THE USE OF ARTIFICIAL INTELLIGENCE FOR PRECISION MEDICINE IN METABOLIC SYNDROME

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

ERA KIM

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

ADVISOR: Dr. GYORGY J. SIMON, PHD

CO-ADVISOR: DAVID S. PIECZKIEWICZ, PHD

FEBRUARY 2019

Acknowledgements

This thesis becomes a reality with the kind support and enormous help of many special individuals. I would like to extend my sincere thanks to all of them.

First of all, I would like to express my deepest appreciation to my advisor, Dr. Gyorgy J. Simon for imparting expertise and knowledge, inspiring me with creative ideas of research topics and methods, excellent supervision of all my work, invaluable guidance with great patience and understanding, and finally extensive time and tremendous effort to finish this dissertation. I am greatly honored to complete this work under his supervision.

I would like to express special thanks of gratitude to my co-advisor, Dr. David S. Pieczkiewicz for his persistent encouragement and unsurpassed knowledge of data visualization. His genuine kindness, generosity, and empathy truly help me sustain positive thinking throughout this long journey.

Besides my advisors, I would like to express my sincere gratitude to the rest of my dissertation committee members, Dr. Karen A. Monsen and Dr. Rui Zhang, for agreeing to serve on my committee and sharing their valuable time reviewing my dissertation.

I am highly indebted to Pedro J. Caraballo, MD and M. Regina Castro, MD for their valuable time reviewing every manuscript carefully, giving constructive feedback, enthusiastically providing necessary clinical information, and their great support in completing this endeavor.

I would like to thank my lab mates for their continuous support and friendship: Wonsuk Oh, Jia Li, Mengdie Wang, Zhen Hu, and Pranjul Yadav.

My thanks also extends to Mr. Ravi Narayanan for originally inspiring me to continue my education to the PhD level and giving me endless support. Thank you to Mary Hopkins, Ohsook Kim, Julee Park, Dooyoung Choi, and Greg Lee for the constant fellowship and prayers.

My thanks and appreciation also go to the staffs at the Institute of Health Informatics, especially Jessica Tetzlaff and Melissa Malikowski.

Finally, I would like to acknowledge Fairview Health Services, Mayo Clinic, and OptumLabs for the providing the invaluable and necessary data to finish my dissertation.

Dedication

This I dedicate to God the LORD of my life and death, who inspires me with the Spirit and leads me besides still waters all the time, and to my parents and my younger brother who support me mentally and spiritually all the time.

Abstract

Type 2 Diabetes Mellitus (T2DM) is a chronic, progressive metabolic disorder, associated with an increased risk of developing micro- and macrovascular complications in many different organ systems, and it is one of the leading causes of death[1,2]. The management of T2DM is complex, difficult, and it requires considerations of the heterogeneity of the population, interactions among the diseases in metabolic syndrome, and the overlap of the risk factors across multiple diabetic complications. For the successful management of T2DM, individualized and evidence-based clinical guidelines are necessary[3,4].

Randomized controlled trials (RCTs) are considered the gold standard for clinical research. RCTs can produce the best unbiased evidence, but occasionally the results from the trials can offer limited practical value[5]. They can be inconclusive, contradictory, or incomplete, leaving many aspects of T2DM management unaddressed[6]. Thus, there exists a critical gap between the optimal individualized and the current patient care.

With the recent accumulation of electronic health record (EHR) data, machine learning holds great promise for the advancement of precision medicine[7] with the capability of offering a new way to generate evidence that enhances clinical practice guidelines with more personalized recommendations[8]. Therefore, the approach of precision medicine harnessing machine learning and large amounts of EHR data can immensely benefit the prevention and management of T2DM.

My overarching goal is to build *clinically useful* and t*ransferable* machine learning models on big data that can influence individual T2DM patient care towards the implementation of precision medicine. The term *clinically useful* refers to addressing some of the complexities of T2DM management, such as the heterogeneity of the patient population, interaction among risk factors, and the overlap of risk factors across various diabetes complications. *Transferable* refers to the model's ability to be ported (transferred) to a different health system without significant loss of performance. If clinically useful models can be transferred to different healthcare systems, their potential use in clinical decision support for the T2DM management will be undoubtedly more assured, contributing to the realization of precision medicine.

Three studies are conducted and reported in this dissertation. The first study aims to develop a semi-supervised divisive hierarchical clustering algorithm for a subpopulation-based T2DM risk model. In this work, I demonstrate that my proposed algorithm successfully identifies subpopulations (clusters) with higher or lower risks of T2DM than the general population; seamlessly incorporates interactions among the risk factors; and handles non-proportional hazards. Even when compared to recently developed state-of-the-art association rule mining (ARM)-based models used as modern diabetes indices[9–12], my model offers excellent predictive performance as well as improved interpretation.

The second study aims to develop a Multi-Task Learning (MTL)-based methodology, which learns six micro- and macrovascular complications simultaneously by extracting a variance component that is common across all outcomes and isolating

variance components that are specific to the individual complications. I call the common component *General Progression* model and the complication-specific components as *Differential Progression* models. These complication-specific components allow for identifying the risk factors that determine the most likely complication that the patient is going to progress to. Importantly, by comparing the MTL-based methodology to a reference methodology which learns each complication separately in isolation, I show that the MTL-based methodology does not compromise its predictive performance due to the improved interpretation.

The third study aims to demonstrate the transferability of my model constructed using a modeling approach, in which a model is learned from a nationally representative cohort and is successfully validated on two local health systems' data. Three datasets are utilized for the model development and validation strategy: i) national claims and EHR data (N=80,091 T2DM patients) from the OptumLabs® Data Warehouse (OLDW), ii) local EHR data from the University of Minnesota Medical Center (UMMC) (N=8,091 T2DM patients), and iii) local EHR data from the Mayo Clinic, Rochester (MCR) (N=2,247 T2DM patients). I show that with my proposed modeling approach, even a complex machine learning model can be transferable, but models constructed on a local healthcare system's data are difficult to be transferred to a different healthcare system.

All these studies have made significant contributions both to health informatics and medicine. Firstly, against the backdrop of contradictory evidence from RCTs, these studies can improve hypothesis generation for clinical research. Secondly, these studies

can improve the state of the art in clinical research. Lastly, these study results have significant translational potential for evidence-based practice and precision medicine.

With respect to contributions to medicine, the new way of generating clinical knowledge through big data and machine learning techniques could complement current evidence-based approach to medicine and enhance clinical practice guidelines.

If my work is transformed into intelligent clinical decision support tools, by combining clinical expertise, scientific evidence, and patient preference, it will help in ensuring that patients and clinicians have the information and tools they need to make the best informed and right decisions in the patient care.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Problem and Significance

Type 2 Diabetes Mellitus (T2DM) is a chronic, progressive metabolic disorder, associated with an increased risk of developing micro- and macrovascular complications in many different organ systems, and it is one of the leading causes of death[1,2]. The management of T2DM is complex, difficult, and it requires considerations of the heterogeneity of the population, complicated interactions among the diseases in metabolic syndrome, and the overlap of the risk factors across various individual complications. For the successful management of T2DM, individualized and evidence-based clinical guidelines are necessary[3,4].

Randomized controlled trials (RCTs) are considered the gold standard for clinical research. RCTs can produce the best unbiased evidence, but occasionally the results from the trials can offer limited practical value[5]. For example, the ACCORD[13], VADT[14], ADVANCE[15], and UKPDS[16] trials are the largest RCTs in T2DM, which studied the impact of intensive glycemic control on micro- and macrovascular complications. Unfortunately, they had contradictory and inconclusive results. With the small number of participants and/or short follow-up periods, these trials only represented subpopulations of T2DM but not the heterogeneous T2DM population as a whole. Also, although T2DM requires multi-factorial risk reduction strategies, these trials focused on only a single intervention (glycemic control) and a single aspect of its effects (to reduce cardiovascular

events). In fact, as increased mortality was observed in the intervention group, the ACCORD trial was terminated earlier. This shows that the results from RCTs can be inconclusive, contradictory, or incomplete, leaving many aspects of T2DM management unaddressed[6]. Thus, there exists a critical gap between the optimal individualized and the current patient care.

With the recent accumulation of electronic health record (EHR) data, machine learning holds great promise for the advancement of precision medicine[7] with the capability of offering a new way to generate evidence that enhances clinical practice guidelines with more personalized recommendations[8]. Therefore, the approach of precision medicine harnessing machine learning and large amounts of EHR data can immensely benefit the prevention and management of T2DM.

In recent years, there has been an increase in the number of studies using machine learning techniques in diabetes research[17–19]. However, most of these studies focus primarily on the technical aspects of modeling and the *clinical usefulness* and *transferability* of these model are very limited[20]. For example, existing risk models are predominantly population-average, additive models that do not fully take the heterogeneity of T2DM population as well as the interactions among the risk factors into consideration. Although their ability to stratify patients by risk seems adequate, their ability to accurately quantify the risks and suggest risk factors for individual patients can be in doubt.

As another example of clinical usefulness, various diabetes complications stem from “the common soil” (metabolic deterioration)[21]. Most studies identified prognostic

factors, which are associated with subsequent clinical outcomes, by examining each outcome in isolation. Due to the "common soil", these studies tended to discover the same risk factors for different outcomes, not only masking which prognostic factor is specific to which outcome, but also often leading to unsatisfactory estimation of prognosis.

The other key limitation of existing machine learning-based risk scores is *transferability*. These scores (models) usually achieve very high predictive performance, but only in a single healthcare system. The expectation today is that even simple machine learning models cannot be transferred, and thus researchers do not even try to externally validate their models. With external validation being the gold standard for observational studies, the lack of external validity in machine learned models is concerning.

My overarching goal is to build *clinically useful* and t*ransferable* machine learning models on big data that can influence individual T2DM patient care towards the implementation of precision medicine. If clinically useful models can be transferred to different healthcare systems, their use in clinical decision support for the management of T2DM will undoubtedly increase, contributing to the realization of precision medicine.

1.2 Related Work

1.2.1 Early Identification of patients at high risk of Type 2 Diabetes

Numerous risk indices (risk scores) aiming at early identification of patients at high risk of T2DM have been developed[22–24]. The Framingham score[25] is the most popular such index which has gained wide acceptance in clinical practice. The Diabetes

Complication Index (DCI)[26] and the Diabetes Complication Severity Index (DCSI)[27] particularly focus on T2DM complications and examine the association with quality of life and utilization of health services. However, none of these indices takes the interactions among the diseases (or risk factors) into account; they only compute the risks in an additive manner by assuming that risk factors act independently even if evidence of interactions among risk factors exists[11,28–30]. These indices may accurately stratify patients' risks (into low, medium, and high), but their quantified risks could be less accurate.

Most early risk models for T2DM (or micro- and macrovascular complications) have been developed using logistic or Cox regressions[24,31], but, recently, data mining techniques such as decision tree (DT), association rule mining (ARM), and clustering are being increasingly employed[32,33].

DTs are one of the most commonly used techniques to predict the risk of T2DM and identify patients at high risk of T2DM[34–36]. Through recursive partitioning, DTs can identify both risk factors (rules) and subpopulations (nodes). Additionally, since the identified subpopulations on the same level are mutually exclusive and collectively exhaustive (i.e., the subpopulations encompass the entire population), interpretation of risk and risk factors for individual patients are straightforward.

As weaknesses of using DTs, it is known that these techniques perform poorly in the presence of complicated interactions among features, resulting in "the replication problem of decision trees (a tree contains two copies of the same sub-trees)"[37]. Also, since DTs are supervised classification techniques, the resultant subpopulations could

possibly be clinically irrelevant. For instance, suppose two T2DM subpopulations who have different clinical characteristics have a considerable overlap in their outcomes. Because DT classifiers that use class labels for partitioning, many of the patients in the subpopulations could possibly be assigned incorrectly. So, risk predictions could be accurate; however, the inherent clinical difference between these two subpopulations may not be captured by DT classifiers. To be clinically more useful, not only predictive performance but also improved interpretation (representative characteristics of a subpopulation) should be considered.

ARM has been predominately used to address interactions among risk factors for T2DM[11,28,29]. With the ability to seamlessly incorporate interactions and straightforward interpretability, these ARM models discover sets of associated risk factors (rules) along with the affected subpopulations at high or low risk of T2DM. However, when the ARM models are sufficient detailed to offer novel insights, the large number of redundant rules quickly erode the interpretability (e.g., since an individual patient can be described by large numbers of rules, it is difficult to determine which subpopulation is the most representing the patient).

Clustering techniques[38] are popular for identifying patients at high risk or distinct gene expressions that are responsible for complications of T2DM[39–41]. However, clustering techniques are traditionally considered unsupervised; thus, the resultant clusters are not necessarily associated with an outcome. To overcome this issue, supervised and semi-supervised clustering techniques have been proposed[42–45]. These techniques incorporate class labels or constraints that are derived from partially labeled

data or background knowledge by domain experts; thus, resultant clusters become relevant to a particular outcome. Huo et al.[42] developed a constraint-based K-means clustering algorithm to predict the risk of coronary heart disease, a macrovascular complication of T2DM, and demonstrated that their proposed semi-supervised K-means outperformed traditional K-means.

1.2.2 Prognostic models for complications of Type 2 Diabetes

Because T2DM progresses to multiple serious complications, for more accurate patient prognosis, evidence should be generated comprehensively by systematically examining multiple risk factors and multiple complications. Although deterioration of the overall metabolic health underlies all of these complications[46–49], and thus commonality among risk factors exists, almost all existing studies focus on a single or occasionally a few complications and model each of them independently (Table 1-1). The studies listed in Table 1-1 used linear regression to build models. Subsequently, in Table 1-2, I show the list of studies that utilized machine learning techniques, making comparisons between the studies and my work.

Table 1-1 Mini Literature Review on Risk Factors for Complications of Type 2 Diabetes: Use Regressions

Author	N	Study Sample	Risk factor of interest	Outcome(s)	Association	Adjusted variables	Study Design
Ten Brinke et al. 2008[50]		T2DM patients	HbA1c	CHD, All-cause mortality	Both positive		Systematic review
Stratton et al. 2000[51]	3,642	T2DM patients	HbA1c	MI, Stroke, HF, PVD	All positive	Age, Sex, Race, FPG, HbA1c, BMI, SBP, LDL, HDL, TG, Albuminuria	Prospective observational study
Nichols et al. 2001[52]	1,131	T2DM patients under poor glycemic control (mean HbA1c:7.7%, 83% patients were < 9% of HbA1c)	HbA1c	CHF	No effect	Age, Sex, Duration of DM, HbA1c, creatinine, SBP, weight, DM med use, presence of IHD	Retrospective study using EHR data
A I Adler et al. 1999[53]	5,063	T2DM patients (25-65yr, newly diagnosed)	HbA1c	MI, Angina, Stroke	All positive	Age, Sex, Race, HbA1c, Smoking, BMI, DBP, physical activity, Occupation, Tchol, HDL, TG, Hx of MI or Stroke	Prospective observational study
Zhao, Katzmarzyk, Horswell, Wang, et al. 2014[54]	30,154	T2DM patients	HbA1c, stratify by sex	Stroke	Female sex: positive	Age, Race, HbA1c, Smoking, Income, Type of insurance, BMI, SBP, LDL, GFR, HTN/DM/Chol med use	Prospective cohort study
Elizabeth Selvin et al. 2005[55]	1,635	T2DM patients w/o Cardiovascular Disease	HbA1c	Ischemic stroke	Positive	Age, Sex, Race, HbA1c, Smoking, BMI, Waist-to-hip ratio, Education, SBP, DBP, LDL, HDL, HTN med use	Prospective cohort study
Muntner P et al. 2005[56]	1,575	T2DM patients	HbA1c	PAD	Positive	Age, Sex, HbA1c, Race, Smoking, Alcohol intake, Physical activity, WC, Tchol, CKD, C-reactive protein	Cross-sectional

E. Selvin et al. 2006[57]	15,792	Diabetic adults (45-64yr)	HbA1c	PAD	Positive	Age, Sex, Race, HbA1c, Smoking, Study center, BMI, Waist-to-hip ratio, SBP, LDL, HDL, Education, HTN/DM med use	Prospective cohort study
Amanda I. Adler et al. 2002[58]	2,398	T2DM patients (25-65yr)	HbA1c, SBP, HDL, Current smoker	PVD	Positive: HbA1c, SBP, and Current smoker. Protective: HDL	Age, Sex, Race, HbA1c, Smoking, BMI, SBP, DBP, Alcohol intake, Aspirin use, Tchol, LDL, HDL, TG, Physical activity, Hx of Cardiovascular Disease or retinopathy.	Prospective cohort study
Bash et al. 2008[59]	1,871	DM patients (45-64yr) w/o albuminuria and retinopathy	HbA1c	CKD (eGFR < 60ml/min/1.73 m^2)	Positive	Age, Sex, Race, HbA1c, Smoking, Study center, eGFR, BMI, LDL, HDL, TG, presence of HTN/CHD, HTN med use	Prospective cohort study
Ravid M, Brosh D, Ravid-Safran D, Levy Z 199[60]	574	T2DM patients (40-60yr,) with normal renal function and recent onset of T2DM	HbA1c, BP, Tchol, HLD, LDL, BMI, Smoking, Male sex	Microalbuminuria	Protective: HDL All others: Positive	Age, Sex, HbA1c, Smoking, BMI, BP, Tchol, LDL, HDL	Prospective cohort study
Retnakaran et al. 2006[61]	4,031	T2DM patients	Sex, Smoking, WC Retinopathy, neuropathy	Albuminuria (micro or macroalbuminuria), Renal impairment (creatinine clearance < 60 ml/min or doubling of plasma creatinine)	Positive for both: SBP, Urinary albumin, Creatinine Positive for Albuminuria: Male sex, WC, TG, LDL, HbA1c, Smoking,	Age, Sex, Race, HbA1c, FPG, Smoking, Weight, WC, SBP, DBP, LDL, HDL, TG, White cell count, Urine albumin, Plasma creatinine, presence of HTN/retinopathy/sensory neuropathy /Cardiovascular Disease	Prospective cohort study

					previous retinopathy Positive for Renal impairment: Female sex, age, previous sensory neuropathy Protective for Renal impairment: WC		
N. Li et al. 2014[62]	30,434	T2DM patients with low income and uninsured (30-95yr, w/o CHD and Stroke)	BMI stratified by sex	CHD	All positive	Age, Race, HbA1c, Smoking, Income, Type of insurance, SBP, LDL, HDL, TG, eGFR, HTN/DM/Chol med use.	Prospective cohort study
W. Li, Katzmarzyk, Horswell, Zhang, Wang, et al. 2015[63]	31,155	T2DM patients	BMI	HF	Positive	Age, Sex, Race, HbA1c, BMI, SBP, LDL, eGFR, HTN/DM/Chol med use	Prospective cohort study
W. Li, Katzmarzyk, Horswell, Zhang, Zhao, et al. 2015[64]	29,554	T2DM patients	BMI	Stroke	Protective	Age, Sex, HbA1c, Smoking, BMI, SBP, Income, Type of insurance, LDL, eGFR, HTN/DM/Chol med use	Prospective cohort study

Kokkinos et al. 2012[65]	4,013	T2DM patients	BMI	All-cause mortality	Protective	Age, Sex, Race, Smoking, BMI, Hx of Cardiovascular Disease, HTM/Chol med use, presence of HTN/ dyslipidemia	Use prospective observational study data
McEwen et al. 2007[66]	8,733	T2DM patients	BMI, Income, Age, Male sex, Duration of DM, Smoking	All-cause mortality	Protective: BMI, Income All others: Positive	Age, Sex, Race, Smoking, BMI, Education, Income, Duration of diabetes, DM med use, the presence of Micro-and Macrovascular complications, Charlson index	Multicenter-prospective, observational study
Doehner et al. 2012[67]	5,202	T2DM patients w/ Cardiovascular Disease	BMI	All-cause mortality, Cardiovascular outcome (Hospitalization/MI/Stroke)	All protective	Age, Creatinine, HbA1c, Smoking, BMI, LDL, HTN/DM/Chol med use, Hx of MI/Stroke/PCI/CABG/PVD	Use PROactive RCT data
Huang et al. 2014[68]	105	T2DM patients with CKD stage 3 or 4	BMI	GFR	protective	Age, Sex, HbA1c, Smoking, BMI, LDL, HDL, Hx of Cardiovascular Disease, Mean arterial pressure, Creatinine, Daily protein intake	Prospective cohort study
Mohammedi et al. 2018[69]		T2DM patients (55yr or older) with diabetes diagnosed at 30yr or older with pre-existing cardiovascular disease or at least one risk factor for cardiovascular disease	BMI	A composite of macroalbuminuria, doubling of the serum creatinine level to at least 200 mmol/l, ESRD (defined as the need for renal-replacement therapy), or death due to renal disease	positive	Age, Sex, Region, HbA1c, Smoking, SBP, Tchol to HDL ratio, TG, Hx of Cardiovascular Disease, GFR, Urinary albumin to Creatinine ratio, Duration of diabetes	ADVANCE RCT

Shepherd et al. 2006[70]	1,501	T2DM patients w/ CHD	lowering LDL-intensive atorvastatin therapy (80mg vs.10 mg)	CVD, CHF, PVD	Positive: CVD No effect: CHF, PVD	Baseline characteristics (LDL, HDL, Tchol, TG, Apolipoprotein B, DM med use) were similar between the treatment groups.	Randomized controlled trial
Hayashi 2009[71]	4,014	T2DM patients (elderly Japanese)	HDL, LDL	IHD, CVD	HDL: protective for IHD and CVD LDL: no effect	Age, Sex, HbA1c, LDL, HDL, TG, SBP, DBP	Prospective cohort study
Deedwania et al. 2016[72]	21,727	T2DM patients	SBP, LDL	CHD, Stroke	SBP: no effect LDL: positive for CHD	Age, Sex, dbp, BMI, TG, HDL, smoking, Hx of HTN/Stroke/IHD/TIA	Used data pooled from 3 RCTs (TNT, CARDS, IDEAL)
Despres et al. 2000[73]	2,103	Middle-aged men	LDL, HDL, Smoking	CHD	Positive: LDL, Smoking Protective: HDL	Age, Smoking, SBP, LDL, HDL, TG, presence of DM, Med use, Family history of IHD	Prospective cohort study
Miselli et al. 2014[74]	1,917	T2DM patients	Lipid-lowering drugs, TG (Male sex is harmful)	All-cause mortality	Protective: Lipid-lowering drugs Positive: TG	Age, Sex, HbA1c, SBP, DBP, LDL, HDL, TG, HTN/DM/Chol med use	Longitudinal Retrospective observational study
S. Franklin et al. 2001[75]	6,539	Men and women aged 20-79yr, w/ CHD and w/o taking HTN meds	SBP, DBP, Pulse, stratified by age < 50, age 50-59, age 60 or older	CHD	DBP: Positive (age <50) SBP: positive (age 60 or older) DBP	Age, Sex, BMI, Smoking, Presence of DM, Tchol/HDL ratio	Used Framingham Heart Study data (prospective cohort study)

