IDENTIFYING AND VALIDATING TYPE 1 AND TYPE 2 DIABETIC CASES USING ADMINISTRATIVE DATA: A TREE-STRUCTURED MODEL

by

Wei-Hsuan Lo

BS, National Taiwan University, Taiwan, 2003

MS, CP, National Cheng-Kung University, Taiwan, 2005

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Wei-Hsuan Lo

It was defended on

April 19th, 2010

and approved by

Thesis Advisor:
Roslyn A. Stone, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Vincent C. Arena, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Janice C. Zgibor, RPh, PhD
Assistant Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Kristine Ruppert, RN, MSN, DrPH
Research Associate
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

Roslyn A Stone, PhD

IDENTIFYING AND VALIDATING TYPE 1 AND TYPE 2 DIABETIC CASES USING ADMINISTRATIVE DATA: A TREE-STRUCTURED MODEL

Wei-Hsuan Lo, MS

University of Pittsburgh, 2010

ABSTRACT

Background: Planning, implementing, monitoring, temporal evolution and prognosis differ between type 1 diabetes (T1DM) and type 2 diabetes (T2DM). To date, few administrative diabetes registries have distinguished T1DM from T2DM, reflecting the lack of required differential information and possible recording bias.

Objective: Using a classification tree model, we developed a prediction rule to distinguish T1DM from T2DM accurately, using information from a large administrative database.

Methods: The Medical Archival Retrieval System (MARS) at the University of Pittsburgh Medical Center from 1/1/2000-9/30/2009 included administrative and clinical data for 209,642 unique diabetic patients aged ≥ 18 years. We identified 10,004 T1DM and 156,712 T2DM patients as probable or possible cases, based on clinical criteria. Classification tree models were fit using TIBCO Spotfire S+ 8.1 (TIBCO Software). We used 10-fold cross-validation to choose model size. We estimated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of T1DM.

***Results*:** The main predictors that distinguished T1DM from T2DM include age < 40 vs. ≥ 40 years, ICD-9 codes of T1DM or T2DM diagnosis, oral hypoglycemic agent use, insulin use, and

episode(s) of diabetic ketoacidosis diagnosis. History of hypoglycemic coma, duration in the MARS database, in-patient diagnosis of diabetes, and number of complications (including myocardial infarction, coronary artery bypass graft, dialysis, neuropathy, retinopathy, and amputation) are ancillary predictors. The tree-structured model to predict T1DM from probable cases yields sensitivity (99.63%), specificity (99.28%), PPV (89.87%) and NPV (99.71%).

Conclusion: Our preliminary predictive rule to distinguish between T1DM and T2DM cases in a large administrative database appears to be promising and needs to be validated. The ***public health significance*** is that being able to distinguish between these diabetes subtypes will allow future subtype-specific analyses of cost, morbidity, and mortality. Future work will focus on ascertaining the validity and generalizability of our predictive rule, by conducting a review of medical charts (as an internal validation) and applying the rule to another MARS dataset or other administrative databases (as external validations).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my advisors, Dr. Stone, Dr. Zgibor, Dr. Ruppert and Dr. Vincent for all their time and support while preparing this thesis.

I would especially like to express my gratitude to John McMichael for taking time to obtaining the MARS data for me. Without his efforts, all of the analyses and this thesis could not be accomplished.

1.0 INTRODUCTION

Diabetes is one of the most costly and burdensome chronic diseases of our millennium. According to the latest World Health Organization report, more than 220 million people worldwide suffer from diabetes.[1] In 2005, an estimated 1.1 million people died from diabetes.[2] This number will very likely double by 2030.[1] As of 2007 in the United States, 23.6 million people (7.8% of the population) have diabetes [3] and total national associated costs with diabetes are exceeded $218 billion, including $174.4 billion for diagnosed diabetes, $18 billion for undiagnosed diabetes, $25 billion for pre-diabetes, and $636 million for gestational diabetes.[4] Diabetes and its complications represent a major burden and pose a major challenge to healthcare systems. Therefore, public health prevention and intervention are urgently needed.

An increasing amount of scientific literature is now available on producing accurate information from administrative data. Advantages of use of administrative data to determine disease incidence include feasibility, accessibility and low cost. However, straightforward use of administrative data can produce biased information on cases of chronic disease like diabetes. Other challenges of using administrative data include representativeness of the population and multiple episodes of health services utilization by a single patient.

There are four types of diabetes: Type 1 diabetes (T1DM), Type 2 diabetes (T2DM), gestational diabetes mellitus (GDM), and diabetes from other causes. This thesis will focus on

T1DM and T2DM. For outcomes evaluation, it is important to distinguish T1DM from T2DM. Planning, implementing, monitoring of appropriate interventions, temporal evolution, complications and prognosis differ between T1DM and T2DM. Health care utilization and outcome also differ by types of diabetes. Accurate information about the magnitude, distribution, and types of diabetes are needed in order to inform policy and support health care evaluation. Investigators in health services research often turn to the International Classification of Diseases (ICD) diagnosis codes in administrative records to study the effect of health care delivery on disease outcomes. However, the ability of administrative records to distinguish T1DM and T2DM is limited due to the lack of required information, definitive diagnosis in clinical practice, and possible recording bias (e.g., coders enter a non-specific ICD code).[5-6] A critical step in using administrative data to study T1DM and T2DM is to develop and validate methods to distinguish T1DM and T2DM accurately.

Decision tree methods, also called recursive partitioning methods, are analytic strategies that were developed to classify or segment a target population for the purpose of clinical diagnosis and/or prognosis in public health. Classification and regression trees (CART), one type of decision tree methodology, is nonparametric procedure developed by Brieman and colleagues.[7] CART identifies easily defined, mutually exclusive population subgroups whose members share characteristics that are important predictors of the outcome of interest. CART can efficiently segment a population into meaningful subsets, which allows researchers to easily identify segments of a population that are more or less likely to exhibit the outcome. Advantages of tree-based methods are that they do not require a parametrical specification of the relationship between the predictors and the outcome. Additionally, assumptions of linearity that are made in

linear and logistic models are not required for tree-based methods. A major strength is that tree-based methods are adept at identifying important interactions between predictor variables.

In this thesis, we used administrative data consisting of known T1DM and T2DM patients, including two sub-cohorts of probable and possible cases, from the Medical Archival Retrieval System (MARS) at the University of Pittsburgh Medical Center (UPMC) and built a tree-structured model to distinguish T1DM and T2DM cases using TIBCO Spotfire S+ 8.1 (TIBCO Software).[8] We used V-fold cross validation method to choose the model size. Distinguishing variables between T1DM and T2DM were obtained from the tree-based model, and sensitivity and specificity were estimated. In the future, we aim to apply the preliminary predictive model to another MARS dataset, and conduct a review of medical charts as an internal validation. This will ultimately lead to separate analyses on processes and outcomes of T1DM and T2DM cases.

1.1 INTRODUCTION TO DIABETES MELLITUS

The classification of diabetes includes four clinical classes: (1) T1DM results from β-cell destruction, leading to absolute insulin deficiency; (2) T2DM results from a progressive insulin secretory defect on the background of insulin resistance; (3) other specified types of diabetes due to other causes, e.g., genetic defects in β-cell function, genetic defects in insulin action, disease of the exocrine pancreas (such as cystic fibrosis), and drug- or chemical-induced (such as in the treatment of AIDS or after organ transplantation); (4) GDM is diabetes diagnosed during

pregnancy. We do not consider GDM or other specified types of diabetes as diabetes cases in this thesis.

1.2 POTENTIAL IDENTIFYING FACTORS BETWEEN TYPE 1 (T1DM) AND TYPE 2 DIABETES (T2DM)

T2DM account for approximately 90% to 95% of prevalent diabetes, and T1DM about 5% to 10%. A step in understanding associations between T1DM and T2DM and disease outcomes is to develop methods to accurately identify individuals with T1DM or T2DM in administrative databases. Potential differential characteristics or risk factors between T1DM and T2DM are summarized in **[Table 1]**.

T1DM (previously known as insulin-dependent or childhood-onset diabetes) is characterized by a lack of insulin production. In the United States, the annual incidence is 15 to 18 cases per 100,000 of the childhood population.[9] Males and females appear to be almost equally affected.[10] There is no apparent correlation with socioeconomic status. Data on age-standardized incidence in relation to racial or ethnic background indicate a range of more than 35 new cases annually per 100,000 population in Finland (45/100,000/year) and Sardinia (38.8/100,000/year) to less than 1 per 100,000 in China and parts of South America.[9-11] In the United States, the occurrence of T1DM in blacks had previously been reported to be only between one-third and two thirds of that in whites. More recent data suggest that the incidence of T1DM in African Americans may be as high as in White Americans.[12] However, it is not clear this new increase in incidence among African Americans is exclusively T1DM or includes

cases of T2DM presenting in ketoacidosis and thus misclassified. Peaks of presentation occur in two age groups: at 5 to 7 years of age and at the time of puberty. The first peak corresponds to the time of increased exposure to infectious agents coincident with the beginning of school. The latter corresponds to the pubertal growth spurt induced by increased pubertal growth hormone secretion and gonadal steroids that antagonize insulin action. Emotional stresses that accompany puberty also have been implicated. Most T1DM cases are younger than 20 years at diagnosis and present in diabetic ketoacidosis (DKA).[13] Symptoms include excessive excretion of urine (polyuria), thirst (polydispsia), constant hunger, weight loss, vision change and fatigue. Ketoacidosis is responsible for the initial presentation of many (about 25 to 40%) diabetic children and likely to be present more often in children younger than 5 years of age. Diabetic ketoacidosis exists when there is hyperglycemia (glucose 300 mg/dl), ketonemia (ketones strongly positive at greater than 1:2 dilution of serum), acidosis (pH 7.30 or less and bicarbonate 15 mEq/L or less), glucosuria, and ketonuria in addition to the clinical features of tachypnea (Kussmaul respiration) and cerebral obtundation.[10] These symptoms may occur suddenly. Measurement of C-peptide kinetics or of urinary excretion of C-peptide can be used as an index of endogenous insulin secretion.[14] T1DM has long been known to have an increased prevalence among persons with such autoimmune disorders as Addision disease, celiac disease and Hashimoto thyroiditis.[15] Evidence from T1DM prevention studies suggests that measurement of islet autoantibodies identifies individuals who are at risk for developing T1DM.[16] T1DM is known to be associated with an increased frequency of certain histocompatibility loci antigens (HLAs), in particular, DR3 and DR4. Although no presently available single marker or test can accurately predict T1DM, a combination of immune and genetic markers may provide predictability.[17-18] Without daily administration of insulin, T1DM is rapidly fatal.

Recommended therapy for T1DM consists of the following components: (1) use of multiple dose insulin injections (3-4 injections per day of basal and prandial insulin) or continuous subcutaneous insulin infusion (CSII, or insulin pump therapy); (2) matching of prandial insulin to carbohydrate intake, pre-meal blood glucose, and anticipated activity; and (3) for many patients (especially if hypoglycemia is a problem), use of insulin analogs.[10, 16] Another synthetic analogue of amylin, pramlintide, also is available as an injectable agent combined with insulin for treating T1DM.

T2DM (formerly called non-insulin-dependent or adult-onset diabetes) results from the body's ineffective use of insulin. T2DM is largely the result of excess body weight and physical inactivity, which cause insulin resistance. About 80 to 90% of persons with T2DM are overweight or have metabolic syndrome characteristics, but some who are leaner and more active do not have the metabolic syndrome. Symptoms may be similar to those of T1DM, but are often less marked. As a result, the disease may be diagnosed several years after onset, once complications have already arisen. About one half of patients with newly diagnosed T2DM have established chronic complications.[19] T2DM is more common in women, and in certain racial and ethnic groups including African Americans, Hispanics and Native Americans.[19] T2DM has been viewed as a disorder of aging, with an increasing prevalence with age. This remains true today, even though a disturbing trend has become apparent in which T2DM in children is rising dramatically.

The classification of diabetes into its two most prominent types (T1DM and T2DM) seems straightforward in theory but in practice is increasingly confusing as more Americans become overweight. Although T1DM patients are traditionally lean, many are now overweight and have metabolic syndrome characteristics. C-peptide measurements are not very helpful for

those who are difficult to classify, but measuring three antibodies, including IA-2 (islet antigen 512), anti-GAD_{65} (glutamic acid decarboxylase), and anti-insulin antibodies in high titers, can clarify a diagnosis of latent T1DM. Younger age at onset, ideal or lean body habitus, severe loss of glycemic control with or without ketonemia, and weight loss all suggest insulin deficiency but might not be definitive. ICD-9 codes with 250.x1 or 250.x3 seems to assign to the T1DM specifically, while 250.x0 or 250.x2 to T2DM or other unspecified diabetes. However, ICD-9 codes might not be accurately entered; therefore using ICD-9 codes to distinguish types of diabetes may be unreliable. Currently, no methods are available to accurately distinguish types of diabetes in administrative data.

Table 1: Potential Differential Characteristics or Risk Factors of Type 1 and Type 2 Diabetes

Potential Differential Characteristics or Risk factors	Type 1 Diabetes (T1DM)	Type 2 Diabetes (T2DM)
Age of onset	Mainly in childhood (5-7 years) or puberty	Adult; pubertal in some children
Onset	Acute; Severe	Mild-severe; often insidious
Insulin secretion	Deficiency	Variation
C-peptide can be used as an index of endogenous insulin secretion[§] (Normal value: fasting: 0.78-1.89 ng/mL or 0.26-0.62 nmol/L)	Decreasing or lack	Normal or higher than normal
Insulin sensitivity	Normal	Decreased or resistant
Symptoms for diagnosis of diabetes*	• Usually the symptoms (polyuria, polydipsia, polyphagia and weight loss) over a several-week period are common • Plasma glucose concentration usually in the range 300 to 500 mg/dL. • A sodium value less than 120 mm/L is usually associated with severe hypertriglycemia that can lead to spurious hyponatremia.	• The presentation of T2DM can be more subtle and sometimes clinical silent.
Islet antigen /auto-antibodies / genetics	• High titers of islet cell, GAD, IA2 • Type 1A (immune-mediated): 90% with positive islet autoantibodies o Genetics: 30-50% DR3 and DR4 in 90% non-Hispanic white children o Genetics: 90% DR3 or DR4 in 50% black children o Genetics: <3% DQB1*0602 in Latin American children • Type 1B (other forms of diabetes with severe insulin deficiency): with negative islet autoantibodies	Negative islet auto-antibodies; with unknown genetics
Insulin dependence	Permanent	Temporary; may occur later

Table 1 continued

Potential Differential Characteristics or Risk factors	Type 1 Diabetes (T1DM)	Type 2 Diabetes (T2DM)
Medication	Mainly insulin therapy; may or may not combined with insulin analogs or amylin agonist	Mainly single or combined OHA and/or combined with insulin or amylin agonist
Association with obesity	Most not	More likely to be overweight (BMI >25 kg/m^2)
Diabetic ketoacidosis (DKA): based on the blood pH (< 7.3), serum bicarbonate (< 10 mg/dL or <15-18 mmol), and the presence of significant ketosis (ketonemia, ketouria), glucosuria. In prolonged and severe cases, Kussmaul respiration is present (an odor of acetone on the breath).	• 25-40 % of children with newly diagnosed T1DM present with DKA (children who are younger (less than 5 years), without a first-degree relative with T1DM, and from a family of lower socioeconomic status are at higher risk of DKA at onset of T1DM). • Majority of DKA episode occur in patients with established diabetes	Occasionally develop: • May occur in hyperglycemic hyperosmolar state of T2DM • ¼ adolescents with T2DM have ketoacidosis at presentation.
Hypoglycemic coma: mainly occurs in those being treated with insulin (differential diagnosis from hyperglycemic and ketoacidotic diabetec coma: hypoglycemic coma lacks of ketoacidotic hyperpnea or dehydration and can be improve by glucose injection)	More common situation in patients with T1DM has injected too much insulin in relation to the amount of carbohydrates consumed, or has not reduced the insulin dose despite increase physical activity etc.	Rarer among T2DM patients being treated with insulin alone or in combination with other medication
Associated autoimmune Disease	Most patients with Type 1A diabetes have one or more additional auto-immune disease. The most common associated disorders are thyroid autoimmunity (Grave disease or Hashimoto's thyroditis) or cliac disease or Addison disease.	Not associated
Environmental risk factor	Congenital rubella infection is related with Type 1A diabetes	NA

Table 1 continued

Potential Differential Characteristics or Risk factors	Type 1 Diabetes (T1DM)	Type 2 Diabetes (T2DM)
Specific diagnostic ICD-9 code	***ICD-9 codes (250.X1 or 250.X3) may include T1DM*** • **250.01**: Diabetes mellitus without mention of complication, type1 [juvenile type], not stated as uncontrolled • **250.03**: Diabetes mellitus without mention of complication, type 1 [juvenile type], uncontrolled • **250.11**: Diabetes with ketoacidosis, type 1 [juvenile type], not stated as uncontrolled • **250.13**: Diabetes with ketoacidosis, type 1 [juvenile type], uncontrolled • **250.21**: Diabetes with hyperosmolarity, type 1 [juvenile type], not stated as uncontrolled • **250.23**: Diabetes with hyperosmolarity, type 1 [juvenile type], uncontrolled • **250.31**: Diabetes with other coma, type 1 [juvenile type], not stated as uncontrolled • **250.33**: Diabetes with other coma, type 1 [juvenile type], uncontrolled • **250.41**: Diabetes with renal manifestations, type 1 [juvenile type], not stated as uncontrolled • **250.43:** Diabetes with renal manifestations, type 1 [juvenile type], uncontrolled • **250.51**: Diabetes with ophthalmic manifestations, type 1 [juvenile type], not stated as uncontrolled • **250.53**: Diabetes with ophthalmic manifestations, type 2 [juvenile type], uncontrolled • **250.61**: Diabetes with neurological manifestations, type 1 [juvenile type], not stated as uncontrolled • **250.63**: Diabetes with neurological manifestations, type 1 [juvenile type], uncontrolled • **250.71**: Diabetes with peripheral circulatory disorders, type 1 [juvenile type], not stated as uncontrolled	***ICD-9 codes (250.X0 or 250.X2) may include T2DM*** • **250.00**: Diabetes mellitus without mention of complication, type 2 or unspecified type, not stated as uncontrolled • **250.02**: Diabetes mellitus without mention of complication, type 2 or unspecified type, uncontrolled • **250.10**: Diabetes with ketoacidosis, type 2 or unspecified type, not stated as uncontrolled • **250.12**: Diabetes with ketoacidosis, type 2 or unspecified type, uncontrolled • **250.20**: Diabetes with hyperosmolarity, type 2 or unspecified type, not stated as uncontrolled • **250.22**: Diabetes with hyperosmolarity, type 2 or unspecified type, uncontrolled • **250.30**: Diabetes with other coma, type 2 or unspecified type, not stated as uncontrolled • **250.32**: Diabetes with other coma, type 2 or unspecified type, uncontrolled • **250.40**: Diabetes with renal manifestations, type 2 or unspecified type, not stated as uncontrolled • **250.42**: Diabetes with renal manifestations, type 2 or unspecified type, uncontrolled • **250.50**: Diabetes with ophthalmic manifestations, type 2 or unspecified type, not stated as uncontrolled • **250.52**: Diabetes with ophthalmic manifestations, type 2 or unspecified type, uncontrolled • **250.60**: Diabetes with neurological manifestations, type 2 or unspecified type, not stated as uncontrolled

Table 1 continued

Potential Differential Characteristics or Risk factors	Type 1 Diabetes (T1DM)	Type 2 Diabetes (T2DM)
	• **250.73**: Diabetes with peripheral circulatory disorders, type 1 [juvenile type], uncontrolled • **250.81**: Diabetes with other specified manifestations, type 1 [juvenile type], not stated as uncontrolled • **250.83**: Diabetes with other specified manifestations, type 1 [juvenile type], uncontrolled • **250.91**: Diabetes with unspecified complication, type 1 [juvenile type], not stated as uncontrolled • **250.93**: Diabetes with unspecified complication, type 1 [juvenile type], uncontrolled	• **250.62**: Diabetes with neurological manifestations, type 2 or unspecified type, uncontrolled • **250.70**: Diabetes with peripheral circulatory disorders, type 2 or unspecified type, not stated as uncontrolled • **250.72**: Diabetes with peripheral circulatory disorders, type 2 or unspecified type, uncontrolled • **250.80**: Diabetes with other specified manifestations, type 2 or unspecified type, not stated as uncontrolled • **250.82**: Diabetes with other specified manifestations, type 2 or unspecified type, uncontrolled • **250.90**: Diabetes with unspecified complication, type 2 or unspecified type, not stated as uncontrolled • **250.92**: Diabetes with unspecified complication, type 2 or unspecified type, uncontrolled

***:** Diagnosis of diabetes: (1) *Symptoms (polyuria, polydipsia, unexplained weight loss) and a casual plasma glucose (any time of day without regard to time since last meal) ≥ 200 mg/dL (11.1 mmol/L) or (2) Fasting (no caloric intake for at least 8 hours) plasma glucose ≥ 126 mg/dL (7.0 mmol/L) or (3) 2-hour plasma glucose ≥200 mg/dL (11.1 mmol/L) during an oral glucose tolerance test (glucose load of 75 g anhydrous glucose dissolved in water or 1.75 g/kg body weight if weight < 43 kg). The last two criteria should be confirmed on a second day if child /adolescent are asymptomatic.*

§: The level of C-peptide in the blood must be read with the results of a blood glucose test. Both these tests will be done at the same time.

Abbreviations: ***BMI***: body mass index; ***DKA***: diabetic ketoacidosis; ***GAD***: glutamic acid decarboxylase ***; HLAs***: histocompatibility loci antigens; ***IA2***: islet antigen 512; ***NA***: Not available; ***OHA***: oral hypoglycemic agents

2.0 REVIEW OF THE REVELENT LITERATURE

In this section, the current diabetic registries and their limitations were briefly reviewed. We mainly focused on the statistical methods in the thesis.

2.1 DIABETES REGISTRY AND ITS LIMITATIONS

Information systems containing data on the level of care, including both processes and outcomes, offer a valuable tool for health systems. These information systems allow for continuous quality improvement, practice change, and improved outcomes. However traditional systems used to track quality of care, such as chart audit or patient and provider self-report, lack internal validity, are expensive, and inefficient.[6] Administrative data also pose challenges. The outcome as well as the explanatory variables may be continuous or discrete. Relationships between variables often are nonlinear and involve complex interactions. Missing values for both explanatory and outcome variables are fairly common, and outliers usually exist. In addition, administrative data applied in clinical fields usually demand methods that are both easy to implement and easy to interpret.

