disease, J. Alzheimer's Dis. 28 (2012) 369-375.

[36] S. Horstmann, L. Budig, H. Gardner, J. Koziol, M. Deuschle, C. Schilling, S. Wagner, Matrix metalloproteinases in peripheral blood and cerebrospinal fluid in patients with Alzheimer's disease, International psychogeriatrics. 22 (2010) 966-972.

[37] J. Marksteiner, W.A. Kaufmann, P. Gurka, C. Humpel, Synaptic proteins in Alzheimer's disease, Journal of Molecular Neuroscience. 18 (2002) 53-63.

[38] R.L. Bowen, G. Verdile, T. Liu, A.F. Parlow, G. Perry, M.A. Smith, R.N. Martins, C.S. Atwood, Luteinizing hormone, a reproductive regulator that modulates the processing of amyloid-beta precursor protein and amyloid-beta deposition, J. Biol. Chem. 279 (2004) 20539-20545.

[39] M. Fandrich, V. Forge, K. Buder, M. Kittler, C.M. Dobson, S. Diekmann, Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 15463-15468.

[40] J. Harrison, S.L. Minassian, L. Jenkins, R.S. Black, M. Koller, M. Grundman, A neuropsychological test battery for use in Alzheimer disease clinical trials, Arch. Neurol. 64 (2007) 1323-1329.

[41] D.A. Llano, V. Devanarayan, A.J. Simon, Alzheimer's Disease Neuroimaging Initiative (ADNI), Evaluation of plasma proteomic data for Alzheimer disease state classification and for the prediction of progression from mild cognitive impairment to Alzheimer disease, Alzheimer Dis. Assoc. Disord. 27 (2013) 233-243.

[42] P. Robert, S. Ferris, S. Gauthier, R. Ihl, B. Winblad, F. Tennigkeit, Review of Alzheimer's disease scales: is there a need for a new multi-domain scale for therapy evaluation in medical practice, Alzheimers Res Ther. 2 (2010) 24.

[43] W.T. Hu, D.M. Holtzman, A.M. Fagan, L.M. Shaw, R. Perrin, S.E. Arnold, M. Grossman, C. Xiong, R. Craig-Schapiro, C.M. Clark, E. Pickering, M. Kuhn, Y. Chen, V.M. Van Deerlin, L. McCluskey, L. Elman, J. Karlawish, A. Chen-Plotkin, H.I. Hurtig, A. Siderowf, F. Swenson, V.M. Lee, J.C. Morris, J.Q. Trojanowski, H. Soares, Alzheimer's Disease Neuroimaging Initiative, Plasma multianalyte profiling in mild cognitive impairment and Alzheimer disease, Neurology. 79 (2012) 897-905.

[44] D. Inekci, D.S. Jonesco, S. Kennard, M.A. Karsdal, K. Henriksen, The potential of pathological protein fragmentation in blood-based biomarker development for dementia–with emphasis on Alzheimer's disease, Frontiers in neurology. 6 (2015).

[45] S. Ray, M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L.F. Friedman, D.R. Galasko, M. Jutel, A. Karydas, Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins, Nat. Med. 13 (2007) 1359-1362.

[46] J.B. Toledo, E. Toledo, M.W. Weiner, C.R. Jack, W. Jagust, V.M. Lee, L.M. Shaw, J.Q. Trojanowski, Alzheimer's Disease Neuroimaging Initiative, Cardiovascular risk factors, cortisol, and amyloid-β deposition in Alzheimer's Disease Neuroimaging Initiative, Alzheimer's & Dementia. 8 (2012) 483-489.

[47] M.M. Breteler, Vascular risk factors for Alzheimer's disease:: An epidemiologic perspective, Neurobiol. Aging. 21 (2000) 153-160.

[48] N.E. Shepardson, G.M. Shankar, D.J. Selkoe, Cholesterol level and statin use in Alzheimer disease: I. Review of epidemiological and preclinical studies, Arch. Neurol. 68 (2011) 1239-1244.

[49] M. Britschgi, K. Rufibach, S.L. Huang, C.M. Clark, J.A. Kaye, G. Li, E.R. Peskind, J.F. Quinn, D.R. Galasko, T. Wyss-Coray, Modeling of pathological traits in Alzheimer's disease based on systemic extracellular signaling proteome, Mol. Cell. Proteomics. 10 (2011) M111.008862.