					Protective (age 60 or older)		
Vaccarino, Holford, and Krumholz 2000[76]	2,152	Men and women aged 65 or older	Pulse	CHD, CHF	Positive for both outcomes	Age, Sex, Education, Marital status, Hx of Angina/DM/Stroke, BMI, Smoking, Heart rate, Body height, HTN classification based on DBP and SBP.	Used the New Haven cohort of the Establishment of Populations for the Epidemiologic Studies of the Elderly (EPESE)
S. S. Franklin et al. 1999[77]	1,924	Men and women aged 50-79yr, w/ CHD and w/o taking HTN meds	SBP, DBP, Pulse,	CHD	Positive: SBP, Pulse Protective: DBP	Age, Sex, Smoking, BMI, Glucose intolerance, Tchol/HDL ratio.	Used Framingham Heart Study data (prospective cohort study)
Said et al. 2018[78]	69,613	UK biobank participants (mean age 56.8yr,45,8% males)	Pulse	Overall Cardiovascular Disease /MI/CHD/HF/ Stroke/Death	All positive	Age, Sex, Smoking, BMI, presence of DM, Hx of Cardiovascular Disease/MI/CHD/HF/Stroke	UK biobank-prospective cohort study
Amada et al. 2000[79]	4,801	T2DM patients	SBP	Overall micro- and macrovascular complications of T2DM/MI/HF/ Stroke/PVD	All positive	Age, Sex, Race, Smoking, LDL, HDL, TG, Albuminuria, HbA1c	Prospective observational study (UKPDS)
Zhao, Katzmarzyk, Horswell,	30,154	T2DM patients	Aggressive BP control (<120/70mmHg)	CHD	Positive	Age, Sex, Smoking, SBP, DBP, Income, Type of insurance, HbA1c, LDL, eGFR, HTN/DM/Chol med use	Prospective cohort study

Wang, Li, et al. 2013[80]							
Zhao, Katzmarzyk, Horswell, Wang, Johnson, et al. 2013[81]	30,154	T2DM patients	Aggressive BP control (<110/65mmHg), High BP(≧160/100mmHg)	Stroke	Both positive (U-shape)	Age, Sex, BMI, LDL, HbA1c, GFR, Type of insurance, Income, Smoking, HTN/DM/Chol med use.	Prospective cohort study
Zhao, Katzmarzyk, Horswell, Li, et al. 2014[82]	29,627	T2DM patients	Low BP(<120/70 mmHg), High BP (≧160/100 mmHg)	HF	Both positive (U-shape)	Age, Sex, Race, BMI, LDL, HbA1c, GFR, Hx of obstructive sleep apnea, Type of insurance, Income	Prospective cohort study
Al-Delaimy et al. 2002[83]	6,547	US female registered nurses w/ T2DM	Cigarette smoking	CHD	Positive	Age, Alcohol intake, Duration of DM, Postmenopausal hormone use, DM med use, BMI, Family history of MI, Physical activity, Hx of High cholesterol/ high BP	Prospective cohort study
Price et al. 1999[84]	1,592	Men and women aged 55-74yrs	Smoking	PAD, CHD	2-3 times more likely to cause PAD than CHD	Age, Sex, Alcohol intake, SBP, DBP, BMI, LDL, HDL, TG, FPG, other cardiovascular risk factors	Prospective cohort study
Fowkes et al. 1993[85]	1,592	Men and women aged 55-74yrs, UK	Smoking	PAD, IHD	2 times more likely to cause PAD than IHD	Age, Sex, Height, BMI, non-HDL, HDL, TG, SBP, DBP, Smoking, Glucose tolerance status, Alcohol intake, Social class	Cross-sectional study

Table 1-2 Mini Literature Review on Risk Factors for Complications of Type 2 Diabetes: Use Machine Learning Techniques

Author	Study Sample(N) /Designs	Outcome(s)	Study Objective(s)	Methods	Comparison to my work
Lagani, Vincenzo, et al.[19]	1441 T1DM patients with 6.5yrs follow-up/ DCCT (RCT)	7 outcomes (CVD, Hypoglycemia, Ketoacidosis, Microalbuminuria, Proteinuria, Neuropathy, Retinopathy)	1) Predict risks of each of the outcomes 2) Identify the minimal set of clinical parameters maximally predictive.	Used 4 feature selection methods and 5 methods (Cox regression, ridge Cox regression, accelerated failure time models, SVM, random survival forest) and found the best combination for each outcome through cross validation.	1) outcome CVD was defined as an aggregate outcome but not individually defined. 2) small sample size 3) model was built for each outcome independently. 4) no model interpretability. 5) external validation-yes but small N (N=393)
Lagani, Vincenzo, et al.[86]	1441 T1DM patients with 6.5yrs follow-up/ DCCT (RCT)	7 outcomes (CVD, Hypoglycemia, Ketoacidosis, Microalbuminuria, Proteinuria, Neuropathy, Retinopathy)	1) Predict risks of each of the outcomes 2) Identify the minimal set of clinical parameters maximally predictive. 3) Impute missing values 4) Compare its performance with that of UKPDS risk engine	used 4 feature selection methods and 5 methods (Cox regression, ridge Cox regression, accelerated failure time models, SVM, random survival forest) and found the best combination for each outcome through cross validation. Bayesian network was used as missing information module.	1) outcome CVD was defined as an aggregate outcome but not individually defined. 2) small sample size 3) model was built for each outcome independently. 4) no model interpretability. 5) missing values were imputed vs. in our study, pts with missing values were dropped when we couldn't determine patient's health status properly.
Sacchi, Lucia, et al.[87]	953 T2DM patients/Retr ospective observationa l study	clinical parameters are BMI and HbA1c, and complications are IHD, Fat liver	Identify groups of patients based on temporal patterns related to drug purchase and compare the	Temporal abstraction	1) this study was not to predict an outcome but generate knowledge 2) small sample size

		disease, nephropathy, neuropathy, retinopathy, PVD, occlusion stenosis of carotid artery.	difference in clinical parameters and complications of T2DM between the groups		
Marini, Simone, et al.[88]	1441 T1DM patients with 6.5yrs follow-up/ DCCT (RCT)	2 complications (CVD and nephropathy) and clinical variables (WHR, HbA1c, SBP, LDL, HDL, TRIG, BMI)	Build trajectories of the 2 outcomes and clinical variables for t2dm pts using dynamic Bayesian network.	dynamic Bayesian network	1) outcome CVD was defined as an aggregate outcome but not individually defined. 2) small sample size 3) model interpretability - Maybe yes (I think model itself is interpretable but it's hard to see clinical usefulness and authors did not explain about that point clearly) 4) no external validation
Kazemi, Maryam, et al.[89]	600 T2DM patients from a diabetes registry	1 outcome with 4 different severity levels (diabetic peripheral neuropathy with healthy, mild, moderate, severe condition)	Predict diabetic peripheral neuropathy severity level	MSVM (multicategory SVM)	1) only one outcome 2) small sample size 3) no model interpretability 4) no external validation
DuBrava, Sarah, et al.[90]	323,378 T2DM patients with 1-yr follow-up	1 outcome (diabetic neuropathy)	1) Predict risk 2) Identify variables correlated with a diagnosis of diabetic peripheral neuropathy	Random forest	1) only one outcome. 2) sample size is big but cohort selection is not concrete (what is only required is data for 1-yr pre and post from the index). 3) short follow-up: 1yr 4) variables consisting of rules were related to utilization rather than clinical variables.

					5) model interpretability-yes (decision tree) 5) no external validation
Huang, Guan-Mau, et al.[91]	345 T2DM patients	1 outcome (diabetic nephropathy)	1) Predict risk 2) Identify variables (clinical and genetic) correlated with a diagnosis of diabetic nephropathy	Used decision tree, SVM, random forest, and naïve bayes; For each classifier, feature selection was conducted. Through cross validation, selected the best model with the highest predictive performance (decision tree was authors' choice)	1) only one outcome. 2) small sample size 3) model interpretability-yes (decision tree) 4) no external validation
Leung, Ross KK, et al.[92]	673 T2DM patients with 7.8yrs follow-up	1 outcome (diabetic nephropathy)	1) Predict risk 2) Compare performance of different machine learning methods that predict diabetic nephropathy	partial least square regression, regression tree, C50 decision tree, random forest, naïve bayes, neural network, and SVM (random forest and SVM were authors' choice)	1) only one outcome. 2) small sample size 3) no model interpretability 4) no external validation
Gulshan, Varun, et al.[93]	128,175 images from patients/Retrospective observational study	2 outcomes (diabetic retinopathy and diabetic macular edema)	Detect diabetic retinopathy and diabetic macular edema in retinal fundus photographs using deep learning	Deep learning (deep convolutional neural network)	This study worked on images. 1) no model interpretability (but results can be interpretable by human, it's image) 2) no external validation
Sudharsan, Bharath, Malinda Peeples, and Mansur Shomali.[94]	56,000 self-monitored blood glucose from patients	1 outcome (hypoglycemia)	Predict hypoglycemia from self-monitored blood glucose measurements	Used random forest, SVM, KNN, naïve bayes (random forest was authors' choice)	1) only one outcome. 2) no model 3) external validation-yes

1.3 Aims and Hypotheses

The overarching goal of my work included in this dissertation is to build *clinically useful* and t*ransferable* machine learning models on big data that can influence individual T2DM patient care towards the implementation of precision medicine.

1.3.1 Hypothesis and Specific Aim 1

I aim to develop a semi-supervised divisive hierarchical clustering algorithm to model the heterogeneity of T2DM patients that allows for interactive exploration of the cohort and patient risks. I hypothesize that there are subpopulations with substantially higher or lower risks of T2DM than the general population. Identifying such subpopulations allows to focus preventive or therapeutic resources onto patients most in need of such efforts offering a better balance between available resources and patient need.

1.3.2 Hypothesis and Specific Aim 2

I hypothesize that each risk factor has two roles: first, describing the extent of the overall deterioration of metabolic health; and second, signaling a specific complication the patient is progressing towards. The first role is common across all complications (common effect), while the second role is specific to each individual complication (outcome-specific effect). By separating these two roles, I am better positioned to identify what condition a patient is most likely to progress to and also paint a clearer picture of the risk factors that are specific to each particular complication. Specifically, I aim to

develop a Multi-Task Learning (MTL)-based methodology that can separate the overall deterioration of metabolic health from progression to specific complications.

1.3.3 Hypothesis and Specific Aim 3

Given the expectation that even simple machine learned models cannot be transferred from one health system to another, healthcare researchers often view results from machine learned models with much skepticism. Indeed, transferability of a model closely parallels the concept of external validation which is the gold standard validation in observational studies. I hypothesize that the lack of transferability is not innate to machine learning models; it is a consequence of *how* these models are constructed. These models are often constructed on data from a single institution that are not nationally representative. Since machine learned models excel at incorporating small details of the population on which they are built, the resulting models are not nationally representative; they are very specific to the population on which they were trained. Thus, I hypothesize that machine learning models built on a nationally representative data can be more easily transferred to local healthcare systems than models built on local data. I aim to demonstrate that it is possible even a complex machine learning model on a nationally representative data can be transferred to two local health systems without significant loss of predictive performance. I also aim to address the question of how to define external validity.

1.4 Outline of Dissertation

Chapter 2 provides comprehensive background of all my work, on which my hypotheses are based, and specific aims are established. It describes T2DM as a major public health problem due to high prevalence, high costs, and its chronic progressive nature. The concept and importance of evidence-based medicine and precision medicine in T2DM are discussed. It describes three challenges of managing T2DM, emphasizing the gap between the promise and the reality of precision medicine. This Chapter ends, stressing the importance of harnessing big data and machine learning techniques for successful implementation of precision medicine.

Chapters 3-5 correspond to the three aims in order. They are all manuscripts that are either published or are considered for publication.

Chapter 6 contains a validation of the models from Chapter 5 with respect to the literature. The strict word count restriction prevented me from including this section into the manuscript presented in Chapter 5.

Finally, Chapter 7 summarizes my achievements, and describes the contributions of my work to health informatics and medicine.

Chapter 2 Background

2.1 Diabetes as a Major Public Health Problem

Diabetes Mellitus (DM) is pandemic, affecting 29.1 million people (9.3% of the US population) in the US. DM is a chronic progressive disease associated with an increased risk of developing serious complications in many different organ systems and is one of the leading causes of death[1,2]. DM has been a major public health problem, imposing considerable health and financial burdens to the communities[95]; patients with DM have substantial reduction in quality of life and life expectancy[96], and the medical costs incurred by them is 3-4 times higher than those without DM[95].

DM is a metabolic disorder characterized by chronic hyperglycemia (high blood glucose). Traditionally, DM is classified into Type 1 Diabetes Mellitus (T1DM, previous known as early-onset or insulin-dependent DM) and Type 2 Diabetes Mellitus (T2DM, previously known as adult-onset or non-insulin-dependent DM).

T1DM occurs due to an autoimmune attack on the beta cells in the pancreas that produce insulin, resulting in absolute insulin deficiency; thus, T1DM patients must take insulin. T2DM starts with insulin resistance; the cells in the body do not absorb glucose properly and are unable to use it for energy effectively, therefore accumulating glucose in the bloodstream. Although the management of T2DM usually involves both lifestyle modifications and oral medications, which are usually effective for a period of time,

because of the progressive nature of the disease, the beta cells eventually begin to wear out, and insulin therapy becomes necessary for T2DM patients.

In all my work, I focus on T2DM and its complications since it is more prevalent accounting for 90-95% of all diagnosed cases[97], genetically more heterogeneous[98–100], and more complex to manage than T1DM. Generally, T1DM develops quickly over days or weeks, and the complications occur many years after the diagnosis. In contrast, T2DM has a gradual onset, so the complications are often present at the time of diagnosis; therefore, the management of T2DM is considered more complex.

2.2 Lift-threatening Complications of Diabetes

With no cure per se, DM progresses to serious complications in many different organ systems. It is common for the patients with DM to develop multiple complications over time, and the number of complications significantly increases the risk of mortality[27]. There are two types of complications of DM: microvascular complications (due to damage to small blood vessels) and macrovascular complications (due to damage to large blood vessels[101,102].

The microvascular complications include diabetic retinopathy, nephropathy, and neuropathy. Firstly, diabetic retinopathy is a leading cause of blindness, affecting 28.5% of the patients ($\geq$ 40 years old)[1,103]. Also, a large population-based cohort study demonstrated 33% of patients with T2DM developed diabetic retinopathy within 5-years after the diagnosis[104]. Secondly, diabetic nephropathy is the primary cause of kidney failure in 44% of all new cases[1]. Lastly, diabetic peripheral neuropathy is a major cause

of lower-limb amputations. It affects 60-70% of the patients (≥ 20 years old), and 15% of the patients develop at least one foot ulcer during their lifetime[105].

Macrovascular complications include ischemic heart disease (IHD), congestive heart failure (CHF), myocardial infarction (MI), cerebrovascular disease (CVD), and peripheral artery disease (PAD). DM increases the risk of developing cardiovascular disease by 2-4 times, accounting for 80% of the deaths in patients with T2DM[106].

2.3 Metabolic Syndrome and Type 2 Diabetes

Metabolic syndrome has already reached epidemic proportions in the US, affecting approximately 24% of US adults[107]. Notably, over 80% of T2DM patients present with metabolic syndrome[108], a cluster of interrelated conditions that include high blood pressure (BP), chronically elevated fasting plasma glucose (FPG), abdominal obesity, and lipids imbalance including elevated triglycerides (TG), and low high-density lipoprotein (HDL)[109]. Each one of these metabolic abnormalities is found to be a significant predictor of cardiovascular disease and total mortality[47,48], and these abnormalities occur together more often than alone, considerably increasing the risk of cardiovascular disease.

The atherosclerotic process (thickening of the arteries) is the main pathological mechanism in cardiovascular disease. And, the presence of the metabolic syndrome in patients with T2DM tends to accelerate the basic atherosclerotic process, increasing the risk of developing cardiovascular diseass[110–112]. Consequently, patients with T2DM have

a greater burden of atherogenic risk factors than patients without T2DM (or non-diabetic individuals)[74].

2.4 Challenges in Type 2 Diabetes Management

The management of T2DM is complex and difficult. Firstly, T2DM is profoundly influenced by combinations of genetic and environmental risk factors (e.g., aging, obesity, dietary patterns, physical inactivity), which result in heterogeneity of treatment response in T2DM patients[113,114]. For example, in the largest 4 randomized controlled trials (RCTs)[13–16], to date, the efficacy of intensive glycemic control on decreasing the risks of micro- and macrovascular complications varied, and even one of the RCTs terminated earlier because some participants responded adversely.

Secondly, T2DM is not an independent disease. The diseases in metabolic syndrome interact with each other, substantially increasing the risk of T2DM and its complications. For instance, hyperglycemia alone seems not a significant risk factor for IHD, while hyperglycemia with the presence of metabolic syndrome remarkably increases the risk of IHD[115]. Thus, even a minor adjustment on a single risk factor can dramatically influence the patient's overall health status and clinical outcomes[2,116,117].

Thirdly, the heterogeneity of T2DM and complicated interactions among the diseases in metabolic syndrome yield enormous variability in disease progression; therefore, a substantial overlap among risk factors for the complications exists, and this makes the complications difficult to distinguish. For example, a large body of evidence has shown that HbA1c is a significant risk factor for all the complications[51–61] and LDL is

a significant risk factor for IHD[72,73], CVD[70], and CKD[60,61]. Therefore, making an accurate prognosis (def. "the likelihood of future outcomes in patients with a given disease or health condition"[118]) is often unsatisfactory[119], and optimal individualized therapeutic strategies remain incomplete[102,120,121]. For the successful management of T2DM, individualized and evidence-based clinical guidelines are necessary[3].

2.5 Evidence-based Medicine

Traditionally, clinical decisions are made based on individual clinical expertise. Its weakness is, when clinicians see a patient who exhibits common clinical conditions but follows unexpected, unusual patterns or who has very rare clinical case, their medical reasoning only based on their past experience could fail. Evidence-based medicine (EBM) is defined as "the integration of best research evidence with clinical expertise and patient values"[122], which aims to optimize clinical decision making by stressing the use of evidence generated from well-designed and well-conducted study.

2.6 Contradictory Evidence from Randomized Controlled Trials

RCT is considered as the gold standard for clinical research. The Action to Control Cardiovascular Risk in Diabetes (ACCORD)[13], the Veterans Affairs Diabetes Trial (VADT)[14], the Action in Diabetes and Vascular Disease Preterax and Diamicron Modified Release Controlled Evaluation (ADVANCE)[15], and the UK Prospective Diabetes Study (UKPDS)[16] are the largest RCTs that studied the impact of intensive

glycemic control on micro/macrovascular complications. However, their results were not consistent, rendering the evidence contradictory and inconclusive.

Table 2-1 Contradictory Evidence from RTCs

Predictors	ACCORD	VADT	ADVANCE	UKPDS
N	10,251	1,791	11,140	1,704
Follow-up (years)	3.5	5.6	5	10
Age (years)	62	60	66	53
Duration of Diabetes (years)	10	11.5	8	
Sex (%male)	39	97	42	46
History of cardiovascular disease	35	40	32	
Baseline HbA1c(%)	8.1	9.4	7.2	7.2
CVD benefit with intensive glycemic control	Yes, in a subgroup (w/o cardiovascular disease)	Yes, in a subgroup (duration < 12) No (adverse events) in a subgroup (duration ≥ 12)	No	Yes
Increased mortality with intensive therapy	Yes	Yes (but insignificant)	No	N/A
Association of hypoglycemia with increased mortality	Both the intensive and standard group	Only the standard	N/A	N/A

The UKPDS demonstrated that intensive glycemic control for participants who were newly diagnosed with T2DM significantly reduced microvascular complications, MI, and all-cause mortality during the 10-year follow-up period, generating evidence that intensive glycemic control can decrease T2DM complications[16].

On the contrary, the ADVANCE and the VADT failed to show significant reduction in cardiovascular events with intensive glycemic control[123,124]. The ACCORD even terminated early since increased mortality was observed in the intervention group (an intensive glycemic therapy group)[13].

Concerning the contradictory evidence, a review article[6] emphasized different participant characteristics. UKPDS participants were relatively young (25-65 years) and newly diagnosed with T2DM, whereas ADVANCE, VADT, and ACCORD participants were aged 60 years or older and had longstanding T2DM (mean duration: 8-11 years). The latter also had either a history of cardiovascular disease or multiple risk factors for cardiovascular disease, indicating the presence of established atherosclerosis. For patients with T2DM and high prevalence of cardiovascular risk factors, only achieving normoglycemia is insufficient for the reduction of cardiovascular events[121]; rather, a multi-drug therapy is recommended since it has shown sustained advantageous effects on decreasing the CVD risk and cardiovascular mortality[121,125,126].

Although being a dominant method to create clinical evidence, RCTs are expensive, extremely time-consuming, and administratively burdensome. Therefore, in RCTs, interventions are narrow focus and participants are highly selected, often resulting in limited generalizability[5].

T2DM is a heterogeneous and complex disease. Therefore, to create reliable, generalizable, and clinically useful evidence for T2DM management, it is critical to examine not only multiple risk factors but also multiple potential clinical outcomes using a large number of T2DM patients with a sufficiently long follow-up period.

2.7 Benefits of using Electronic Health Record

Electronic Health Record (EHR) adoption has been accelerated since The Health Information Technology for Economic and Clinical Health (HITECH) Act was enacted.

It offers substantial financial incentives through Medicare and Medicaid for healthcare providers to encourage the adoption and meaningful use of EHRs[127], resulting in a 69% adoption rate in primary care settings[128].