[Table 2] is adapted from Zgibor et al.[6] and summarized a number of diabetes registries which are currently available in the United State and Canada. Most of the databases are derived from homogenous populations from one health plan or insurer, which limit the translation of methodologies to other populations. For example, the Veterans Health Administration (VA)

database includes only veterans,[20] while the Center for Medicare and Medicaid Services serves mainly the elderly.[21] Therefore, extrapolating results beyond these populations may not be appropriate. Zgibor et al. developed a registry including more general population in a variety of settings, and used two or more indicators or an out-patient diagnosis to identify diabetic patients with 99% to 100% sensitivity and 96% to 97% positive predictive value (PPV).[6] Zgibor et al. pointed out using only one criterion would likely incorrectly classify patients (PPV: 21% to 95%), except for outpatient diagnosis code for diabetes.[6] However, none of the registries has further distinguished T1DM from T2DM, reflecting the lack of definitive diagnosis in clinical practice and possible recording bias. This research will focus on the methodology to accurately distinguish T1DM from T2DM in an administrative and clinical diabetic database.

Table 2: Summary of Populations, Case Definitions, and Validation Criteria for Existing Diabetes Database

Organization, Year of Publication	Population	Case Definition	Validation Process	Positive Predictive Value	Sensitivity	Comment/Biases / Limitations
Kaiser CA, 1997[22]	85,209	Administrative database Diabetes medications/supplies (since 1994) or HbA1c ≥ 6.7 % (since 1991) or Primary or Secondary discharge diagnosis of diabetes (since 1971) or ER visit (since 1991)	Mailed survey to ~20,000 HMO enrollees (self-reported diabetes)	Not reported	90%	One insurer
Managed Care Organizations' Diabetes Surveillance System (3 HMO's), 1998[23]	16,363	Administrative database (1993) Primary or secondary inpatient diagnosis of diabetes or Diabetes medications or 2 or more outpatient visits or 2 HbA1c test	Not reported	Not reported	≥2: 82%; ≥3: 94%; 4: 100%	No length of stay data for out of HMO hospitalizations, or race
Veteran's Administration, 1998[20]	139,646	Pharmacy derived database (1994) Diabetes medication or supplies	Pharmacy department checked drug nomenclature	Not reported	Not reported	Homogeneous population
Medicare, 1999[24]	1,941,517	Medicare claims data (1992-1993) ≥1 indicator of an ICD-9 code for diabetes (including complications)	1,135 of 7,562 individuals that responded to a survey (self-reported)	Not reported	1 year period: up to 71.6% 2 year period: up to 79.1%	One insurer Using different Medicare claims file can construct a claims-based algorithm for identifying persons with diabetes with adequately sensitivity and high specificity, and good reliability.
Puget Sound, 1999 (detection of diabetes complications)[25]	8,905	Administrative database Diabetes medications/supplies or GHb ≥ 7.5% or Discharge diagnosis of diabetes or random glucose > 200 mg/dl FPG >140 mg/dl	Random selection of 471 individuals with diabetes	90.4%	Not reported	One insurer

Table 2 continued

Organization, Year of Publication	Population	Case Definition	Validation Process	Positive Predictive Value	Sensitivity	Comment/Biases / Limitations
Kaiser OR, 2000[26]	13,099	Diabetes medication, testing supplies, discharge notes, diabetes education contacts	Random selection of 425 medical charts	Not reported	99%	One insurer
Indian Health Service, 2001[27]	6,870	Electronic health record > 1 ICD-9 code for diabetes Medication /supplies ≥ 2 glucoses > 200 mg/dl	Chart review on 462 patients	94%	92%	Homogeneous population
Kaiser CA, 2001[28]	57,222	From 1994-1995 Pharmacy prescriptions for diabetes medications Abnormal HbA1c values (≥6.7%) in laboratory files Primary hospital diagnoses of diabetes Emergency department records of diabetes as the reason for visit	Matched with >1,500 diabetes self-reported (through mailing)	97.5%	90%	One insurer
Zgibor JC, 2007 (UPMC) [6]	99,144	UPMC electronic clinical, administrative, and financial databases ICD-9 code 250 for either inpatient, emergency room or outpatient visits (treated as three indicators) or Any HbA1c result (regardless value) or A blood glucose > 200 mg/dl or Use of any diabetes medication	Two validation studies: three outpatient clinics (validation population n = 254) and general internal medicine clinic (validation population n = 95)	Single: 94% , 95%; ≥ 2: 96%, 97%	≥ 2: 100%	Advantages: 80 insurers and heterogeneous population Disadvantages: No independent source or gold standard to quantify the cases might be missing; Missing laboratory data if samples were sent to non-UPMC laboratories; Medications data were only available for inpatients

Table 2 continued

Organization, Year of Publication	Population	Case Definition	Validation Process	Positive Predictive Value	Sensitivity	Comment/Biases / Limitations
Asghari S, PRIMUS group (Quebec, Canada), 2009[5]	263,213	RAMQ and registry Med-Echo database (2002)- for incident diabetes cases NDDS definition: Two physician claims with a diagnosis on 2 different days within 2 years or one hospital discharge with a diabetes diagnosis code in any field among 16 diagnosis codes; To differentiate between prevalent and incident cases, a minimum diabetes-free retrospective observation (clearance) period was used.	Not reported	Not reported	Not reported	Advantage: Incidence of diabetes and kappa agreement by exclusion criteria and clearance period for 5 years and 10 years are 1.3% and 0.957, and 1.24% and 1.00, respectively. A clearance period of 5 years or more is sufficient to improve performance to estimate of incident diabetes.

This table is adapted from Zgibor et al. Diabetes Research & Clinical Practice 2007; 75:313-319, Table 1

2.2 THE CLASSIFICATION AND REGRESSION TREES (CART) METHOD OR TREE-STRUCTURED METHODS

Tree-structured methods were introduced by social scientists. The use of trees in regression dates back to the Automatic Interaction Detection (AID) program developed by Morgan and Sonquist in 1963 as a sequential procedure for the analysis of survey data.[29] AID was suitable only for small to medium size data sets, and it could generate only regression trees. Kass extended the methodology of AID to categorical data called CHAID in 1980. The classic CART algorithm was introduced by Breiman, Friedman, Olshen, and Stone[7] and is similar to the AID in that it uses binary splits to achieve the final classification. However, the splitting mechanism used in CART is different from that of AID or CHAID. CART uses the Gini-index to measure the homogeneity of cases at a leaf node, while AID uses the unexplained sum of squares and CHAID uses a Chi-square measure to evaluate splits based on the significance of differences in response distributions between groups.

A classification or regression tree is the collection of many rules determined by a procedure known as "***recursive partitioning***", while ***"linear combinations***" are the primary mode of expressing relationships between variables in classical logistic and linear regression analyses. A tree is typically shown growing upside down, beginning at its top node of the tree (called the root). Trees are typically fit the "binary" recursive partitioning. In the binary tree, each group or subgroup (parent node) within the scheme can potentially be further subdivided into two groups (two child nodes). The term recursive is used to indicate that each child node

will, in turn, become a parent node. At minimum, constructing a tree involves making decisions about three major issues. The first choice is how splits are to be made: which explanatory variables will be used and where the split will be imposed. A tree is grown that overfits the data. The second choice involves determining appropriate tree size from an overfitted tree, generally using a pruning process. The third choice is to determine how application-specific costs should be incorporated. This might involve decisions about assigning varying misclassification costs or accounting for the cost of model complexity. Then, a final tree is selected that represents the best estimate of the tree function for the data. Throughout this thesis, we use the notation that Breiman et al.[7] used to describe the original CART methodology.

2.2.1 Notation and terminology

A collection of M measurements (e.g. age, sex…., etc.), which is referred to as a measurement vector, is denoted by $\boldsymbol{x} = (x_1, x_2, \ldots\ldots, x_M)$. The measurement space X is defined as containing all possible measurement vectors. Suppose that the cases or objects can be partitioned into J classes. Number the classes 1, 2, ……, J and let C be the set of all disjoint and exhaustive classes; i.e., $C = \{1, 2, \ldots..., J\}$. Therefore, given any $\boldsymbol{x} \in \boldsymbol{X}$, each possible measurement vector x will be uniquely assigned to a single class in C. A classifier (or classification rule) is a function $d(\boldsymbol{x})$ defined on $\boldsymbol{X}$ so that every $\boldsymbol{x}$, $d(\boldsymbol{x})$ is equal to one of the particular classes $1, 2, \ldots\ldots\ldots, J$. In other words, $d(\boldsymbol{x}) \in \{1, 2, \ldots..., J\} \; \forall \; \boldsymbol{x} \in \boldsymbol{X}$. A classifier is constructed based on past experience or data, summarized by a *learning sample (or training dataset).* This consists of the measurement data on N cases observed in the past along with their actual classification. Thus, the learning sample is denoted by $\boldsymbol{L}$, (1)

$$L = \{(x_1, j_1), \ldots\ldots, (x_N, j_N)\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

An important criterion for a good classification procedure is that it not only produce accurate classifiers (within the limits of the data), but that it also provide insight and understanding into the predictive structure of the data.

Given a classifier, that is, given a function d(x) defined on X taking values in C, we denoted by **R*(d)** as its "***true misclassification rate***". Two questions arise: does this classifier express the "truth" and how accurate is the estimate? Four types of internal estimates were developed to determine the accuracy. These methods are ***resubstitution estimate, test sample estimation, cross-validation, and bootstrap estimation.*** However, Breiman, et al.[7] found that the bootstrap estimation might not work well when applied to tree structure classifiers.

The first method utilizing ***"resubstitution estimate"*** is the easiest, most commonly used, but least accurate. After constructing a classifier *d(x)* and all of the cases in ***L*** are run through *d(x)*, the proportion of cases misclassified is the resubstitution estimate. Define the indicator function *X(·)* to be *1* if the statement inside the parentheses is true, otherwise zero. The resubstitution estimate, denoted R(d), is (2)

$$R(d) = \frac{1}{N}\sum_{n=1}^{N} X(d(X_n) \neq j_n) \ldots\ldots\ldots\ldots\ldots (2)$$

However, the problem with the resubstitution estimate is that it is computed using the same data used to construct d, instead of an independent sample. All classification procedure attempts to minimize R(d), which tends to underestimate the true misclassification rate R*(d) of *d(x)*.

The ***"test sample method"*** randomly divides the learning sample *L* into two parts L_1 and L_2. Only the cases in L_1 are used to construct *d(x)*. Then the cases in L_2 are used to estimate

R*(d). If N_2 is the number of cases in L_2, the test sample estimate, $R^{ts}(d)$, is given by (3)

$$R^{ts}(d) = \frac{1}{N_2} \sum_{(x_n, j_n) \in L_2} X(d(X_n) \neq j_n) \dots\dots\dots\dots\dots (3)$$

Frequently, L_1 consists of 2/3 of the data in the learning sample and L_2 consists of the other 1/3, but we do not know of any theoretical justification for this 2/3, 1/3 split. The classifier *d(x)* is constructed by using L_1 and the misclassification rate is estimated by finding the proportion of cases misclassified by *d(x)* in L_2. Despite reducing the bias found in the resubstitution estimates, the disadvantage of test sample approach is that it reduces effective sample size. This disadvantage is a minor difficulty if the sample size is large because a more accurate estimate is obtained.

The last method, ***"cross-validation"***, uses the entire sample to construct *d(x)*. This method works by randomly dividing the data into V equal-sized subsets and holding out one subgroup at a time to construct an independent *d(x)*. We used cross-validation to estimate the accuracy in this thesis, and the details related to cross-validation will be discussed further in Section 2.2.5.

2.2.2 Constructing a tree

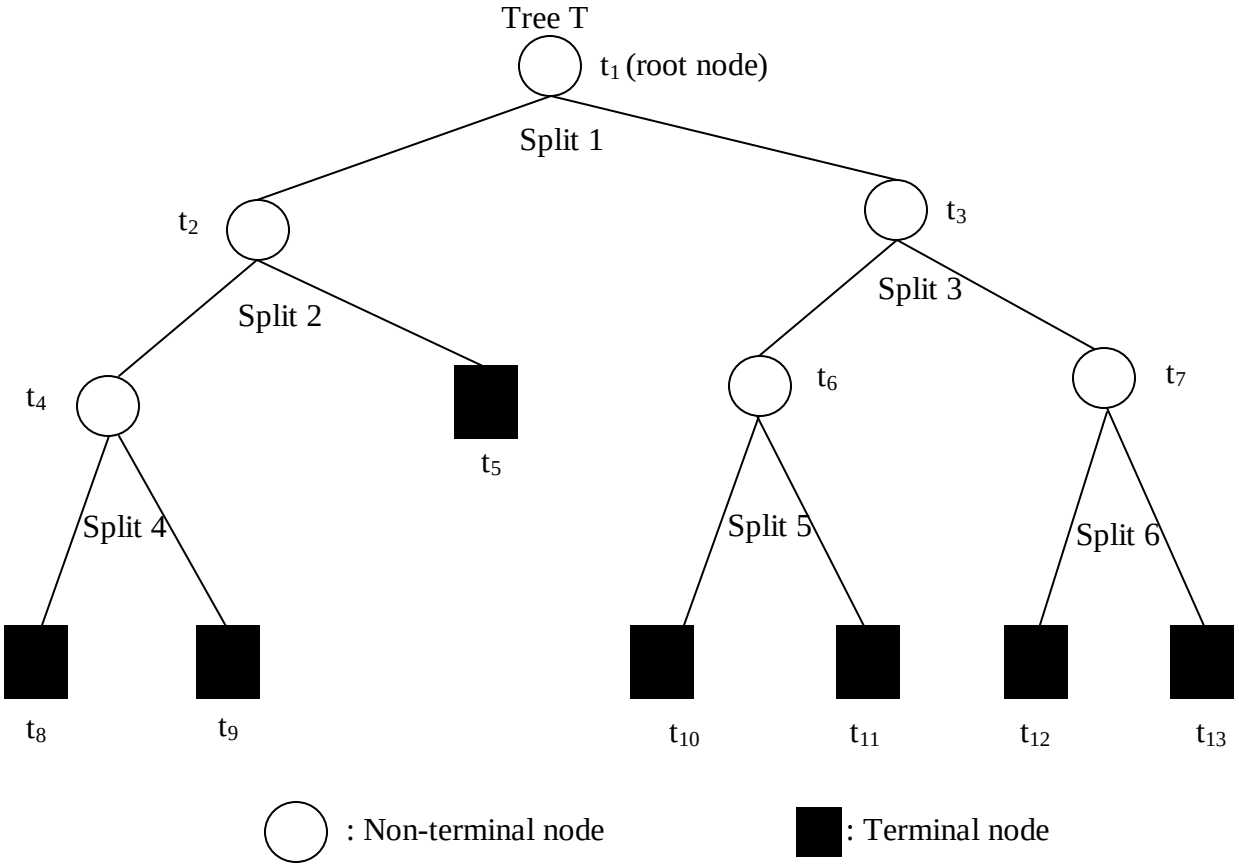

Figure 1. An Example of a Basic Tree

The entire of construction of a tree revolves around three elements:

(1) The selection of the splits.

(2) The decision to declare a node as terminal or to continue splitting it.

(3) The assignment of each terminal class to a node.

Both Breiman et al.'s CART method and the S functions represent tree structured classifiers graphically by a binary tree, denoted by *T* (see Figure 1).[7] Tree structured classifiers are constructed by splitting the dataset into two subsets. All of the observations in a dataset start

in a “root node (or called parent node)”. For a continuous variable, the allowed splits are of the form “$x_j < t$ versus $x_j \geq t$”. For a categorical variable, the splits are of the same type “$x \in$ subset i”. The fundamental idea is to select each split of a subset so that the data in each of the descendant subsets are “purer” than the data in the parent subset. These splits are generated in the following fashion. Starting with the first variable, x_1, CART splits a variable at all of its possible split points. At each possible split point of the variable, the sample splits into two binary child nodes. Observations with a "yes" response to the question posed are sent to the left node t_L and the "no" responses are sent to the right node, t_R. Some algorithms allow a linear combination of continuous variables to be split, and Boolean combinations to be formed of binary variables.[30] CART then applies its goodness of split criteria to each split point and evaluates the decrease in impurity (or heterogeneity) due to the split.

In ***Figure 1***, T will be referred to as a tree and each element of T will be referred to as a node. The minimum element of a tree T is called the root node of T, denoted by root (T). If $p, t \in T$ and t = left(p) or t = right(p), then p is called the parent of t. The root of T has no parent, but every other node has a unique parent. The data in the “root node” (t_1) is first split into two subsets, t_2 and t_3. The subsets are split into smaller subsets at the largest decrease of the node impurity (this will be described in detail later). The node impurity is the largest when all classes are equally mixed together within a node and smallest when the node contains only one class. This process is repeated for each of the remaining variables at the root node. CART ranks all of the best splits on each variable according to the reduction in impurity achieved by each split. It selects the variable and its split point that most reduced impurity of the root or parent node. CART then assigns classes to these nodes according to a rule that minimizes misclassification costs. Each subset is then considered for an additional binary split so that t_2 could be split into t_4

and t_5 and t_3 could be split into t_6 and t_7. This process is repeated for each subset until the process can no longer be continued. The goal is to have subsets of $\boldsymbol{X}$ that are "purer", in that the majority of the measurement vectors in each subset belong to the same class.

If the subset can be further split into two more subsets to achieve a more accurate classification tree then the subset is referred to as a ***nonterminal node***, denoted with a circle in **Figure 1**. When a node *t* was reached such that no significant decrease in impurity was possible, then node *t* was not split and became a ***terminal node***, denoted with a square in **Figure 1**. $\tilde{T}$ denotes the collection of terminal nodes of *T*. The elements in *T*- $\tilde{T}$ are called non-terminal nodes. When a terminal node, such as t_5 in **Figure 1**, has been reached then all of the measurement vectors that belong to this node, $\{\boldsymbol{x} : \boldsymbol{x} \in t_5\}$, are then designated as the same class.

2.2.3 The splitting and stop-splitting rule

Choosing the splits of the measurement space $\boldsymbol{X}$ is the first step to build a tree. The fundamental idea is to have the majority of the subjects belong to the same class in each terminal node. In developing a methodology to evaluate and compare potential splits, Breiman, et al. developed the goodness of fit criterion, which was originally derived from the impurity function, to evaluate and compare potential splits. [7] An impurity function Φ is defined on the set of all J-tuples (p_1, p_2,…. p_J) satisfying $p_j \geq 0$, j=1, 2,……J, $\sum_j p_j = 1$ and with properties:

i. Φ achieves its maximum only at the point (1/*j*, 1/*j*,…, 1/*j*)

ii. Φ achieves its minimum only at the points (1, 0, …,0), (0, 1, 0,…,0),…, (0,…, 0, 1)

iii. Φ is a symmetric function of $p_1, p_2, \ldots p_J$.

Given an impurity function and the conditional probabilities for the *J* classes at any node *t*, an impurity measure *i(t)* at any node *t* can be defined as (4)

$$i(t) = \Phi\big(p(1|t), p(2|t), \ldots\ldots\ldots, p(J|t)\big) \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

A candidate split *s* will be selected based on its reduction of the impurity in the node *t*. Now, consider splitting the node *t* into nodes t_L and t_R by split *s*. The proportions of the cases, p_L and $p_{R,}$ in *t* will be sent into nodes t_L and t_R, respectively. Thus, our measure of the decrease in impurity in node *t* due to split *s* is simply defined as (5),

$$\Delta i(s,t) = i(t) - P_R i(t_R) - P_L i(t_L) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Then we consider the goodness of split *Φ (s, t)* to be $\Delta i(s,t)$. We select the split s that maximizes Δ*i*(s, t) for each node. Once a node is split, the children nodes are evaluated to determine if they can be split. This process is repeated until every node contains a small number of subjects and to minimize overall tree impurity.

There are many splitting criteria by which node impurity is minimized in a classification problem, but four commonly used metrics include ***"Misclassification error", "Gini index", "Entropy index", and "Twoing'***. The ***misclassification error*** is simply the proportion of observations in the node that are not members of the majority of class in that node. ***Gini index*** supposes that there are a total of K classes, each indexed by *k*. Let $\hat{p}_{mk}$ be the proportion of class *k* observations in node m. The ***Gini inde*x** can then be written as $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) = 1 - \sum_k \hat{p}_{mk}^2$. This measure is frequently used in practice, and is more sensitive than the misclassification error to changes in node probability. ***Entropy index*** is also called the information index, cross-entropy or deviance measure of impurity. The entropy index can be written as $\sum_{k=1}^{K} \hat{p}_{mk} log\hat{p}_{mk}$. This is more sensitive than misclassification error to change in node

probability. ***Twoing***, designed for multiclass problems, favors separation between classes rather than node impurity (heterogeneity). Every multiclass split is treated as a binary problem. Splits that keep related classes together are favored. The approach offers the advantage of revealing similarities between classes and can be applied to ordered classes as well. Breiman et al.[7] preferred the ***Gini index***. Breiman, et al.[7] concluded that within a wide range of splitting criteria the properties of the final tree selected are surprisingly insensitive to the choice of splitting rule. The criterion used to prune or recombine upward is much more important.

A stopping rule is to set a criterion for determining a terminal node. An early stopping rule was simple but unsatisfactory. Set a threshold $\beta > 0$, and declare a node t terminal if the split that maximizes $\Delta i(s,t) < \beta$. The concept of this stopping rule was plagued with problems. For example, the threshold β could be set too low that every terminal node has only one case and misclassification rate is zero. In general, misclassification rate decreases while the number of terminal nodes increases. These problems lead to the development of pruning. The pruning process begins with a tree that is split until every node contains a small number of cases, forming the tree *Tmax*. Then children nodes are selectively pruned into the single parent upward toward the root node, creating more general trees. Another stop-splitting rule is to declare a terminal node if the number of cases assigned to the node is less than some value or contains only identical measurement vectors. Measurement vectors in a terminal node are typically assigned to the class with the largest conditional probability, $p(j|t)$. Note that if the class **prior probability** is determined from the training data, i.e., $\pi(j)=N_j/N$, this rule assigns a terminal node to the class with the largest number of measurement vectors falling into the node. The tree construction continues until the number of cases reaching each leaf is small (by default, $N_j(t) < 20$ in *rpart*, $N_j(t) < 10$ in *tree;* $N_j(t)$ is the number of class j cases with $x \in t$).[8]

2.2.4 Class assignment for terminal nodes and resubstitution estimates

The justification for the original S methodology is to view the tree as providing a probability model (hence the title "tree-based models" of Clark and Pregibon, 1992). Of the three elements of tree construction, the assignment of classes to terminal nodes is the easiest to perform. In the learning sample L, let N be the total number of cases and N_j be the number of cases in class j. Often the ***prior probability, $\pi(j)$***, are estimated to be N_j/N or supplied by the analyst. In a node t, let $N(t)$ be the total number of cases in L with $x_n \in t$, and $N_j(t)$ be the number of class j cases with $x \in t$. The proportion of the class j cases in L falling into node t is $N_j(t)/N_j$. For a given set of priors, $\pi(j)$ is interpreted as the probability that a class j will be presented to the tree. Thus, the resubstitution estimate for the probability that a case will be both in class j and fall into node t will be $p(j,t) = \pi(j)*N_j(t)/N_j$. The resubstitution estimate $p(t)$ of the probability that any case falls into node t is defined by $p(t) = \sum_j p(j,t)$. The resubstitution estimate of the conditional probability that a case is in class j given that it falls into node t is given by $p(j|t) = p(j,t)/p(t)$ and satisfies $\sum_j p(j|t) = 1$. When $\pi(j) = N_j/N$, so the $p(j|t)$ are the relative proportions of class j cases in node t.