[50] D.C. Lee, J. Rizer, J.B. Hunt, M. Selenica, M.N. Gordon, D. Morgan, Review: experimental manipulations of microglia in mouse models of Alzheimer's pathology: activation reduces amyloid but hastens tau pathology, Neuropathol. Appl. Neurobiol. 39 (2013) 69-85.

[51] A.P. Reiner, E.M. Lange, N.S. Jenny, P.H. Chaves, J. Ellis, J. Li, J. Walston, L.A. Lange, M. Cushman, R.P. Tracy, Soluble CD14: genomewide association analysis and relationship to cardiovascular risk and mortality in older adults, Arterioscler. Thromb. Vasc. Biol. 33 (2013) 158-164.

[52] A.L. Chin, S. Negash, R. Hamilton, Diversity and disparity in dementia: the impact of ethnoracial differences in Alzheimer disease, Alzheimer Dis. Assoc. Disord. 25 (2011) 187-195.

[53] L.A. Farrer, L.A. Cupples, J.L. Haines, B. Hyman, W.A. Kukull, R. Mayeux, R.H. Myers, M.A. Pericak-Vance, N. Risch, C.M. van Duijn, Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis, JAMA. 278 (1997) 1349-1356.

[54] L.E. Hebert, L.A. Beckett, P.A. Scherr, D.A. Evans, Annual incidence of Alzheimer disease in the United States projected to the years 2000 through 2050, Alzheimer Disease & Associated Disorders. 15 (2001) 169-173.

[55] D.E. Barnes, K. Yaffe, The projected effect of risk factor reduction on Alzheimer's disease prevalence, The Lancet Neurology. 10 (2011) 819-828.

[56] G.G. Potter, B.L. Plassman, J.R. Burke, M.U. Kabeto, K.M. Langa, D.J. Llewellyn, M.A. Rogers, D.C. Steffens, Cognitive performance and informant reports in the diagnosis of cognitive impairment and dementia in African Americans and whites, Alzheimer's & Dementia. 5 (2009) 445-453.

[57] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, E.M. Stadlan, Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease, Neurology. 34 (1984) 939-944.

[58] J. Sundelof, J. Arnlov, E. Ingelsson, J. Sundstrom, S. Basu, B. Zethelius, A. Larsson, M.C. Irizarry, V. Giedraitis, E. Ronnemaa, M. Degerman-Gunnarsson, B.T. Hyman, H. Basun, L. Kilander, L. Lannfelt, Serum cystatin C and the risk of Alzheimer disease in elderly men, Neurology. 71 (2008) 1072-1079.

[59] J. Oh, H. Lee, J. Song, S.I. Park, H. Kim, Plasminogen activator inhibitor-1 as an early potential diagnostic marker for Alzheimer's disease, Exp. Gerontol. 60 (2014) 87-91.

[60] T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel, K. Lange, Genome-wide association analysis by lasso penalized logistic regression, Bioinformatics. 25 (2009) 714-721.

[61] J.J. Manly, D.M. Jacobs, M. Sano, K. Bell, C.A. Merchant, S.A. Small, Y. Stern, Effect of literacy on neuropsychological test performance in nondemented, education-matched elders, Journal of the International Neuropsychological Society. 5 (1999) 191-202.

[62] J.L. Gatz, S.L. Tyas, P. St John, P. Montgomery, Do depressive symptoms predict Alzheimer's disease and dementia? J. Gerontol. A Biol. Sci. Med. Sci. 60 (2005) 744-747.

[63] L.L. Barnes, D.A. Bennett, Alzheimer's disease in African Americans: risk factors and challenges for the future, Health. Aff. (Millwood). 33 (2014) 580-586.

[64] C. Choi, J. Jeong, J.S. Jang, K. Choi, J. Lee, J. Kwon, K. Choi, J. Lee, S.W. Kang, Multiplex analysis of cytokines in the serum and cerebrospinal fluid of patients with Alzheimer's disease by color-coded bead technology, Journal of Clinical Neurology. 4 (2008) 84-88.

[65] R.L. Nussbaum, Genome-wide association studies, Alzheimer disease, and understudied populations, JAMA. 309 (2013) 1527-1528.