Recently, EHR data have become abundantly available, containing years of medical history for a large number of patients. Especially, the benefits of using EHR data (e.g., low cost, large volume and easy access[129]) enable scientists to generate clinical evidence efficiently[5].

EHR data collected from real-world clinical practice include a large number of heterogeneous T2DM patients; thus, harnessing EHR data (especially, from multiple healthcare systems) could possibly alleviate the generalizability problem. Additionally, it offers an opportunity to identify and analyze clinically meaningful subpopulations. Besides, as longitudinal patient records, EHR data are invaluable for studying progressive diseases where time is critical. Once EHR data are converted into useful clinical information by intelligent clinical decision support (CDS) tools, substantial clinical benefits can be expected.

2.8 Need for Advanced Clinical Decision Support Tools

Recently published studies have shown positive effects of using EHRs over paper-based records on T2DM care and patient outcomes[130–133]. However, most studies demonstrating the success of using EHRs on T2DM care used clinical guideline adherence (e.g., screening or conducting laboratory tests) or medication adherence as measures of success[134,135] even if they are not necessarily translated into improved

healthcare outcomes. For example, receiving the HbA1c test on a recommended schedule does not necessarily mean that the patient's glucose levels are being controlled. Additionally, it is not expected for the patients with comorbidities, polypharmacy issues, and/or limited life expectancy to meet all-or-none bundled measures[135]. What this suggests is simply adopting EHRs does not guarantee improved T2DM care and the patients' outcomes[136–138].

EHR adoption is an important first step. But considering the limited time for the physician office visit, the complexity of T2DM management, and information and cognitive overload[139], CDS at the point of care is imperative to support clinicians with full benefits of EHRs[136,140].

Unfortunately, current existing EHR-based CDS systems exhibit restricted functionality (e.g., *non-specific action* prompts and simply translating existing clinical guidelines into rules). Such simple functions often frustrate physicians who need immediate and actionable answers with detailed advice about treatment at the point of care[141]. Therefore, to achieve improved T2DM care and the patients' outcomes, the evolution of CDS functionality from simple data capture to intelligent data analysis is critical.

2.9 EHR Data Mining

The increased volume, variety, and complexity of EHR data require advanced techniques to automatically understand, process, and summarize the data. Data mining is an approach to data analysis and knowledge discovery that emerged in the mid-1990's[142].

Data mining is defined as "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner"[143]. Thus, successful application of data mining to EHR data can transform the massive amount of EHR data into useful clinical information. Accordingly, data mining techniques are becoming more widely used in clinical applications[17,19,142,144–148]. Some examples of data mining applications in T2DM will be shown in "1.2 Related work" section.

2.10 Precision Medicine

The definition of precision medicine by the National Institutes of Health (NIH) is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person"[7]. This new model of customized (tailored, personalized) healthcare requires transformative technologies based on the increasingly available computational power, the use of big data (EHR, genome, wearable devises, etc.), and machine learning algorithm to reach clinically applicable knowledge able to influence individual patient care.

T2DM is an important topic in precision medicine due to the complexity of the management. And, as we have seen, a huge gap exists between the promise and reality of precision medicine in the management of T2DM. Big data analytics and machine learning can provide a new way to generate evidence that enhances clinical practice guidelines with more personalized recommendations[8]. Therefore, the approach of precision medicine will immensely benefit the prevention and management of T2DM.

Chapter 3 Divisive Hierarchical Clustering towards Identifying Clinically Significant Pre-Diabetes Subpopulations

3.1 Introduction

Type 2 Diabetes Mellitus is one of the fastest growing chronic diseases in the United States, with a profound influence on public health quality and cost[149,150]. It is a progressive disease, associated with an increased risk of developing serious cardiac, vascular, renal and ophthalmological complications, and it is one of leading causes of death[150]. With no cure per se, prevention and management are of paramount importance. As effective preventive measures such as lifestyle change and drug therapy exist[151,152], early identification and management of patients at high risk is an important healthcare need.

Numerous diabetes risk indices aimed at early identification of patients at high risk have been developed[24]. Arguably, the most popular such index is the Framingham score[25], which has gained wide acceptance in clinical practice. The Framingham model assigns a risk to a patient based on the risk factors the patient presents with and the resulting score can be used to stratify patients into low, moderate, or high-risk groups. Almost all indices, the Framingham score included, estimate the risk of diabetes in an additive fashion, assuming that the risk factors act independently.

Interactions among risk factors are known to exist[30,153–156]. Recent work[30,153–155] aimed to address interactions, most prominently through the application of association rule mining (ARM)[9–12]. ARM was specifically designed to discover sets of associated

risk factors, along with the affected subpopulations. While association does not always translate into (non-additive) interaction, it often does. Given its ability to seamlessly incorporate interactions, ARM has successfully identified patient subpopulations that face significantly increased or decreased risk of diabetes[153,156]. Another beneficial characteristic of the ARM model lies in the straightforward interpretability of the individual rules. Thus, the ARM model does not just provide a risk estimate, but it also offers a "justification" in the form of the associated risk factors in the rule.

ARM has its own shortcomings. While interpretability is one of the hallmarks of the ARM modeling approach, ARM algorithms tend to extract combinatorially large sets of redundant rules, which quickly erodes interpretability. Under these conditions, it is necessary to offer fine control of the amount of details the ARM model extracts; however, this is precisely where ARM falls flat. When ARM discovers a manageable number of rules, they tend to be too general to be useful; when the model is sufficiently detailed to give the user new insights, the sheer number of rules impedes interpretation. There is a reason for this phenomenon. The ARM rule set is highly redundant: the same subpopulation is described by an exponential number of rules, each rule associating the subpopulation in question with a different set of risk factors. This unfortunate property obfuscates the disease mechanism.

In this work, we propose the use of a novel divisive hierarchical clustering[157] technique, which retains most of the advantages of ARM, while it alleviates the interpretability issues. From a hierarchical clustering, depending on the desired amount of detail, many clusterings can be extracted. Each clustering consists of a varying number of

clusters, is complete (they include all patients) and non-overlapping (each patient belongs to exactly one cluster).

Our proposed approach retains all advantageous properties of ARM and alleviates its primary shortcoming: interpretability is enhanced through the elimination of redundancy and through lending the user fine control over the amount of details the clustering should incorporate.

3.2 Methods and Materials

3.2.1 Data

In this study we utilized a large cohort of Olmsted County, Minnesota, residents identified by using Rochester Epidemiology Project resources. The Rochester Epidemiology Project (REP)[158] is a unique research infrastructure that follows residents of Olmsted Co., MN over time. The baseline of our study was set at Jan. 1, 2005. We included all adult Mayo Clinic patients with research consent, who are part of the REP, resulting in a study cohort of 69,747 patients. From this cohort, we excluded all patients with a diagnosis of diabetes before the baseline (478 patients), missing fasting plasma glucose measurements (14,559 patients), patients whose lipid health could not be determined (1,023 patients) and patients with unknown hypertension status (498 patients). Our final study cohort consists of 52,139 patients (overlaps between the groups exist) who were followed until the summer of 2013.

We collected demographic information (age, gender, body mass index BMI), laboratory information (primarily fasting plasma glucose and lipid panel), vital signs

(blood pressure and pulse), relevant diagnosis diagnoses (obesity, hyperlipidemia, hypertension, renal failure and various cardiac and vascular conditions), aspirin use, and medications used to treat hypertension and hypercholesterolemia. Additional known risk factors for diabetes (such as tobacco usage) were also included.

3.2.2 Features

To enhance the interpretability of our results, the variables were transformed into binary variables to indicate the presence and severity of risk factors. These variables are typically constructed as a meaningful combination of diagnoses, abnormal vital signs, abnormal laboratory results, and use of medications by drug class. Laboratory results were considered abnormal when they exceeded the cutoffs published in the American Diabetes Association (ADA)[159] guidelines. Table 3-1 shows the definitions of the variables used henceforth.

Table 3-1 Predictors and their definitions

Predictors	Definitions
Demographics	
age.18+	Age > 18 and < 45
age.45+	Age ≥ 45 and < 65
age.65+	Age ≥ 65
genderM	Male
Comorbidities	
obese	Obesity (BMI ≥ 30 or diagnosis)
tobacco	Current smoker
renal	Renal disease
chf	Congestive Heart Failure
ihd	Ischemic Heart Disease
Major risk factors and their severities	
ifg.no	Normo-glycemic patients: fasting plasma glucose (FPG) ≤ 100
ifg.pre1	Impaired Fasting Glucose level 1: FPG > 100 and ≤ 110
ifg.pre2	Impaired Fasting Glucose level 2: FPG > 110 and ≤ 125

htn.no	No indication of Hypertension: no diagnosis of HTN, no hypertensive drugs are described and blood pressure results (if present) are normal.
htn.any	Indication of Hypertension exists in the form of either a HTN diagnosis or abnormal blood pressure measurement
htn.tx	Hypertension required therapeutic intervention; however, at most 3 HTN drugs were prescribed.
htn.pers	Persistent Hypertension. Patients present with abnormal blood pressure measurements despite having been prescribed 3 or more drugs; or they are prescribed 4 or more drugs (regardless of blood pressure results).
hyperlip.no	No indication of Hyperlipidemia: no diagnosis of hyperlipidemia, no cholesterol drugs and no abnormal lipid panel results are present.
hyperlip.any	Indication of Hyperlipidemia exists in the form of diagnosis or abnormal laboratory results.
hyperlip.tx	Hyperlipidemia with therapeutic intervention: a diagnosis code or abnormal laboratory result indicates hyperlipidemia and a single cholesterol drug is prescribed.
hyperlip.multi	Hyperlipidemia requiring multi-drug intervention: multiple cholesterol drugs are prescribed.

3.2.3 Patient Clusterings

The purpose of clustering is to partition patients into groups (clusters), such that patients within the same cluster are more similar to each other than to patients in a different cluster. Formally, in our application, a **cluster** is a set of patients, who share risk factors relevant to diabetes progression and have similar diabetes risk. A **clustering** is a non-overlapping complete set of clusters. A clustering is *complete* in the sense that all patients in the population are assigned to a cluster, and it is *non-overlapping*, as each patient is assigned to a single cluster in a clustering. Our goal is to create a patient clustering, where the clusters correspond to clinically meaningful patient subpopulations.

To identify such subpopulations, we applied bisecting divisive hierarchical clustering. The algorithm iteratively constructs a hierarchy of clusters in a top-down (divisive) fashion, in each iteration bisecting a cluster into two new (child) clusters. A

cluster is bisected using a *splitting variable.* One of the two child clusters contains all patients from the parent cluster for whom the splitting variable is true, and the other child contains all patients for whom the splitting variable evaluates to false. For example, if the parent cluster (cluster to split) consists of patients with hypertension (htn is true) and the splitting variable is ifg.pre2 (fasting plasma glucose FPG > 110), one of the child clusters is comprised of hypertensive patients having high FPG (ifg.pre2=true) and the other cluster is comprised of hypertensive patients with lower FPG (ifg.pre2 is false).

The algorithm proceeds by recursively bisecting each cluster into two child clusters starting with a cluster that represents the entire population. The algorithm terminates when no cluster can be bisected without having insufficient number of patients in the resultant child clusters; or when the patients in the cluster are sufficiently similar to each other.

The splitting variable is selected on the basis of how much variability in the diabetes outcome it can explain; bisections that explain a large amount of variability are preferred. Let t_j denote the follow-up time (in days) and δ_j the diabetes status at the end of follow-up for patient *j*. This patient is censored when the diabetes outcome is negative ($\delta_j = \text{false}$) at the end of follow-up. The martingale residual $M_j(t)$ for a patient *j* at time t_j is computed as the difference between the observed number $\delta_j(t)$ of event (1 if a patient *j* had developed diabetes before (or exactly at) time t_j, 0 if censored) and the estimated number $H_j(t)$ of events (cumulative hazard)

$$M_j(t) = \delta_j(t) - H_j(t).$$

To calculate the cumulative hazard, we use the Nelson-Aalen estimator,

$$H_j(t) = \sum_{t_i \le t} h_j(t_i) = \sum_{t_i \le t} \sum_k \frac{dN_k(t_i)}{Y_k(t_i)},$$

where $h_j(t_i)$ denotes the (non-cumulative) hazard of patient *j* at time t_i, *k* iterates over all patients, $dN_k(t_i)$ denotes the number of diabetes incidents that patient *k* suffers exactly at time t_i (0 or 1) and $Y_k(t_i)$ indicates whether patient *k* is at risk at time t_i. The formula for the non-cumulative hazard can be thought of as the number of events occurring exactly at time t_i divided by the number of patient at risk at that time. When multiple patients suffer events at exactly the same time, these events are arbitrarily serialized.

Suppose we have a cluster C_1, which we need to bisect into clusters C_{11} and C_{12} using a particular splitting variable. Further, let SSR(*C*) denote the sum of squared martingale residuals for any cluster *C*. Bisecting C_1 will decrease the total SSR by

$$G = SSR(C_1) - [SSR(C_{11}) + SSR(C_{12})].$$

Each splitting variable produces a different G value. Among the possible splitting variables, we select the one that reduces the SSR the most, or equivalently, maximizes G. This is the splitting variable that explains the diabetes outcome in C_1 the best, thus it can be thought of as 'most relevant' to diabetes in the subpopulation corresponding to cluster C_1.

Once the cluster hierarchy has been constructed, the final clustering can be extracted. A **leaf cluster** is a cluster that is not bisected. Our hierarchical clustering algorithm ensures that each patient falls into exactly one leaf cluster, thus the collection of leaf clusters form a non-overlapping complete clustering of the patient population.

We wish to make two notes. First, our clustering algorithm is similar to the survival tree construction algorithm[160]; in fact, one can think about it as an adaptation of

the recursive partitioning algorithm[161] for censored outcomes to a clustering application. Indeed, we follow Therneau et al.[162] in broad strokes and adapt their ANOVA criterion for censored outcome: we use sum squared martingale residuals instead of sum squared error. Second, the analogy between recursive partitioning and our algorithm goes deeper. The martingale residual can be rescaled into a deviance residual. Just as the sum squared error relates to the "deviance" of two nested Gaussian models, the sum squared deviance residuals relate to the deviance of two nested survival models, enabling the use of likelihood ratio tests for significance testing. Since our purpose is to construct the full hierarchy of clusters, we do not perform significance testing and use the martingale residuals instead of the deviance residuals.

Clustering and statistical analysis were conducted with the use of R version 3.0.1.

3.3 Results

In what follows, we demonstrate that our clustering algorithm successfully identified potentially interesting clusters that consist of patients with substantially higher or lower risk of diabetes than the general patient population. The clustering (the collection of leaf clusters) assigns each patient to exactly one leaf cluster and each leaf cluster possesses a cumulative hazard curve (specific to the subpopulation that the leaf consists of). Thus, the clustering can be used to estimate a patient's risk of diabetes and can hence serve as a diabetes index. We will also demonstrate that our clustering used in this fashion outperforms the popular Framingham score. Thirdly, we will also show by constructing the entire hierarchy of clusters, that we can extract clusterings that

encompass varying amounts of detail. Finally, we will demonstrate that our clustering can model non-proportional hazards as well as interactions among the risk factors.

3.3.1 Identifying high and low risk subpopulations

We performed hierarchical clustering of our patient population under the user-defined constraint that clusters with less than 50 patients are not bisected further. We identified 275 leaf clusters. From these leaves, we selected two: one indicative of very high risk and one indicative of very low risk. We then compared these leaves with the general population.

In Figure 3-1, we display the Kaplan-Meier survival curve and the cumulative hazard curve for the entire population (blue dotted line) and for the above two clusters: red solid line is used for the high-risk cluster and green dotdash line for the low-risk one. The patients (n = 61) in the high-risk cluster have fasting plasma glucose greater than 110 mg/dL (ifg.pre2), hyperlipidemia that requires therapeutic interventions (hyperlip.tx), and are current smokers (tobacco). The low-risk cluster consists of the patients (n = 498) characterized by fasting plasma glucose level equal to or lower than 100 mg/dL, no indication of hypertension, no indication of hyperlipidemia, no obesity, no renal disease, no congestive heart failure, no ischemic heart disease, no aspirin use, male non-smokers being between 45 and 65 years of age.

From Figure 3-1, the difference in diabetes progression among these three subpopulations is obvious and the risk factors for these patients are consistent with our clinical expectation.

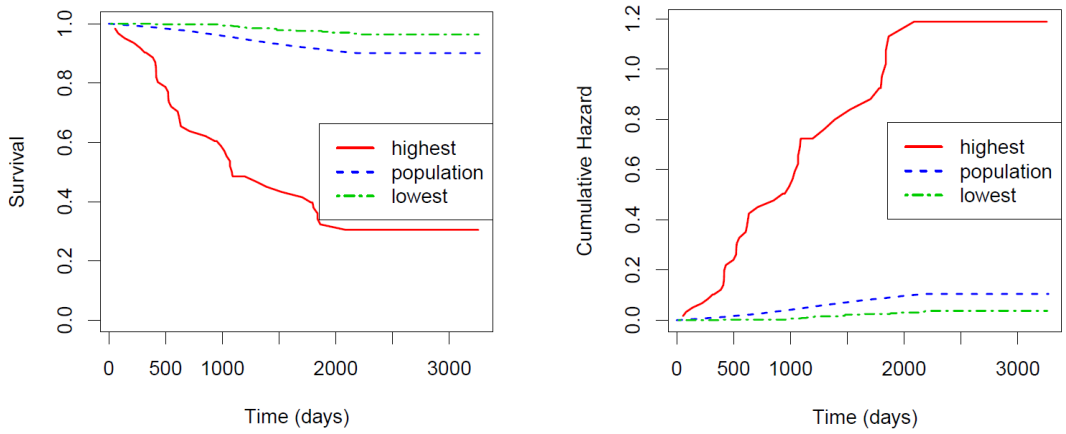


Figure 3-1 Kaplan-Meier survival curve (Left) and Cumulative incidence of diabetes (Right) for two pre-diabetic subpopulations (red sold and green dotdash) and the entire population (blue dotted)

3.3.2 Clustering as a diabetes index

As we described earlier, the leaf clusters form a non-overlapping clustering, where each patient belongs to exactly one leaf cluster. Since each leaf cluster contains a survival function of the corresponding subpopulation, it is possible to use the clustering as a diabetes index, providing a risk estimate for each patient. In this section, we compare the clustering as a diabetes index against the popular Framingham score, an actual diabetes index in clinical use. Specifically, we use concordance as our evaluation measure. Concordance is the probability that for any two patients where one progressed to diabetes earlier than the other, the one that progressed earlier has a higher predicted risk. The clustering achieved a concordance of 0.78, while the Framingham score achieved a lower concordance of 0.70, signifying the clustering has imporved discriminatory power.

3.3.3 Controlling the amount of detail

An important advantage of the proposed clustering technique over alternative methods, such as association rules, is that it offers fine control over the amount of detail it presents to the user. This control can be achieved by **cutting** the hierarchy at a particular level. To illustrate this point, we depict the entire cluster hierarchy in Figure 3-2. The leaf clusters are listed along the horizontal axis and the vertical axis indicates the SSR of the cluster. The hierarchy is represented by a dendrogram, which can be interpreted as follows. The root of the dendrogram is at the top (SSR=4645) and it represents a cluster that includes all patients. The root is split into two clusters (on ifg.pre2; not shown) one with SSR 730 (ifg.pre2=true) and one with SSR 3914 (ifg.pre2=false). The subpopulation having ifg.pre2=true is split on htn.pers into two clusters, one with SSR 41 (persistent hypertension present) and one with SSR 688 (htn.pers=false). In short, the dendrogram allows us to trace the bisections our algorithm performed and it also depicts the SSR of the resultant clusters.

We can cut the dendrogram at any SSR of our choice. Cutting the hierarchy produces a new set of leaf clusters, which in turn forms a non-overlapping complete clustering of the patient population. For example, if we cut the dendrogram at SSR of 4000 (which is very close to the top), we obtain only two leaf clusters: ifg.pre2=true with SSR of 730 and ifg.pre2=false with SSR of 3914. If we cut the dendrogram at a lower SSR, say at 500, we will obtain a larger set of leaf clusters (26 in this example) each having lower SSR. This particular cut is shown in Figure 3-2 as a magenta line. Larger number of leaf clusters offers a larger amount of detail. Selecting an SSR for cutting the

dendrogram is what allows us to control the amount of detail (number of leaf clusters) in a predictable fashion. The SSR of the resulting leafs also shows the within-cluster similarity of the patients.

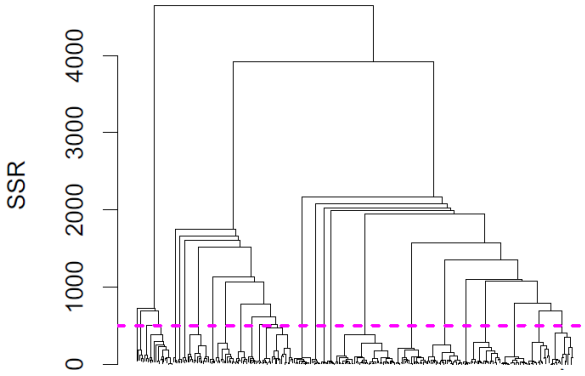


Figure 3-2 Dendrogram of the entire hierarchy of clusterings

Naturally, a tradeoff exists between the amount of detail and the predictive capability of a clustering. In Figure 3-3, we visualize this tradeoff. The horizontal axis represents the SSR at which the dendrogram was cut and the vertical axis represents the resultant clustering with the number of clusters depicted in left pane and the predictive capability (as measured by concordance) in the right pane. The figure shows that as we increase the SSR (move right on the horizontal axis), we decrease the amount of detail (number of clusters) and along with the decreased amount of detail, the predictive capability of the clustering decreases, as well.