Now, we attempt to develop a classification assignment rule. Suppose a tree T has been constructed and has a set of terminal nodes $\tilde{T}$. A class assignment rule assigns a class $j \in \{1, \ldots, J\}$ to each terminal node $t \in \tilde{T}$. The class assigned to $t \in \tilde{T}$ is denoted by $j(t)$. The joint probability of a case being from class j and falling into node t, $p(j|t)$, is estimated from the data

as $p(j,t) = \pi(j)N_j(t)/N_j$. By extension, the resubstitution estimate for the probability of any case falling into node *t*, $p(t)$ is (6),

$$p(t) = \sum_j \frac{\pi_j N_j(t)}{N_j} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

The resubstitution estimate for the probability of misclassification given that a case falls into node *t* is given by (7),

$$P(j|t) = \frac{p(j,t)}{p(t)} \quad with \sum_j p(j|t) = 1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

When the $\pi(j)$ are estimated from the data using $N_j(t)/N(t)$, $p(j|t)$ can be estimated by (8),

$$P(j|t) = \frac{N_j(t)}{\sum_j N_j} = \frac{N_j(t)}{N(t)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

For any class assignment rule $j(t)$, $\sum_{j \neq j(t)} p(j|t)$ is the resubstitution estimate of the probability of misclassification given that a case falls into node *t*. We take as our class assignment rule $j^*(t)$ that minimize the resubstitution estimate, that is, if $p(j|t) = \max_i p(i|t)$ then $j^*(t)=j$. If two or more classes achieve the maximum, then assign $j^*(t)$ arbitrarily as any one of the maximizing classes.

Using this class assignment rule, we get the resubstitution estimate $r(t)$ of the probability of misclassification, given that a case fall into node *t* as $r(t) = 1 - \max_j p(j|t)$. When we define the joint probability $R(t)$ that a case falls into node *t* and is misclassified as $R(t) = r(t)p(t)$, then the resubstitution estimate for the overall misclassification rate $R^*(T)$ of the tree classifier *T* is (9)

$$R(T) = \sum_{t \in \tilde{T}} R(t) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)$$

Since misclassifying a case might be worse in a realistic setting, the idea of including a set of ***misclassification costs C(i|j)*** was introduced. The misclassification costs refer to the penalty that one assigns to the different possible misclassifications. For example, when trying to classify patients into whether they are at high risk or low risk for a certain type of cancer, we may feel that there is little penalty for identifying someone as high risk when in fact they are low risk. However, if a patient is classified as low risk when they are really high risk, the repercussion of this misclassification is much worse. Thus, *C(i|j)* is the cost of misclassifying a class j object as a class i and satisfies in (10) and (11):

$$(i)\ C(i|j) \geq 0, i \neq j, \ldots\ldots\ldots\ldots\ldots (10)$$

$$(ii)\ C(i|j) = 0, i = j \ldots\ldots\ldots\ldots\ldots (11)$$

Given a node *t* with estimated node probability *p(j|t)*, if a randomly selected object of unknown class falls into *t* and is classified as class *i*, then the estimated expected misclassification cost is shown in (12).

$$Estimated\ expected\ misclassification\ cost = \sum_{j=1}^{J} C(i|j)\, p(j|t) \ldots\ldots\ldots\ldots\ldots\ldots (12)$$

A natural node rule to develop our class assignment rule *j*(t)* is to select *i* to minimize the estimated expected misclassification cost. Therefore, the resubstitution estimate *r(t)* of the expected misclassification cost for a node *t* is (13)

$$r(t) = \min_i \sum_{j=1}^{J} C(i|j)\, p(j|t), \ldots (13)$$

and the resubstitution estimate *R(T)* of the misclassification cost of the tree *T* is (14)

$$R(T) = \sum_{t \in \tilde{T}} r(t)p(t) = \sum_{t \in \tilde{T}} R(t), \qquad where\ R(t) = r(t)p(t). \ldots\ldots\ldots\ldots\ldots (14)$$

2.2.5 Selecting the best tree

The best tree is one that is small, easily interpretable but still retains the ability to classify correctly. Appropriate tree size can be determined in a number ways. One way is to set a threshold for the reduction in impurity measure, below which no split will be made. Depending on the threshold, the splitting was either too soon at some terminal nodes or continued too far in other parts of the tree. A preferred approach is to grow an overly large tree until some minimum node size is reached, and then prune the tree back to an optimal size. Cross-validation or test sample estimates can be used to select the subtree with the lowest estimated misclassification rate.

The pruning process results in a decreasing sequence of subtrees, $T_1 > T_2 > \ldots\ldots > \{t_1\}$, where $T_k = T(\alpha_k)$ *and* $\alpha_1 = 0$. We have to select one of these as the optimum-sized tree. Each tree in the sequence is best for some range of the ***complexity parameter α*** in that it minimizes the cost-complexity function. Optimal size can be determined using an ***independent test sample*** or ***cross-validation***. If the sample size is sufficiently large, the data can be divided into two subsets randomly, namely, one for training and other for testing. Defining sufficiently large is problem specific, but one rule of thumb in classification problem is to allow a minimum of 200 observations for a binary classification model, with an additional 100 observations for each additional class. An overly large tree is grown on the training data. Then, using the test set, error rates are calculated for the default tree as well as all smaller subtrees. The subtree with the smallest error rate based on the independent test set is then chosen as the best tree.

The method that we will focus on to determine the best tree is ***cross-validation***. If the sample size is not large, it is necessary to retain all the data for training purposes. However,

pruning and testing must be done using independent data. Cross-validation involves randomly dividing the data into V roughly equal groups. One of the V portions is left out while the remaining portions all are used to build a model. The portion not used in building the model is used to assess the accuracy of the current model. This process is repeated for each of the other $V-1$ portions and then the V estimates are averaged to get the final cross-validation estimate of model accuracy. The most common cross-validation is the 10-fold cross-validation which subtrees of different sizes are constructed with 90% of the data and their misclassification rates on the remaining are computed. This process is done 10 times with each 1/10 of the data held out one time. The misclassification rates are aggregated over the replications. Finally, the optimal tree size is the one whose aggregated misclassification rate is smallest.

Cross-validation can be implemented in the complex sequence of trees $T_1 > T_2 > \ldots\ldots > \{t_1\}$ in the following way. In V-fold cross-validation, the learning sample L is divided randomly into V equal (nearly) subsets. Let L_v be the v^{th} portion of the learning sample and $L^{(v)}$ represent the entire learning sample missing only the v^{th} portion. We already have our sequence of trees and critical values of α based on the entire learning sample. Now we must repeat the tree growing and cost-complexity pruning steps for $L^{(1)}, L^{(2)}, \ldots, L^{(v)}$. For $L^{(1)}$, we will obtain a sequence of trees $T_1^{(1)} > T2^{(1)} > \ldots.. > \{t_1\}$ and a sequence of critical values $\alpha_1^{(1)} < \alpha_2^{(1)} < \ldots < \alpha_{k1}^{(1)}$ where k_1 is the number of subtrees in the previous sequence. A similar sequence of trees and critical values will be obtained for the other $V-1$ sets.

Let $T^{(1)}(\alpha'_j)$ be the minimal cost-complexity tree for complexity parameter α'_j based on $L^{(1)}$. The minimal cost-complexity tree can be found for the other $V-1$ sets giving us $T^{(1)}(\alpha'_j)$, $T^{(2)}(\alpha'_j), \ldots, T^{(V)}(\alpha'_j)$. Now define

$N_{ij}^{(v)}$ = the number of class j cases in L_v classified as i by $T^{(V)}(\alpha'_j)$, and (15)

$$N_{ij} = \sum_{v=1}^{V} N_{ij}^{(v)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (15)$$

Each case in L appears in one and only one test sample L_v. Therefore, the total number of class j observations in L is N_j. The $T^{(v)}(\alpha)$ should have about the same classification accuracy as $T(\alpha)$ for V large. Hence, we make the cross-validation estimate for the probability of classifying a case as i given that it is j for $T(\alpha)$ as $Q^{CV}(i|j) = N_{ij} / N_j$. For the prior probability $\{\pi(j)\}$ is given or estimated, set the cross-validation estimate for the cost associated with class j by (16)

$$R^{CV}(j) = \sum_{i=1} C(i|j) Q^{CV}(i|j) \ldots\ldots\ldots\ldots\ldots (16)$$

and the cross-validation estimate for the cost of $T(\alpha)$ by (17)

$$R^{CV}(T(\alpha)) = \sum_{i=1} R^{CV}(j)\pi(j) \ldots\ldots\ldots\ldots\ldots\ldots (17)$$

If the data from L is used to estimate the priors, then $\pi(j) = N_j/N$ and the cross-validation estimate reduces to (18)

$$R^{CV}(T(\alpha)) = \frac{1}{N} \sum_{i,j} C(i|j) N_{ij} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (18)$$

In the unit cost case, (18) simply shows the proportion of the test set cases misclassified. Selection of the right sized tree is now obtained by finding $T(\alpha'_{j0})$ such that (19).

$$R^{CV}\left(T(\alpha'_{j0})\right) = \min_{j=1 \text{ to } k} R^{CV}\left(T\left(\alpha'_j\right)\right) \ldots\ldots\ldots\ldots (19)$$

Then use $R^{CV}(T(\alpha'_{j0}))$ as an estimate of the misclassification cost. If the sample is small, cross-validated trees tend to be conservative in the direction of overestimating misclassification costs. Now, the question is how large to take V? Breiman et al. stated that taking V=10 gives adequate accuracy.[7] In some examples, smaller values of V also give sufficient accuracy. However, they did not come across any situations where taking V larger than 10 gave a significant improvement in accuracy for the tree selected.

As mentioned before, one of the advantages of CART is their ability to accommodate missing values. If an outcome variable is missing, that observation can be excluded from the analysis, or in the case of classification problem, treated as a new class to identify any potential patterns in the loss of information. If explanatory variables are missing, tree can use surrogate variables in their place to define the split. Alternatively, an observation can be passed to the next node using a variable that is not missing for that observation. The easiest approach is to treat the missing attribute as a distinct value and to assign all samples with missing values to the same node. Breiman, et al.[7] introduced ***surrogate splits*** to deal with missing attributes. Suppose that a measure of similarity between any two splits *s, s'* of a node *t*. If the best split of node *t* is the split *s* on the variable x_m, find the split *s'* on the variable other than x_m that is most similar to *s*. We call *s'* the best surrogate for *s*. Similarly, we can define the second best surrogate, third best, and so on. In other words, if an observation is missing a value for the best split, then it is classified using the first surrogate split. If that value is missing then the second surrogate split is used, and so on. If an observation is missing all the surrogate splits then the default rule *max* $(p_L;p_R)$, where the observation is sent to the child with the largest relative frequency at that node, is used. The surrogate splits have advantages because they use other available information to make the split. Breiman, et al.[7] also proposed ranking the importance of variables through

surrogate splits. The ranking of a variable that does not appear in a split on the final tree, indicates that this variable is masked by other variables.

2.2.6 Regression tree

The underlying ideas for classification and regression trees are quite similar, but terminology differs. However, several things become simpler because there are no priors so each case is weighted equally. The prediction for a regression tree is constant over each cell of the partition of **X** induced by the leaves of the tree. We define that the predictor *d(x)* is a real-value function on ***X***. Thus *d(x)* produces real numbers not classes. A learning sample *L* consists of the observations (x_1, y_1), (x_2, y_2),……, (x_N, y_N) where ***x*** is a vector of measurements that hopefully explain the real valued response *y*. The misclassification rate of the classifier *d(x)* is denoted by *R(d)*. In a regression tree, the estimated mean squared error for *d(x)* will be represented as misclassification rate *R(d)* to measure the accuracy of the predictor. For a given rule *d(x)*, the resubstitution estimate for the mean squared error is (20).

$$R(d) = \frac{1}{N}\sum_{n}(y_n - d(X_n))^2 \ldots\ldots\ldots\ldots\ldots\ldots (20)$$

$\overline{y}(t)$, the average of y_n for all cases that fall into node *t*, is used to minimize *R(d)*. Therefore, our estimate for the mean squared error of the tree *T* is (21).

$$R(T) = \frac{1}{N}\sum_{t\in\tilde{T}}\sum_{n:X_n\in t}(y_n - \overline{y(t)})^2 \ldots\ldots\ldots\ldots\ldots (21)$$

If *R(t)* is the within node sum of squares divided by *N*, then *R(T)* can be written as (22).

$$R(T) = \sum_{t\in\tilde{T}} R(T) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (22)$$

Finding the split of t into t_L and t_R which decreases $R(T)$ the most is the way to determine what the best split of a current terminal node t. Let s be a candidate split in the set of possible splits S. Let (23) be

$$\Delta R(s,t) = R(t) - R(t_L) - R(t_R) \dots\dots\dots\dots (23)$$

and find the best split s^* satisfying in (24).

$$S^* = \underset{s \in S}{\operatorname{argmax}} \Delta R(s,t) \dots\dots\dots\dots\dots\dots\dots\dots (24)$$

We used the same method to grow a regression tree as for a classification tree. T_{max} is created by splitting to minimize $R(T)$. The tree is declared T_{max} once each terminal node contains at most some small number of observations (usually 5). Minimal error-complexity pruning is performed the same way as in classification trees. Define the error-complexity measure as (25).

$$R_\alpha(T) = R(T) + \alpha\left\|\tilde{T}\right\| \dots\dots\dots\dots\dots\dots\dots (25)$$

We obtain our sequence of trees and critical values based on the entire learning sample. Let $d_k(x)$ be the prediction rule associated with the tree T_k. L is randomly divided into V groups. In cross-validation, we find the sequence of trees and critical values for each of the V possible groups formed by leaving out one of the V portions. For each of the V sequence of trees and critical values, we can form the function $T^{(V)}(\alpha)$ that is the minimal error-complexity tree for parameter α in the vth sequence of trees. The cross-validation estimate for the mean squared error of tree T_k is (26)

$$R^{CV}(T_k) = \frac{1}{N}\sum_{v=1}^{V}\sum_{n:(x_n:y_n)\in L_v}(y_n - d_k^{(v)}(x_n))^2 \dots\dots\dots\dots\dots (26)$$

which leads to the cross-validation estimate, (27).

$$RE^{CV}(T_k) = \frac{R^{CV}(T_k)}{R(\bar{y})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (27)$$

2.3 S-PLUS AND "RECURSIVE PARTITIONING (RPART)" FUNCTIONS

S-PLUS is an implementation of the language S developed at Bell Laboratories by Becker, Chamberlain and Wilks.[8] The ***"recursive and partitioning and regression tree (rpart)"*** library and affiliated packages is part of the R public domain statistical software and can be downloaded into S-Plus. In most details, ***"rpart"*** function follows many ideas quite closely found in CART by Breiman et al.[7] Because CART™ is the trademarked name of a particular software implementation of the CART ideas and "***tree"*** has been used for the S-Plus routines, different acronym ***"rpart"*** was chosen to call the library connected in the S-Plus. The introduction within S of tree-based models described by Clark and Pregibon in 1992 made the methods much more freely available through function ***"tree"*** and its support functions.[30] The library section ***" (rpart)"*** differs from the ***"tree"*** function mainly in its handling of surrogate variables.[8] The ***"tree"*** function has some advantages, mainly in showing some process in more detail. We used the "***rpart"*** function in this thesis because it provides a faster for fitting trees to a large dataset.[30]

How is the ***"rpart"*** package in conjunction with TIBCO Spotfire S+ 8.1 (TIBCO Software)? A powerful feature of S-PLUS is to allow users to extend its functionality and interface with other languages, namely, C and Fortran. By using other languages, users can combine the speed and efficiency of compiled code with the robust, flexible programming environment of S-PLUS. Although S is a versatile language with many routines already included, same procedures are not covered. Fortunately, the core routines are easily augmented

with additional user-written routines, which can be loaded into S-PLUS. These routines are usually provided in what S-PLUS calls a ***"library"***. We can easily package up user-created functionality in order to share it with other users through the S-PLUS libraries.[8] The ***"rpart"*** package is written in C language and implements classification and regression trees defined by Breiman, et al.[7] and Therneau and Atkinson[31]. The ***"rpart"*** is loaded and called from within S-PLUS and returns a standard S-PLUS tree object, which can then be manipulated using other S-PLUS visualization and statistical functions. Hence, the ***"rpart"*** package contains both C and S code. The purpose of the C code is to improve on the cpu time required for lengthy computations that are performed during the recursive partitioning process.

The ***"rpart"*** algorithm that we mainly used in this thesis was developed by Therneau and Atkinson.[31] The ***"rpart"*** programs build classification and regression trees of a very general structure using a two stage procedure; the resulting models can be represented as binary trees.[31] First, the single variable is found which best splits the data into two groups. The data is separated and then this process is applied to each of sub-groups recursively, and so on recursively until the subgroups either each a minimum size or no improvement can be made. The initial default tree may be too large and complex. The second stage of the ***"rpart"*** program consists of using ***cross-validation*** to prune back the default tree. A cross validated estimate of risk was computed for a nested set of subtrees. The final model is the subtree with the lowest estimated risk. Two types of decision trees are distinguished: *regression trees* for *continuous* responses and *classification tree* for *categorical* responses. The types of endpoints that ***"rpart"*** handles include classification (such as yes/no), continuous variables (such as blood pressure), Poisson counts (such as the numbers of fractures in Medicare patients), and survival information (such as time to death).

2.3.1 Building a tree and splitting criteria

The ***"rpart"*** uses a measure of impurity of a node to build a binary tree. Let ***f*** be some impurity function and define the impurity of a node *A* as (28).

$$I(A) = \sum_{i=1}^{C} f(p_{iA}) \dots\dots\dots\dots\dots\dots\dots\dots (28)$$

where p_{iA} is the proportion of those in *A* that belong to class *i* for future observations. Since we would like *I(A)* =0 when *A* is pure, *f* must be concave with *f*(0) = *f*(1) = 0. Two candidates for *f* in the "rpart" program are the ***entropy (or information) index, f(p) = -p log(p)*** and the ***Gini index, f(p) = p(1-p)***. We then use the split with maximal impurity reduction in (29).

$$\Delta I = \ P(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R) \dots\dots\dots\dots\dots\dots (29)$$

2.3.2 Pruning the tree

The initial tree possibly too large or complex, and we have to decide how much of that model to retain. The **"rpart"** uses a ***likelihood-based*** pruning criterion as a way to prune branches of the overfitted tree. Let $T_1, T_2, \dots, T_k$ be the terminal nodes of a tree *T*, $|T|$ be the number of terminal nodes, *R(T)* be the risk of T and $R(T_0)$ be the risk for the zero split tree. Define the cost for the tree to be (30).

$$R_\alpha(T) = R(T) + \alpha|T| \dots\dots\dots\dots\dots\dots (30)$$

where $R(T) = \sum_{i=1}^{k} P(T_i)R(T_i),\ P(T_i)\ is\ the\ probabilty\ of\ node\ T_i$, and α is a complexity parameter between 0 and ∞. Therefore the risk of *T* is the sum of all the probabilities of each split tree multiplied by the risk of each split tree. T_α is the subtree of the full model that has

minimal cost. Therefore, we can define T_α as the smallest tree T for which $R_\alpha(T)$ is minimized. In comparison to regression, $|T|$ is analogous to the model degrees of freedom and $R(T)$ to the residual sum of squares.

The ***"rpart"*** then uses ***V-fold cross-validation*** to choose the best value for α by the following step:

i. Fit the full model on the data set and compute measures of impurity

ii. Divide the data randomly into roughly equal V sized groups where V is an integer usually between 5 and 10. V-*1* groups, the training set, are used to generate the model separately and the remaining portion, the test set, is used to evaluate the model. This step is repeated until all test sets have been used in model evaluation.

iii. The results of these trees are averaged for all the combinations where one of the groups is withheld. For that complexity parameter with the smallest risk is chosen as the best-pruned tree. The default for the ***"rpart"*** is V=10, called 10-fold cross-validation.

In actual practice, we may use instead the ***1-SE rule***.[7, 31] A plot of complexity parameters versus risk often has an initial sharp drop followed by a relative flat plateau and then a slow rise. The choice of complexity parameter among those models on the plateau can be essentially random. To avoid this, both an estimate of the risk and its standard error are computed during the cross-validation. Any risk within one standard error of the achieved minimum is marked as being equivalent to minimum which is considered to be part of the flat plateau. Then the simplest model among all those "tied" on the plateau is chosen.

2.3.3 Missing data and surrogate variables

Most statistical procedures deal with missing data by excluding incomplete observations in the analysis. The ***"rpart"*** handles missing data differently. Any observation with values for the dependent variable and at least one independent variable will be included in the modeling. In other words, the ***"rpart"*** by default, removes only those rows for which either the response *y* or all of the independent variables are missing. This ability to retain partially missing observations is perhaps one of the most useful features of the ***"rpart"*** function. With missing data, the object is still to maximize the reduction of impurity in (32).

$$\Delta I = P(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R) \dots\dots\dots\dots\dots (32)$$

The first term stays the same for all variables and splits irrespective of missing data because it is calculated over all observations in node *A*, but the last two terms (dealing with the right and left children nodes) are somewhat modified. Firstly, the impurity indices of both child nodes, $I(A_L)$ *and* $I(A_R)$, are calculated only over observations that are not missing a particular predictor. Secondly, the two probabilities for both the left and right child nodes, $p(A_L)$,$p(A_R)$, are also calculated only over observations that are not missing a particular variable, but they are later adjusted so that they sum to the probability of node *A* $(p(A))$. This takes some extra effort while building a tree, but ensures that the probabilities of the terminal node will sum to one.

Once a splitting variable and it's split cut point have been decided, what is to be done with observations missing that variable? The ***"rpart"*** uses surrogate variables. As mentioned before, the missing values are ignored and the probabilities and impurity measures are calculated from the non-missing values of that variable. For any observations with a missing value for the split variable, the surrogate variables are ranked and surrogate splits are used to allocate the

missing cases to the child nodes. Any observation that is missing the first surrogate split variable is then classified using the first surrogate variable, or if missing that, the second surrogate is used, etc. If an observation is missing all the surrogates the blind rule (i.e., go with the majority) is used. In addition, the "rpart" routines impose one more constraint during the construction of surrogates: a candidate surrogate must send at least two cases down each branch (i.e., at least two to right and two to left).

3.0 METHODS

Our goal is to develop a set of criteria to distinguish T1DM from T2DM in a large administrative database. The several steps involved in developing a clinical prediction rule to distinguish T1DM from T2DM in an administrative database include (1) assemble the cohort; (2) decide which predictors to obtain from those available in an administrative database; (3) identify cases of T1DM and T2DM (outcome); and (4) build a tree-model and estimate the misclassification rate of the rule.