[66] T.O. Obisesan, R.F. Gillum, S. Johnson, N. Umar, D. Williams, V. Bond, J. Kwagyan, Neuroprotection and neurodegeneration in Alzheimer’s disease: role of cardiovascular disease risk factors, implications for dementia rates, and prevention with aerobic exercise in African Americans, International Journal of Alzheimer’s Disease. 2012 (2012).

[67] S. Greenland, M. Gago-Dominguez, J.E. Castelao, The value of risk-factor (“black-box”) epidemiology, Epidemiology. 15 (2004) 529-535.

[68] L.K. Evans, Knowing the patient: the route to individualized care, J. Gerontol. Nurs. 22 (1996) 15-9; quiz 52.

[69] L. Radwin, K. Alster, Individualized nursing care: an empirically generated definition, Int. Nurs. Rev. 49 (2002) 54-63.

[70] P. Ryan, D.R. Lauver, The efficacy of tailored interventions, Journal of Nursing Scholarship. 34 (2002) 331-337.

[71] R. Whittemore, Consequences of Not" Knowing the Patient", Clinical Nurse Specialist. 14 (2000) 75-81.

[72] M. Susser, Does risk factor epidemiology put epidemiology at risk? Peering into the future, J. Epidemiol. Community Health. 52 (1998) 608-611.

[73] P. Skrabanek, Risk factor epidemiology: Science or non-science, Social Affairs Unit, Health, Lifestyle and Environment: Countering the Panic, London: Social Affairs Unit. (1991).

[74] D.G. Seigel, S.W. Greenhouse, Multiple relative risk functions in case-control studies, Am. J. Epidemiol. 97 (1973) 324-331.

[75] K.R. Petronis, J. Samuels, E.K. Moscicki, J.C. Anthony, An epidemiologic investigation of potential risk factors for suicide attempts, Soc. Psychiatry Psychiatr. Epidemiol. 25 (1990) 193-199.

[76] S.B. Johnson, H. Lee, J. Baxter, B. Lernmark, R. Roth, T. Simell, The Environmental Determinants of Diabetes in the Young (TEDDY) study: predictors of early study withdrawal among participants with no family history of type 1 diabetes, Pediatric diabetes. 12 (2011) 165-171.

[77] S.B. Johnson, K.F. Lynch, H. Lee, L. Smith, J. Baxter, B. Lernmark, R. Roth, T. Simell, TEDDY Study Group, At high risk for early withdrawal: using a cumulative risk model to increase retention in the first year of the TEDDY study, J. Clin. Epidemiol. 67 (2014) 609-611.

[78] O. Maimon, L. Rokach, Data mining with decision trees: theory and applications, (2008).

[79] L. Breiman, Random forests, Mach. Learning. 45 (2001) 5-32.

[80] J.H. Friedman, B.E. Popescu, Predictive learning via rule ensembles, The Annals of Applied Statistics. (2008) 916-954.

[81] TEDDY Study Group, The Environmental Determinants of Diabetes in the Young (TEDDY) study: study design, Pediatr. Diabetes. 8 (2007) 286-298.

[82] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society.Series B (Methodological). (1996) 267-288.

[83] J. Liu, J. Chen, J. Ye, Large-scale sparse logistic regression, (2009) 547-556.

[84] S. Kim, K.A. Sohn, E.P. Xing, A multivariate regression approach to association analysis of a quantitative trait network, Bioinformatics. 25 (2009) i204-12.

[85] S. Ma, X. Song, J. Huang, Supervised group Lasso with applications to microarray data analysis, BMC Bioinformatics. 8 (2007) 60.

[86] J. Leclere, G. Weryha, Stress and auto-immune endocrine diseases, Hormone Research in Paediatrics. 31 (1989) 90-93.

[87] F. Saravia-Fernandez, S. Durant, A.E. Hasnaoui, M. Dardenne, F. Homo-Delarche, Environmental and experimental procedures leading to variations in the incidence of diabetes in the nonobese diabetic (NOD) mouse, Autoimmunity. 24 (1996) 113-121.

[88] P.F. Slawson, W.R. Flynn, E.J. Kollar, Psychological factors associated with the onset of diabetes mellitus, JAMA. 185 (1963) 166-170.