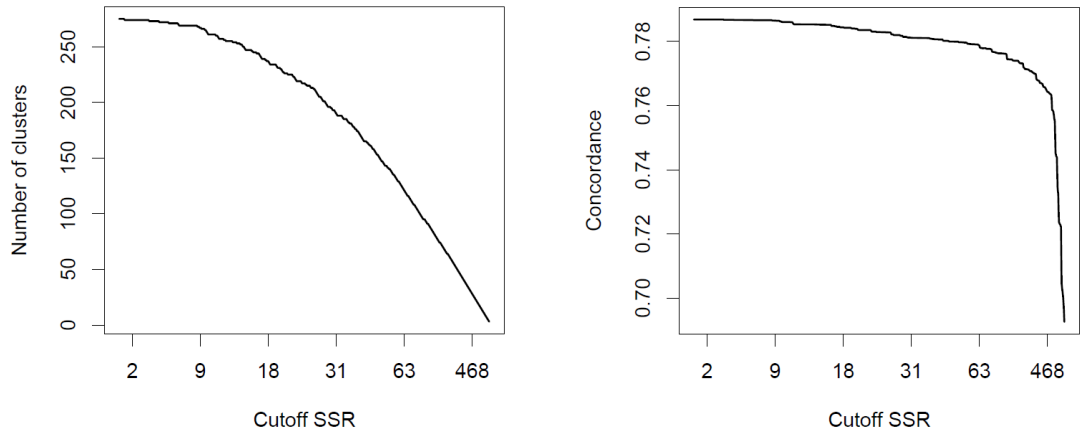


Figure 3-3 Tradeoff between the amount of detail (number of clusters) and the predictive capability (concordance)

3.3.4 Non-proportional hazard

We plot the cumulative hazard functions for the 26 clusters we extracted earlier (by cutting the hierarchy at SSR 500) in Figure 3-4. To avoid overcrowding the image and preserve good visibility of the lines, we separated the 26 clusters into four panes essentially at random. The IDs in the legend refer to their original IDs (IDs before cutting), thus they can exceed 26. Cumulative hazards across the clusters are not proportional because the LOGLOGS plots of the 26 clusters (shown in Figure 3-5) do not appear as parallel lines, indicating interactions between time and subpopulations. This non-proportionality was correctly captured by our approach. To show these clusters are clinically relevant, we selected the 12 highest risk clusters out of the 26 and described in Table 3-2.

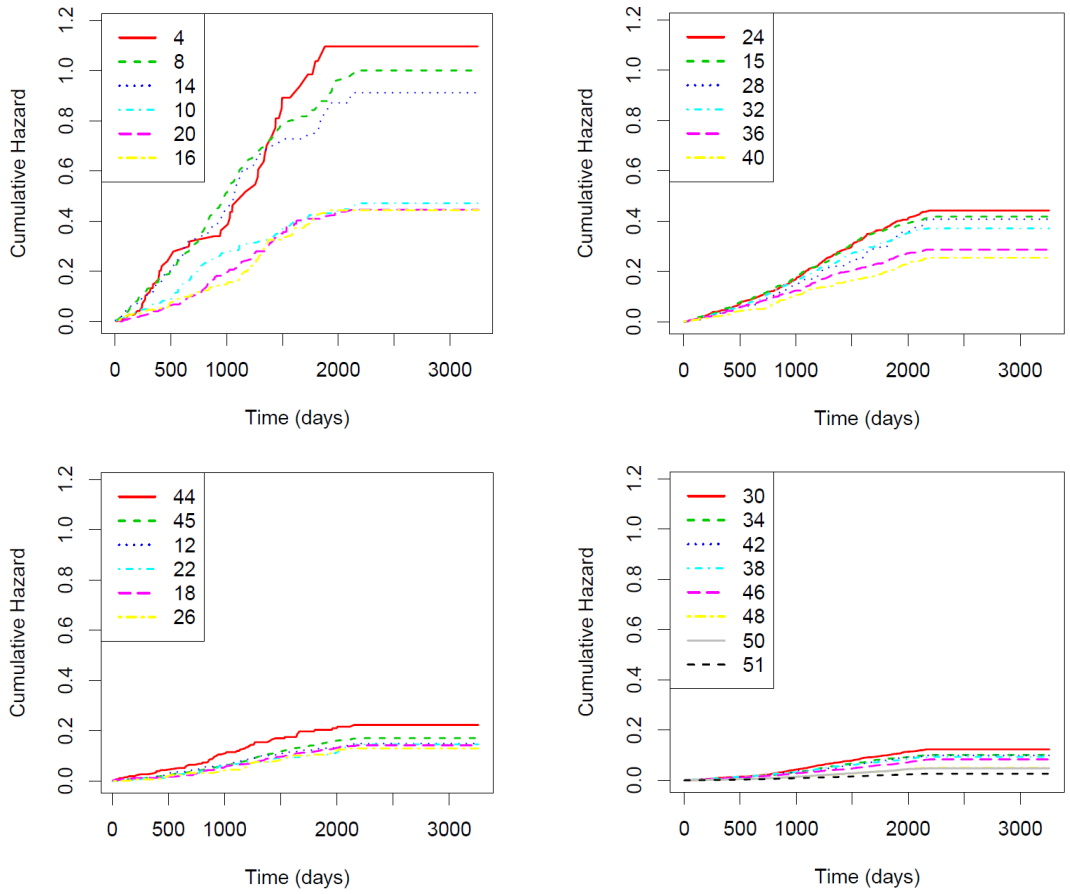


Figure 3-4 Identified pre-diabetic subpopulations based on cumulative hazard after infinite follow up time

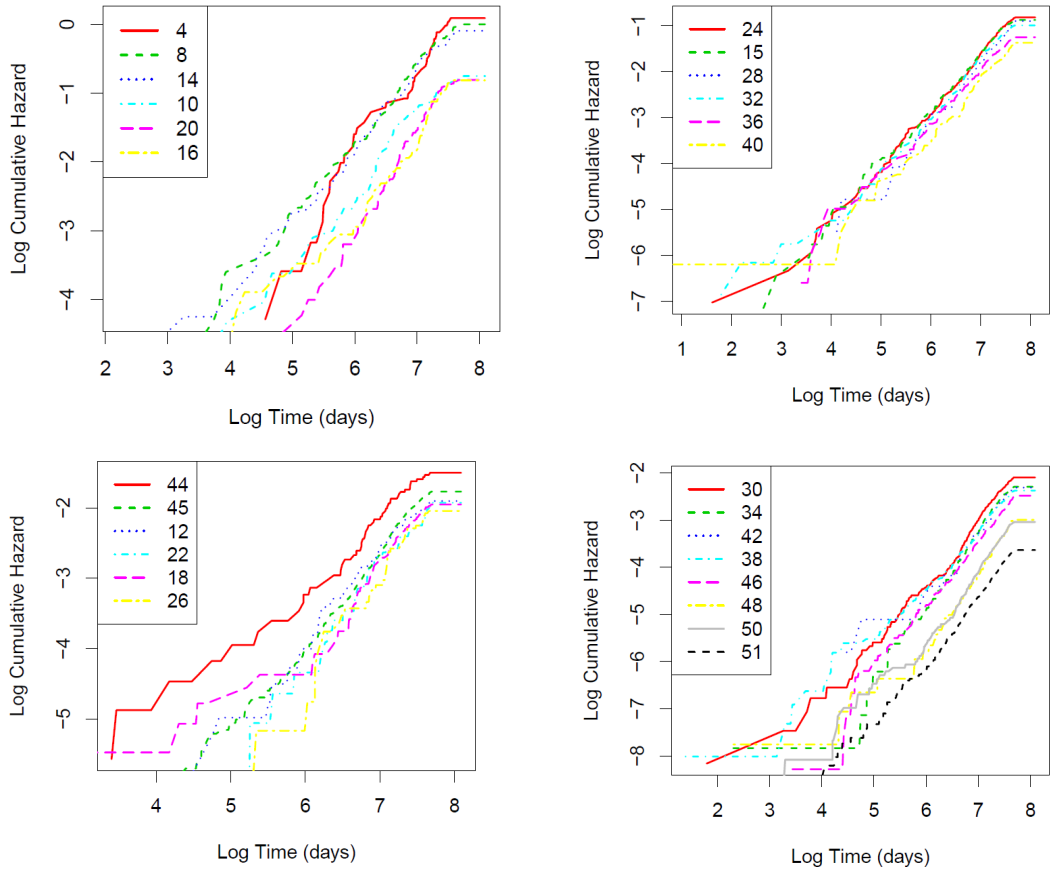


Figure 3-5 LOGLOGS plot of cumulative hazard

Table 3-2 Subpopulation summarization with cumulative hazard at the end of the study

Cluster ID	Patient Count	SSR	Cumulative Hazard	Risk Factors
4	74	40	1.02	ifg.pre2=true, htn.pers=true
8	297	180	1.00	ifg.pre2=true, htn.pers=false, obese=true
14	212	121	0.91	ifg.pre2=true, htn.pers=false, obese=false, hyperlip.tx=true
10	227	81	0.47	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=true
20	280	88	0.45	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=true
12	736	88	0.15	ifg.pre2=false, ifg.pre1=false, htn.pers=true

24	1130	380	0.44	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=true
15	1276	384	0.42	ifg.pre2=true, htn.pers=false, obese=false, hyperlip.tx=false
28	241	68	0.41	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=false, ihd=true
32	949	277	0.37	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=false, ihd=false, obese=true
36	735	166	0.28	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=false, ihd=false, obese=false, htn.tx=true
40	493	102	0.25	ifg.pre2=false, ifg.pre1=true, hyperlip.multi=false, renal=false, htn.pers=false, hyperlip.tx=false, ihd=false, obese=false, htn.tx=false, aspirin=true

3.3.5 Interactions among risk factors

Cluster 4 consists of patients who have ifg2.pre2 and htn.pers, and the estimate of the cumulative hazard at the end of the study is 1.02. To estimate the hazard under the assumption that the two risk factors are additive (act independently), we fit a Cox regression model to the entire population using only the above two variables as predictors. We used this Cox model to make a prediction for the subpopulation represented by cluster 4 and their cumulative hazard is 1.44. The difference between this prediction (of 1.44) and the prediction of 1.02 by the clustering strongly suggests that an interaction between these two risk factors (FPG and HTN) exists. While we do not know the exact risk of diabetes (true value for the cumulative hazard) in this subpopulation, it is between the observed prevalence of diabetes in this subpopulation, which is 0.57, and 1.0 (each patient can only experience at most one event of diabetes). 1.02 is closer to this range than 1.44, thus the additive Cox regression model overestimated the risk.

3.4 Discussion

In this paper, we presented a novel bisecting divisive hierarchical clustering algorithm to identify clinically relevant patient subpopulations using type 2 diabetes as the endpoint. In a good clustering, patients within the same cluster are more similar to each other than to patients in a different cluster. Patients in our clusters are similar to each other because they share the same risk factors that are most relevant to diabetes and also have similar risk of developing diabetes. We have shown that our clustering can be used as a diabetes index: when the clustering is sufficiently detailed, it outperformed the Framingham score in terms of concordance (ability to distinguish high-risk patients from low-risk patients). While ARM models have also shown excellent predictive performance, their high level of redundancy leads to unnecessary computational cost. In the following discussion, we examine the beneficial properties of the clustering particularly compared to ARM models and the potential of overfitting.

3.4.1 Comparison to Association Rule Mining

Recent developments of ARM[163], including survival association rule mining[154] have demonstrated its applicability in the EHR mining domain and its appropriateness to serve as a diabetes index. The key advantage of the ARM methodology lies in its interpretability: individual rules are straightforward to interpret, and the interpretation provides a context around the risk estimate (e.g. the high risk is due to persistent hypertension and severe hyperlipidemia).

As previously discussed, the disadvantage of ARM is that it generates an exponentially large, redundant rule set. With many rules applying to the same patient, making prediction for an individual becomes non-trivial[153,154], making a direct comparison between ARM and clustering leave room for arguments. Just to show that the predictive performance of ARM and clustering are similar, we performed a simple, albeit admitted imperfect, comparison.

We largely followed the methodology outlined in the studies using ARM to assess the risk of type 2 diabetes[11,153]: we built a Cox model with age and gender as the predictors, and extracted distributional association rules[12,153] indicating an association between the martingale residual and the major risk factors (IHD, hypertension and hyperlipidemia as defined in Table 3-1 that covered at least 50 patients (same coverage as used for clustering). For each patient, we made a prediction using the most specific rule. The concordance of the resultant model was .7601 with a standard error of 004. This is comparable to the performance of the clustering model. Additionally, both the ARM-based models and our clustering have the ability to automatically discover interactions among risk factors and seamlessly incorporate them into the model or clustering.

Our proposed method goes beyond the state of the art by allowing the user to control the amount of details the clustering should incorporate. This is particularly beneficial, because the amount of detail can be adjusted to the needs of the consumer of the model. For example, when the user of the clustering is an automated clinical decision support system, a highly detailed clustering may be desirable. Computational systems can handle

complex models, even as complex as the ARM models, and thus a clustering that incorporates a large amount of details (without overfitting) can be most appropriate. Another potential use of the clustering produced by our method concerns clinical investigation, where clinicians, rather than computers, view the clustering results. Presenting excessively detailed complex models to investigators can be more distracting than useful, thus a moderately complex clustering may be most desirable. Our method constructs the entire cluster hierarchy upfront allowing investigators to drill down for further details. This can be achieved through further clustering a specific subpopulation (leaf), as needed.

3.4.2 Overfitting

Models as flexible as the clustering-based model or the association rule set are susceptible to overfitting the data. In this application, we were not particularly concerned with the predictive performance of the model as it is secondary to its interpretability. To avoid overfitting, we required the presence of at least 50 patients in each node (or association rule), which is sufficient to reliably estimate their risk. Also, Figure 3-3 shows no sign of overfitting: increased number of nodes have consistently led to improved performance on a validation set. Nonetheless, when the clustering is used as a predictive modeling tool, the number 50 needs to be tuned more carefully and attention must be paid to the potential overfitting.

In summary, we have demonstrated that our clustering method retains the benefits of existing diabetes risk models and adds its own advantage through allowing for fine

control of detail that is presented to the user. This promises great potential of contributing to clinical practice.

Chapter 4 Multi-Task Learning to Identify Outcome-Specific Risk Factors that Distinguish Individual Micro and Macrovascular Complications of Type 2 Diabetes Subpopulations

4.1 Introduction

Type 2 Diabetes Mellitus (T2DM) is an irreversible chronic disease. It is associated with the metabolic syndrome, a cluster of interrelated conditions that include high blood pressure (BP), chronically elevated fasting plasma glucose (FPG), abdominal obesity, and lipids imbalance including elevated triglycerides (TG), and low high-density lipoprotein (HDL)[109]. Since complicated interactions among these conditions exist, even a minor adjustment on a single risk factor can dramatically influence the patient's health status and clinical outcomes[2,116,117]. Hence, a comprehensive understanding of the effects of these risk factors on various complications is necessary for the successful long-term management of T2DM patients.

Studies that identify risk factors for complications of T2DM abound[61,164,165], however they fail to paint an accurate picture of the patient's health status and progression to the most likely next complication. In these studies, regardless of which complication they focus on, the risk factors tend to be largely the same (e.g., BP, FPG, lipids, and kidney function). The reason for this large overlap is that the above risk factors capture the effect of deteriorating overall metabolic health that underlies all these outcomes rather than capturing the effects that differentiate among the outcomes. This

suggests that the risk factors have two roles: first, they describe the extent to which the patient's overall metabolic health has deteriorated and second, they signal a particular complication that the patient is progressing to. Given that existing studies have focused on a single or occasionally a few complications and modeled them independently, they have not separated these two roles. Hence, it is difficult to know whether a risk factor is significant in progression to a particular complication or whether it merely describes the deterioration of overall metabolic health. To understand the direction of progression, namely, which of the many possible complications the patient is most likely to develop next, separating these two roles is critical.

The deterioration of underlying metabolic health is a commonality across all the complications. If we identify the commonality and remove it from the entirety of a risk factor's effect, all that remains is outcome-specific effect. To model this, we had two challenges. First, to correctly capture the commonality, we needed to examine a wide range of complications using sufficient amounts of patient data. Because, if we study a single complication, the commonality is not identifiable, so the distinction is lost. In this study, we used two independent datasets. As the primary, we had EHR data (N=9,793) collected from the University of Minnesota Medical Center (UMMC) and used them for model training and internal validation. As the secondary, we had claims and EHR data from the OptumLabs Data Warehouse (OLDW) (N=72,720)[166] and used them for external validation. Because these datasets contained years of medical history of a large number of patients, they offered sufficient amounts of patient data and allowed us to examine multiple complications simultaneously.

Second, it was methodologically challenging to isolate the commonality from the entirety of a risk factor's effect because, it is not distinguishable. Multi-Task Learning (MTL) is a technique to model multiple related outcomes by exploiting their commonality[167,168]. In our case, modeling progression to each individual complication is a modeling *task*, and these tasks are related because deterioration in overall metabolic health underlies them all. We used MTL to integrate these tasks and identify the commonality among them. This approach is tantamount to applying MTL in reverse: rather than exploiting the commonality across the outcomes towards improved predictive performance, we discard the commonality to reveal *differential markers*, risk factors that are specific to each complication.

Considering that an accurate and timely prognosis for the patients often remains unsatisfactory[119] and there is limited evidence available for clinical decision support, our methodology that improves the interpretation of predictions and generates more understandable clinical information will help prioritizing the outcomes and developing optimal individualized T2DM management.

4.2 Methods and Materials

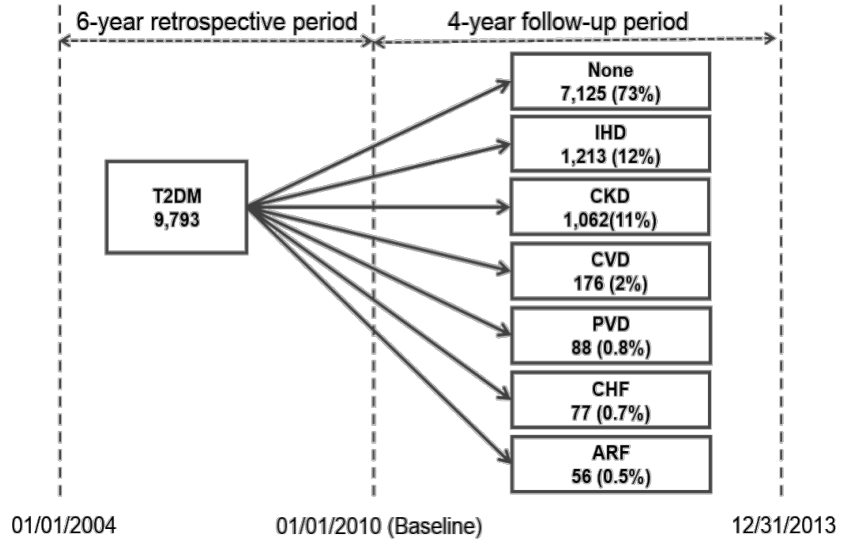
4.2.1 Primary Dataset for Training and Internal Validation

We used 10-year, de-identified EHR data (Jan 1, 2004-Dec 31, 2013) including inpatient, outpatient, and emergency department visits from the University of Minnesota Medical Center (UMMC), a main university hospital, located in Minneapolis, MN. From

the EHR data, we extracted patient demographics (age, gender), smoking status, vital signs (BP, pulse, and Body Mass Index BMI), lab results (HbA1c, lipid panel, Glomerular Filtration Rate GFR), three diagnoses comorbid to T2DM (dyslipidemia, hypertension, obesity), and six diagnoses of complications of interest: chronic kidney disease (CKD), acute renal failure (ARF), ischemic heart disease (IHD), congestive heart failure (CHF), peripheral vascular disease (PVD), and cerebrovascular disease (CVD). ARF is not usually associated with T2DM but involves organs or functions that are affected by T2DM. To demonstrate our proposed methodological validity, we intentionally included ARF as one of outcomes. We used 80% and 20% of UMMC data for model training and internal validation, respectively.

4.2.2 Study Design and Cohort Selection

We conducted retrospective cohort study. In UMMC data, we set up the study baseline at Jan. 1, 2010, collected patients' 6-year medical history to create baseline patient characteristics, and followed them from baseline to Dec. 31, 2013, determining whether or not they developed any complication of interest (Figure 4-1).



None: No complication is developed; IHD: Ischemic Heart Disease; CKD: Chronic Kidney Disease; CVD: Cerebrovascular Disease; PVD: Peripheral Vascular Disease; CHF: Congestive Heart Failure; ARF: Acute Renal Failure

Figure 4-1 Study Design

Initially, we identified 22,946 adult T2DM patients based on ICD-9 codes. These patients were at least 18 years old at baseline, and they were generally diagnosed with T2DM within the 6-year period. When patients develop multiple complications, the effects of risk factors become conflated. Thus, we excluded 8,979 patients who already developed any of the complications before baseline and 914 patients who developed multiple complications during the follow-up period. This was because we wanted to start with simple data without such conflated effects and achieve our goal of examining the feasibility of our methodology able to separate the two roles. We excluded 1,152 patients who had no HbA1c measurements at all, 1,611 patients who had no BP, pulse, or BMI

measurements at all, 494 patients who had no lipids information at all, and 3 patients without any known smoking status, resulting in 9,793 patients.

4.2.3 Second Dataset for External Validation

For external validation, we used claims and EHR data from the OptumLabs Data Warehouse (OLDW), which includes de-identified claims data for privately insured and Medicare Advantage enrollees in a large, private, U.S. health plan, as well as de-identified EHR data from a nationwide network of provider groups. The database contains longitudinal health information on enrollees, representing a diverse mixture of ages, ethnicities and geographical regions across the United States. The health plan provides comprehensive full insurance coverage for physician, hospital, and prescription drug services. The EHR data sourced from provider groups reflects all payers, including uninsured patients[166]. We extracted 10-year data (Jan 1, 2006-Dec 31, 2015) from the OLDW and identified 72,720 T2DM patients using the same study design (Figure 4-1) and selection procedure.

4.2.4 Baseline Patient Characteristics in UMMC and OLDW Datasets

Table 4-1 shows baseline patient characteristics in UMMC and OLDW datasets with variables available in this study. These variables represent risk factors and are used henceforth. UMMC patients had similar HbA1c but higher SBP and DBP compared to US adults with diabetes (HbA1c, SBP, and DBP are 7.2%, 131.5mmHg, and 69.4mmHg,

respectively.)[169], and they had signs of established CKD based on GFR[170]. Compared to UMMC patients, OLDW patients were older and had better HbA1c, better lipids, better kidney function, but higher SBP and DBP.