3.1 THE SETTING AND ASSEMBLE THE COHORT

Because we sought a potentially generalizable and accurate predictive rule to identify people with T1DM and T2DM that could applied in heterogeneous clinical settings, we used the database from the UPMC. UPMC is one of the largest non-profit integrated health care systems in the United States. UPMC integrates 20 academic, community, and specialty hospitals and 400 outpatient sites, employs 2,700 physicians and also offers rehabilitation, retirement and long-term care facilities, pharmacy services, a managed care health plan, and a community health division.[32] Each year, UPMC hospitals have more than 187,000 inpatient admissions, 4.5 million outpatient visits and 480,000 emergency visits.[32] Of the 23.6 million people with diabetes

nationwide, Pennsylvania ranked sixteenth for the percentage of the adults who had ever been told by a doctor that they had diabetes (8%, about 764,000 people).[33] UPMC provides services for the large majority of people with diabetes throughout the Western Pennsylvania.[34] A large diverse cohort maximizes the chance that the clinical prediction rule will generalize to other populations of diabetic patients. One empirical rule of the size of the training dataset is to include five patients in the smallest outcome category for every clinical predictor in the rule.

3.2 DIABETES REGISTRY FROM THE MEDICAL ARCHIVAL RETRIEVAL SYSTEM (MARS) DATA

UPMC has used an electronic medical record called MARS since 1987. MARS is a repository of information forwarded from the health system's electronic clinical, administrative, and financial databases. Laboratory results and patient demographics, visits, and charges are captured into MARS. MARS is indexed on every word in the medical records, and is capable of recovering all encounters for a given patient between specified dates. MARS-based data sources used to establish the diabetes registry included: (1) Medical Records Discharge Abstract file, consisting of all visits coded by the Medical Records Department, with up to 25 ICD-9 diagnosis codes assigned per patient; (2) the Hospital Laboratory Information System that includes inpatient, emergency room, hospital-based clinics, outpatient surgery, and mail-in specimens. Laboratory data are sent to MARS using the Misys® laboratory information system, which supports all of the UPMC hospitals and clinics; (3) the Hospital Pharmacy Information System which includes inpatient information on medication dispensed in the emergency room, hospital-based clinics,

and outpatient surgery settings. The MARS database is continually audited by Internal Audit at UPMC as well as part of the Capability Maturity Model process that is required of all UPMC databases and part of the UPMC policies and procedures for maintaining data integrity.[6]

We examined data for the period 1/1/2000-09/30/2009. In total, 140,781,751 lab reports, 5,720,470 visits and 139,750,158 charges records representing approximately 290,552 patients with unique accounts were searched. First, the initial source population was identified by the presence of any one of the following six criteria: ICD-9 code 250 (diabetes) for either inpatient, emergency room or outpatient visits (treated as three separate indicators); any HbA1c result (regardless of value, reference range 4.3-6.1); a blood glucose >200 mg/dl; or use of any diabetes medication (acarbose, acetohexammide, chlorpropamide, glimepiride, glipizide, glucagon, glyburide, insulin, metformin, miglitol, pioglitazone, repaglinide, rosiglitazone, tolazamide, tolbutamide, troglitazone).[6] According to the crteria from Zgibor et al.,[6] we used two or more of the above indicators or an out-patient-diagnosis to identify potential diabetic patients. 80,145 patients who did not meet criteria and 807 patients younger than 18 years of age were excluded. Each patient is represented only once in the database. We identified 209,642 potential diabetic patients with age ≥ 18. Data cleaning, formatting, and recoding were done using SAS 9.2 (SAS Institute Inc., Cary, NC).

3.3 POTENTIAL PREDICTORS AND RULES TO IDENTIFY POTENTIAL T1DM AND T2DM CASES IN THE MARS DATASET

Unfortunately, not all of the potential variables in the **Table 1** can be obtained from an MARS database. As shown in the **Table 3,** we were able to obtain 32 variables from the MARS database including ICD-9 codes for T1DM (250.x1 or 250.x3) and T2DM (250.x0 or 250.x2), insulin use, pramlintide use, other oral hypoglycemia agents (OHA) use (acarbose, acetohexamide, chlorpropramide, **exenatide**, glimepiride, glipizide, glucagon, glyburide, metformin, miglitol, **nateglinide**, pioglitazone, repaglinide, rosiglitazone, **sitagliptin**, tolazamide, tolbutamide, troglitazone), DKA, hypoglycemic coma, comorbidities (Addison, thyroid, and celiac diseases), and complications (myocardial infarction (MI), coronary artery bypass graft (CABG), dialysis, amputation, retinopathy, and neuropathy) and age at having the diagnosis of each complication for the first time . These variables were defined mainly by ICD-9 codes or searching parsing notes.

The response variable was types of DM (i.e., T1DM or T2DM). We ascertained T1DM and T2DM cases through these predictors exclusively. Briefly, T1DM patients compared to T2DM patients were more likely to be diagnosed diabetes at their early ages, to take insulin only, to have DKA or hypoglycemic coma diagnosis, to have other autoimmune–related comorbidities, and to have complications occurred earlier (especially have cardiovascular events before age of 40 years); T2DM patients compared to T1DM patients were more likely to be diagnosed diabetes at their middle ages, to take OHA and/or with insulin, to have complications occurred later (especially have cardiovascular events before age of 65 years), and less likely to have DKA, hypoglycemic coma diagnosis, or other autoimmune–related comorbidities.

Indicators for T1DM cases included (1) in-patient insulin use only; (2) specific ICD 9 code for T1DM (250.x1 or 250.x3); (3) parsing notes for T1DM-related diagnosis including childhood-onset DM, juvenile DM, insulin-dependent DM etc.; (4) DKA diagnosis; (5) hypoglycemia diagnosis; (6) other autoimmune-related comorbidities; (7) any complications diagnosis (especially before age of 40 years); (8) age; (9) diabetes diagnosis at emergency room or inpatient visits. Indicators for T2DM cases included (1) in-patient use of OHA with or without insulin; (2) ICD 9 code for T2DM or unspecified type DM (250.x0 or 250.x2); (3) parsing notes for T2DM-related diagnosis including adult DM, non-insulin-dependent diabetes etc.; (4) no DKA diagnosis; (5) without hypoglycemic coma diagnosis; (6) no other autoimmune-related comorbidities; (7) with or without other complications (evaluated by ages, e.g., especially without after age of 40 years or with after age of 65 years); (8) age. We considered ICD-9 codes diagnosis, medication use, and presence of DKA as relatively stronger indicators and the rest of these indicators as ancillary.

To minimize misclassification between T1DM and T2DM cases, we identified **probable, possible** T1DM and T2DM cases and **unknown** types of DM cases by using different combinations of the above indicators in the MARS dataset. "Probable" T1DM cases were assigned when there are more obvious indicators for T1DM than T2DM, while "probable" T2DM cases are with more obvious indicators for T2DM than T1DM. "Possible" T1DM cases were assigned when there are about equally strong indicators for both T1DM and T2DM with one or more ancillary indicators for T1DM. Unknown types of diabetes were assigned to the patients when all of the indicators are with missing values or only one or two ancillary indicators are not missing. The clinical detailed algorithms to define these exclusive categories identifying probable cases, unknown types of diabetes, and possible cases are presented in the **Table 4**,

Table 5, and Table 6, respectively. As shown in **Table 4, Table 5, and Table 6**, we designated probable cases as categories 1 to 39, unknown types of diabetic cases as categories 41 to 53, and possible case as categories 56 to 92. In total, we identified 126,097 probable cases (T1DM: 7,857, T2DM: 118,240), 40,619 possible cases (T1DM: 2,147, T2DM: 38,472), and 42,926 unknown types of diabetic cases. We excluded the 42,926 unknown types of diabetic cases, and used probable and possible cases to construct the tree models (**Figure 2 and Figure 3**).

Table 3: Potential Predictors Available in the MARS Dataset

Variable*	Code and definition
Age	Age (years) of patient at first entry into the MARS database
Gender	F= Female, M= Male
Race	0 = White, 1 = Black, 2= Asian, 3 = Hispanic, 4 = Others
ICD_DM	1= ICD-9 codes with T1DM specific diagnosis (250.1 or 250.3); 2= ICD-9 codes with T2DM or unspecified types diagnosis (250.0 or 250.2); 3= with both ICD-9 codes for T1DM, and T2DM or unspecified type diagnosis ((250.x1 or 250.x3) AND (250.x0 or 250.x2)) Missing values = without any ICD-9 codes for DM
Text_DM1	Identified by parsing notes with T1DM or insulin-dependent diabetes mellitus (IDDM), or childhood-onset diabetes
Text_DM2	Identified by parsing notes with type 2 diabetes or noninsulin-dependent diabetes mellitus (NIDDM), or adult-onset diabetes
Insulin	With in-patient medication charges of insulin use
Oralagent	With in-patient medication charges include acarbose, acetohexamide, chlorpropramide, exenatide, glimepiride, glipizide, glucagon, glyburide, metformin, miglitol, nateglinide, pioglitazone, repaglinide, rosiglitazone, sitagliptin, tolazamide, tolbutamide, troglitazone
DKA	ICD-9 codes with 250.11 OR 250.13 OR 250.10 OR 250.12 OR Text that states DKA (diabetic ketoacidosis) OR Text that states Kussmaul respiration OR Text that states an odor of acetone on the breath
Hypo_coma	ICD9 codes with 250.31 OR 250.33 OR 250.30 OR 250.32 OR text that states hypoglycemic coma
Thyroid	Text that states Grave disease OR Hashimoto's thyroditis OR ICD9 code with 242.X
Celiac	Text that states Celiac disease OR ICD9 code with 579.0
Addison	Text that states Addison disease OR ICD9 code with 255.4
Rubella	Text that states congenital rubella infection OR rubella infection OR ICD-9 code with 056.XX
MI	ICD-9 code with 410-414
CABG	ICD-9 code with 414.04
CVE	ICD-9 code of MI or CABG (410-414 or 414.04)
Dialysis	ICD-9 code with v45.11
Amputation	ICD-9 code with 89X
Retinopathy	ICD-9 code with 362
Neuropathy	ICD-9 code with 357.2

Table 3 continued

Variable*	Code and definition
Age_MI	Age at having the first MI diagnosis
Age_CABG	Age at having the first CABG
Age_CVE	Age at having the first MI or CABG
Age_Dialysis	Age at having the first dialysis diagnosis
Age_Amputation	Age at having the first amputation diagnosis
Age_Retinopathy	Age at having the first retinopathy diagnosis
Age_Neuropathy	Age at having the first neuropathy diagnosis
Age_Complication	Age at having any first complication diagnosis
ERpt	ICD-9 code 250 for emergency room visit
Inpt	ICD-9 code 250 for inpatient visit
Outpt	ICD-9 code 250 for outpatient visit

Abbreviation: CABG: coronary artery bypass graft; ; **CVE:** cardiovascular events (including MI and CABG); **DM:** diabetes mellitus**; ICD-9 codes**: International Classification of Diseases diagnosis codes, version 9; **MI:** myocardial infarction

*: Except age, age_MI, age_CABG, age_CVE, age_dialysis, age_amputation, age_retinopathy, age_neuropathy are continuous variables, the rest of the variables are categorical variable. Except gender, race, and ICD_DM coded differently, the rest of the categorical variables are coded as 1 when patients met the rules or criteria, otherwise, coded as 0.

Table 4: Defined Clinical Rules to Identify "Probable" Type 1 and Type 2 Diabetic Cases in the MARS Data

Category	ICD_DM[¶]	Insulin	OHA	Cofactors[¥]	Complications[φ] or age at having first complications	Age	Other factors	DM type (n)
1	1	Yes	No					T1DM (1677)
2	2	No	Yes					T2DM (8915)
3	≠1	Yes	No		Age_comp ≤ 40	≠ .		T1DM (181)
4	≠ 2	No	Yes	DKA=0 OR hypocoma = 0				T2DM (1850)
5	≠ 2	Yes	No	DKA=1 OR hypocoma = 1		≠ .		T1DM (653)
6	≠ 1	Yes	Yes	DKA=0 OR hypocoma = 0				T2DM (39714)
7	1	No	No	DKA=1 OR hypocoma =1	OR (complication>0 AND age_comp≤ 40)			T1DM (65)
8	2	No	No	DKA=0AND hypocoma=0	Complication = 0	≥ 65		T2DM (19064)
9	3	No	No	DKA=1 OR hypocoma=1	OR (complication>0 AND age_comp ≤ 40)	≠ .	Text_T2DM=0	T1DM (83)
10	2	No	No	Cofactor=0	Complication = 0	≠ .	Text_T1DM=0	T2DM (22792)
11	1	No	No			≠ .		T1DM (2273)
12	2	Yes	Yes			≠ .		T2DM (196)
13	3	Yes	Yes	DKA=1 OR hypocoma=1	Complication>0 AND age_comp ≤ 40		Text_T1DM=1	T1DM (24)
14	.	No	No	Cofactor = 0	CVE = 0	≥ 65		T2DM (2982)
15	.	No	No	DKA =1 OR hypocoma=1	OR (complication>0 AND age_comp ≤ 40)	≠ .		T1DM (26)
16	2	Yes	No	DKA=0 AND hypocoma=0	Age_comp ≥ 65		Text_T1DM=0	T2DM (871)
17	≠ 2	Yes	No			≠ . , ≤ 40	Text_T2DM=0	T1DM (2209)
18	3 or .	No	Yes			≠ .		T2DM (18)
19	2	Yes	No	Cofactor=0	CVE = 0 AND dialysis=0 AND amputation=0	>40	Text_T1DM=0	T2DM (14777)
20	1	Yes	Yes	Cofactor=0	Complication>0	> 40		T2DM (4)
21	3	Yes	No		CVE = 1	≠ .	Text_T1DM=1 & Text_T2DM=0	T1DM (33)
22	3 or .	Yes	Yes	DKA =1 OR hypocoma=1		≠ .	Text_T1DM=1	T1DM (85)

Table 4 continued

Category	ICD_DM[¶]	Insulin	OHA	Cofactors[¥]	Complications[φ] or age at having first complications	Age	Other factors	DM type (n)
23	2 or .	No	No	DKA=0 AND hypocoma=0		≠ .	Text_T1DM=0 & Text_T2DM=1	T2DM (206)
24	2	No	No	DKA= 0 AND hypocoma=0	Complication=0		Text_T1DM=0 & Text_T2DM=1	T2DM (330)
25	1	Yes	Yes	DKA=1 OR hypocoma=1	OR complication=1	OR ≠ . , ≤ 40	OR Text_T1DM=1	T1DM (29)
26	Any	Any	Any		CVE=1 AND age_comp≤ 40	≠ .		T1DM (7)
27	Any	No	Yes					T2DM (2)
28	2	Yes	No		CVE=1 AND age_comp> 40			T2DM (293)
29	3	Yes	No	DKA=1				
30	.	Yes	No	DKA=0	CVE=1	> 40		T2DM (2453)
31	3	Yes	No		CVE=1	≠ ., <65		T1DM (145)
32	3 or .	Yes	Yes	DKA=1		>40		T2DM (205)
33	2	Any	Yes	DKA=0		≠ .		T2DM (1533)
34	3 or .	Yes	Yes	DKA=1		≠ . , ≤ 40		T1DM (43)
35	3	No	No	DKA=1 OR hypocoma=1				T1DM (44)
36	3	Yes	No	DKA=0 AND hypocoma=0	Complication=0	≥ 65		T2DM (1773)
37	3	Yes	No	DKA=0 AND hypocoma=0	Complication≥2 And age_comp <65	≠ .		T1DM (51)
38	3	No	No	DKA=0 AND hypocoma=0	Complication>0 AND age_comp>40	≠ .		T2DM (147)
39	.	Yes	No		CVE=1	≠ . , ≤ 40		T1DM (1)

Abbreviation: OHA: oral hypoglycemic agents; **age_comp**: age at having first complication; **age_CVE**: age at the first cardiovascular event including myocardial infarction (MI) and coronary artery bypass graft (CABG); **DKA**: diabetic ketoacidosis; **ER**: ICD-9 code 250 for emergency room diagnosis; **hypocoma**: hypoglycemic coma; **inpt**: ICD-9 code 250 for in-patient diagnosis; **outpt**: ICD-9 code 250 for out-patient diagnosis; **T1DM**: type 1 diabetes mellitus; **T2DM**: type 2 diabetes mellitus; **text_T1DM**: with any related T1DM diagnosis in the parsing notes; **text_T2DM**: with any related T2DM diagnosis in the parsing notes

¶: ICD_DM: **1**= only with ICD-9 code for T1DM specific diagnosis (250.x1 or 250.x3); **2**= only with ICD-9 code for T2DM or unspecified type DM diagnosis (250.x0 or 250.x2); **3**= with both ICD-9 code for T1DM and T2DM or unspecified diagnosis [(250.x1 or 250.x3) AND (250.x0 or 250.x2)]; **missing values**= without any ICD-9 code diagnosis for DM.
¥: Cofactors include diabetic ketoacidosis (DKA), hypoglycemic coma, thyroid, Addison, and celiac diseases
φ: Complications include coronary artery bypass graft (CABG), myocardial infarction (MI), dialysis, retinopathy, neuropathy, and amputations.

Table 5: Defined Clinical Rules to Identify "Unknown" Type 1 and Type 2 Diabetic Cases in the MARS Data

Category	ICD_DM¶	Insulin	OHA	Cofactors¥	Complications$^{\varphi}$ or age at having first complications	Age	Other factors	DM type (n)
41	.	No	No	Cofactor=0	Complication=0	= .	Er=0 & inpt=0 & outpt=0	DK (2008)
42	3	No	No	Cofactor=0	Complication=0	= .	Er=0 & inpt=0 & outpt=0	DK (1275)
43	.	Yes	No	Cofactor=0	Complication=0	= .	Er=0 & inpt=0 & outpt=0	DK (820)
44	.	No	No	Cofactor>0 AND DKA=0 AND hypo_coma=0	Complication=0	= .	Er=0 & inpt=0 & outpt=0	DK (2)
45	.	No	No	Cofactor=0	Complication>0	= .	Er=0 & inpt=0 & outpt=0	DK (13)
46	.	No	No	Cofactor=0	Complication=0	≠ .	Er=0 & inpt=0 & outpt=0	DK (52)
47	.	No	No	Cofactor=0	Complication=0	= .	Er=1 & inpt=0 & outpt=0	DK (1698)
48	.	No	No	Cofactor=0	Complication=0	= .	Er=0 & inpt=1 & outpt=0	DK (3988)
49	.	No	No	Cofactor=0	Complication=0	= .	Er=0 & inpt=0 & outpt=1	DK (1190)
50	.	No	No	DKA=0 AND hypo_coma=0	Complication≥0	= .		DK (12170)
51	.	Yes	No	Cofactor=0	Complication=0	= .		DK (19190)
52	3	No	No	Cofactor=0	Complication=0	= .		DK (594)
53	.	No	No	(Cofactor>0	OR Complication>0)	=.		DK (24)

Abbreviation: OHA: oral hypoglycemic agents; **age_comp**: age at having first complication; **age_CVE**: age at the first cardiovascular event including myocardial infarction (MI) and coronary artery bypass graft (CABG); **DK**: unknown types of diabetes; **DKA**: diabetic ketoacidosis; **ER**: ICD-9 code 250 for emergency room diagnosis; **hypocoma**: hypoglycemic coma; **inpt**: ICD-9 code 250 for in-patient diagnosis; **outpt**: ICD-9 code 250 for out-patient diagnosis; **T1DM**: type 1 diabetes mellitus; **T2DM**: type 2 diabetes mellitus; **text_T1DM**: with any related T1DM diagnosis in the parsing notes; **text_T2DM**: with any related T2DM diagnosis in the parsing notes

¶: ICD_DM: **1**= only with ICD-9 code for T1DM specific diagnosis (250.x1 or 250.x3); **2**= only with ICD-9 code for T2DM or unspecified type DM diagnosis (250.x0 or 250.x2); **3**= with both ICD-9 code for T1DM and T2DM or unspecified diagnosis [(250.x1 or 250.x3) AND (250.x0 or 250.x2)]; **missing values**= without any ICD-9 code diagnosis for DM.

¥: Cofactors include diabetic ketoacidosis (DKA), hypoglycemic coma, thyroid, Addison, and celiac diseases

φ: Complications include coronary artery bypass graft (CABG), myocardial infarction (MI), dialysis, retinopathy, neuropathy, and amputations.

Table 6: Defined Clinical Rules to Identify "Possible" Type 1 and Type 2 Diabetic Cases in the MARS Data

Category	ICD_DM[¶]	Insulin	OHA	Cofactors[¥]	Complications[φ] or age at having first complications	Age	Other factors	DM type (n)
56	3	Yes	Yes	DKA=1	CVE=1 & age_CVE > 40	≠ .		T2DM (13)
57	3	Yes	No		CVE=1 & age_CVE > 40	≠ .		T2DM (138)
58	1	Yes	Yes	DKA=1 OR hypocoma=1		= .		T1DM (1)
59	1	No	No	Cofactor>0		= .		T1DM (3)
60	1	No	No	Cofactor=0	Complication=0	= .	ER=1 OR inpt=1 OR outpt=1	T1DM (58)
61	1	No	No	Cofactor=0	Complication=0	= .	ER=0 & inpt=0 & oupt=0	T1DM (125)
62	1	Yes	Yes			= .		T2DM (12)
63	2	Yes	No	DKA=1	Age_complication >40	≠ .		T2DM (15)
64	2	Yes	No	Cofactor>0	Complication>0	≠ .		T2DM (193)
65	2	No	No	Cofactor>0		≠ .		T2DM (71)
66*	2					= .		T2DM (12014)
67	3	Yes	Yes	Cofactor>0		>40		T2DM (185)
68	3	Yes	No	Cofactor>0	Complication>0	>40		T1DM (8)
69	3	Yes	No	DKA=0 AND hypocoma=0		≠ . , ≤ 40		T1DM (97)
70	3	Yes	No	Cofactor>0	Complication=0	≠ .		T1DM (21)
71	3	Yes	No	Cofactor=0	Complication=0	40<age<65		T2DM (1772)
72	3	Yes	No	Cofactor=0	Complication>0 AND age_comp>40	≠ .		T2DM (160)
73	3	No	No	DKA=0 AND hypocoma=0	Complication=0	≠ . , ≤ 30		T1DM (290)
74	3	No	No	DKA=0 AND hypocoma=0	Complication=0	≠ . , > 30		T2DM (2679)
75	3	Yes	Yes	DKA=1 OR hypocoma=1	Complication=0	= .		T2DM (61)
76	3	Yes	Yes	DKA=1 OR hypocoma=1	Complication>0	= .		T2DM (32)
77	.	Yes	Yes	DKA=1 OR hypocoma=1		≠ .		T2DM (9)
78	.	Yes	No		Complication=0	≠ . , ≤ 30		T1DM (24)
79	.	Yes	No	DKA=0 AND hypocoma=0	Complication=0	≠ .		T2DM (9720)
80	.	Yes	No		Complication>0	≠ . , >40		T2DM (180)
81	.	No	No	Cofactor>0	Complication=0	≠ . , ≤ 30		T1DM (13)
82	.	No	No		CVE=1 AND age_CVE>40	≠ .		T2DM (1111)
83	.	No	No		Complication>0	≠ .		T2DM (26)
84	.	No	No	DKA=0 AND hypocoma=0	Complication =0	≠ .		T2DM (3483)

Table 6 continued

Category	ICD_DM[¶]	Insulin	OHA	Cofactors[¥]	Complications[φ] or age at having first complications	Age	Other factors	DM type (n)
85	.	Yes	Yes	DKA=1 OR hypocoma=1		=.	ER=1 OR text_T1DM=1	T1DM (148)
86	.	Yes	Yes			= .		T2DM (40)
87	3	Yes	No	DKA=0 AND hypocoma=0	Complication>0	= .	ER=1 OR text_T1DM=1	T1DM (328)
88	3	Yes	No			= .		T1DM (745)
89*	3	No	No	(Cofactor>0	OR complication >0)	= .		T2DM (98)
90	.	Yes	No	DKA=1 OR hypocoma=1		= .		T1DM (291)
91	.	Yes	No	DKA=0 AND hypocoma=0	Complication=0	= .		T2DM (738)
92*	.	Yes	No	DKA=0 AND hypocoma=0	Complication>0	= .		T2DM (5722)

Abbreviation: OHA: oral hypoglycemic agents; **age_comp**: age at having first complication; **age_CVE**: age at the first cardiovascular event including myocardial infarction (MI) and coronary artery bypass graft (CABG); **DK**: unknown types of diabetes; **DKA**: diabetic ketoacidosis; **ER**: ICD-9 code 250 for emergency room diagnosis; **hypocoma**: hypoglycemic coma; **inpt**: ICD-9 code 250 for in-patient diagnosis; **outpt**: ICD-9 code 250 for out-patient diagnosis; **T1DM**: type 1 diabetes mellitus; **T2DM**: type 2 diabetes mellitus; **text_T1DM**: with any related T1DM diagnosis in the parsing notes; **text_T2DM**: with any related T2DM diagnosis in the parsing notes

¶: ICD_DM: **1**= only with ICD-9 code for T1DM specific diagnosis (250.x1 or 250.x3); **2**= only with ICD-9 code for T2DM or unspecified type DM diagnosis (250.x0 or 250.x2); **3**= with both ICD-9 code for T1DM and T2DM or unspecified diagnosis [(250.x1 or 250.x3) AND (250.x0 or 250.x2)]; **missing values**= without any ICD-9 code diagnosis for DM.