[89] G.M. Thernlund, G. Dahlquist, K. Hansson, S.A. Ivarsson, J. Ludvigsson, S. Sjoblad, B. Hagglof, Psychological stress and the onset of IDDM in children, Diabetes Care. 18 (1995) 1323-1329.

[90] R.J. Marshall, The use of classification and regression trees in clinical epidemiology, J. Clin. Epidemiol. 54 (2001) 603-609.

[91] A. Suwa, K. Nishida, K. Utsunomiya, S. Nonen, M. Yoshimura, Y. Takekita, M. Wakeno, A. Tajika, M. Yoshino, Y. Koshikawa, Neuropsychological Evaluation and Cerebral Blood Flow Effects of Apolipoprotein E4 in Alzheimer's Disease Patients after One Year of Treatment: An Exploratory Study, Dementia and geriatric cognitive disorders extra. 5 (2015) 414-423.

[92] M. El Haj, P. Antoine, P. Amouyel, J. Lambert, F. Pasquier, D. Kapogiannis, Apolipoprotein E (APOE) ε4 and episodic memory decline in Alzheimer’s disease: A review, Ageing Research Reviews. 27 (2016) 15-22.

[93] F.B. Baker, S. Kim, Item Response Theory: Parameter Estimation Techniques, CRC Press, 2004.

[94] S.J. Sajjadi, X. Qian, B. Zeng, A.A. Adl, Network-based methods to identify highly discriminating subsets of biomarkers, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 11 (2014) 1029-1037.

[95] C. Barnhart, E.L. Johnson, G.L. Nemhauser, M.W. Savelsbergh, P.H. Vance, Branch-and-price: Column generation for solving huge integer programs, Oper. Res. 46 (1998) 316-329.

[96] M.E. Lübbecke, J. Desrosiers, Selected topics in column generation, Oper. Res. 53 (2005) 1007-1023.

APPENDIX A: COPYRIGHT PERMISSIONS

Below is permission for the use of material in Chapter 2.

Mona Haghighi <monahaghighi@mail.usf.edu>

Request for permission

Carry Koolbergen <C.Koolbergen@iospress.nl> Mon, Apr 25, 2016 at 4:06 AM
To: Mona Haghighi <monahaghighi@mail.usf.edu>

Dear Mona Haghighi,

We hereby grant you permission to reproduce the below mentioned material in **print and electronic format** at no charge subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.

2. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol number, Author(s), Title of article, Pages No., Copyright (Year), with permission from IOS Press".

The final publication is available at IOS Press through http://dx.doi.org/[insert DOI]

3. This permission is granted for non-exclusive world **English** rights only. For other languages please reapply separately for each one required.