Table 4-1 Baseline Patient Characteristics in UMMC and OLDW Datasets

Variable	Description	UMMC (N=9,793)	OLDW (N=72,720)
male	Male	51	46
age	Age (years)	58±13	60±12
never_smoker	Non-smoker	56	45
a1c	HbA1C	7.2±1	7.0±1
ldl	LDL-cholesterol (mg/dL)	103±28	101±28
hdl	HDL-cholesterol (mg/dL)	44±12	46±12
trigl	Triglycerides (mg/dL)	172±90	169±117
tchol	Total-cholesterol (mg/dL)	181±34	179±34
gfr	Glomerular Filtration Rate ($ml/min/1.73m^2$)	58±32	76±27
gfr_norm	Normal Glomerular Filtration Rate	22	7
bmi	Body Mass Index (kg/m^2)	34±7	34±8
sbp	Systolic Blood Pressure (mmHg)	127±11	131±11
dbp	Diastolic Blood Pressure (mmHg)	75±7	77±7
pls	Pulse (bpm)	76±9	77±9
hyperlip	Hyperlipidemia	81	86
htn	Hypertension	71	81
obese	Obesity (BMI > 30)	70	67

4.2.5 Developing Multi-Task Learning Methodology

Under our hypothesis, each risk factor played two roles. The first role quantified the extent to which the patient's metabolic health had deteriorated, and the second role signaled which complication the patient was most likely to develop next. The first role was common across all complications (common effect), and the second role was specific to each complication (outcome-specific effect).

Formally, given a design matrix *X* that contained patients as rows and variables as columns, and *t* measuring time to event (complication or censoring), we simultaneously built the following six models, one for each complication *c*

$$D^c: \quad \lambda^c(t) = \lambda_0^c(t)\exp(\mathrm{X}\alpha)\exp(X\beta^c) \quad \text{subject to} \quad \|\alpha\|_1 \leq C_1 \text{ and } \|\beta^c\|_1 \leq C_2 \qquad \text{Eq. (1)}$$

where $\lambda^c(t)$ was the patient's hazard of developing complication *c* at time *t*, $\lambda_0^c(t)$ was a complication-specific baseline hazard, C_1 and C_2 were user-defined thresholds, chosen via cross validation, and $\|.\|_1$ denoted the L-1 norm (LASSO-penalty). X contained all variables in Table 4-1 except T2DM-comorbidities (hyperlipidemia, hypertension and obesity) since their defining factors (lab results and vital signs) were included.

Conceptually, each D^c model could be separated into two submodels as

$$\lambda^c(t) = \{\kappa_0(t)\exp(X\alpha)\}\{\kappa_0^c(t)\exp(X\beta^c)\} \qquad \text{Eq. (2)}$$

where the first submodel (with coefficients α) was a Cox model[171] capturing the common effects, and the second submodel (with coefficients β^c) was a Cox model capturing outcome-specific effects for each complication *c*. We called the first submodel General Progression Model and the second submodel Differential Progression Model.

Since these two models used the same set of variables, coefficients α and β^c were generally not identifiable (for each complication *c*, the effects of α and β^c were not distinguishable). While the first of the two LASSO constraints in Eq. (1) simply induced sparsity in General Progression Model with the purpose of performing variables

selection, the second LASSO-penalty made Differential Progression Models identifiable as it shrunk β^ccoefficients towards 0 forcing General Progression Model to explain as much of the variability as possible. We iteratively updated α and β^c coefficients until they were stabilized (squared differences of coefficients between previous and current iterations were effectively zero).

If the entirety of a variable's effect was only general deterioration of the metabolic health, its β^c would be exactly 0. Conversely, if $\beta^c > 0$, the variable increased the risk of complication c by β^c from α (harmful); and if $\beta^c < 0$, it decreased the risk of complication c by β^c from α (protective). Therefore, non-zero β^c coefficients identified differential markers; these were the risk factors that had effects beyond General Progression and enabled improved interpretation of progression to the most likely next complication.

4.2.6 Internal and External Validation

To determine the significance of α and β^c coefficients, we performed 1,000 permutation tests and calculated empirical p-values[172]. The key idea of permutation test was that variables were independent of randomly permuted labels; thus, coefficients of permuted labels were expected to have weaker associations than those of true labels. Then, the p-value of a coefficient could be calculated as the ratio of the number of permutation tests resulting in a stronger association to the total number of permutation tests.

We internally evaluated predictive performance of our models in 20% of UMMC data and externally evaluated it in OLDW data using concordance index (c-index), typically used to assess predictive performance of Cox models. In internal validation, we also performed 1,000 bootstrapping with sample size of 100% UMMC patients to obtain 95% confidence intervals (95CIs).

To demonstrate that we did not suffer a loss of performance due to our proposed MTL-based methodology, we compared predictive performance between ours and a reference methodology that built six independent models (LASSO-penalized Cox regression) for the six complications at a time.

4.3 Results

In this section, we are presenting results from our proposed methodology focusing on improved interpretation of risk factors and predictive performance in comparison with the reference methodology.

4.3.1 Coefficients from Multi-Task Learning Methodology

Figure 4-2 presents α and β^c coefficients from General Progression and Differential Progression Models. The rows are the variables. The first column corresponds to α coefficients from General Progression Model and the remaining columns correspond to β^c coefficients from Differential Models for each complication *c*. The interpretation of the coefficients is analogous to the regular Cox models: the exponent of a coefficient is the hazard ratio (HR) that the variable confers on the patient.

Association between variable and complication
Harmful Protective More important Less important

P-value
P < 0.001 P < 0.01 P < 0.05

Variable	Coefficient						
	General	CKD	ARF	IHD	PVD	CHF	CVD
a1c	0.0385	0.0407	0.0000	-0.0544	-0.1026	-0.0368	0.0000
ldl	0.0000	-0.0003	0.0000	0.0034	0.0000	0.0000	0.0000
hdl	-0.0023	0.0000	0.0000	-0.0010	0.0002	0.0277	0.0000
trigl	0.0007	0.0000	0.0000	0.0006	0.0003	-0.0032	0.0000
tchol	-0.0020	-0.0014	0.0000	-0.0026	0.0062	0.0018	0.0000
gfr	-0.0261	-0.0385	0.0000	0.0421	0.0614	0.0000	0.0000
gfr_norm	-2.3385	-3.4771	0.0000	3.5403	4.6379	0.6186	0.0000
bmi	0.0038	0.0059	0.0000	-0.0076	-0.0079	0.0463	0.0000
pls	0.0002	0.0000	0.0000	-0.0024	0.0037	0.0324	0.0000
sbp	0.0044	0.0033	0.0000	-0.0151	0.0146	0.0256	0.0000
dbp	-0.0083	0.0000	0.0000	0.0113	-0.0515	0.0000	0.0000
never_smoker	-0.1401	0.0626	-0.1632	-0.0482	-0.3956	-0.3487	0.0000
age	0.0137	0.0000	0.0000	-0.0043	0.0408	0.0576	0.0188
male	0.0292	0.0564	0.0000	-0.1467	0.5626	-0.2188	0.0000

Figure 4-2 Coefficients from Multi-Task Learning Methodology

For most variables (e.g., HbA1c, LDL) which higher values are associated with higher risks, if $\alpha > 0$, it indicates a harmful association; and if $\alpha < 0$, it indicates a protective association with General Progression. In Differential Progression, if $\beta^c > 0$, the variable increases the risk of complication c by β^c from α, making the variable more important (harmful effect becomes larger); and if $\beta^c < 0$, the variable decreases the risk of complication c by β^c from α, making the variable less important (harmful effect becomes smaller). There are also variables (HDL, GFR, normal GFR, and never smoker) which higher values are associated with lower risks; thus, the interpretation of their coefficients is opposite. For example, if α of GFR > 0, it means that higher GFR is protective of General Progression; if β^c of GFR > 0, it indicates that higher GFR is more important in progression to complication c (prospective effect becomes larger).

To help detecting significant α and β^c coefficients, we visualized associations between variables and complications (harmful, protective, more important, and less important) and p-values (Figure 4-2). Coefficients with a circle are statistically significant, and those without are insignificant. Larger circles indicate smaller p-values (more significant). The exact p-values can be found in appendix Table A-1.

As expected, most variables significantly predicted General Progression: HbA1c, triglycerides, total cholesterol, GFR, normal GFR, BMI, SBP, DBP, non-smoker, and age (Figure 4-2). Traditionally, higher DBP is known to be harmful. But, several recent studies showed that DBP was protective of cardiovascular disease especially for older adults[75]. We also found that DBP is protective of General Progression. General Progression Model was a latent model in the sense that it did not have an observable outcome; it described the extent of deterioration in overall metabolic health. These variables of General Progression were those that many studies found to be significantly associated with an increased risk of micro and macrovascular complications and all-cause mortality[108,173].

4.3.2 What is General Progression Model?

We defined General Progression mathematically as the effects that were common across all the complications and explained that General Progression captured deteriorating overall metabolic health. As an alternative, we interpreted α coefficients as the log HR of progression to *any* complication. To illustrate this, we built a LASSO-penalized Cox model that predicted the development of *any* complication (this model had

an event if a patient developed *any* complication) and compared coefficients from this model (Figure 4-3) with α coefficients from General Progression Model (Figure 4-2). We found that they were similar with respect to effect size, sign, and significance, and this suggested that General Progression could be indicative of progression to *any* complication.

Variable	Coefficient
a1c	0.0305
ldl	0.0000
hdl	-0.0023
trigl	0.0009
tchol	-0.0030
gfr	-0.0279
gfr_norm	-2.5260
bmi	0.0065
pls	-0.0017
sbp	0.0040
dbp	-0.0091
never_smoker	-0.1259
age	0.0154
male	0.0160

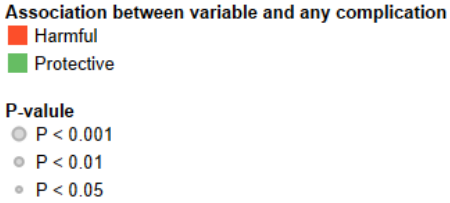


Figure 4-3 Coefficients for Risk of Development *Any* Complication

4.3.3 Interpretation of Coefficients from Multi-Task Learning Methodology

After achieving the overarching goal of our proposed methodology to separate common effects (α) and outcome-specific effects (β^c) of a risk factor, we examined if results and their interpretations from our models clinically made sense. Especially, we wanted to have some of them consistent with known facts because if they were not, the utility of our methodology could be in doubt. To demonstrate this, let us consider the role of HbA1c in progression to CKD and IHD as an example as it is commonly accepted

facts in practice: hyperglycemia is a key driver of microvascular complications (e.g., CKD), while dyslipidemia is a key driver of macrovascular complications (e.g., IHD)[101].

General Progression showed that a unit increase in HbA1c conferred a HR of 1.039 (exp(.0385)) on all complications uniformly (Figure 4-2, row1, column1). However, higher levels of HbA1c ultimately affect the different complications differently. As mentioned, it is well-known that HbA1c is more predictive of CKD than IHD. Indeed, Differential Progression for CKD showed that a unit increase in HbA1c conferred an additional log HR of .0407 on patients, increasing the HR of CKD from 1.039 to 1.0824 (exp(.0385+.0407) (Figure 4-2, row1, column2).

It is also known that HbA1c is not as important in IHD as in CKD. Differential Progression for IHD showed that patients with higher HbA1c tended to suffer other (microvascular) complications. The log HR of IHD that a unit increase in HbA1c conferred on patients was negative, which decreased the HR of IHD from 1.039 to 0.9842 (exp(.0385-.0544)) (Figure 4-2, row1, column4). What it means is that patients with higher HbA1c are more likely to progress to a complication than patients with lower HbA1c, and that complication is less likely to be IHD but more likely to be a microvascular complication such as CKD. That is, General Progression described the patient's tendency to progress to a complication, and the Differential Progression helped to target which complication the patient is more likely to develop next.

4.3.4 Differential Markers of CKD, IHD, PVD and CHF

To easily detect distinguishing patterns of differential markers, we visualized each of CKD, IHD, PVD and CHF as a series of spider plots[174] in Figure 4-4. In a spider plot, variables are arranged as axes extending radially from a central point, and each observation makes a closed polygon connecting points on all of the axes. Emphasis is upon discerning the characteristic shapes of these polygons among observations, rather than extracting specific values.

The interpretation of our plots is as follows. Each plot corresponds to a complication. Ten variables for vital signs and lab results construct individual axes, radially arranged around a center point. The β^c cofficient of each variable is depicted by an anchor (node) on an axis. As higher values of HDL, GFR, and DBP are protective, the sign of coefficients of them are reversed only for visualization purposes. The same color encoding was used to identify significantly more or less important differential markers.

For each variable, distance from the center indicates an increased risk. In each plot, a navy line connecting the β^c coefficients represents Differential Progression, while a green line connecting zero on each axis conceptually represents General Progression, a reference for Differential Progression. By comparing these two lines on each axis, differential risk for complication *c* beyond or below General Progression is easily distinguished.

As we focused on straightforward interpretation, we did not perform normalization; thus, the scales of the variables are not comparable with each other. What

is important is whether the navy line (Differential Progression) is outside or inside the green line (General Progression) on each axis.

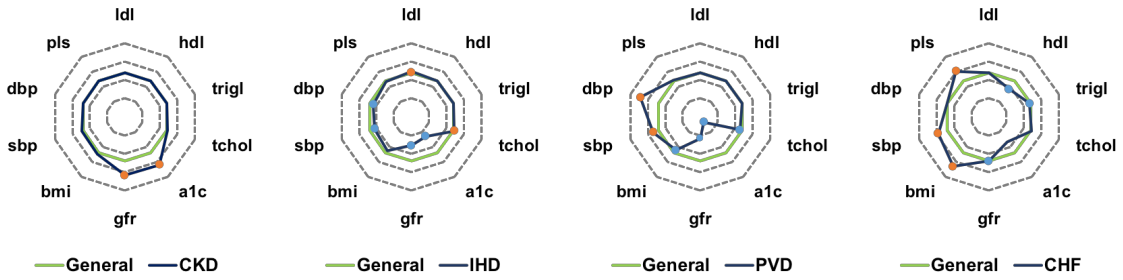


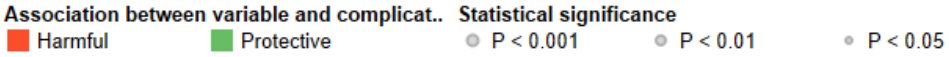
Figure 4-4 Characteristic Shapes of Differential Markers for CKD, IHD, PVD, CHF

IHD, PVD, and CHF are well-known concomitant macrovascular complications. They share similar pathophysiology and are believed to have similar risk factors[85]. Given these facts, distinguishing among them without a methodology like ours is more difficult. In Figure 4-4, spider plots show distinguishing patterns of differential markers among these very similar diseases. In progression to IHD, LDL was more important; SBP, and lower DBP were less important[75]. In progression to PVD, SBP and lower DBP were more important; BMI was less important. In progression to CHF, lipid abnormalities were less important; BMI, pulse, and SBP were more important (irregular or fast pulse is one of the symptoms of CHF).

4.3.5 Coefficients from Reference Methodology

Figure 4-5 shows coefficients from reference models. The rows are the variables, and the columns are complications. If a coefficient > 0, it indicates a harmful association; and if a coefficient < 0, it indicates a protective association with a complication *c*. In

reference models, the two roles of a variable (α and β^c coefficients) were not identifiable; thus, only the entirely of a variable's effect was estimated, and the outcome-specific effect was masked. This was the motivation of our study and the key difference from our proposed methodology. To help detecting significant coefficients, we visualized the association between variables and complications (harmful and protective) and p-values. The exact p-values can be found in appendix Table A-2.



Variable	Coefficient					
	CKD	ARF	IHD	PVD	CHF	CVD
a1c	0.0511	0.1938	-0.0161	-0.0630	-0.0043	0.0799
ldl	-0.0033	-0.0264	0.0034	0.0011	0.0000	0.0034
hdl	-0.0038	-0.0113	-0.0032	0.0000	0.0249	-0.0083
trigl	0.0000	-0.0019	0.0013	0.0013	-0.0026	0.0000
tchol	0.0000	0.0180	-0.0046	0.0030	0.0000	0.0000
gfr	-0.0539	-0.0280	0.0161	0.0351	-0.0048	0.0086
gfr_norm	-4.9076	-3.2846	1.2006	2.2757	0.0000	0.4309
bmi	0.0050	-0.0041	-0.0038	-0.0039	0.0539	0.0006
pls	0.0000	0.0213	-0.0022	0.0038	0.0323	-0.0045
sbp	0.0053	0.0123	-0.0107	0.0190	0.0290	0.0119
dbp	-0.0002	-0.0212	0.0030	-0.0597	-0.0006	-0.0003
never_smoker	0.0000	-0.7203	-0.1885	-0.5354	-0.5033	-0.0472
age	0.0140	0.0190	0.0092	0.0543	0.0834	0.0658
male	0.0000	0.4370	-0.1178	0.5967	-0.2899	0.1543

Figure 4-5 Coefficients from Baseline Methodology

4.3.6 Utility of Our Proposed Multi-Task-Learning Methodology

To demonstrate clinical utility of our methodology in comparison with reference methodology, let us take ARF as an example. Although a major cause of ARF is not diabetes, reference models identified virtually *all the variables* to be predictive of ARF (Figure 4-5). While, Differential Progression Model for ARF showed that progression to

ARF was only associated with the underlying advanced metabolic deterioration (General Progression), and all the variables were not specific to ARF (Figure 4-2).

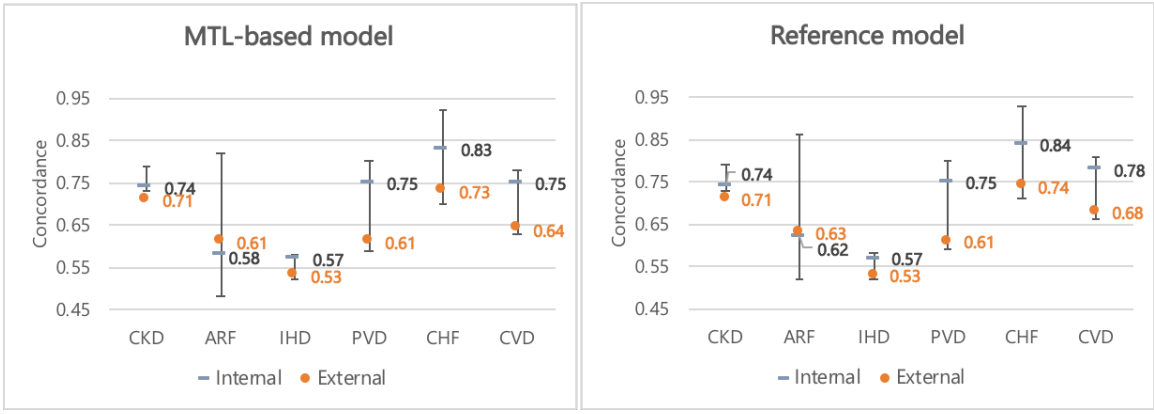
Another example is CKD. Risk factors of CKD are well-understood. The reference model for CKD identified HbA1c (barely), GFR and age as significant risk factors, and they are indeed known risk factors. In fact, reference models identified age as a risk a factor *for every complication*; however, it is not that a patient is more likely to develop CKD just because he is older. Whereas, General Progression Model and Differential Progression Model for CKD suggested that older patients were more likely to have their metabolic health deteriorated than younger patients; and, age played no role in progression to CKD beyond General Progression.

4.3.7 Internal and External Validation

Table 2 presents predictive performance in C-Index of our MTL-based models and reference models. Generally, they achieved similar predictive performance. Minimal albeit statistically significant differences were only observed in complications with small number of progressing patients.

For both, predictive performance was lower in external validation. UMMC data consisted of smaller number of patients from one healthcare system. Thus, they might be less representative of T2DM population than OLDW patients, or they might be subpopulation of OLDW patients. Also, patient characteristics differed fundamentally between them (Table 4-1). However, except CKD, C-Indices were still within 95CIs.

Table 4-2 Predictive Performance in C-Index (95CIs)



Dataset	Methodology	CKD	ARF	IHD	PVD	CHF	CVD
Internal (UMMC)	MTL	.74 (.73-.79)	.58 (.48-.82)	.57 (.52-.58)	.75 (.59-.80)	.83 (.70-.91)	.75 (.63-.78)
	Reference	.74 (.73-.79)	.62 (.48-.79)	.57 (.52-.58)	.75 (.60-.81)	.84 (.67-.91)	.78 (.65-.80)
External (OLDW)	MTL	.71	.61	.53	.61	.73	.64
	Reference	.71	.63	.53	.61	.74	.68

4.4 Discussion

Given that the effect of deteriorating overall metabolic health is common across all the complications, we hypothesized each risk factor had two roles: describing the extent of deteriorating overall metabolic health and signaling a particular complication the patient is progressing towards. We have successfully demonstrated that our proposed methodology separated these two roles of risk factors and revealed distinguishing patterns of differential markers. Also, we modeled multiple complications simultaneously by sharing their information; thereby, generating systematic and comprehensive interpretation of different roles of risk factors among various complications.

Our study has important strengths. First, we made more understandable predictions for clinicians by focusing on improved interpretability. Usually, high predictive accuracy is of key importance in prediction models. However, lack of clarity in the interpretation of predictions limits their usefulness in practice. Second, we externally evaluated predictive performance of our models, which were rarely done in other studies. Although the reference model did well or slightly better than our proposed model, the difference was minimal. So, we would say that we did not compromise predictive performance due to our proposed methodology. Third, our methodology is of high utility because it can be applied to other clinical conditions in which comorbidities matter.

We have several limitations. When identifying cohorts, we excluded patients who developed multiple complications. Although this action limits the generalizability of our work, our primary interest was to demonstrate the feasibility and utility of our proposed methodology. Additionally, we excluded patients with unknown vital signs, lab results, and/or smoking status. We tested differences between final study cohort and these excluded patients. All variables except hdl, tchol, dbp, never_smoker, and age were significantly different, and the excluded patients were generally sicker. Thus, our study is subject to selection bias. But, ours was not to estimate the effect of a risk factor, in which addressing selection bias using imputation methods is critical, but to separate the entirety of the effect into common and outcome-specific effects. Lastly, we used variables easily obtained from EHR data. Many studies have found that T2DM is disproportionally affected by race, ethnicity and/or socioeconomic status[175]. Although they are very important risk factors, we mainly focused on modifiable risk factors.