¥: Cofactors include diabetic ketoacidosis (DKA), hypoglycemic coma, thyroid, Addison, and celiac diseases

φ: Complications include coronary artery bypass graft (CABG), myocardial infarction (MI), dialysis, retinopathy, neuropathy, and amputations.

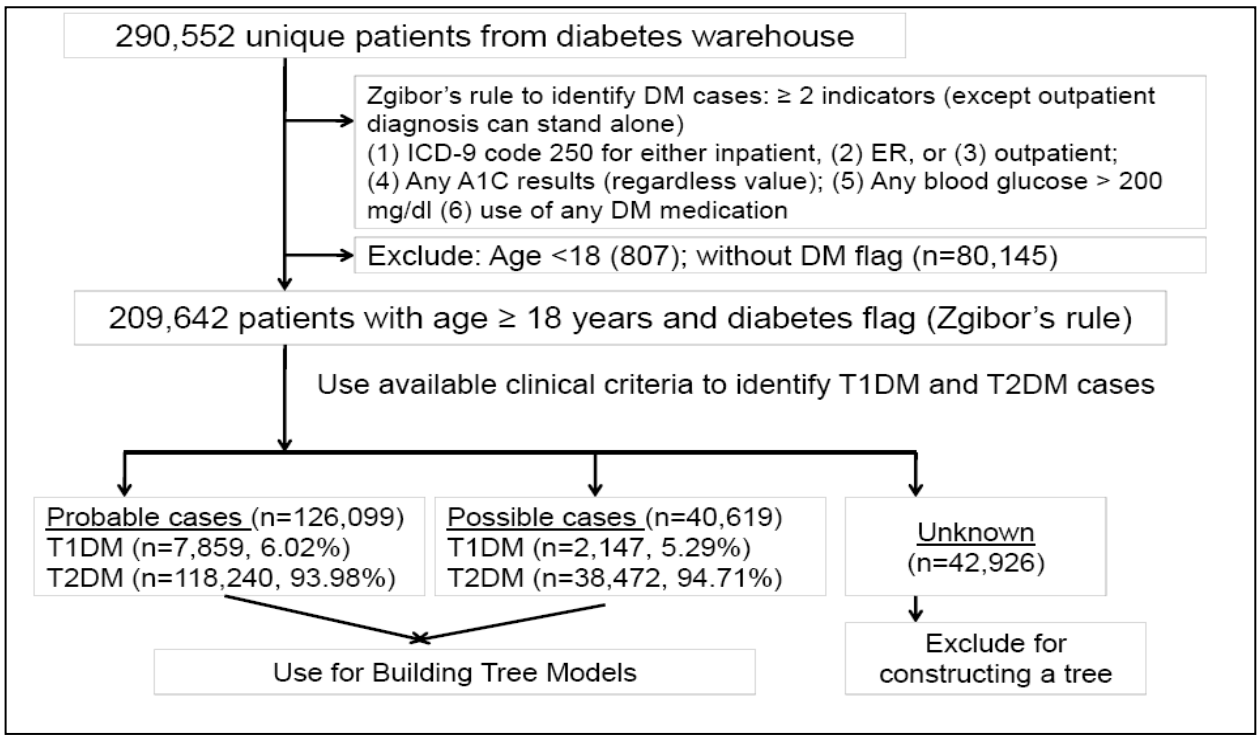

Figure 2. Obtaining the Final Dataset: Probable and Possible Cases

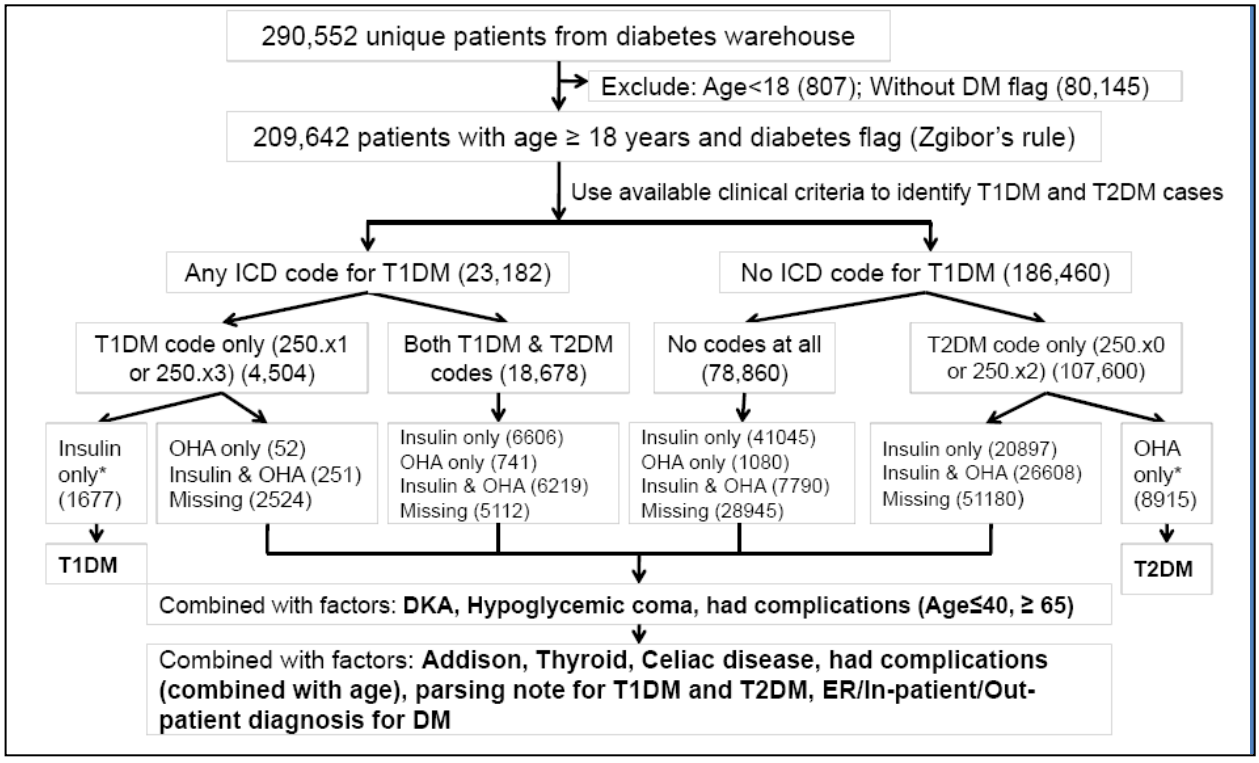

Figure 3. Obtaining the Final Dataset: Criteria to Distinguish T1DM and T2DM

3.4 STATISTICAL ANALYSIS AND DEVELOP TREE-STRUCTURED MODELS

A wide of variety of software package are available for implementing classification and regression trees. Popular commercial packages include Salford Systems CART, Rulequest's See5, R, TIBCO Spotfire S^+, and DTREG, to name a few. We used the ***"rpart"*** functions loaded in the Spotfire S+® 8.1 (TIBCO Software Inc.) to construct tree models. To ensure that we developed the best predictable model from the available data, we built tree models for probable cases and possible cases, separately. We also examined the tree models without ICD-9 codes diagnosis for T1DM and T2DM. The tree-model that best served our objectives of simplicity and accuracy will be selected as the best model. The details of the programming codes were included in the ***Appendix A.***

The basis of the decision tree algorithms is the binary recursively partitioning of the data into homogenous subsets. Starting at the tree root, the most discriminative variable is first selected to partition the data set into leaf nodes. The partitioning is repeated until the nodes are homogeneous enough to be terminal. The successive split was chosen such that the binomial deviance was minimized, and the nodes became increasingly homogeneous with respect to the proportion of individuals who are either T1DM or T2DM. The ***"rpart"*** functions in the S-Plus uses 10-fold cross-validation, cost-complexity pruning and surrogate splits.[8, 31] The classification tree is grown to its maximum size and then pruned on the basis of a criterion that balances the number of terminal nodes (complexity) against the accuracy of the tree in classifying people, sometimes termed misclassification cost. Cost-complexity pruning is performed to stop generating new split nodes when subsequent splits only result in very little overall improvement of the prediction. This prevents overfitting. Surrogate splits technique is

used to analyze the data of the subjects whose values for any variable are missing.[8, 30-31] If the primary splitter variable field is missing, a surrogate splitter variable is used instead. During the cross-validation process, all possible combinations of trees were developed using nine of the subsets and these were then tested on the 10^{th} subset. This result provides cross-validation error rate, which gives an equitable evaluation of the predictive precision of tree models of different sizes. Using this value, the optimal tree with the minimum error is selected. We evaluated predictive accuracy of the classification trees by using **sensitivity, specificity, positive predictive value (PPV)**, and **negative predictive value (NPV)** of T1DM. Ultimately, the classification rules will be validated internally through randomly chart reviews as our future work. The purpose of the chart review is to obtain an independent ascertainment of a diagnosis of T1DM and T2DM.

4.0 REASERCH QUESTIONS AND RESULTS

Among 209,642 diabetic patients with age of 18 years or older, 126,097 probable cases (T1DM: 7,857 and T2DM: 118,240), 40,619 (T1DM: 2,147 and T2DM: 38,472), and 42,926 patients with unknown diabetic type were identified. The sample consists of 103,502 females (49.37%) and 97,077 males (46.31%); 160,291 whites (76.46%), 25,433 blacks (12.13%), 335 Asians (0.16%), 496 Hispanics (0.24%), and 2,009 other races (0.96%). 9,063 (4.32%), 21,080 (10.06%), and 79,522 (37.93%) patients had missing values in gender, race, and age, respectively. **Table 7** shows the distribution of probable cases, possible cases, and patients with unknown diabetic types by demographic factors, and variables that contribute to identifying T1DM from T2DM. T1DM cases were significantly more likely than T2DM cases to be younger (mean ages: probable T1DM cases = 44.92, probable T2DM case = 65.00; possible T1DM cases = 27.30, and possible T2DM cases = 60.32), to be black, to use insulin, to have ICD-9 codes for T1DM diagnosis, to have complications earlier, and to have autoimmune-related comorbidities, to have DKA or hypoglycemic coma diagnosis, and to have in-patient, out-patient, or ER diagnosis of diabetes. T2DM cases were more likely than T1DM cases to be older, to be white, to have OHA use, and to have complications later. For each variable related to age, the distribution of the individuals with probable and possible T1DM and T2DM cases are displayed in **Figure 4**. The distributions of the various predictors defined by age differ substantially, the lower quartiles

of those with T2DM are just below the upper quartile of those with T1DM. Of all these age-related variables, “age” appears to be the best single predictor, because all other ages variables were calculated from “age”.

Table 7: Descriptive Distributions of Each Predictor Variable by Probable, Possible, and Unknown Cases in the MARS dataset

Variable*	Probable T1DM (n=7,857)	Probable T2DM (n=118,240)	Possible T1DM (n=2,147)	Possible T2DM (n=38,472)	Unknown types (n=42,926)
Age at first entry into MARS database (years, SD)	44.92 (18.91)	65.00 (14.88)	27.30 (8.45)	60.32 (15.18)	47.19 (11.58)
Missing values (%)	433 (5.51)	15,799 (13.36)	1,699 (79.13)	18,717 (48.65)	42,874 (99.88)
Gender					
Male (%)	4,012 (51.06)	56,432 (47.73)	975 (45.41)	17,505 (45.50)	18,153 (42.29)
Female (%)	3,702 (47.12)	60,491 (51.16)	1108 (51.61)	17,841 (46.37)	20,360 (47.43)
Missing values	143 (1.82)	1,317 (1.11)	64 (2.98)	3,126 (8.13)	4,413 (10.28)
Race (%)					
White	5,902 (75.12)	93,168 (78.80)	1,528 (71.17)	28,330 (73.64)	31,363 (73.06)
Black	1,093 (13.91)	14,922 (12.62)	466 (21.70)	4,318 (11.22)	4,634 (10.80)
Asian	8 (0.10)	187 (0.16)	3 (0.14)	69 (0.18)	68 (0.16)
Hispanics	27 (0.34)	303 (0.26)	6 (0.28)	74 (0.19)	84 (0.20)
Other	75 (0.95)	1,214 (1.03)	20 (0.93)	329 (0.86)	371 (0.86)
Missing value	752 (9.52)	8,446 (7.14)	124 (5.78)	5,352 (13.91)	6,406 (14.92)
In-patient insulin use (%)	5,361 (68.23)	61,419 (51.94)	1,658 (77.22)	22,645 (58.86)	20,010 (46.62)
In-patient OHA use (%)	183 (2.33)	50,907 (43.05)	149 (6.94)	417 (1.08)	0 (0.00)
ICD code (%)[¶]					
T1DM only	4,044 (51.47)	261 (0.22)	187 (8.71)	12 (0.03)	0 (0.00)
T2DM or unspecified only	30 (0.38)	95,277 (80.58)	0 (0.00)	12,293 (31.95)	0 (0.00)
Both T1DM and T2DM	1,837 (23.38)	8,448 (7.14)	1,484 (69.12)	5,138 (13.36)	1771 (4.13)
Missing value	1,946 (24.77)	14,254 (12.06)	476 (22.17)	21,029 (54.66)	41,155 (95.87)
Age at having first complication[φ] in MARS database (years, SD)	48.53 (17.78)	70.52 (11.92)	***	52.5 (9.70)	72.45 (12.13)
History of CVE (%)	437 (5.56)	10,006 (8.46)	260 (12.11)	7,565 (19.66)	2,558 (5.96)
CABG	246 (3.13)	6,136 (5.19)	156 (7.27)	4,647 (12.08)	1,259 (2.93)
MI	305 (3.88)	6,143 (5.20)	172 (8.20)	4,764 (12.38)	1,871 (4.36)

Table 7 continued

Variable*	Probable T1DM (n=7,857)	Probable T2DM (n=118,240)	Possible T1DM (n=2,147)	Possible T2DM (n=38,472)	Unknown types (n=42,926)
Age at having first CVE (years, SD)	53.92 (14.28)	71.71 (11.32)	***	***	74.98 (10.98)
CABG	57.07 (13.83)	72.33 (10.38)	***	***	74.98 (9.75)
MI	52.03 (14.01)	71.16 (11.92)	***	***	***
History of dialysis (%)	171 (2.18)	594 (0.50)	65 (3.03)	778 (2.02)	211 (0.49)
Age at having first dialysis (years, SD)	42.98 (13.69)	66.53 (13.14)	52.5 (9.19)	65.44 (12.95)	***
History of amputation (%)	73 (0.93)	478 (0.40)	45 (2.10)	350 (0.91)	114 (0.27)
Age at having first amputation (years, SD)	48.05 (17.03)	68.60 (12.61)	***	48.00 (7.00)	67.52 (13.00)
History of retinopathy (%)	295 (3.75)	868 (0.73)	330 (0.86)	163 (7.59)	17 (0.04)
Age at having first retinopathy (years, SD)	49.63 (13.03)	65.76 (11.89)	***	68.84 (10.37)	***
History of neuropathy (%)	397 (5.05)	2,976 (2.52)	293 (13.65)	716 (1.86)	24 (0.06)
Age at having first neuropathy (years, SD)	50.14 (12.84)	66.97 (12.51)	55.2 (11.90)	67.19 (11.78)	***
History of DKA (%)	1,414 (18.00)	254 (0.21)	426 (19.84)	726 (1.86)	23 (0.05)
History of hypoglycemic coma (%)	740 (9.42)	170 (0.14)	148 (6.89)	81 (0.21)	8 (0.02)
History of thyroid disease (%)	53 (0.67)	566 (0.48)	45 (2.10)	786 (2.04)	338 (0.79)
History of celiac disease (%)	22 (0.28)	63 (0.05)	12 (0.56)	96 (0.25)	28 (0.07)
History of Addison disease (%)	66 (0.84)	313 (0.26)	35 (1.63)	951 (2.47)	153 (0.36)
With in-patient-DM diagnosis (%)	5,144 (65.47)	77,512 (65.55)	1,520 (70.80)	27,214 (70.74)	33,291 (77.55)
With out-patient-DM diagnosis (%)	1,292 (16.44)	11,241 (9.51)	609 (28.37)	8438 (21.93)	13,065 (30.44)
With ER-DM diagnosis (%)	2,826 (35.97)	37,943 (32.09)	1,288 (59.99)	15,153 (39.39)	18,211 (42.42)

Abbreviation: OHA: oral hypoglycemic agents; **CABG**: coronary artery bypass graft; **CVE**: cardiovascular events including myocardial infarction (MI) and coronary artery bypass graft (CABG); **DKA**: diabetic ketoacidosis; **ER**: emergency room; **MI**: myocardial infarction; **T1DM**: type 1 diabetes mellitus; **T2DM**: type 2 diabetes mellitus

*: All the variables are significantly different cross different groups.

***: All of the patients in these groups are with missing values in age or only one case with age value.

¶: ICD_DM: **T1DM only**= only with ICD-9 code for T1DM specific diagnosis (250.x1 or 250.x3); **T2DM or unspecified only**= only with ICD-9 code for T2DM or unspecified type DM diagnosis (250.x0 or 250.x2); **Both T1DM and T2DM**= with both ICD-9 code for T1DM and T2DM or unspecified diagnosis [(250.x1 or 250.x3) AND (250.x0 or 250.x2)]; **missing values**= without any ICD-9 code diagnosis for DM.
φ: Complications include coronary artery bypass graft (CABG), myocardial infarction (MI), dialysis, retinopathy, neuropathy, and amputations.

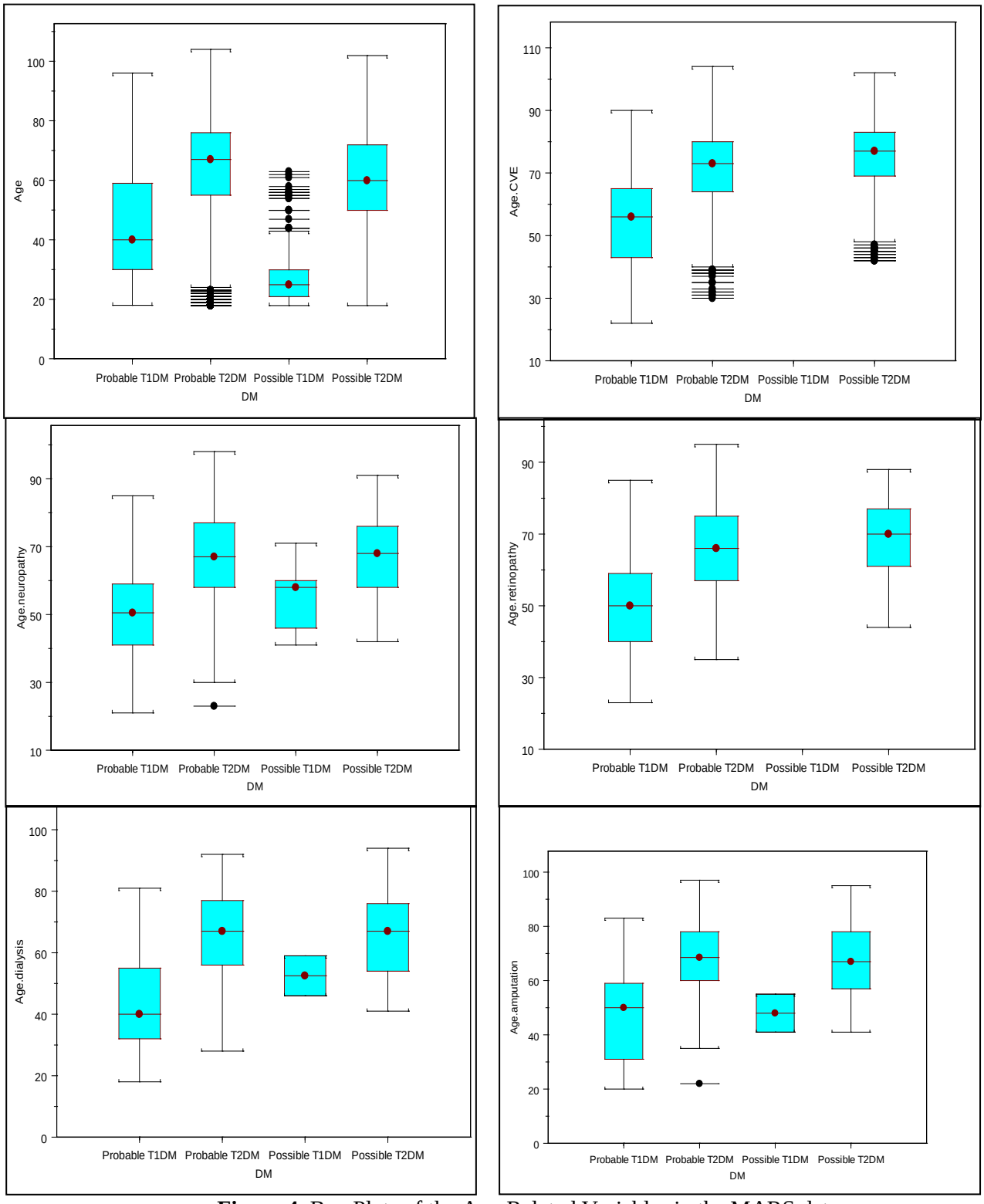

Figure 4. Box Plots of the Age-Related Variables in the MARS data.