4. Reproduction of this material is confined to the purpose for which permission is hereby given.

Yours sincerely

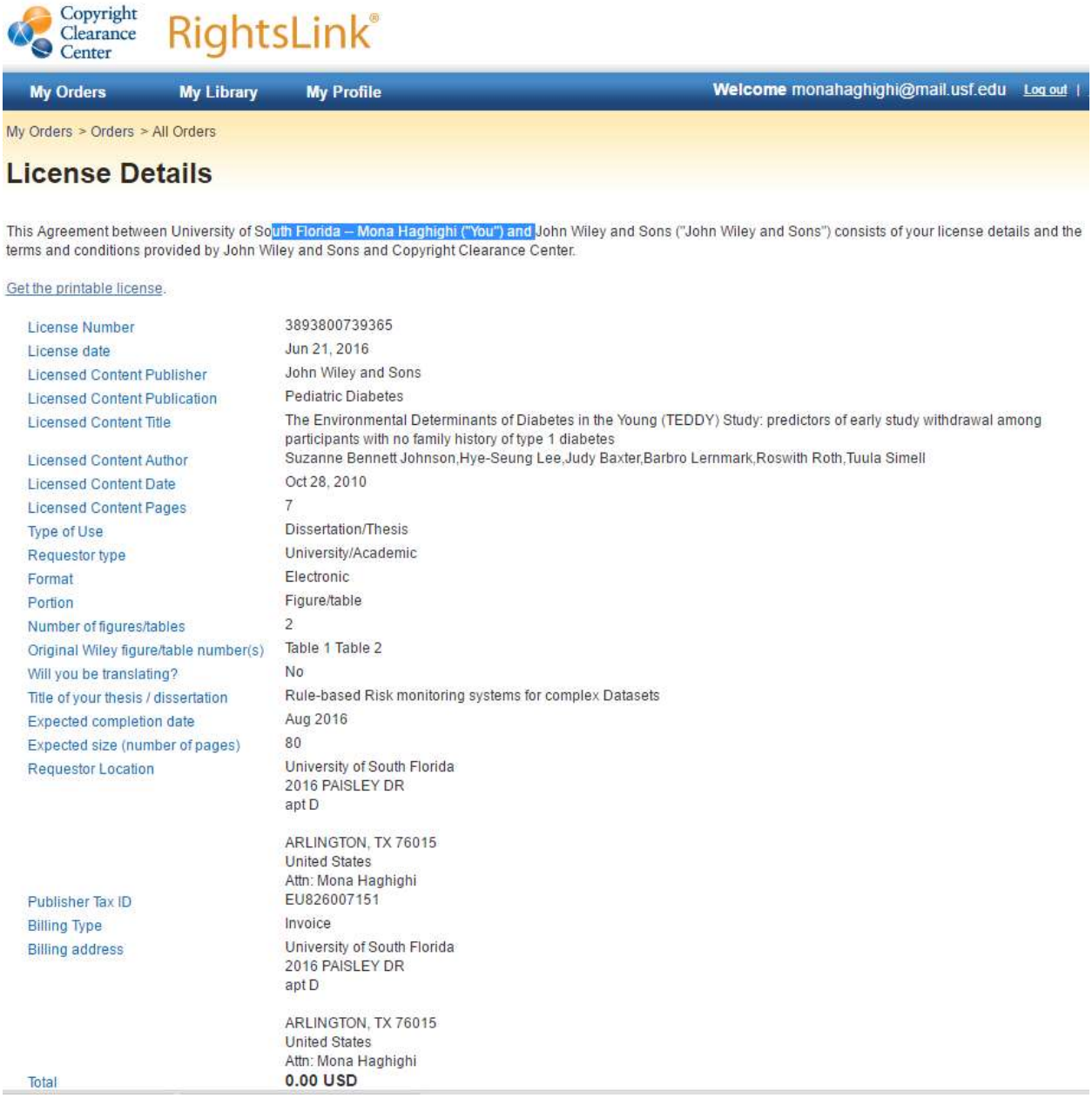
Carry Koolbergen (Mrs.)

Contracts, Rights & Permissions Coordinator

Not in the office on Wednesdays

IOS Press BV

Below is permission for the use of Table 5 and Table 6.



Copyright Clearance Center RightsLink®

My Orders | My Library | My Profile | Welcome monahaghighi@mail.usf.edu | Log out

My Orders > Orders > All Orders

License Details

This Agreement between University of South Florida – Mona Haghighi ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

Get the printable license.

License Number	3893800739365
License date	Jun 21, 2016
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Pediatric Diabetes
Licensed Content Title	The Environmental Determinants of Diabetes in the Young (TEDDY) Study: predictors of early study withdrawal among participants with no family history of type 1 diabetes
Licensed Content Author	Suzanne Bennett Johnson,Hye-Seung Lee,Judy Baxter,Barbro Lernmark,Roswith Roth,Tuula Simell
Licensed Content Date	Oct 28, 2010
Licensed Content Pages	7
Type of Use	Dissertation/Thesis
Requestor type	University/Academic
Format	Electronic
Portion	Figure/table
Number of figures/tables	2
Original Wiley figure/table number(s)	Table 1 Table 2
Will you be translating?	No
Title of your thesis / dissertation	Rule-based Risk monitoring systems for complex Datasets
Expected completion date	Aug 2016
Expected size (number of pages)	80
Requestor Location	University of South Florida 2016 PAISLEY DR apt D ARLINGTON, TX 76015 United States Attn: Mona Haghighi
Publisher Tax ID	EU826007151
Billing Type	Invoice
Billing address	University of South Florida 2016 PAISLEY DR apt D ARLINGTON, TX 76015 United States Attn: Mona Haghighi
Total	**0.00 USD**