When we build a model on large amounts of data, most variables become statistically significant; however, they may be clinically irrelevant. To impact individualized patient care, it is critical to develop enabling technologies that extract clinically useful information from the large amounts of data.

Our future work is to extend this work to larger cohorts and overcome the limitations. If we can obtain reasonable levels of generalizability, we believe that our methodology will have significant potential to help clinicians prioritizing outcomes and making a more accurate prognosis for T2DM patients.

Chapter 5 Towards More Accessible Precision Medicine: Building a More Transferable Machine Learning Model to Support Prognostic Decisions for Micro- and Macrovascular Complications of Type 2 Diabetes Mellitus

5.1 Introduction

Machine learning holds great promise for precision medicine, providing a new way to generate evidence that enhances clinical practice guidelines with more personalized recommendations[8]. One disease that would immensely benefit from this approach is Type 2 Diabetes Mellitus (T2DM), a complex chronic disease that requires multifactorial risk-reduction strategies to prevent and manage clinically significant micro- and macrovascular complications[176]. In recent years, machine learning models have been increasingly developed for clinical decision support for T2DM patients, yet the adoption of these models into clinical practice remains limited[17,19,20,177].

Most of these machine learning models are learned from a population in a single healthcare system, which narrows their applicability to that particular population. Machine learning models are developed on a *training* (or development) cohort and are expected to be applicable to the population from which the training cohort was drawn. If the model performs well on the population, the model is said to *generalize* well to that population; if it does not, the model is said to *overfit* the training cohort. *Internal validation* ensures that the model generalizes from the training cohort to the population

from which the training cohort was drawn. Most models today are trained on a cohort from a single healthcare system, are internally validated, and thus generalize well to the patients in the population of that healthcare system. However, when these models encounter patients who are not typical for that healthcare system but may be common in other parts of the nation, these models invariably fail. The uncertainty caused by the fact that the model may fail for a (possibly atypical) patient erodes the applicability of the model to clinical decision support, hindering the adoption of machine learning in clinical practice.

In observational studies, external validation is considered the strongest evidence of the generalizability of a model. A model is said to be *transferable* to a different healthcare system if the model generalizes to the patient population in the target healthcare system. Thus, *external validation* ensures not only generalizability but also transferability. If the models were transferable, they would reliably apply to a much greater variety of patients, reducing the above uncertainty. Today's models are rarely transferable; in fact, it has been shown that even the simplest decision-tree-based models require retraining when they are applied to patients in a different healthcare system[178]. Given the expectation that complex machine learning models will not be transferable, they are rarely externally validated[144–146].

In this paper, we aim to (1) demonstrate that even a complex machine learning model built on a nationally representative sample can be transferred to two local healthcare systems, (2) while a model constructed on a local healthcare system's data will

be difficult to transfer to a different healthcare system; and (3) we discuss criteria for external validity.

5.2 Methods and Materials

5.2.1 Dataset

We used three, independent, retrospectively collected datasets. The first dataset was nationally representative. It included 10-year claims and EHR data (Jan 1, 2006-Dec 31, 2015) from the OptumLabs® Data Warehouse (OLDW)[179,180], a database which includes de-identified claims data for privately insured and Medicare Advantage enrollees in a large, private, U.S. health plan, as well as de-identified EHR data from a nationwide network of provider groups. The database contains longitudinal health information on enrollees, representing a diverse mixture of ages, ethnicities and geographical regions across the United States. The health plan provides comprehensive full insurance coverage for physician, hospital, and prescription drug services. The EHR data sourced from provider groups reflects all payers, including uninsured patients (N=951,793 patients diagnosed with T2DM before Jan 1, 2011, an index date).

The second and third datasets were from local healthcare systems. The second dataset was 8.5-year EHR data (Jan 1, 2008-Jun 30, 2016) from the University of Minnesota Medical Center (UMMC) in Minneapolis, MN (N=12,797 patients diagnosed with T2DM before Jun 1, 2011), a local healthcare system. The third dataset was 8-year EHR data (Jan 1, 2007-Dec 31, 2014) from the Mayo Clinic, Rochester (MCR), MN (N=5,479 patients diagnosed with T2DM before Jan 1, 2010), another local healthcare

system. All three datasets contained patients' demographics, smoking status, diagnoses, lab results, vital signs including BMI, and prescription drug information.

5.2.2 Cohort selection and study design

The cohort selection criteria used to identify study cohorts is shown in Table 5-1, which resulted in 81,091 OLDW, 8,091 UMMC, and 2,247 MCR primary care patients with T2DM.

This study was a retrospective cohort study. We established baseline characteristics of the OLDW cohort using 5 years of medical history (4 years for the UMMC and MCR cohorts) before the index date. From the index date, we followed the cohort for 5 years, determining their outcomes. When patients developed multiple complications during the 5-year follow-up period, we censored them after their first event occurred, so that we focused on the most likely complication that individual patients could develop next.

Table 5-1 Cohort selection criteria (OLDW cohort)

Description	Patient (N)	(%)
Inclusion		
Patients w/ at least two encounters w/ diagnosis or history of T2DM during the 2-years before the index date (Jan 1, 2011)	915,739	100
Exclusion		
Patients w/ ICD-9 codes for T1DM, gestational diabetes, secondary diabetes, poisoning by hormones and synthetic substitutes, or other specified disorder of pancreatic internal secretion (source: HEDIS 2015 Diabetes Exclusions Value Set).	-208,938=706,801	77
Patients w/ multiple DOBs or sex.	-22,037=684,764	75
Patients whose age < 18 at as of the index date	-178,581=506,183	55
Patients w/o any record since the index date.	-36,359=469,824	51
Patients w/o any HbA1c measurements at all before the index date[a].	-129,813=340,011	37

Patients w/o any SBP and DBP measurements at all before the index date[b].	-28,795=311,216	34
Patients w/o any pulse measurements at all before the index date.	-65,150=246,066	27
Patients w/o any BMI measurements at all before the index date.	-41,920=204,146	22
Patients w/o any known smoking status before the index date.	-30,576=173,570	19
Patients w/ ICD-9 codes for hyperlipidemia but no LDL, HDL, and Triglycerides measurements at all before the index date.	-8,504=165,066	18
Patients w/ ICD-9 codes for chronic kidney disease or chronic renal failure but no creatinine and GFR measurement at all before the index date.		
Patients w/ indication of HTN (i.e. ICD-9 codes for HTN or abnormal SBP and DBP) but no drug records at all before the index date.	-8,664=156,402	17
Patients w/ indication of HLD (i.e., ICD9 codes for HLD or abnormal lipids) but no drug records at all before the index date.		
Patients w/ shorter than 4-year follow-up[c].	-31,075=125,327	14
Patients whose minimum gap of two adjacent HbA1c measurements was larger than the 3-years before the index date[d].	-44,236= 81,091	9
Patients whose maximum gap of two adjacent HbA1c measurements was less than the 1-year before the index date[d].		

Abbreviations: T2DM, Type 2 Diabetes Mellitus; T2DM, Type 1 Diabetes Mellitus; DOB, date of birth; SBP, Systolic Blood Pressure; DBP, Diastolic Blood Pressure; BMI, Body Mass Index; LDL, Low Density Lipoprotein; HDL, High Density Lipoprotein; GFR, Glomerular Filtration Rate; HTN, Hypertension; HLD, Hyperlipidemia;

[a] If patients are diabetic, their HbA1c should be measured regularly.

[b] If patients are diabetic, their blood pressure should be measured at every routine visit.

[c] Although survival analysis effectively handles the censoring of patients, unequal follow-up among different complications may produce biased results. Also, since complications of T2DM develop over many years, we preferred observing a patient for at least 4 years. We found that the excluded 31,075 patients had more pre-existing complications, worse lab results, and worse vital signs (long-standing T2DM and/or less controlled health conditions) than the remaining 126,327 patients.

[d] In an effort to identify active primary care patients, we loosely applied recommendations for the HbA1c test[181]. Specifically, we considered a diabetic patient as a primary care patient if he had at least two HbA1c tests with a maximum gap of at most 3-years, and a minimum gap of at least 1-year during the baseline period. The excluded 44,236 patients were not different from the remaining 81,091 patients in terms of mean baseline characteristics and outcome rates. This exclusion only reduced the variance of baseline characteristics.

5.2.3 Outcomes

We considered six outcomes relevant to T2DM: ischemic heart disease (IHD), congestive heart failure (CHF), cerebrovascular disease (CVD), peripheral vascular disease (PVD), chronic kidney disease, with CKD representing earlier stages of chronic kidney disease and chronic renal failure (CRF) representing advanced chronic kidney disease or end-stage renal disease. Although CKD and CRF are generally considered the same clinical entity, the patient can take many years to progress from the earlier to the more advanced stages of CKD; thus, we defined them as separate outcomes. Outcomes were determined using billing diagnoses listed in Table 5-2.

Table 5-2 ICD9-codes of outcomes (micro- and macro complications of T2DM

Outcome	ICD-9 Code
Ischemic Heart Disease (IHD)	410.*, 411.*, 412, 413.*, 414.0x, 414.2, 414.3, 414.4, 414.8, 414.9, V45.81, V45.82
Congestive Heart Failure (CHF)	428.0, 428.1, 428.2, 428.3, 428.4, 428.9
Cerebrovascular Disease (CVD)	431, 432.9, 436, 437.0, 437.1, 437.8, 437.9, V12.54, 433.*, 434.*, 435.*, 438.*
Peripheral Vascular Disease (PVD)	440.2x, 440.4, 443.9, 445.89, 444.2x, 444.8x, 445.0x, 557.*
Chronic Kidney Disease (CKD)	585.1, 585.2, 585.3, 585.9
Chronic Renal Failure (CRF)	585.4, 585.5, 585.6, 586

5.2.4 Overview of model development and validation

Available variables and measurement methods could differ across healthcare systems. For example, HbA1c was only available in the UMMC dataset, while fasting plasma glucose (FPG) was only available in the MCR dataset. To externally validate our model against these two local datasets, we built model variants based on the different

datasets and measurement methods: $variant_{OL.A1c}$, $variant_{OL.FPG}$, $variant_{UMMC}$, and $variant_{MCR}$ (Figure 5-1). For example, the $variant_{OL.A1c}$ is built on data from OptumLabs Data Warehouse (OLDW) using HbA1c. Due to the above availability constraints, the UMMC variant uses HbA1c and the MCR variant uses FPG.

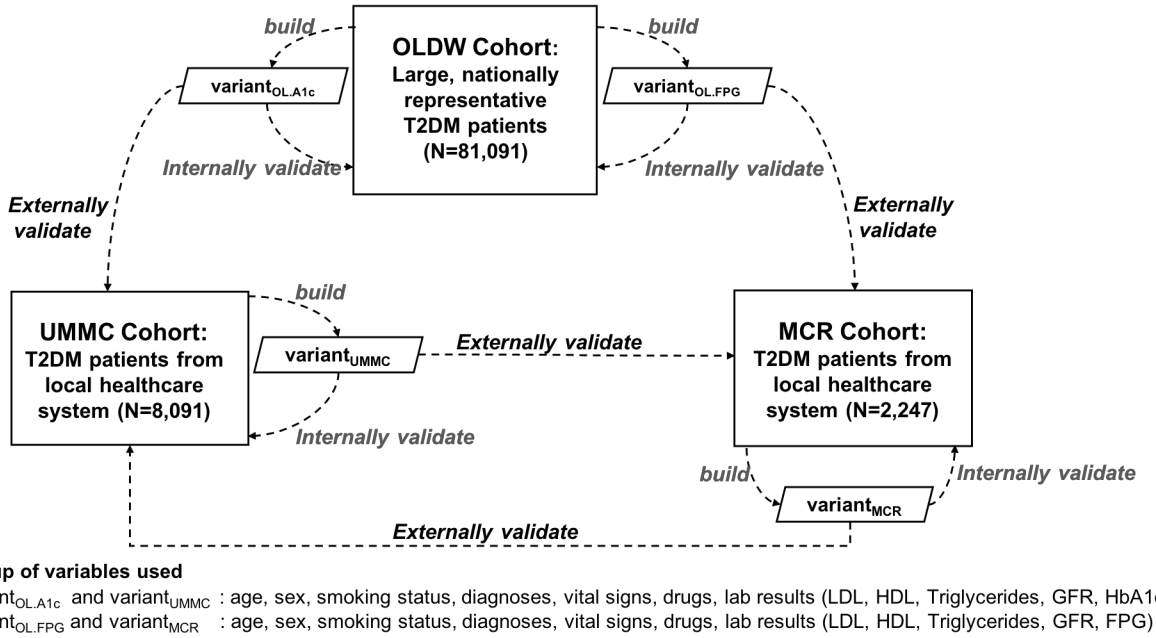


Figure 5-1 Overview of model development and validation

The $variant_{OL.A1c}$ and $variant_{OL.FPG}$ were built on 80% of the nationally representative OLDW cohort, internally validated against the remaining 20% of the OLDW cohort, and externally validated against the entire UMMC and MCR cohorts, respectively. The $variant_{UMMC}$ was built on 80% of the local UMMC cohort, internally validated against the remaining 20%, and externally validated against the entire MCR cohort. Likewise, the $variant_{MCR}$ was built on 80% of local MCR cohort, internally

validated against the remaining 20%, and externally validated against the entire UMMC cohort. We used the concordance index (C-index)[182] as a performance measure and followed [183] to convert between HbA1c and FPG.

5.2.5 Internal Validation

We calculated the C-index for each model variant and complication pair on its internal validation cohort (20% cohort). To construct 95% confidence intervals (95CIs) of the C-index, we performed bootstrapping on its entire cohort with 500 iterations. In each iteration, we built the model variant on the bootstrap sample and calculated the C-index for each of the outcomes on the Out-of-bag (OOB) sample (which was not used for building the model). Using the C-indices from the 500 OOB samples, we finally constructed 95CIs for each outcome. Validation was considered successful if the C-index of the model variant on the internal validation cohort fell into the 95CIs.

5.2.6 External Validation

In external validation, we examined whether a model variant built on one cohort was *validated* on a different (external) cohort. We defined successful validation as (i) the model variant achieved non-random C-index (C-index is statistically significantly different from .5), (ii) the C-index of the model variant on the external cohort fell into the 95CI of the training C-index (computed during internal validation), and (iii) the significant coefficients in the model variant being validated had the same effect directions as a model variant built on the external cohort.

5.2.7 Modeling Method

The model variants we evaluated in this study were developed using our previously published Multi-Task Learning (MTL)-based methodology[184]. This methodology learns the six outcomes simultaneously by extracting a variance component that is common across all outcomes and isolating variance components that are specific to the individual complications. We call the common component the *General Progression* model and the individual components as *Differential Progression* models. The result of the procedure is seven Cox proportional hazard regression models: one for the General Progression and one for each of the six outcomes. We will discuss the interpretation of the coefficients in the results section.

We deliberately chose a relatively complex model with a latent outcome (the General Progression) to demonstrate that even models of this complexity could be externally validated.

5.3 Results

5.3.1 Baseline characteristics of the OLDW, UMMC, and MCR cohorts

Table 5-3 shows baseline characteristics of the OLDW, UMMC, and MCR cohorts. Univariate cohort differences were tested using ANOVA for continuous variables or Chi-square tests for categorical variables. Although p-values were adjusted by the Bonferroni correction, significant differences among the cohorts were found in most variables. Significant coefficients (Bonferroni corrected $P \leq .001$) are marked with an asterisk '*'.

Table 5-3 Baseline characteristics of the OLDW, UMMC, and MCR cohorts

Variable	OLDW (N=81,091)	UMMC (N=8,091)	MCR (N=2,247)	P-value
Age (year)	60.4±9.7[a]	62.6±11.3	61.9±12.9	<.001*
Male (%)	48	49	51	.006
Census Region (%)				
South (reference)	43			N/A
Midwest	41	100	100	N/A
Northeast	6			N/A
West	8			N/A
Unknown	2			N/A
Smoking Status (%)				
Never Smoker	22	43	12	<.001*
Former Smoker	61	25	71	<.001*
Current Smoker	17	32	17	<.001*
Lab Results				
HbA_{1c} (%)	7.1±1.2	6.9±1.1	N/A	1.000
FPG (mg/dl)	116.34±43.0	N/A	131.9±19.0	<.001*
LDL (mg/dl)	96.4±29.2	96.6±28.7	103.9±23.9	0.145
HDL (mg/dl)	44.7±12.4	42.9±12.3	43.9±10.6	<.001*
Triglycerides (mg/dl)	173.7±118.2	178.7±120.5	199.3±98.5	<.001*
Missing LDL	0.6	1.4	0.0	<.001*
Missing HDL	0.6	1.5	0.0	<.001*
Missing Triglycerides	0.6	1.3	0.0	<.001*
GFR (ml/min/1.73m^2)	93.1±20.6	74.1±16.5	54.8±14.3	<.001*
Normal GFR (%)[b]	0.4	0.4	4	<.001*
Vital Signs				
BMI (kg/m^2)	34.7±7.5	34.0±7.0	34.1±7.0	<.001*
SBP (mmHg)	130.1±11.3	127.1±10.3	133.4±12.4	<.001*
DBP (mmHg)	75.8±7.2	73.8±6.8	73.4±8.6	<.001*
Pulse (bpm)	76.7±9.0	75.7±8.8	77.7±9.7	<.001*
Pre-Existing Complications[c]				
IHD	22	22	37	<.001*
CHF	7	8	5	<.001*
CVD	11	10	14	<.001*
PVD	8	7	20	<.001*
CKD	9	15	18	<.001*
CRF	3	2	8	<.001*
Severity of HTN, HLD, and DM (%)				
HTN				
No indication of HTN (No Dx, normal SBP and DBP, and no HTN drug)	7	8	8	.008

Untreated HTN (presence of Dx, or abnormal SBP or abnormal DBP but HTN drug)	5	5	5	.655
At most two HTN drugs	49	44	58	<.001*
At least three HTN drugs and controlled HTN (normal SBP and DBP)	28	36	17	<.001*
At least three HTN drugs but uncontrolled HTN (abnormal SBP or DBP)	11	7	12	<.001*
HLD				
Cholesterol drug use	78	87	83	<.001*
DM				
Metformin only	33	41	25	<.001*
Monotherapy except Metformin or combination therapy without insulin	37	32	35	<.001*
Insulin	30	26	40	<.001*
Drug (%)				
Aspirin	29	38	NA	
HTN Drug				
ACEI or ARB	76	76	79	.009
Diuretic	55	57	13	<.001*
Beta Blocker	43	53	54	<.001*
Calcium Channel Blocker	29	29	31	.062
Other	6	5	3	<.001*
Cholesterol Drug				
Statin	76	86	79	<.001*
Fibrate	15	18	17	<.001*
Other	7	24	15	<.001*
Diabetes Drug				
Metformin	72	69	71	<.001*
Alpha-Glucosidase Inhibitor	0.4	0.4	0	.408
Amylin	0.4	0.3	0	.750
Meglitinide	1	2	2	.001
Incretin	7	7	4	<.001*
Sulfonylurea	41	41	52	<.001*
Thiazolidinedione	25	26	21	<.001*
DDP4 Inhibitor	14	5	5	<.001*
Insulin	30	26	39	<.001*
Other	2	1	1	<.001*
Drug Missingness (%)				

No HTN, Cholesterol and DM drug info	0.1	0.1	0.1	.062

Abbreviations: FPG, Fasting Plasma Glucose; LDL, Low Density Lipoprotein; HDL, High Density Lipoprotein; GFR, Glomerular Filtration Rate; BMI, Body Mass Index; SBP, Systolic Blood Pressure; DBP, Diastolic Blood Pressure; IHD, Ischemic Heart Disease; CHF, Congestive Heart Failure; CVD, Cerebrovascular Disease; PVD, Peripheral Vascular Disease Chronic Kidney Disease; CRF, Chronic Renal Failure; HTN, Hypertension; HLD, Hyperlipidemia; DM, Diabetes; Dx, Diagnosis; ACEI, Angiotensin-Converting Enzyme Inhibitor; ARB, Angiotensin Receptor Blocker; DDP4, Dipeptidyl Peptidase 4.

[a] Data are presented as mean ± standard deviation unless otherwise indicated.

[b] N/A indicates not applicable

[c] This indicates % patients with no Dx of CKD, creatinine or GFR measurements at all prior to index date. We considered them to have normal kidney function. We imputed zero for these patients' GFR.

[d] When patients had already presented with outcome complication(s) on the index date, we called them pre-existing complications.

5.3.2 Cumulative hazard curves

Figure 5-2 presents the cumulative probabilities for developing each of the complications in the OLDW (median follow-up: 4.95-year), UMMC (median follow-up:4.83-year), and MCR (median follow-up:4.81) cohorts. As for major differences, the UMMC cohort showed higher incidence of CKD and IHD; and the MCR cohort showed higher incidence of CKD, PVD, and CHF, compared to the OLDW cohort.

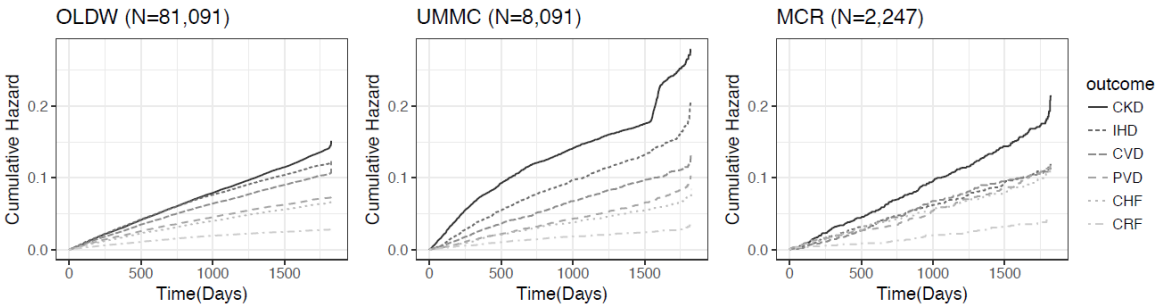


Figure 5-2 Kaplan-Meier curves of the Nelson-Aalen Estimator for the OLDW, UMMC, and MCR cohort

5.3.3 Description of the Model

Table 5-4 presents coefficients in log hazard ratio (HR) resulting from $variant_{OL.A1c}$. The first column corresponds to the General Progression model and the remaining columns correspond to the Differential Progression models. Significant coefficients (Bonferroni corrected $P \leq .001$) are in the colored cells (red for harmful, blue for protective, orange for more important, and light blue for less important) . To show how to interpret them, let us consider HbA1c. HbA1c increases the risk of progressing to any complication by a HR of 1.08(=exp(0.0740)) (Table 4, row 4, column 1), and it further increases the risk of progressing to CKD by a HR of 1.03(=exp(0.0360)) (Table 4, row 4, column 6). Most effect directions coincide with what is reported in the relevant literature (RCTs and prospective cohort studies)[61,63,72,73,75,185]. A more detailed description of how to interpret the coefficients can be found in our previous work[184].