4.1 WHAT ARE THE MAIN CHARACTERISTICS TO DISTINGUISH T1DM AND T2DM?

Constructing a good classifier whose performance will stand up under cross-validation and that is useful and practical is our first priority. All the predictors available in the MARS data (**Table 3**) were used in the estimation. In this section, we describe the results from the default tree models briefly and further addressed results shown in the pruned trees in more detail. For the **"probable cases"**, **Figure 5** presents the default tree-based model whereby successive partitions of the data into homogenous subsets are shown with the rule labeling each split. This form of tree display is primarily for presentation purpose, as it conceals the details of the tree-growing process. The numbers under each terminal node are the observations of T1DM and T2DM, separately; for example, in the leftmost terminal node in the **Figure 5,** which labeled T1DM, 4,015 patients were actually probable T1DM cases and 0 patients were actually probable T2DM cases. The predictors selected in the default tree model were ICD codes for DM diagnosis, OHA use, history of DKA diagnosis, in-patient insulin use, age of 40.5 years, duration in the MARS database, and in-patient diagnosis of DM. The algorithm underlying tree-based prediction determines the cutoff point of age more objectively (by optimization) as 40.5 years. The result of repeated "recursive partitioning" also leads to the tree displayed in the **Table 8**. The semigraphical representation is different from those used in **Figure 5**. It is most useful when the details of the fitting procedure are of interest. In the **Table 8**, the first number after the split is the number of observations, the second number is the loss number from target category, and the third number is the predicted target category. The "yprob" gives the probabilities of T1DM and T2DM individually in each node.

Note that the tree has not been pruned to the final size yet. The pruning of the classification tree in **"rpart"** defaults impurity criterion the **Gini index**. The pruining criterion is the predicted loss, or normally called error rate. We then pruned the default tree to an optimally sized tree based on the cross-validation results in the **Table 9 and Figure 6**. In **Table 9**, the columns xerror and xstd are random, depending on the random partition used in the 10-fold cross-validation that has been computed within **"rpart"**. The table is printed from the smallest tree (no splits) to the largest one. The number of splits is listed, rather than the number of nodes. The number of nodes is always 1 plus the number of splits. For easier reading, the error columns have been scaled so that the first node has an error of 1. All the errors are the proportions of the error for the root tree. In **Table 9**, the model with no splits must make 7,860/126,097 misclassifications, multiply columns 3 by 7,860 to get a results in terms of absolute error (computations are done on the absolute error scale, and printed on relative scale). The **complexity parameter (cp)** may then be chosen to minimize xerror. A practical procedure is to use the **1-SE rule**; i.e., pick up the value no larger than minimum xerror within one standard deviation. In **Table 9**, the 1-SE rule gives 0.157+0.00445, so we chose the line 6, a tree with 7 splits and 8 terminal nodes. This is easier to see graphically in **Figure 6**, where we take the leftmost pruning point with value below the line.

Now, we explain the pruned tree explicitly (**Figure 7 and Table 10**). The split on "ICD.DMtype" (i.e., if ICD.DM type=1, go to the left node; if ICD.DMtype = 2 or 3 or missing, go to the right node) partitions the 126,097 observations into group of 4,305 and 121,792 individuals (nodes 2 and 3), with the probability of T1DM of 0.93937 and 0.03130, respectively. The first group is then partitioned into groups of 4,015 and 290 individuals (nodes 4 and 5), depending on whether they use OHA or not. The former group, with probability of T1DM of 1.0,

is not subdivided further. The latter group, with probability of T1DM of 0.1 (i.e., probability of T2DM of 0.9), is also a terminal node. The group of the node 3 is subdivided into group of 1,399 and 120,393 individuals (nodes 6 and 7), depending on whether or not having a DKA diagnosis. The retrospective probabilities of T1DM for these groups are 0.81844 and 0.02216. The group of the node 6 is subdivided into group of 1,146 and 253 individuals (nodes 12 and 13), depending on ICD codes for DM (i.e., if ICD.DM type=3, go to the left node; if ICD.DMtype = 2 or missing, go to right node). Both of the groups are terminal nodes. The probabilities of T1DM are 0.98778 and 0.05138 in the former and latter groups, separately. The group of the node 7 is partitioned into group of 8,757 and 111,636 individuals (nodes 14 and 15); depending on age at first entry into the MARS database is < 40.5 or ≥ 40.5 years. The latter group, with probability of T1DM of 0.00236 (i.e., probability of T2DM of 0.99764), is not subdivided further. The former group continued partitioning process based on in-patient insulin and OHA use. This procedure continues, yielding 8 distinct probabilities of T1DM ranging from 0 to 1. Clearly, as the partitioning continues, our trust in the individual estimated probabilities decreases as they are based on less and less data.

Considering ICD-9 codes for T1DM or T2DM as a strong predictor presumably, we evaluated the tree model leaving the variable "ICD.DMtype" out of the formula. In **Figure 8 and Table 9**, the pruned tree is as same as the full-sized tree since selected cp is 0.01 as same as the default of cp in the **"rpart"**. The predictors selected in the pruned tree model were age of 40.5 years, insulin use, history of DKA diagnosis and OHA use. Obviously, it became more heterogeneous in the terminal nodes labeled as T1DM with probabilities of T2DM from 0 to 0.25.

For the **"possible cases"**, **Figure 9 and Table 12** present the default tree-based model. Selected predictors include ICD codes for DM diagnosis, age of 30.5 years, history of DKA diagnosis, history of hypoglycemic coma, in-patient insulin use, in-patient OHA use, age of 40.5 years, duration in the MARS database, and numbers of complications recorded. The overplotting of labels is a common occurrence with this type of display when the tree is big with many splits. Then, we chose the pruned tree in size of 12 based on cross-validation results in the **Table 13 and Figure 10**. In **Table 9**, we selected 0.644+0.0170 based on the 1-SE rule. We also can take the leftmost pruning point with value below the line shown in **Figure 10.** The pruned tree is exactly equal to the original default tree (**Figure 11 and Table 12**). The split on "ICD.DMtype" (i.e., if ICD.DM type=1 or 3, go to the left node; if ICD.DMtype = 2 or missing, go to the right node) partitions the 40,619 observations into group of 7,018 and 33,601 individuals (nodes 2 and 3), with the probability of T1DM of 0.25919 and 0.00976, respectively. The first group is then partitioned into groups of 315 and 6,703 individuals (nodes 4 and 5), depending on whether age < 30.5 or ≥ 30.5 years. The former group, with probability of T1DM of 1.0, is not subdivided further. The group of node 5 continued subdivided into group of 3,740 and 2,963 individuals (nodes 10 and 11), depending on whether or not having insulin use. The retrospective probabilities of T1DM for these groups are 0.35241 and 0.06277. The group of the node 10 is subdivided into group of 69 and 3,671 individuals (nodes 20 and 21), depending on age <40.5 or ≥ 40.5 years. The former group became a terminal node with T1DM probability of 1.0. The latter group further partitioned into the groups of 2,188 and 1,483 depending on their duration in the MARS database less or longer than 1468.5 days (about 4 years). The latter group is a terminal node with the probability of T1DM of 0.19218. The group of the node 42 is further subdivided according to the durations in the MARS database (less or longer than 2672.5 days, or

about 7 years), in-patient OHA use and numbers of complications occurred. The group of node 3 continued subdivided into group of 633 and 32,938 individuals (nodes 6 and 7), depending on whether or not having DKA diagnosis. The latter group became a terminal node with the probability 0.00154 of T1DM. The group of the node 6 is partitioned into group of 108 and 555 individuals (nodes 12 and 13) based on having hypoglycemic coma or not. Both of them became terminal nodes with the predicted probabilities of T1DM being 0.96296 and 0.31171. This procedure yielding 12 distinct probabilities of T1DM ranging from 0 to 1. Clearly, as the partitioning continues, our trust in the individual estimated probabilities decreases as they are based on less and less data. The heterogeneity became more obvious in the nodes with smaller splits. Similarly, in **Figure 8 and Table 9**, the pruned tree without considering ICD-9 code as a predictor in the model selected age of 30.5 years, in-patient diagnosis of DM, duration in the MARS database, in-patient insulin use, history of DKA diagnosis, history of hypoglycemic coma, and in-patient OHA use. Compared with the pruned tree with ICD-9 code as a predictor, it became more heterogeneous in the terminal nodes labeled as T1DM with probabilities of T2DM from 0.02 to 0.33.

In summary, the main characteristics to distinguish T1DM and T2DM include ICD-9 codes for T1DM and T2DM, in-patient insulin use, in-patient OHA use, history of DKA diagnosis, and age of 40.5 years across all of the pruned tree-based models. History of hypoglycemic coma, duration in the MARS database, number of complications, and an-inpatient-diagnosis of DM may be considered as ancillary factors to predict T1DM cases.

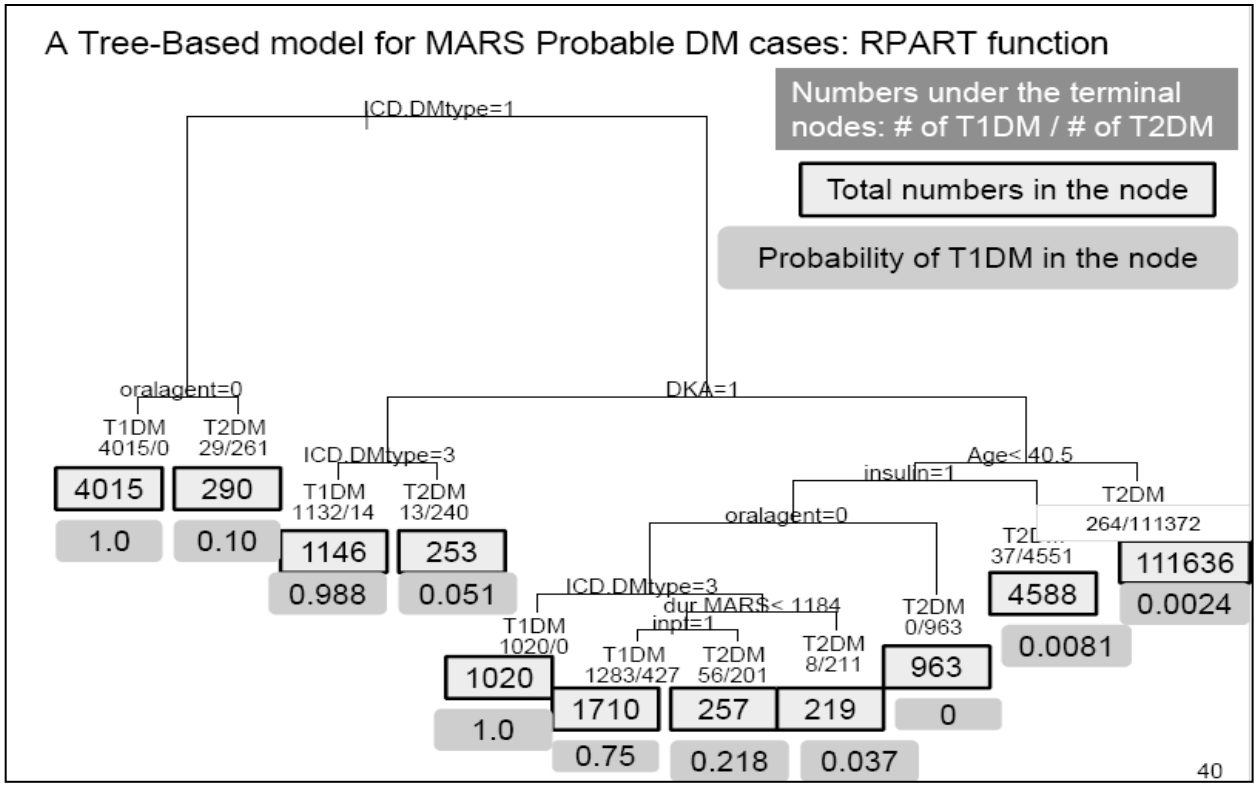

Figure 5. A Default Tree-Based Model for Predicting "Probable" T1DM and T2DM Cases

Table 8: A Default Tree-Based Model for Predicting "Probable" T1DM and T2DM Cases

```
node), split, n, loss, yval, (yprob)            * denotes terminal node
1) root 126097 7857 T2DM (0.06230 0.93769)
   2) ICD.DMtype=1 4305  261 T1DM (0.93937 0.06062)
     4) oralagent=0 4015    0 T1DM (1.00000 0.00000) *
     5) oralagent=1 290   29 T2DM (0.10000 0.90000) *
   3) ICD.DMtype=2,3 121792 3813 T2DM (0.03130 0.96869)
     6) DKA=1 1399  254 T1DM (0.81844 0.18156)
      12) ICD.DMtype=3 1146   14 T1DM (0.98778 0.01221) *
      13) ICD.DMtype=2 253   13 T2DM (0.05138 0.94862) *
     7) DKA=0 120393 2668 T2DM (0.02216 0.97784)
      14) Age< 40.5 8757 2404 T2DM (0.27452 0.72548)
        28) insulin=1 4169 1802 T1DM (0.56776 0.43224)
          56) oralagent=0 3206  839 T1DM (0.73830 0.26170)
            112) ICD.DMtype=3 1020    0 T1DM (1.00000 0.00000) *
            113) ICD.DMtype=2 2186  839 T1DM (0.61619 0.38381)
              226) dur.MARS< 1183.5 1967  628 T1DM (0.68073 0.31927)
                452) inpt=1 1710  427 T1DM (0.75029 0.24971) *
                453) inpt=0 257   56 T2DM (0.21790 0.78210) *
              227) dur.MARS>=1183.5 219    8 T2DM (0.03653 0.96347) *
          57) oralagent=1 963    0 T2DM (0.00000 1.00000) *
        29) insulin=0 4588   37 T2DM (0.00806 0.99194) *
      15) Age>=40.5 111636  264 T2DM (0.00236 0.99764) *
```

Table 9: Complexity Parameters from Cross-validations for Predicting "Probable" T1DM and T2DM Cases

Classification tree:
rpart(formula = DM ~ sex + race2 + Age + ICD.DMtype + insulin + oralagent + inpt + outpt + er + DKA + Hypo.Coma + CVE + dialysis + retinopathy + neuropathy + amputation + thyroid + celiac + addison + dur.MARS + Age.CVE + Age.neuropathy + Age.retinopathy + Age.dialysis + Age.amputation + Age.complication + count.cofactor + count.complication, data = mars.031010.dm12.probable, method = "class")
Variables actually used in tree construction:
[1] Age DKA ICD.DMtype dur.MARS inpt insulin oralagent
Root node error: 7860/126097 = 0.0623; n= 126097

	CP	nsplit	rel error	xerror	xstd
1	0.4815	0	1.000	1.000	0.01092
2	0.1134	1	0.519	0.519	0.00799
3	0.0648	2	0.405	0.405	0.00709
4	0.0295	5	0.211	0.211	0.00514
5	0.0289	6	0.181	0.193	0.00493
6	0.0129	7	0.152	0.157	0.00445
7	0.0100	10	0.108	0.169	0.00462

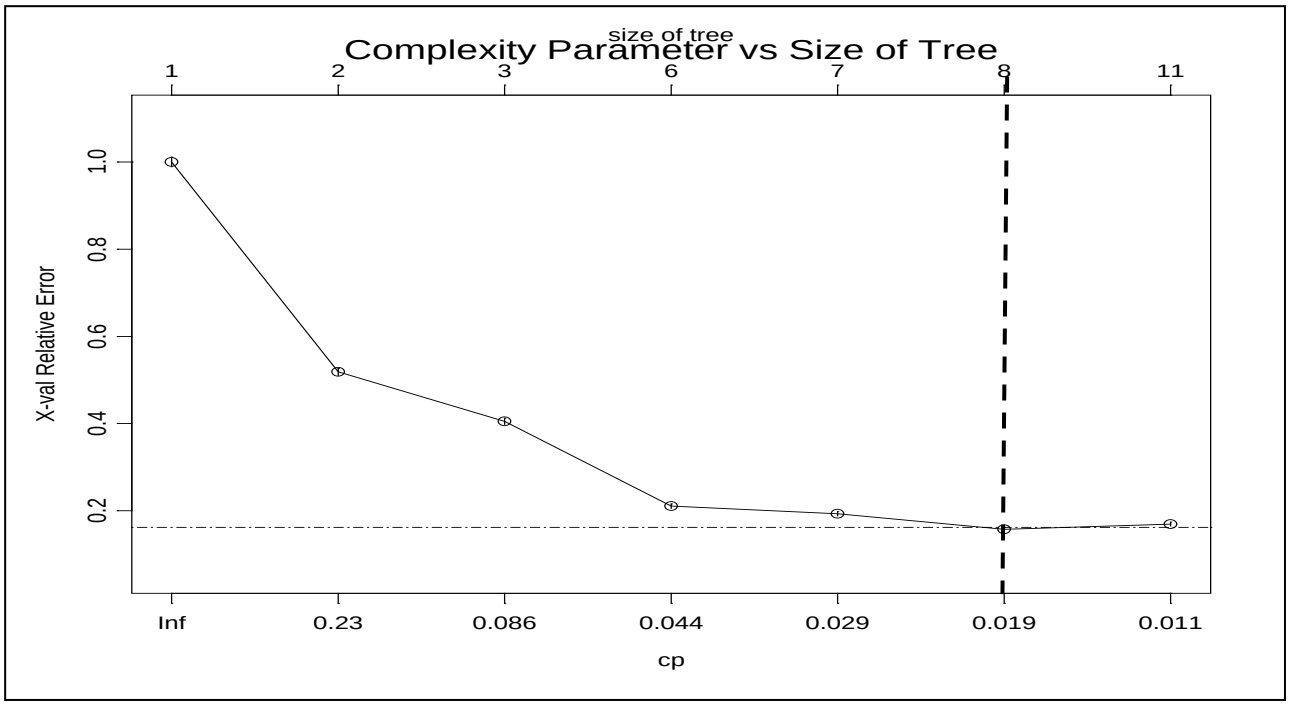

Figure 6. Complexity Parameter Plot for Predicting "Probable" T1DM and T2DM Cases

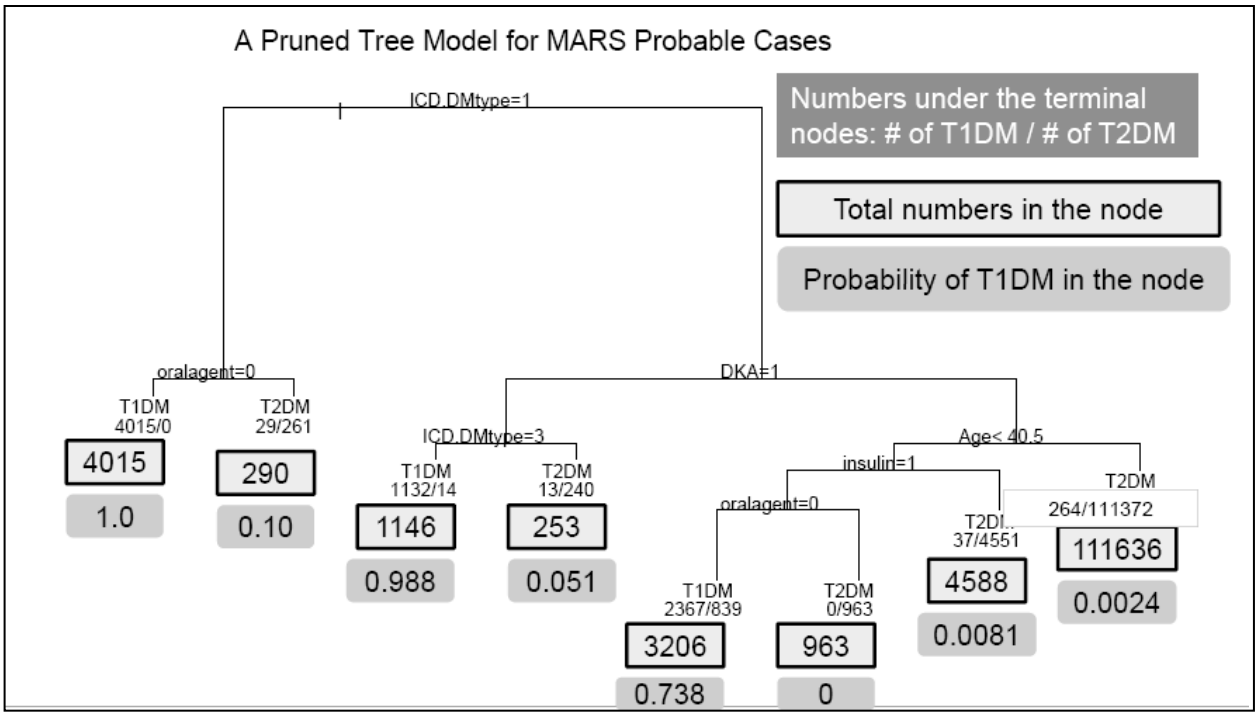

Figure 7. A Pruned Tree for Predicting "Probable" T1DM and T2DM Cases

Table 10: A Pruned Tree-based Model for Predicting "Probable" T1DM and T2DM Cases

```
node), split, n, loss, yval, (yprob)          * denotes terminal node
1) root 126097 7857 T2DM (0.06230 0.93769)
  2) ICD.DMtype=1 4305  261 T1DM (0.93937 0.06062)
    4) oralagent=0 4015    0 T1DM (1.00000 0.00000) *
    5) oralagent=1 290   29 T2DM (0.10000 0.90000) *
  3) ICD.DMtype=2,3 121792 3813 T2DM (0.03130 0.96869)
    6) DKA=1 1399  254 T1DM (0.81844 0.18156)
     12) ICD.DMtype=3 1146   14 T1DM (0.98778 0.01221) *
     13) ICD.DMtype=2 253   13 T2DM (0.05138 0.94862) *
    7) DKA=0 120393 2668 T2DM (0.02216 0.97784)
     14) Age< 40.5 8757 2404 T2DM (0.27452 0.72548)
       28) insulin=1 4169 1802 T1DM (0.56776 0.43224)
         56) oralagent=0 3206  839 T1DM (0.73830 0.26170) *
         57) oralagent=1 963    0 T2DM (0.00000 1.00000) *
       29) insulin=0 4588   37 T2DM (0.00806 0.99194) *
     15) Age>=40.5 111636  264 T2DM (0.00236 0.99764) *
```

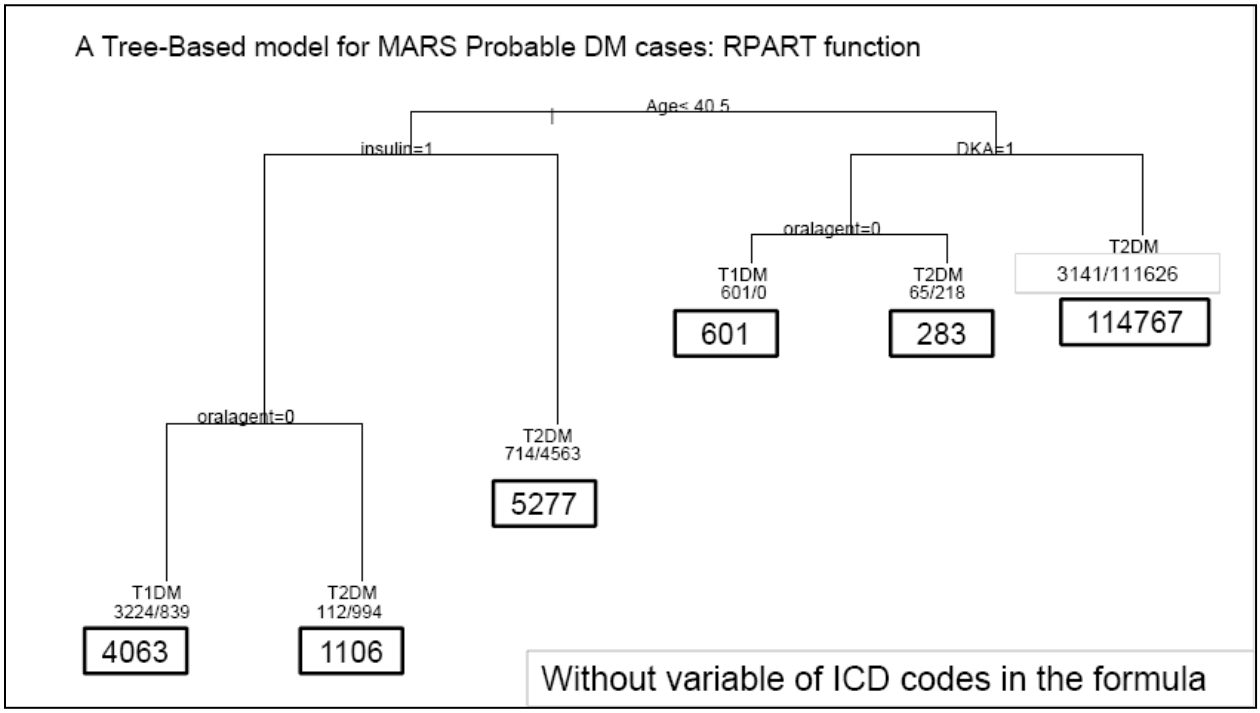

Figure 8. A Pruned Tree for Predicting "Probable" T1DM and T2DM Cases: Without ICD Codes in the Formula

Table 11: A Pruned Tree-Based Model and Complexity Parameters for Predicting "Probable" T1DM and T2DM Cases: Without ICD codes in the Formula

A pruned tree model	Complexity parameters and errors
n= 126097 node), split, n, loss, yval, (yprob) * denotes terminal node 1) root 126097 7857 T2DM (0.06230 0.93769) 2) Age< 40.5 10446 4050 T2DM (0.38771 0.61229) 4) insulin=1 5169 1833 T1DM (0.64539 0.35461) 8) oralagent=0 4063 839 T1DM (0.79350 0.20650) * 9) oralagent=1 1106 112 T2DM (0.10127 0.89873) * 5) insulin=0 5277 714 T2DM (0.13530 0.86470) * 3) Age>=40.5 115651 3807 T2DM (0.03291 0.96708) 6) DKA=1 884 218 T1DM (0.75339 0.24661) 12) oralagent=0 601 0 T1DM (1.00000 0.00000) * 13) oralagent=1 283 65 T2DM (0.22968 0.77032) * 7) DKA=0 114767 3141 T2DM (0.02736 0.97263) *	Variables actually used in tree construction: [1] Age DKA insulin oralagent Root node error: 7860/126097 = 0.0623 n= 126097 CP nsplit rel error xerror xstd 1 0.0956 0 1.000 1.000 0.01092 2 0.0570 3 0.696 0.696 0.00921 3 0.0195 4 0.639 0.639 0.00884 4 0.0100 5 0.620 0.620 0.00871

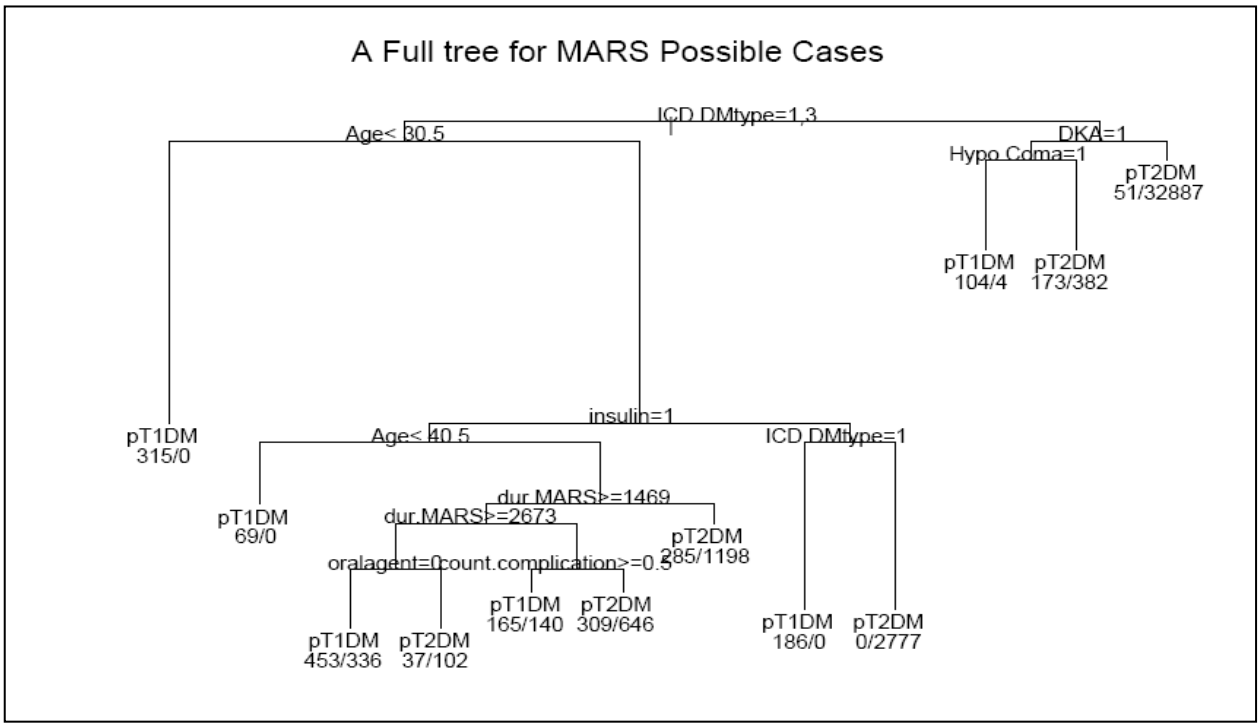

Figure 9. A Default tree-Based Model for Predicting "Possible" T1DM and T2DM Cases

Table 12: A Full and Pruned Tree-Based Model for Predicting "Possible" T1DM and T2DM Cases

node), split, n, loss, yval, (yprob)
*** denotes terminal node**

1) root 40619 2147 pT2DM (0.05285 0.94714
2) ICD.DMtype=1,3 7018 1819 pT2DM (0.25919 0.74081)
4) Age< 30.5 315 0 pT1DM (1.00000 0.00000) *
5) Age>=30.5 6703 1504 pT2DM (0.22438 0.77562)
10) insulin=1 3740 1318 pT2DM (0.35241 0.64759)
20) Age< 40.5 69 0 pT1DM (1.00000 0.00000) *
21) Age>=40.5 3671 1249 pT2DM (0.34023 0.65977)
42) dur.MARS>=1468.5 2188 964 pT2DM (0.44059 0.55941)
84) dur.MARS>=2672.5 928 438 pT1DM (0.52802 0.47198)
168) oralagent=0 789 336 pT1DM (0.57414 0.42586) *
169) oralagent=1 139 37 pT2DM (0.26619 0.73381) *
85) dur.MARS< 2672.5 1260 474 pT2DM (0.37619 0.62381)
170) count.complication>=0.5 305 140 pT1DM (0.54098 0.45902) *
171) count.complication< 0.5 955 309 pT2DM (0.32356 0.67644) *
43) dur.MARS< 1468.5 1483 285 pT2DM (0.19218 0.80782) *
11) insulin=0 2963 186 pT2DM (0.06277 0.93723)
22) ICD.DMtype=1 186 0 pT1DM (1.00000 0.00000) *
23) ICD.DMtype=3 2777 0 pT2DM (0.00000 1.00000) *
3) ICD.DMtype=2 33601 328 pT2DM (0.00976 0.99024)
6) DKA=1 663 277 pT2DM (0.41780 0.58220)
12) Hypo.Coma=1 108 4 pT1DM (0.96296 0.03703) *
13) Hypo.Coma=0 555 173 pT2DM (0.31171 0.68829) *
7) DKA=0 32938 51 pT2DM (0.00154 0.99845) *

Table 13: Complexity Parameters from Cross-Validations for Predicting "Possible" T1DM and T2DM Cases

Classification tree:
rpart(formula = DM ~ sex + race2 + Age + ICD.DMtype + insulin + oralagent + inpt + outpt + er + DKA + Hypo.Coma + CVE + dialysis + retinopathy + neuropathy + amputation + thyroid + celiac + addison + dur.MARS + Age.CVE + Age.neuropathy + Age.retinopathy + Age.dialysis + Age.amputation + Age.complication + count.cofactor + count.complication, data = mars.031010.dm12.possible, method = "class")

Variables actually used in tree construction:
[1] Age DKA Hypo.Coma ICD.DMtype count.complication dur.MARS insulin oralagent

Root node error: 2150/40619 = 0.0529; n= 40619

	CP	nsplit	rel error	xerror	xstd
1	0.0734	0	1.000	1.000	0.0210
2	0.0433	2	0.853	0.853	0.0195
3	0.0321	4	0.767	0.767	0.0185
4	0.0233	5	0.735	0.735	0.0181
5	0.0182	7	0.688	0.677	0.0174
6	0.0116	10	0.633	0.670	0.0174
7	**0.0100**	**11**	**0.622**	**0.644**	**0.0170**

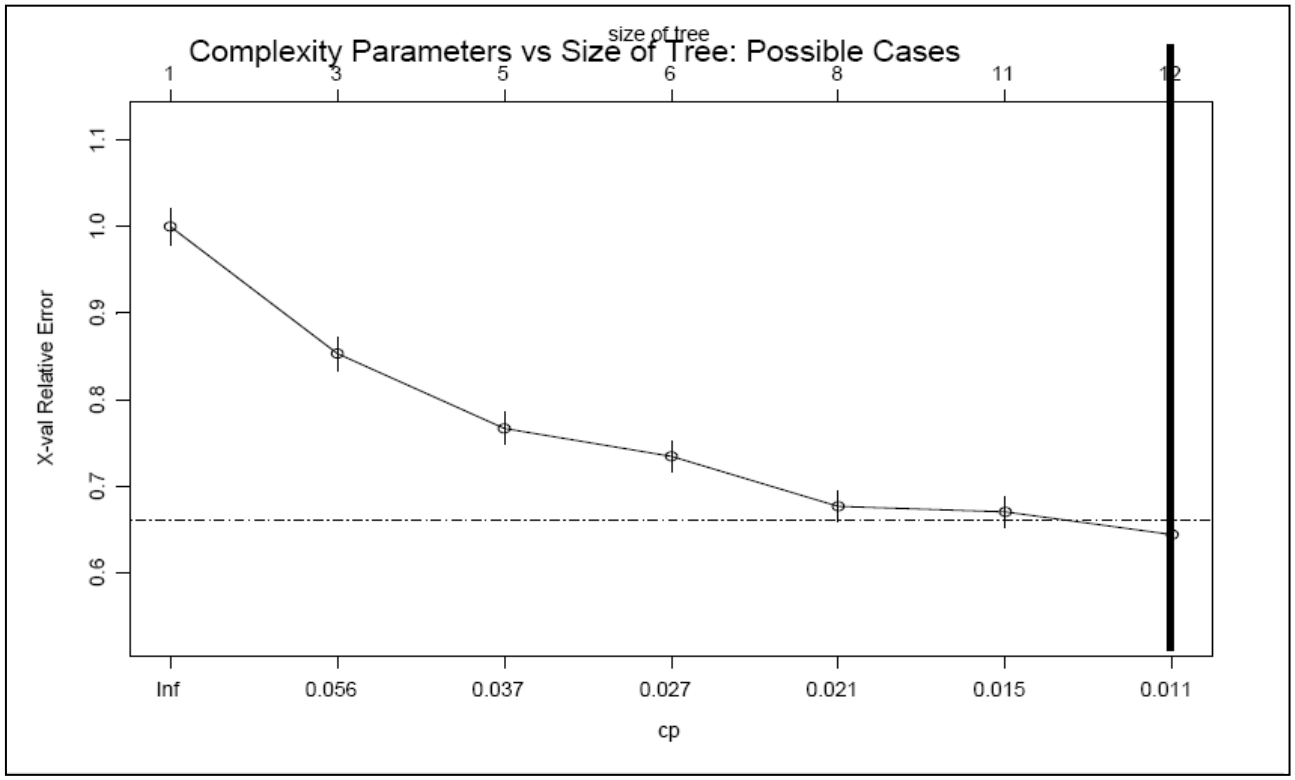

Figure 10. Complexity Parameter Plot for Predicting "Possible" T1DM and T2DM Cases

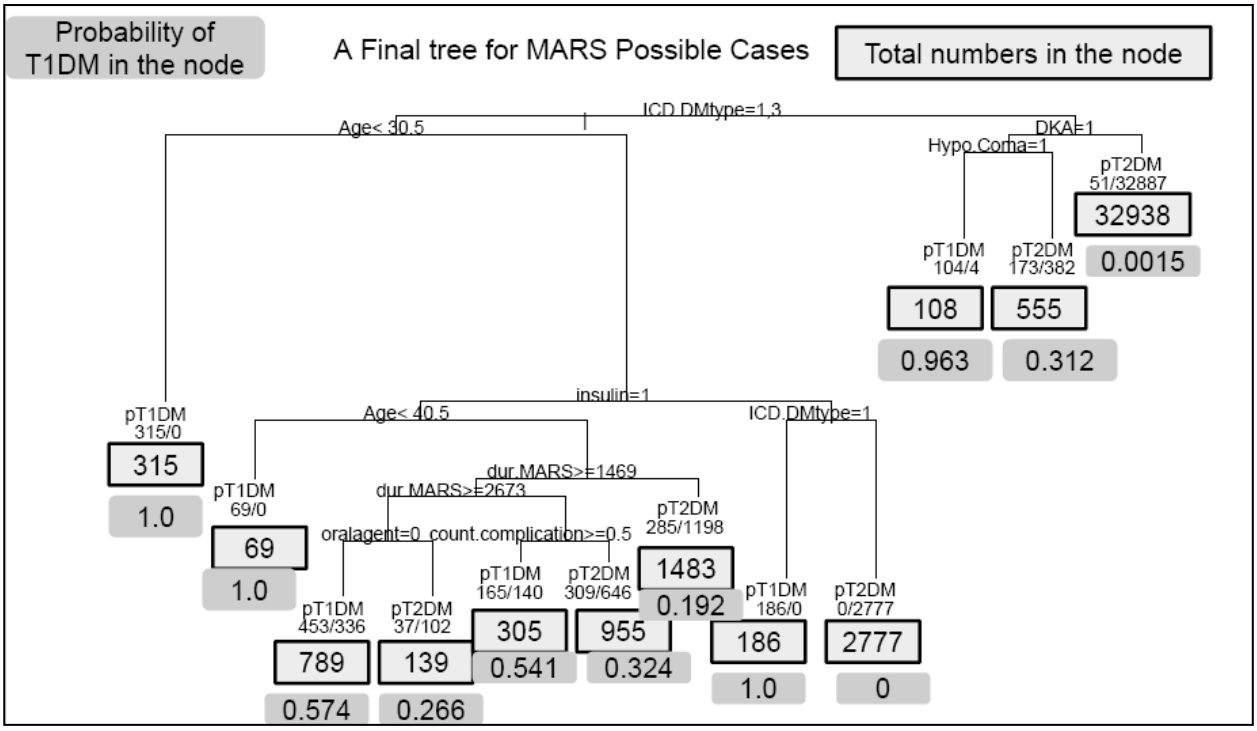

Figure 11. A Pruned Tree for Predicting "Possible" T1DM and T2DM Cases

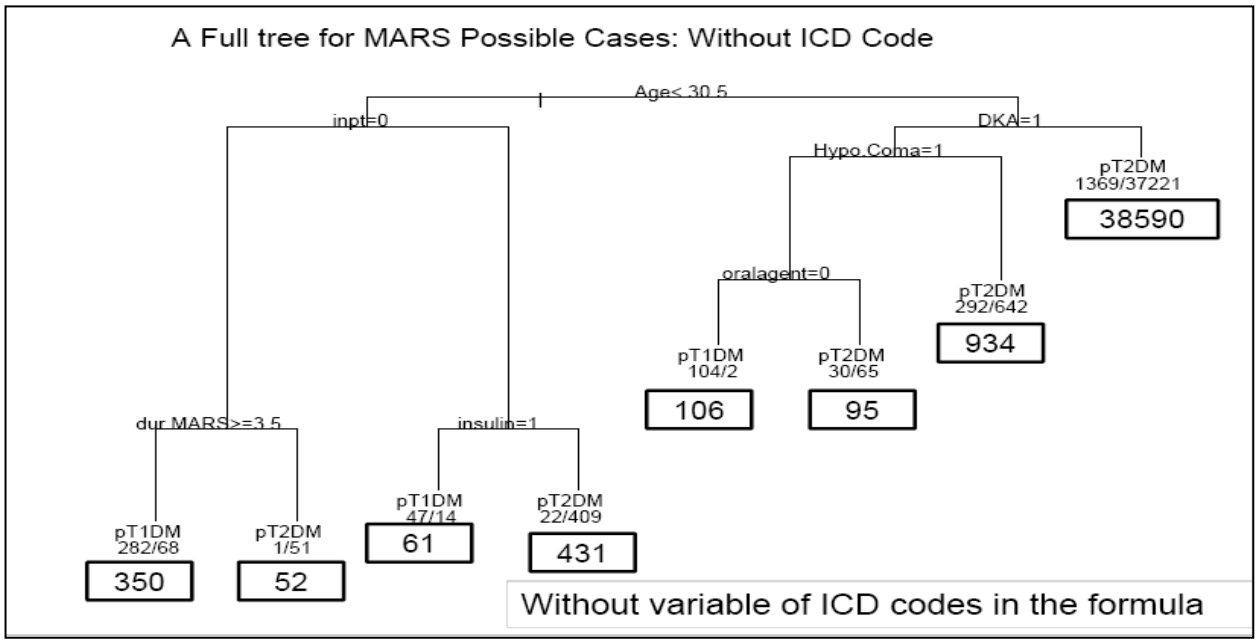

Figure 12. A Pruned Tree for Predicting "Possible" T1DM and T2DM Cases: Without ICD Codes in the Formula

Table 14: A Pruned Tree-Based Model and Complexity Parameters for Predicting "Possible" T1DM and T2DM Cases: Without ICD Codes in the Formula

A pruned tree model	Complexity parameters and errors
node), split, n, loss, yval, (yprob) *** denotes terminal node** 1) root 40619 2147 pT2DM (0.052857 0.94714) 2) Age< 30.5 894 352 pT2DM (0.39374 0.60626) 4) inpt=0 402 119 pT1DM (0.70398 0.29602) 8) dur.MARS>=3.5 350 68 pT1DM (0.80571 0.19429) * 9) dur.MARS< 3.5 52 1 pT2DM (0.019231 0.98077) * 5) inpt=1 492 69 pT2DM (0.14024 0.85976) 10) insulin=1 61 14 pT1DM (0.77049 0.22951) * 11) insulin=0 431 22 pT2DM (0.051044 0.94896) * 3) Age>=30.5 39725 1795 pT2DM (0.045186 0.95481) 6) DKA=1 1135 426 pT2DM (0.37533 0.62467) 12) Hypo.Coma=1 201 67 pT1DM (0.66667 0.33333) 24) oralagent=0 106 2 pT1DM (0.98113 0.018868) * 25) oralagent=1 95 30 pT2DM (0.31579 0.68421) * 13) Hypo.Coma=0 934 292 pT2DM (0.31263 0.68737) * 7) DKA=0 38590 1369 pT2DM (0.035476 0.96452) *	Variables actually used in tree construction: [1] Age DKA Hypo.Coma dur.MARS inpt insulin oralagent Root node error: 2150/40619 = 0.0529 n= 40619 CP nsplit rel error xerror xstd 1 0.0382 0 1.000 1.000 0.0210 2 0.0233 2 0.924 0.924 0.0202 3 0.0156 3 0.900 0.902 0.0200 4 0.0154 6 0.853 0.879 0.0198 **5 0.0100 7 0.837 0.839 0.0193**

4.2 HOW WELL THE TREE-STRUCTURED MODELS CAN DISTINGUISH T1DM FROM T2DM? WHAT IS THE MISCLASSIFICATION RATE, SENSITIVITY, SPECIFICITY, POSITIVE PREDICTIVE VALUE (PPV), NEGATIVE PREDICTIVE VALUE (NPV) OF T1DM CASES?

Considering T1DM as the truth (i.e., positive category) and T2DM as a negative category, we calculated misclassification rates, sensitivity, specificity, PPV and PNV of T1DM cases for each final tree-based model (**Table 15 and Table 16**). As shown in **Table 15**, the tree model using probable identified cases has the highest sensitivity (default tree: 94.82%; pruned tree: 95.63%), specificity (default tree: 99.63%; pruned tree: 99.28%), PPV (default tree: 94.41%; pruned tree: 89.81%) and PNV (default tree: 99.60%; pruned tree: 99.71%). The tree model using possible identified cases has the lowest sensitivity of 60.18%, specificity of 98.75%, PPV of 72.91% and PNV of 97.80%, even though the tree is bigger and with more variables to split the observations. The misclassification rate of T1DM cases as T2DM cases increases from 4.37% to 39.82% while we used possible cases instead of probable cases. In total, 37,705 (22.62%) probable and possible cases had missing ICD-9 codes. As we expected, all of the models lost the accuracy when we removed ICD-9 codes diagnosis from the model (**Table 16**). The misclassification rate of T1DM as T2DM cases was almost 80% in the possible cases. The sensitivity of T1DM was dropped greatly from 48.68% to 20.17%.