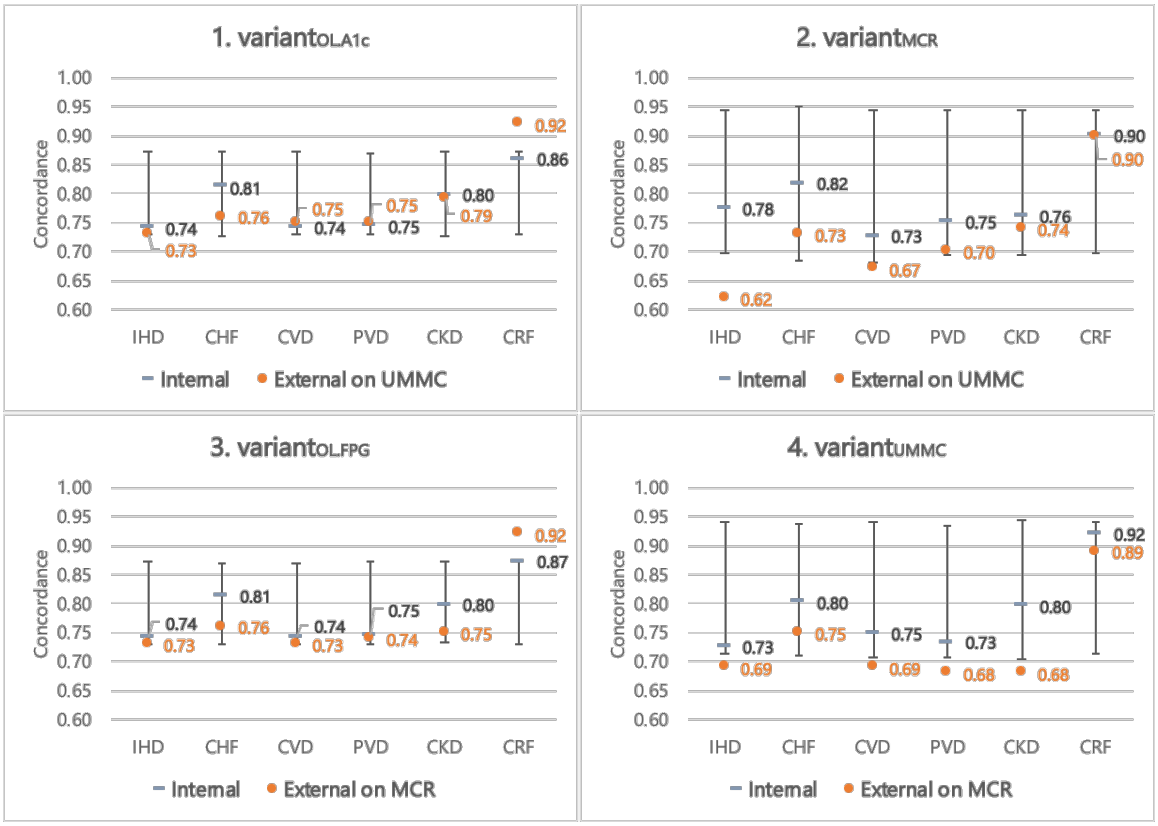
Table 5-4 Coefficients in log hazard ratio from variant$_{OL.A1c}$

Variable	Coefficient						
	General	CKD	CRF	IHD	CHF	CVD	PVD
Low-Density Lipoprotein (LDL)	0.0010	-0.0030	0.0000	0.0020	0.0000	0.0020	0.0000
High-Density Lipoprotein (HDL)	-0.0060	0.0070	0.0060	-0.0090	-0.0010	-0.0010	0.0000
Triglycerides	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
HbA1c	0.0740	0.0360	0.0090	-0.0200	0.0100	-0.0540	0.0220
Glomerular Filtration Rate (GFR)	-0.0200	-0.0270	-0.0280	0.0170	0.0090	0.0170	0.0160
Normal GFR	-2.0910	-2.8310	-1.6150	1.8460	0.9550	1.4120	1.1840
Body Mass Index (BMI)	-0.0070	0.0070	-0.0020	0.0060	0.0490	-0.0110	0.0060
Systolic Blood Pressure (SBP)	0.0100	0.0000	0.0060	-0.0030	0.0000	0.0010	0.0030
Diastolic Blood Pressure (DBP)	-0.0140	0.0000	-0.0130	0.0030	0.0000	0.0080	-0.0110
Pulse	0.0050	-0.0010	0.0070	-0.0060	0.0110	-0.0020	0.0040
Former Smoker	0.1630	-0.0310	-0.0180	0.1210	0.0000	-0.0270	0.0330
Current Smoker	0.2290	-0.1220	-0.0100	-0.1190	0.1480	0.0000	0.3040
Age	0.0270	-0.0190	-0.0440	0.0080	0.0210	0.0190	0.0040
Male	0.2430	0.1760	-0.0440	0.0410	-0.0380	-0.3720	-0.0790
Midwest	-0.0750	0.2490	-0.0730	-0.1310	0.0780	-0.1050	-0.2240
Northeast	-0.0980	0.1130	0.1570	-0.0380	-0.0890	-0.0020	-0.2430
west	-0.2770	0.6550	-0.2000	-0.5000	0.1370	-0.2400	-0.4060
Unknown Region	-0.0550	0.0000	-0.1730	-0.0800	0.1180	-0.0840	-0.0760
Aspirin Use	0.1190	-0.0670	-0.0250	0.1170	0.0360	-0.0050	-0.0250
Untreated Hypertension (HTN)	0.2630	0.1370	0.0000	-0.0310	0.0770	-0.1560	0.1400
HTN Mono Therapy	0.1000	0.1250	0.5070	-0.0810	0.0490	0.0000	-0.1880
HTN Combination Therapy, Controlled HTN	0.2100	0.0010	-0.0420	-0.0570	0.2000	-0.0530	-0.0390
HTN Combination Therapy, Uncontrolled HTN	0.0620	0.0690	0.1560	-0.0640	-0.0050	-0.0370	0.0000
Cholesterol Medication Use	-0.0210	0.0280	-0.1660	0.0920	-0.1850	0.0000	-0.0370
Diabetes Combination Therapy	0.1030	0.0130	0.0110	0.0010	-0.0320	0.0000	-0.0360
Diabetes Insulin Therapy	0.1610	-0.1270	0.0000	0.0040	0.1330	0.0000	0.0650
Normal LDL, No HLD Dx, and No Cholesterol Drugs	0.0000	-0.2640	-0.4220	-0.0850	-0.5610	0.0000	0.2860
Normal HDL, No HLD Dx, and No Cholesterol Drugs	0.0000	0.0000	0.5870	-0.2170	0.7690	0.0000	0.0000
Normal Triglycerides, No HLD Dx, and No Cholesterol Drugs	0.1710	0.0000	-0.2260	0.1890	0.2080	0.0250	-0.3140
Pre-existing Ischemic Heart Disease (IHD)	-1.6650	1.7850	1.6730		2.5150	2.1850	2.1880
Pre-existing Congestive Heart Failure (CHF)	6.2490	-5.8230	-5.6370	-4.9330		-5.9320	-5.8640
Pre-existing CerebroVascular Disease (CVD)	-0.0390	0.1250	0.2110	0.3770	0.2680		0.6360
Pre-existing Peripheral Vascular Disease (PVD)	0.0460	0.1800	0.2400	0.3610	0.3080	0.3580	
Pre-existing Chronic Kidney Disease (CKD)	-0.5160		1.3870	0.4260	0.5450	0.5820	0.5230
Pre-existing Chronic Renal Failure (CRF)	-0.9730	-3.8070		1.0880	1.2410	0.9840	1.2700

5.3.4 Performance Evaluation

Table 5-5 shows that C-indices of the various model variants predicting the outcomes on the three data sets. We report the C-indices on validation cohorts and the 95CIs of the validation performance in parentheses. Rows 1-4 correspond to internal validation and rows 5-8 to external validation.

Table 5-5 Predictive performance measured by the C-index



Row	Validation Type	Validation Dataset	Model Variant	IHD	CHF	CVD	PVD	CKD	CRF
1	Internal	OLDW	$variant_{OL.A1c}$	.74 (.73-.87)	.81 (.73-.87)	.74 (.73-87)	.75 (.73-87)	.80 (.73-87)	.86 (.73-.87)
2			$variant_{OL.FPG}$	.75 (.73-.87)	.81 (.73-.87)	.74 (.73-.87)	.74 (.73-.87)	.81 (.73-.87)	.87 (.73-.87)
3		UMMC	$variant_{UMMC}$	.73 (.71-.94)	.80 (.71-.94)	.75 (.71-.94)	.73 (.71-.94)	.80 (.71-.94)	.92 (.71-.94)
4		MCR	$variant_{MCR}$	.78 (.70-.94)	.82 (.68-.95)	.73 (.68-.94)	.75 (.70-.94)	.76 (.70-.94)	.90 (.70-.95)
5	External	UMMC	$variant_{OL.A1c}$	.73	.76	.75	.75	.79	.92
6			$variant_{MCR}$	.62	.73	.67	.70	.74	.90
7		MCR	$variant_{OL.FPG}$	.73	.76	.73	.74	.75	.92
8			$variant_{UMMC}$	.69	.75	.69	.68	.68	.89

In internal validation, all model variants showed very similar predictive performance (C-indices: .73 to .92) (row 1-4 in Table 5-5). All models were successfully validated: for all variant-outcome pairs, the validation performance fell into the 95CIs, indicating the overfitting did not occur. While all performances were high, $variant_{OL.A1c}$ and $variant_{OL.FPG}$ built on the national cohort showed much narrower 95CIs than $variant_{UMMC}$ and $variant_{MCR}$ built on the local cohorts. Given the large sample size, this is expected.

In external validation, all models achieved non-random performance (first criterion). When $variant_{OL.A1c}$ and $variant_{OL.FPG}$ (built on the nationally representative cohort) were applied to local cohorts, they still performed well (.73-.92) (row 5,7); in fact, they performed as well as $variant_{MCR}$ and $variant_{UMMC}$ in their own populations (row 3,4), satisfying the second criterion of external validation. In contrast, the performance of $variant_{MCR}$ and $variant_{UMMC}$ reduced substantially when applied to each other's healthcare system (.62-.90) (row 6,8) with C-index outside the internal validation CI for IHD, CVD, PVD (only for the $variant_{UMMC}$) and CKD (only for the $variant_{UMMC}$) thus failing external validation.

5.3.5 Consistency of the effect directions for significant coefficients among variants

The third criterion for external validation is that we required the significant variables of the model to have the same effect directions in the validation sets. Specifically, we wanted to observe the same effect directions of variables between $variant_{OL.A1c}$ and $variant_{UMMC}$, and between $variant_{OL.FPG}$ and $variant_{MCR}$. So, we

visualized coefficients of modifiable clinical variables in the Differential Progressions—individual complication- specific effects—identified by each variant (Figure 5-3).

The cells of only significant coefficients (Bonferroni corrected $P \leq .001$) were colored where orange cells indicate the variables that are significantly more important in determining progression to a particular outcome than in the General Progression (i.e. progression to any of the seven diseases) and light blue cells indicate the variables that are significantly less important in progression to a particular outcome than in the General Progression. Unless we observed symbols with different colors for a variable between the variant pair of comparison, we considered the role of the variable to be consistent between the variant pair. Statistically insignificant roles without a symbol were not counted as evidence of being inconsistent. We found that the effect directions of all the variables to be consistent.

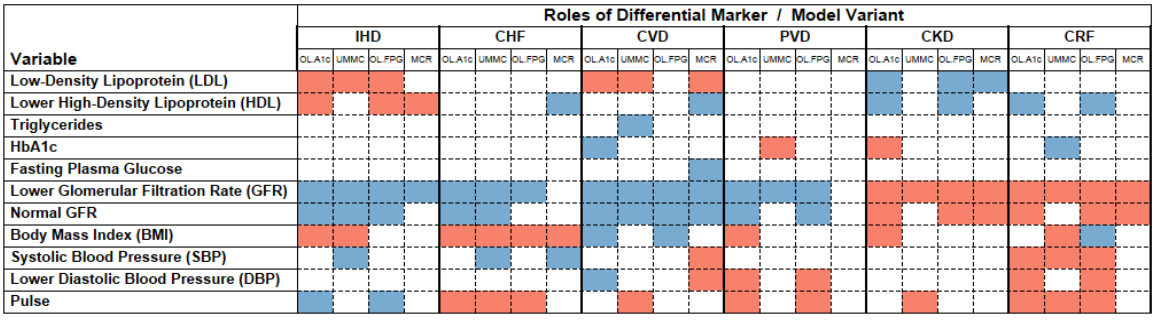


Figure 5-3 Consistency of the effect directions of significant Differential Markers among variants

5.4 Discussion

In healthcare, the generalizability of study results has always been emphasized as evidenced by various reporting standards for clinical trials[186], observational studies[187], diagnostic and prognostic studies[188,189], and meta-analyses[190,191]. Detailed and transparent reporting is very helpful to objectively evaluate the generalizability of models developed for clinical applications; however, merely following them does not guarantee that the resultant model is generalizable or transferable. For example, we followed the reporting guidelines (the TRIPOD statement [189]) with our models, yet the models constructed on local data sets failed external validation. In this work, we stress the importance of external validity.

External validity. The concept of external validity is complex and not easily formalized[192]. The terms external validity, generalizability, transferability, and applicability are often used interchangeably with overlapping meanings. Even the criteria for external validation are undefined. In the vast majority of machine learning literature, predictive performance is predominantly used as a criterion for external validity [193,194], although in clinical applications, external validity could, in theory, include clinical findings such as risk factors. Complex machine learning models tend to be challenging to interpret[195–198]. For such models, making clinical findings part of the validity criterion is impractical, hence the vast dominance of validation based on predictive performance. In our work for the prognosis of complications of T2DM, the model was interpretable and thus we used a compound criterion including predictive performance and common significant risk factors.

Another challenge to incorporating risk factors into the external validation is posed by institutional policies, which can influence available variables for study: some variables may only be available in certain healthcare systems and not in others. For example, health disparity related variables are highly predictive[175,199,200] but not commonly collected in routine practice. Although one healthcare system may build a model using those variables, the unavailability of the variables will render the models impractical in many other healthcare systems.

Dangers of ignoring external validity. It could be argued that as long as a model is only applied to a patient population on which it was developed (e.g. the model is never transferred to a different healthcare system), internal validity is sufficient. This is not necessarily true. We already mentioned that a model that overfits a particular healthcare system will have difficulty making a prediction for a patient who is atypical in that particular healthcare system, but common in other parts of the nation.

Institutional and administrative policies can vary among healthcare systems and change anytime. Most models for clinical applications are based on patient and disease characteristics; however, the policies can influence patient outcomes in a way that is not determined by patient and disease characteristics[192]. Besides physiological factors, models will implicitly incorporate the effects of institutional policy, as well. The institutional policies could suddenly change and render the model inaccurate. External validation can help at least detect the presence of such policies; using a nationally representative cohort collected over multiple provider networks can help marginalize such institution-specific effects.

Large samples. The use of small sample sizes, such as data from a single healthcare provider, may lead to optimistic predictive performance and limited generalizability[22,201]. Another factor that may have contributed to the lack of external validation is the high cost of obtaining a high-quality external validation set. An argument can be made that a better use of high-quality external data sets would be to add them to the training data to increase the sample size and decrease the effects of institutional policies rather than using them for external validation. In our work, with a nationally representative cohort in hand, we prioritized external validation and we used our two external cohorts as validation cohorts. These cohorts have baseline characteristics and outcome incidence rates that are fairly different from our training cohort, which makes our external validation more robust.

Limitations. Many patients were excluded due to missing lab results and vital signs measurements, which we did not impute. If patients are diabetic, they are supposed to receive routine check-ups or have access to primary care in order to have lab tests and their vital signs measured. Hence, for patients who did not have any HbA1c or blood pressure measurements *at all* during the 5-year baseline period, we were uncertain as to whether or not we could establish their baseline characteristics correctly; thus, they were excluded.

Summary. We demonstrated that even a complex machine learning model can be successfully externally validated when it is constructed on a nationally representative large data set collected from multiple providers. Having a nationally representative data set allows the model to capture patients who may be atypical in certain parts of the

nation, but common in others. It can also help marginalize institutional policies that could suddenly change, and render a model trained in a single institute inaccurate. Conversely, we have also demonstrated that models built in a single local healthcare system did not transfer to a different healthcare system. Given the current status quo of machine learning models being built on local healthcare systems and not being externally validated, their robustness can be questioned. We advocate for external validation for machine learning models to make them more robust, which will undoubtedly help in their adoption for clinical decision support.

5.5 Conclusions

Our modeling approach, in which a model is learned from a nationally representative cohort and is externally validated, can facilitate the production of a more generalizable and transferable model, thus allowing for precision medicine to be more accessible and to become a reality.

Chapter 6 Face Validation of the MTL-based Models

As a part of our last study described in Chapter 5, we conducted extensive literature review to examine whether interpretation of modifiable clinical factors identified by our MTL-methodology applied to the national cohort were consistent with the best available clinical evidence or beliefs (i.e., face validation[202]). Due to the strict word count restriction, we were not able to include this section into the manuscript so present it separately in this chapter.

The directions of associations between the variables and General Progression are in agreement with conclusions from other numerous studies[50,66,74,78,79], however *Differential Markers (i.e., outcome-specific risk factors)* are more explicit about explaining the various roles of their corresponding variables. For instance, although an extensive body of evidence has shown that higher LDL is harmful for IHD[72,73], CVD[70], and CKD[60,61], our model shows that LDL is particularly more harmful for IHD and CVD and less harmful for CKD. As for HDL, our model explained that higher HDL protects patients against complications in general (e.g., IHD[73], PVD[79], and CKD[60]) and particularly against IHD.

Also, a large body of evidence has shown that higher HbA1c significantly indicates progression to all the outcome complications[51–61]. However, based on our findings, indeed, HbA1c signals only CKD; its effect on the development of other complications is totally explained by General Progression. Given the different pathogenic mechanisms of micro- and macrovascular complications[101,203], our interpretation of

various roles for LDL, HDL, and HbA1c depending to individual complications seem to be consistent with the current clinical understanding.

The most recent evidence suggests that age-related hemodynamic changes influence the associations between blood pressure indices and cardiovascular risk[75,204,205]. Our model shows that lower DBP, higher SBP, and higher pulse are harmful for General Progression, consistent with recent study findings in patients older than 59-years[75,77,206]. Moreover, our model identifies pulse as a Differential Marker for IHD and CHF but not for CVD, also consistent with existing study results demonstrating such hemodynamic changes occur more frequently in cardiac than cerebrovascular diseases[185,206–209].

Even in cardiovascular diseases, our model further distinguishes the role of pulse between IHD (*less* harmful) and CHF (*more* harmful), suggesting that if patients with higher pulse progress, IHD is *less* likely but CHF is *more* likely to occur. Lower DBP and higher pulse are recognized characteristics of PVD[206], and uncontrolled blood pressure indices are leading indicators of CRF[206], consistent with their distinguishing patterns in Figure 1-1.

The impact of BMI on clinical outcomes in T2DM patients is somewhat controversial in that its impact can be both protective (CVD, advanced CKD stages 3-5[68], renal impairment[61], and all-cause mortality[64–67]) and harmful (IHD[62], CHF[63], and micro- and microalbuminuria[60,61,69]). And the roles of BMI we found are consistent with these results (Figure 6-1).

The interpretation of these plots in Figure 6-1 is as follows. Each plot represents a complication; axes are clinically modifiable variables; a darker polygon is drawn

connecting the coefficients in Differential Progression (Table 5-4) depicted by anchors on axes; as higher HDL, GFR, and DBP are protective, the sign of their coefficients is reversed only for visualization purpose. Thus, the correct readings of the axes of these variables are lower HDL, lower, GFR, and lower DBP.

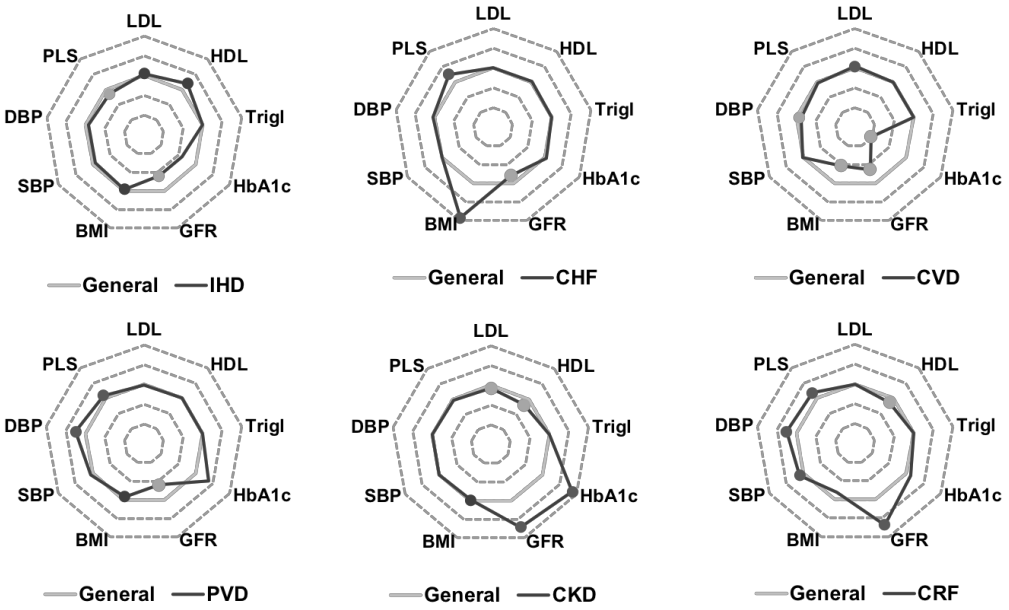


Figure 6-1 Distinguishing patterns of modifiable clinical Differential Markers

Chapter 7 Summary and Conclusions

7.1 Summary of studies

This dissertation documents our effort to reduce the gap between the promise and the reality of precision medicine in the management of T2DM. I successfully achieved all of the specific aims laid out in Chapter 1.3.