Compared the default tree model with the pruned tree model using probable cases, we can see the predictors “duration of MARS” and “in-patient diagnosis of DM” increase the PPV of T1DM. However, they do not improve the misclassification rate and sensitivity of T1DM. The specificity and NPV of T1DM are about the same. Considering the predictor “duration of

MARS" which reflects the duration of follow-up in the MAR database may not be generalizable to other administrative date, either the default tree model without "duration of MARS" or pruned tree model using probable cases can be the predictive rules to distinguish T1DM from T2DM.

Table 15: Summary Results: Predictive Trees of Probable and Possible T1DM and T2DM Cases

	Probable Cases (Default tree)	Probable Cases (Pruned tree)	Possible Cases
Actual cases (T1DM/T2DM)	7,857 / 118,240	7,857 / 118,240	2,147 / 38,472
Predicted cases (T1DM/T2DM)	7,891 / 118,206	8,367 / 117,730	1,772 / 38,847
Size of the tree (no. of terminal nodes)	11	8	12
Variable selected	ICD_DM¶ In-patient OHA use In-patient insulin use DKA Age Duration in MARS In-patient diagnosis	ICD_DM¶ In-patient OHA use In-patient insulin use DKA Age	ICD_DM¶ In-patient OHA use In-patient insulin use DKA Hypocoma Age Duration in MARS No of complications$^{\varphi}$
Misclassification rate of T1DM (%)	5.18	4.37	39.82
Misclassification rate of T2DM (%)	0.37	0.72	1.25
Sensitivity of T1DM (%)	94.82	95.63	60.18
Specificity of T1DM (%)	99.63	99.28	98.75
PPV of T1DM (%)	94.41	89.87	72.91
NPV of T1DM (%)	99.60	99.71	97.80

Abbreviation: OHA: oral hypoglycemic agents; **DK**: unknown types of diabetes; **DKA**: diabetic ketoacidosis; **hypocoma**: hypoglycemic coma; **T1DM**: type 1 diabetes mellitus; **T2DM**: type 2 diabetes mellitus;
¶: ICD_DM: **1**= only with ICD-9 code for T1DM specific diagnosis (250.x1 or 250.x3); **2**= only with ICD-9 code for T2DM or unspecified type DM diagnosis (250.x0 or 250.x2); **3**= with both ICD-9 code for T1DM and T2DM or unspecified diagnosis [(250.x1 or 250.x3) AND (250.x0 or 250.x2)]; **missing values**= without any ICD-9 code diagnosis for DM.
φ: Complications include coronary artery bypass graft (CABG), myocardial infarction (MI), dialysis, retinopathy, neuropathy, and amputations.

Table 16: Summary Results: Final Predictive Trees of Probable and Possible T1DM and T2DM Cases: Without ICD Codes¶ in the Formula

	Probable Cases	Possible Cases
Actual cases (T1DM/T2DM)	7,857 / 118,240	2,147 / 38,472
Predicted cases (T1DM/T2DM)	4,664 / 121,433	517 / 40,102
Size of the tree (no. of terminal nodes)	6	8
Variable selected	Age OHA Insulin DKA	Age OHA Insulin DKA Hypocoma Duration in MARS Inpt
Misclassification rate of T1DM (%)	51.32	79.83
Misclassification rate of T2DM (%)	0.71	0.22
Sensitivity of T1DM (%)	48.68	20.17
Specificity of T1DM (%)	99.29	99.78
PPV of T1DM (%)	82.01	83.75
NPV of T1DM (%)	96.68	95.73

Abbreviation: OHA: oral hypoglycemic agents; **DK**: unknown types of diabetes; **DKA**: diabetic ketoacidosis; **ER**: ICD-9 code 250 for emergency room diagnosis; **hypocoma**: hypoglycemic coma; **inpt**: ICD-9 code 250 for in-patient diagnosis; **T1DM**: type 1 diabetes mellitus; **T2DM**: type 2 diabetes mellitus;
¶: ICD_DM: **1**= only with ICD-9 code for T1DM specific diagnosis (250.x1 or 250.x3); **2**= only with ICD-9 code for T2DM or unspecified type DM diagnosis (250.x0 or 250.x2); **3**= with both ICD-9 code for T1DM and T2DM or unspecified diagnosis [(250.x1 or 250.x3) AND (250.x0 or 250.x2)]; **missing values**= without any ICD-9 code diagnosis for DM.

5.0 DISCUSSION

We have developed a tree-structured model to help distinguish people with T1DM or T2DM. ICD-9 codes of T1DM or T2DM, history of DKA, age of 40, in-patient insulin use, and in-patient OHA use are consistent predictors in the tree models to distinguish T1DM from T2DM cases. Other ancillary predictors may include in-patient diagnosis of DM, history of hypoglycemic coma, episodes of complications including MI, CABG, dialysis, retinopathy, neuropathy and amputation, and duration in the MARS database. The sensitivity, specificity, PPV, and NPV of the default tree to predict T1DM are 94.82%, 99.63%, 94.41% and 99.60% among the probable diabetic cases. To the best of our knowledge, there are no other tools or studies designed to distinguish between T1DM or T2DM using a large administrative/clinical database. Default tree without "duration of MARS" which is not dependent on one insurer or other specific population may be more applicable to another general diabetic population. Either the default or pruned tree can be used to identify a cohort of T2DM well.

Administrative data poses some challenges in statistical analysis. It often involves complex interactions between explanatory variables. Missing values for both explanatory and outcome variables are fairly common, and outliers usually exist. The high levels of missing predictor information in our clinically derived data would be considered a serious limitation in epidemiological analyses. Thus, a tree-structured model provides some advantages of handling a

large administrative data. (1) Using surrogate splits as the treatment of missing values (as NAs in the tree dataset) is more satisfactory compared to the linear or logistic regression model. (2) It naturally handles non-homogeneous or nonlinear relationships. (3) It also does automatic stepwise variable selection and complexity reduction. (4) It is invariant to monotone transformations of predictor variables so that the precise form in which these appear in a model formula is irrelevant. (5) It is robust with respect to outliers and misclassified points in the learning sample. (6) The tree structure gives easily understood and provides intuitive information regarding the predictive structure of the data. Traditional linear or logistic regression models do allow for the testing of statistical interactions among independent variables, which assess differences in the effects of one or more independent variable according to levels of another independent variable. However, statistical interactions can be difficult to interpret and to identify, particularly when three or more variables are assessed at a time.[35] (7) The final classification has a simple form that can be compactly stored and used efficiently to classify new data.

Nevertheless, there are some problems inherent in the tree procedure. Tree stability is sometimes a concern. It is not necessarily stable to small perturbations in the data. Splits that tend to separate a node into one small and one large subset (also called “end cuts”), while all other things are equal and favored over “middle cut”. Also, variable selection is biased in favor of those variables having more values and thus offering more splits. Ideally, a reliable predictive factor or model has a high sensitivity to detect most patients who are destined to have outcome and a specificity that results in a low false-positive rate and the correct identification of most patients who will have the outcome. A factor is adequately predictive based not only on the magnitude of the relative risk but also on the prevalence of both of the factor and the outcome.

A risk factor that has a strong association with a clinical outcome may not be good predictor if its prevalence and the prevalence of the outcome are extremely low.[36] That is probably the reason why some expected factors were not identify in our tree-model, such as cardiovascular complications.

We also acknowledge some limitations inherent in the MARS administrative database. First, the onset of T1DM mainly occurs in childhood or puberty. We were able to obtain the ages of patients seeking UPMC services for the first time between 2001 and 2009, instead of the ages at diagnosis of DM or other complications. Also, the present dataset includes age data only for 65% of patients. We may be able to identify more probable and possible cases when more complete age data become available. Second, factors that are helpful to determine whether patients have T1DM are the presence of islet antibodies, low c-peptide, episode of DKA, and other genetic and environmental factors. Unfortunately, these events are rarely described in the office or hospital charts of patients with DM. Third, T1DM patients depend on insulin permanently and critically. However, medication data of insulin and/or OHA use were only available for inpatients. We cannot depend on inpatient insulin use as the only judgment to distinguish T1DM from T2DM cases because diabetic patients are more likely to use insulin when they are hospitalized, and some patients may only use insulin temporarily after surgery or during hospitalizations. Fourth, we used an adult cohort (age of 18 years or more) in the MARS database. Our results have limited generalizability to children. Fifth, there were some missing laboratory data if samples were sent to non-UPMC laboratories that do not supply data to MARS

Another limitation to this study is that incomplete ascertainment of diabetic cases and types of diabetic cases. It is likely some of the 80,145 patients who had only single indicator and did not meet Zgibor's criteria[6] may have true DM. We used medico-administrative data to

identify the cases. This may underestimate the real cases because most people with T2DM remain asymptomatic for years after onset of the disease, and only patients who received health care services are entered in these databases. These are two possible conditions that could result in non-detection of potential cases of diabetes during the study period: (1) people with T2DM who did not need the services or treated exclusively with diet and exercise that were not on diabetic medications; (2) diabetic patients who use services for other reasons. We may also have some misclassifications between T1DM and T2DM. Some patients cannot be clearly classified as T1DM or T2DM. Clinical presentation and disease progression vary considerably in both types of diabetes. Occasionally, patients who otherwise have T2DM may present with ketoacidosis. Similarly, patients with T1DM may have a late onset and slow, but virulent progression of disease despite having features of autoimmune disease. Such difficulties in diagnosis may occur in children, adolescents, and adults. The true diagnosis may become obvious overtime. We have no independent source or "gold standard" such as surveys or diagnostic testing to quantify the cases we might be missing. Even though the default tree in the probable cases selected the duration of MARS as a predictor, this factor may not be a good predictor generalizing to other administrative databases. Lastly, we did not ascertaining the accuracy of our predictive rule through any internal or external validations in the current thesis. Due to all these inheriting limitations in an administrative database, our goal was to determine a set of criteria as an accurate and valid predictive rule to distinguish T1DM from T2DM instead of depending on a single indicator. The predictors were developed for use that aim to support rather than replace the clinical judgment.

Planning, implementing, monitoring, temporal evolution and prognosis differ between patients with T1DM and T2DM. Accurate information about the magnitude, distribution, and

types of diabetes can inform policy, support health care evaluation and clinical prognosis in public health. Our preliminary predictive rule to distinguish between T1DM and T2DM cases in a large administrative database appears to be promising and needs to be validated. The ***public health significance*** is that being able to distinguish between these diabetes subtypes will allow future subtype-specific analyses of cost, morbidity, and mortality. This study allows researchers to easily identify segments of a population that are more or less likely to exhibit the outcomes. We will not only be able to track different processes and costs of diabetes care delivery and progressions of clinical outcomes, but also will be able to identify characteristics that are important barriers to or facilitators of the health-related behavior of interest among T1DM and T2DM cases. Future work will focus on ascertaining the validity and generalizability of our predictive rule, by conducting a review of medical charts (as an internal validation) and applying the rule to another MARS dataset or other administrative databases (as external validations). Then, we can further modify the preliminary rule to increase the numbers of T1DM and T2DM cases captured. This ultimately will facilitate subtype specific analysis on processes and outcomes of patients with T1DM and T2DM.

APPENDIX A: PROGRAMMING DETAILS

A.1 RPART LIBRARY AND ITS FUNCTIONS

The basic steps to use RPART are addressed as follow.

1) To use RPART, the users will have to load and attach the RPART package into S-Plus using the "***library()***" function.

   library(rpart)

2) Decide what type of endpoint you have. The ***rpart*** system was designed to be easily extended to new types of responses, including ***"anova", "poisson", "class", or "exp".*** If the ***method*** argument is missing, an appropriate type is intelligently inferred from the response variable in the formula. We only described ***"class"*** and ***"anova"*** in more detail here.

   - Categorical→ ***method = 'class'***. A classification tree, with a categorical response and default impurity criterion the Gini index, selected by the argument ***parms = list (split="information").*** The pruning criterion R(T) is the *predicted loss*, normally the *error rate*.
   - Continuous→ ***method = 'anova'***. A regression tree, with the impurity criterion the reduction in sum of squares on creating a binary split of the data at that node. The

criterion R(T) used for pruning is the *mean square error* of the predictions of the tree on the current dataset (that is, the *residual mean square*).

- Poisson process or count→ method = 'poisson'
- Survival→ ***method = 'exp'***

3) Fit the model using the standard S-Plus language. The general usage of ***rpart*** is

   "fit <— rpart(formula, data, weights, subset, na.action = na.rpart, method, model = false, x = false, y = true, parms, control, cost,….)".

   The default action deletes all observations for which y is missing, but keeps those in which one or more predictors are missing.

4) Print a text version of the tree.

   print(fit) #fit is the name of the tree

5) Print a summary which examines each node in depth.

   summary(fit)

6) Plot a standard version of the plot with some basic function.

   plot(fit)

   text(fit, use.n = TRUE)

7) Create a prettier version of the tree.

   post(fit, file = '')

The ***rpart*** that both grows a tree and computes the optimal pruning for all α; although there is a function ***prune.rpart***, it merely further prunes the tree at points already determined by the call to ***rpart***, which has itself done some pruning. It is also possible to print a pruned tree by

giving a pruning parameter to ***print.rpart***. By default ***rpart*** runs a 10-fold cross-validation and the results are stored in the rpart object to allow the user to choose the degree of pruning at a later stage. The default behavior of ***rpart*** is to find surrogate splits during tree construction, and use them if missing values are found during prediction. This can be changed by the option ***usesurrogate = 0*** to stop cases as soon as a missing attribute is encountered. A further choice is what do to if all surrogates are missing: option usesurrogate =1 stops whereas usesurrogate = 2 (the default) send the case in the majority direction. Function ***predict.tree*** allows a choice of case-splitting or stopping (the default) governed by the logical ***split***.

A.2 PROGRAMMING CODES FOR MARS DATA

A. Probable Cases

- **Load RPART library**: *library (rpart)*
- **Fit the tree model**

  mars.rp <- rpart(DM~ sex+ race2 + Age + ICD.DMtype+insulin + oralagent + inpt+ outpt + er + DKA + Hypo.Coma + CVE +dialysis+ retinopathy + neuropathy + amputation + thyroid + celiac + addison+dur.MARS+ Age.CVE+ Age.neuropathy + Age.retinopathy+ Age.dialysis+ Age.amputation + Age.complication + count.cofactor + count.complication , data = mars.031010.dm12.probable, method='class')
- **A Plot and summary of default tree model (using proportional distance by the importance of the splits)**

print(mars.rp);

plot(mars.rp); text(mars.rp, use.n=T, pretty=0)

title('A Tree-Based model for MARS Probable DM cases: RPART function')

- **A summary and plot of complexity parameter through cross-validations**

  plotcp(mars.rp); title('Complexity Parameter vs Size of Tree')

  printcp(mars.rp)

- **Fit the pruned tree according to the complexity parameter**

  mars.rp1 <- prune(mars.rp, cp=0.013)

  plot(mars.rp1); text(mars.rp1, use.n=T, pretty=0)

  title('A Pruned Tree Model for MARS Probable Cases-RPART function')

  print(mars.rp1)

B. Possible Cases

- **Load RPART library**: *library (rpart)*
- **Fit the tree model**

  mars.rp.possible <- rpart(DM~ sex+ race2 + Age + ICD.DMtype+insulin + oralagent + inpt+ outpt + er + DKA + Hypo.Coma + CVE +dialysis+ retinopathy + neuropathy + amputation + thyroid + celiac + addison+dur.MARS+ Age.CVE+ Age.neuropathy + Age.retinopathy+ Age.dialysis+ Age.amputation + Age.complication + count.cofactor + count.complication , data = mars.031010.dm12.possible, method='class')

- **A Plot and summary of default tree model (using proportional distance by the importance of the splits)**

print(mars.rp.possible);

plot(mars.rp.possible); text(mars.rp.possible, use.n=T, pretty=0)

title('A Tree-Based model for MARS Possible DM cases: RPART function')

- **A summary and plot of complexity parameter through cross-validations**

  plotcp(mars.rp.possible); title('Complexity Parameter vs Size of Tree')

  printcp(mars.rp. possible)

- **Fit the pruned tree according to the complexity parameter**

  mars.rp.possible1 <- prune(mars.rp.possible, cp=0.010)

  plot(mars.rp.possible1); text(mars.rp.possible1, use.n=T, pretty=0)

  title('A Pruned Tree Model for MARS Possible Cases-RPART function')

  print(mars.rp.possible1)

BIBLIOGRAPHY

1. Diabetes facts. Fact sheet N^{0}312, 2009
2009; http://www.who.int/mediacentre/factsheets/fs312/en/. Accessed October 20, 2009.

2. Economic consequences of diabetes mellitus in the U.S. in 1997. American Diabetes Association. *Diabetes Care.* Feb 1998;21(2):296-309.

3. *Centers for Disease Control and Prevention. National diabetes fact sheet: general information and national estimates on diabetes in the United States, 2007.* Atlanta, GA:: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention;2008.

4. Dall TM, Zhang Y, Chen YJ, Quick WW, Yang WG, Fogli J. The Economic Burden Of Diabetes. *Health Aff (Millwood).* Jan 14 2010.

5. Asghari S, Courteau J, Carpentier AC, Vanasse A. Optimal strategy to identify incidence of diagnostic of diabetes using administrative data. *BMC Med Res Methodol.* 2009;9:62.

6. Zgibor JC, Orchard TJ, Saul M, et al. Developing and validating a diabetes database in a large health system. *Diabetes Res Clin Pract.* Mar 2007;75(3):313-319.

7. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression trees.* 1 ed. Belmont, CA: Wadsworth International Group; 1984.

8. TIBCO Spotfire S+® 8.1 for Windows® User's Guide. *TIBCO Software Inc.* 2008.

9. Incidence and trends of childhood Type 1 diabetes worldwide 1990-1999. *Diabet Med.* Aug 2006;23(8):857-866.

10. Sperling MA. Diabetes Mellitus. *Sperling: Pediatric Endocrinology.* 3 ed. Philadelphia: Saunders, an imprint of Elsevier Inc.; 2008:374-421.

11. Libman IM, LaPorte RE. Changing trends in epidemiology of type 1 diabetes mellitus throughout the world: how far have we come and where do we go from here. *Pediatr Diabetes.* Sep 2005;6(3):119-121.

12. Libman IM, LaPorte RE, Becker D, Dorman JS, Drash AL, Kuller L. Was there an epidemic of diabetes in nonwhite adolescents in Allegheny County, Pennsylvania? *Diabetes Care.* Aug 1998;21(8):1278-1281.

13. Daneman D. Type 1 diabetes. *Lancet.* Mar 11 2006;367(9513):847-858.

14. Polonsky KS, Licinio-Paixao J, Given BD, et al. Use of biosynthetic human C-peptide in the measurement of insulin secretion rates in normal volunteers and type I diabetic patients. *J Clin Invest.* Jan 1986;77(1):98-105.

15. Hattersley A, Bruining J, Shield J, Njolstad P, Donaghue K. ISPAD Clinical Practice Consensus Guidelines 2006-2007. The diagnosis and management of monogenic diabetes in children. *Pediatr Diabetes.* Dec 2006;7(6):352-360.

16. Standards of medical care in diabetes--2009. *Diabetes Care.* Jan 2009;32 Suppl 1:S13-61.
17. Wilson DM, Buckingham B. Prevention of type 1a diabetes mellitus*. *Pediatr Diabetes.* Mar 2001;2(1):17-24.
18. Lipton RB. Is now the time for an intervention to prevent autoimmune type 1 diabetes? *Pediatr Diabetes.* Mar 2001;2(1):12-16.
19. McCall AL, Saunders JT. Diabetes Mellitus in Adults. In: Rakel RE, Bope ET, eds. *Rakel & Bope: Conn's Current Therapy*2009.
20. Pogach LM, Hawley G, Weinstock R, et al. Diabetes prevalence and hospital and pharmacy use in the Veterans Health Administration (1994). Use of an ambulatory care pharmacy-derived database. *Diabetes Care.* Mar 1998;21(3):368-373.
21. Arday DR, Fleming BB, Keller DK, et al. Variation in diabetes care among states: do patient characteristics matter? *Diabetes Care.* Dec 2002;25(12):2230-2237.
22. Selby JV, Ray GT, Zhang D, Colby CJ. Excess costs of medical care for patients with diabetes in a managed care population. *Diabetes Care.* Sep 1997;20(9):1396-1402.
23. Engelgau MM, Geiss LS, Manninen DL, et al. Use of services by diabetes patients in managed care organizations. Development of a diabetes surveillance system. CDC Diabetes in Managed Care Work Group. *Diabetes Care.* Dec 1998;21(12):2062-2068.
24. Hebert PL, Geiss LS, Tierney EF, Engelgau MM, Yawn BP, McBean AM. Identifying persons with diabetes using Medicare claims data. *Am J Med Qual.* Nov-Dec 1999;14(6):270-277.
25. Newton KM, Wagner EH, Ramsey SD, et al. The use of automated data to identify complications and comorbidities of diabetes: a validation study. *J Clin Epidemiol.* Mar 1999;52(3):199-207.
26. Brown JB, Nichols GA, Glauber HS. Case-control study of 10 years of comprehensive diabetes care. *West J Med.* Feb 2000;172(2):85-90.
27. Wilson C, Susan L, Lynch A, Saria R, Peterson D. *Patient with diagnosed diabetes mellitus can be accurately identified in an Indian Helah Service patient registration database. Publice Health Reports.*2001.
28. Selby JV, Karter AJ, Ackerson LM, Ferrara A, Liu J. Developing a prediction rule from automated clinical databases to identify high-risk patients in a large population with diabetes. *Diabetes Care.* Sep 2001;24(9):1547-1555.
29. Morgan JN, Sonquist JA. Problems in the analysis of survey data and a proposal. *JASA.* 1963;58:415-434.
30. Tree-Based Methods. In: Chambers J, Eddy W, Hardle W, Sheather S, Tierney L, eds. *Modern Applied Statistics with S.* 4 ed. New York: Springer-Verlag New York, Inc.; 2002.
31. Therneau TM, Atkinson EJ. *An Introduction to Recursive Partitioning Using the RPART Routines: Tachnical Report 61.* Rochester, MN: Mayo Foundation, Section of Statistics;1997.
32. UPMC. UPMC Fast Facts "http://www.upmc.com/AboutUPMC/Pages/default.aspx" (last access date: March 29, 2010).
2010; http://www.upmc.com/AboutUPMC/Pages/default.aspx. Accessed March 29, 2010, 2010.
33. *The burden of diabetes in Pennsylvania in 2007*: Pennsylvania Department of Health Diabetes Control Program;2007.

34. Siminerio LM, Drab SR, Gabbay RA, et al. Diabetes educators: implementing the chronic care model. *Diabetes Educ.* May-Jun 2008;34(3):451-456.
35. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Ann Behav Med.* Dec 2003;26(3):172-181.
36. Grobman WA, Stamilio DM. Methods of clinical prediction. *Am J Obstet Gynecol.* Mar 2006;194(3):888-894.