The first study aimed to develop a semi-supervised divisive hierarchical clustering algorithm for a subpopulation-based T2DM risk model. In this work, I identified subpopulations (clusters) that possessed different risk factors, risks, and time-to-progression and behaved differently from the general population. Also, the clustering which incorporated interactions among the risk factors, outperformed an additive statistical model (i.e., the Framingham score). By assigning each patient to a single subpopulation in a clustering, it obtained enhanced interpretability.

The second study aimed to develop a Multi-Task Learning (MTL)-based methodology that separated the overall deterioration of metabolic health from progression to specific complications utilizing large amounts of EHR data. My proposed methodology was able to separate the hypothesized two roles of the risk factors, offering an unprecedentedly detailed view on the progression to each individual complication. Also, the improved interpretability did not come at the cost of compromising predictive performance.

The third study aimed to demonstrate the transferability of the models constructed using the proposed approach, in which a model was learned from a nationally

representative cohort and was externally validated on two local health systems data. I systematically quantified and compared the external validity of a national model and locally constructed models with respect to a range of criteria for external validity. By demonstrating that the locally constructed models were difficult to transfer and discussing issues related to external validity, I successfully raised the awareness of the importance of external validity.

7.2 Contribution to Health Informatics and Medicine

My thesis work offers contributions to the fields of health informatics and medicine through (i) improved hypothesis generation, (ii) improvements to clinical research methodologies and (iii) the translational potential of the findings.

Improved hypothesis generation. Against the backdrop of contradictory evidence from RCTs, my studies can improve hypothesis generation for clinical research. As for the first study, among subpopulations identified by my semi-supervised divisive hierarchical clustering, one can identify subpopulations with specific risk factors that may benefit from an alternative management approach. Additionally, since my clustering offers fine control over the granularity of the clustering, researchers can explore the entire cluster hierarchy by drilling down for further details to discover clinically interesting subpopulations. Considering the heterogeneity of T2DM patients, the clustering algorithm's ability to identify clinically meaningful subpopulations will definitely help in improving hypothesis generation to maximize clinical benefits for that targeted subpopulation.

The roles of risk factors elucidated by my MTL-based methodology can also help hypothesis generation. After proving the feasibility of the methodology in the second study, I applied the methodology to a nationally representative dataset and comprehensively examined whether the identified roles of risk factors coincided with the best available clinical evidence or beliefs (Chapter 6) in the third study. My study findings and the results of the face validation can help improving hypothesis generation to develop therapeutic strategies for T2DM patients with multiple potential outcomes.

Improvements to clinical research informatics methodologies. My studies improve the state of the art in clinical research. My proposed algorithms and methodology not only retained key advantages of modern machine learning techniques but also added their own advantages. In the first study, my clustering algorithm demonstrated the ability to automatically discover interactions among risk factors seamlessly incorporating them into the model (clustering) and showed high predictive performance, originally proven as benefits of the ARM-based models. On top of that, the clustering obtained improved interpretability by eliminating the redundancy, known to be a shortcoming of the ARM-based models.

Regarding the second study, most studies aiming to model the progression to complications of T2DM focused on a single complication, thus the commonality was not identifiable and the distinction of the two roles were lost. With the large amounts of EHR data containing survival outcomes of six individual complications, my MTL-based methodology that learned the progression to various individual complications simultaneously was able to extract the commonality; thereby, revealing outcome-specific

effects. Also, the methodology did not compromise predictive performance due to the improved interpretability.

In the third study, I noticed that most machine learning studies tended to focus on model development but omitted external validation, failing to prove that the models capture universal physiological facts that hold true anywhere rather than simply learning local artifacts that are only true at a particular locale and could change at any time. This limits the practical utility of these models. I made considerable effort to present concrete strategies for model development that result in a transferable machine learning model and to demonstrate the external validity of this model using three independent big datasets.

Translational potential of the findings. Results from my studies may have significant translational potential for evidence-based practice and precision medicine. While machine learning models are increasingly being developed for clinical decision support for T2DM patients, the adoption of these models into clinical practice remains limited. I believe that the lack of *clinical usefulness* and evidence on *transferability* are the key impediments to the adoption so attempted to overcome the impediments. First, I focused on useful clinical findings (i.e., subpopulations and outcome-specific risk factors) but not typical findings (e.g., HbA1c increases the risk of T2DM). Second, I emphasized both predictive performance and interpretability in my modeling tasks. Third, I put effort into not only developing machine learning models but also into generating the evidence of external validity or *transferability*, which is another factor that can undoubtedly help the acceptance of machine learned models.

The second point is very pertinent to modern black-box models. Although the predictive performance of some machine learning models is adequate and in some areas may even surpass humans[93], high performing models are typically black-box models. They make decisions in a way that is not interpretable and thus engenders little clinician trust. Recently, the machine learning community itself has raised questions about the interpretability of their methods, and some recommendations on how to make them more interpretable have emerged[195–197,210]. I focused on the interpretability, because I believe that it will help to increase clinicians' acceptance of our models[195–198].

With respect to contribution to medicine, I believe the ability of generating clinically useful hypotheses, and even evidence itself, using big data, and machine learning techniques will complement the current evidence-based approach, enhancing current clinical practice guidelines.

7.3 Conclusion

The availability of reliable clinical evidence is critical for clinicians to make the right decision and produce high-quality results in healthcare delivery. In all of my work, I studied big data and machine learning techniques for the clinical applications in T2DM, which could potentially be transformed into intelligent decision support tools. My methods produce findings that coincide with evidence (when evidence exists), are interpretable, and can be externally validated. Such reliable and intelligible findings when implemented in decision support are the cornerstone of learning health care systems. Incorporating clinical expertise, scientific evidence, and patient preferences, such

intelligent decision tools will help ensuring that patients and clinicians have the information and tools they need to make informed and right decisions in patient care.

BIBLIOGRAPHY

1. Centers for Disease Control and Prevention. National Diabetes Statistics Report. 2014.
2. Saltiel AR, Kahn CR. Insulin signaling and the regulation of glucose and lipid metabolism. *Nature*. 2001;414(6865):799-806.
3. AN M, OK N. An Evidence-Based Medicine Approach to Antihyperglycemic Therapy in Diabetes to Overcome Overtreatment. *Circulation*. 2017;135(2):180-195. doi:10.1016/j.chemosphere.2012.12.037.Reactivity
4. Grant RW, Wexler DJ. Personalized medicine in Type 2 diabetes: what does the future hold? *Diabetes Manag*. 2012;2(3):199-204. doi:10.1016/j.surg.2006.10.010.Use
5. Stewart WF, Shah NR, Selna MJ, Paulus RA, Walker JM. Bridging the inferential gap: the electronic health record and clinical evidence. *Health Aff*. 2007;26(2):w181-w191. doi:10.1377/hlthaff.26.2.w181
6. Skyler JS, Bergenstal R, Bonow RO, et al. Intensive glycemic control and the prevention of cardiovascular events: implications of the ACCORD, ADVANCE, and VA Diabetes Trials. *J Am Coll Cardiol*. 2009;53(3):298-304. doi:10.1016/j.jacc.2008.10.008
7. A New Initiative on Precision Medicine. *Perspective*. 2010;363(1):1-3. doi:10.1056/NEJMp1002530
8. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181
9. Agrawal R, Srikant R, others. Fast algorithms for mining association rules. In: *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*. Vol 1215. ; 1994:487-499.
10. Liu G, Feng M, Wang Y, et al. Towards exploratory hypothesis testing and analysis. *2011 IEEE 27th Int Conf Data Eng*. April 2011:745-756. doi:10.1109/ICDE.2011.5767907

11. Caraballo PJ, Castro MR, Cha SS, Li PW, Simon GJ. Use of Association Rule Mining to Assess Diabetes Risk in Patients with Impaired Fasting Glucose. In: *AMIA 2011 Symposium Proceedings*. ; 2011.

12. Simon GJ, Li PW, Jack CR, Vemuri P. Understanding atrophy trajectories in alzheimer's disease using association rules on MRI images. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*. New York, New York, USA: ACM Press; 2011:369. doi:10.1145/2020408.2020469

13. Gerstein HC, Michael EM, Byington RP et al. Effects of Intensive Glucose Lowering in Type 2 Diabetes. *N Engl J Med*. 2008;358(24):2545-2559.

14. Duckworth W, Abraira C, Moritz T, et al. Glucose control and vascular complications in veterans with type 2 diabetes. *N Engl J Med*. 2009;360(2):129-139. doi:10.1056/NEJMoa0808431

15. Patel A et al. Intensive Blood Glucose Control and Vascular Outcomes in Patients with Type 2 Diabetes. *N Engl J Med*. 2008;358:2560-2572. doi:10.1056/NEJMoa0802987

16. Holman RR, Paul SK, Bethel MA, Matthews DR, Neil HAW. 10-Year Follow-Up of Intensive Glucose Control in Type 2 Diabetes. *N Engl J Med*. 2008;359(15):1577-1589. doi:10.1056/NEJMoa0806470

17. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. *Comput Struct Biotechnol J*. 2017;15:104-116. doi:10.1016/j.csbj.2016.12.005

18. Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol*. 2017:193229681770637. doi:10.1177/1932296817706375

19. Lagani V, Chiarugi F, Thomson S, et al. Development and validation of risk assessment

models for diabetes-related complications based on the DCCT/EDIC data. *J Diabetes Complications*. 2015;29(4):479-487. doi:10.1016/j.jdiacomp.2015.03.001

20. Cichosz SL, Johansen MD, Hejlesen O. Toward Big Data Analytics : Review of Predictive Models in Management of Diabetes and Its Complications. 2016. doi:10.1177/1932296815611680

21. Stern MP. Perspectives in Diabetes Diabetes and Cardiovascular Disease The "Common Soil" Hypothesis. (4):369-374.

22. Collins GS, Mallett S, Omar O, et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*. 2011;9(1):103. doi:10.1186/1741-7015-9-103

23. Abbasi A, Peelen LM, Corpeleijn E. Prediction models for risk of developing type 2 diabetes : systematic literature search and independent. *BMJ*. 2012;5900(September):1-16. doi:10.1136/bmj.e5900

24. Collins GS, Mallett S, Omar O, Yu L. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*. 2011;9(1):103. doi:10.1186/1741-7015-9-103

25. Wilson PWF, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med*. 2007;167(10):1068-1074. doi:10.1001/archinte.167.10.1068

26. Fincke BG, Clark J a, Linzer M, et al. Assessment of Long-term Complications due to Type 2 Diabetes Using Patient Self-report. *Diabetes*. 2005;28(3):262-273.

27. Young B, Lin E, Korff M Von. Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization. *Am J Manag Care*. 2008;14(1):15-23.

28. Simon GJ, Caraballo PJ, Therneau TM, Cha SS, Castro MR, Li PW. Extending

Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus. *Knowl Data Eng IEEE Trans*. 2015;27(1):130-141. doi:10.1109/TKDE.2013.76

29. Simon GJ, Schrom J, Castro MR, Li PW, Caraballo PJ. Survival association rule mining towards type 2 diabetes risk assessment. *AMIA Annu Symp Proc*. 2013;2013:1293.

30. Kim HS, Shin a M, Kim MK, Kim YN. Comorbidity study on type 2 diabetes mellitus using data mining. *Korean J Intern Med*. 2012;27(2):197-202. doi:10.3904/kjim.2012.27.2.197

31. Lagani V, Koumakis L, Chiarugi F, Lakasing E, Tsamardinos I. A systematic review of predictive risk models for diabetes complications based on large scale clinical studies. *J Diabetes Complications*. 2013;27(4):407-413. doi:10.1016/j.jdiacomp.2012.11.003

32. Yadav P, Steinbach M, Kumar V, Simon G. Mining Electronic Health Records: A Survey. *ACM Comput Surv*. 2018;50(6):1-40. doi:1539-9087/2016/04-ART1

33. Contreras I VJ. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. *J Med Internet Res*. 2018;20(5).

34. Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. *Artif Intell Med*. 2002;26(1-2):37-54. doi:10.1016/S0933-3657(02)00051-9

35. Mani S, Chen Y, Elasy T, Clayton W, Denny J. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc*. 2012;2012(Dm):606-615. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540444&tool=pmcentrez&rendertype=abstract.

36. Meng X-H, Huang Y-X, Rao D-P, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung J Med Sci*. 2013;29(2):93-99. doi:10.1016/j.kjms.2012.08.016

37. Rokach L, Maimon O. Decision Trees. In: *Data Mining and Knowledge Discovery*

Handbook. ; 2005:pp 165-192. doi:10.1007/978-0-387-09823-4_9

38. Tan P-N, Steinbach M, Kumar V. *Introduction to Data Mining*. Pearson; 2005.

39. Giacomozzi C, Martelli F. Peak pressure curve: an effective parameter for early detection of foot functional impairments in diabetic patients. *Gait Posture*. 2006;23(4):464-470. doi:10.1016/j.gaitpost.2005.06.006

40. Guttula SV, Allam AR, Al. E. Cluster analysis and phylogenetic relationship in biomarker identification of type 2 diabetes and nephropathy. *Int J Diabetes Dev Ctries*. 2010;30(1):52-56. doi:10.4103/0973-3930.60003

41. Antonelli D, Baralis E, Bruno G, Cerquitelli T, Chiusano S, Mahoto N. Analysis of diabetic patients through their examination history. *Expert Syst Appl*. 2013;40(11):4672-4678. doi:10.1016/j.eswa.2013.02.006

42. Huo Y, Azuaje F, McCullagh P, Harper R. Semi-Supervised Clustering Models for Clinical Risk Assessment. *Bioinforma Bioeng*. 2006:243-250. doi:10.1109/BIBE.2006.253341

43. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*. 2004;2(4):E108. doi:10.1371/journal.pbio.0020108

44. Koestler DC, Marsit CJ, Christensen BC, et al. Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics*. 2010;26(20):2578-2585. doi:10.1093/bioinformatics/btq470

45. Eick CF, Zeidat N. *Using Supervised Clustering to Enhance Classifiers*. Springer Berlin Heidelberg; 2005.

46. Miselli M-A, Nora ED, Passaro A, Tomasi F, Zuliani G. Plasma triglycerides predict ten-years all-cause mortality in outpatients with type 2 diabetes mellitus: a longitudinal observational study. *Cardiovasc Diabetol*. 2014;13:135. doi:10.1186/s12933-014-0135-6

47. Monami M, Lamanna C, Balzi D, et al. Metabolic Syndrome and Cardiovascular Mortality in Older Type 2 Diabetic Patients: A Longitudinal Study. *J Gerontol*. 2008;63(6):646-649. doi:10.1093/gerona/63.6.646

48. Malik S, Wong ND, Franklin SS, et al. Impact of the Metabolic Syndrome on Mortality From Coronary Heart Disease, Cardiovascular Disease, and All Causes in United States Adults. *Circulation*. 2004;110(10):1245-1250. doi:10.1161/01.CIR.0000140677.20606.0E

49. Fox CS, Coady S, Sorlie PD, et al. Increasing cardiovascular disease burden due to diabetes mellitus: the Framingham Heart Study. *Circulation*. 2007;115(12):1544-1550. doi:10.1161/CIRCULATIONAHA.106.658948

50. ten Brinke R, Dekker N, de Groot M, Ikkersheim D. Lowering HbA1c in type 2 diabetics results in reduced risk of coronary heart disease and all-cause mortality. *Prim Care Diabetes*. 2008;2(1):45-49. doi:10.1016/j.pcd.2007.12.004

51. Stratton IM, Adler AI, Neil HAW, et al. Association of glycaemia with macrovascular and prospective observational study (UKPDS 35): prospective observational study. *BMJ*. 2000;321(7258):405-412.

52. Nichols GA, Hillier TA, Erbey JR, Brown JB. Congestive Heart Failure in Type 2 Diabetes: Prevalence, incidence, and risk factors. *Diabetes Care*. 2001;24(9):1614-1619.

53. Adler AI, Neil HA, Manley SE, Holman RR, Turner RC. Hyperglycaemia and hyperinsulinaemia at diagnosis of diabetes and their association to subsequent cardiovascular disease in the UK Prospective Diabetes Study (UKPDS 47). *Am Heart J*. 1999;138(5):S353-9.

54. Zhao W, Katzmarzyk PT, Horswell R, Wang Y, Johnson J, Hu G. Sex differences in the risk of stroke and HbA1c among diabetic patients. *Diabetologia*. 2014;57(5):918-926. doi:10.1007/s00125-014-3190-3

55. Selvin E, Coresh J, Shahar E, Zhang L, Steffes M, Sharrett AR. Glycaemia (haemoglobin A1c) and incident ischaemic stroke: the Atherosclerosis Risk in Communities (ARIC) Study. *Lancet Neurol*. 2005;4(12):821-826. doi:10.1016/S1474-4422(05)70227-1

56. Muntner P, RP W, K R, KB D, J C, V. F. Relationship Between HbA1c Level and Peripheral Arterial Disease. *Diabetes Care*. 2005;28(8):1981-1987.

57. Selvin E, Wattanakit K, Steffes MW, Coresh J, Sharrett AR. HbA1c and Peripheral Arterial Disease in Diabetes. *Diabetes Care*. 2006;29(4):877-882.

58. Adler AI, Stevens RJ, Neil A, Stratton IM, Boulton AJM, Holman RR. UKPDS 59: Hyperglycemia and other potentially modifiable risk factors for peripheral vascular disease in type 2 diabetes. *Diabetes Care*. 2002;25(5):894-899. doi:10.2337/diacare.25.5.894

59. Bash LD, Selvin E, Steffes M, Coresh J, Astor BC. Poor Glycemic Control in Diabetes and the Risk of Incident Chronic Kidney Disease Even in the Absence of Albuminuria and Retinopathy. *Arch Intern Med*. 2008;168(22):2440. doi:10.1001/archinte.168.22.2440

60. Ravid M, Brosh D, Ravid-Safran D, Levy Z RR. Main Risk Factors for Nephropathy in Type 2 Diabetes Mellitus Are Plasma Cholesterol Levels, Mean Blood Pressure, and Hyperglycemia. *Arch Intern Med*. 1998;158(9):998-1004.

61. Retnakaran R, Cull CA, Thorne KI, Adler AI, Holman RR. Risk Factors for Renal Dysfunction in Type 2 Diabetes. *Diabetes*. 2006;55(6):1832-1839. doi:10.2337/db05-1620.

62. Li N, Katzmarzyk PT, Horswell R, et al. BMI and coronary heart disease risk among low-income and underinsured diabetic patients. *Diabetes Care*. 2014;37(12):3204-3212. doi:10.2337/dc14-1091

63. Li W, Katzmarzyk PT, Horswell R, et al. Body Mass Index and Heart Failure among

Patients with Type 2 Diabetes Mellitus. *Circ Hear Fail*. 2015;8(3):455-463. doi:10.1161/CIRCHEARTFAILURE.114.001837

64. Li W, Katzmarzyk PT, Horswell R, et al. Body mass index and stroke risk among patients with type 2 diabetes mellitus. *Stroke*. 2015;46(1):164-169. doi:10.1161/STROKEAHA.114.006718

65. Kokkinos P, Myers J, Faselis C, Doumas M, Kheirbek R, Nylen E. BMI-mortality paradox and fitness in African American and Caucasian men with type 2 diabetes. *Diabetes Care*. 2012;35(5):1021-1027. doi:10.2337/dc11-2407

66. McEwen LN, Kim C, Karter AJ, et al. Risk Factors for Mortality Among Patients With Diabetes. *Diabetes Care*. 2007;30(7):1736-1741. doi:10.2337/dc07-0305.Abbreviations

67. Doehner W, Erdmann E, Cairns R, et al. Inverse relation of body weight and weight change with mortality and morbidity in patients with type 2 diabetes and cardiovascular co-morbidity: an analysis of the PROactive study population. *Int J Cardiol*. 2012;162(1):20-26. doi:10.1016/j.ijcard.2011.09.039

68. Huang W-H, Chen C-Y, Lin J-L, Lin-Tan D-T, Hsu C-W, Yen T-H. High Body Mass Index Reduces Glomerular Filtration Rate Decline in Type II Diabetes Mellitus Patients With Stage 3 or 4 Chronic Kidney Disease. *Medicine (Baltimore)*. 2014;93(7):e41. doi:10.1097/MD.0000000000000041

69. Mohammedi K, Chalmers J, Herrington W, et al. Associations between body mass index and the risk of renal events in patients with type 2 diabetes. *Nutr Diabetes*. 2018;8(1). doi:10.1038/s41387-017-0012-y

70. Shepherd J, Barter P, Carmena R, et al. Effect of lowering LDL cholesterol substantially below currently recommended levels in patients with coronary heart disease and diabetes: The Treating To New Targets (TNT) study. *Diabetes Care*. 2006;29(6):1220-1226.

doi:10.2337/dc05-2465

71. Hayashi T. Low HDL Cholesterol Is Associated With the Risk of Stroke in Elderly Diabetic. *Diabetes Care*. 2009;32(7):10-12. doi:10.2337/dc08-1677.

72. Deedwania PC, Pedersen TR, DeMicco DA, et al. Differing predictive relationships between baseline LDL-C, systolic blood pressure, and cardiovascular outcomes. *Int J Cardiol*. 2016;222:548-556. doi:10.1016/j.ijcard.2016.07.201

73. Despres JP, Lemieux I, Dagenais GR, Cantin B, Lamarche B. HDL-cholesterol as a marker of coronary heart disease risk: The Quebec cardiovascular study. *Atherosclerosis*. 2000;153:263-272. doi:10.1016/S0021-9150(00)00603-1

74. Miselli M-A, Nora ED, Passaro A, Tomasi F, Zuliani G. Plasma triglycerides predict ten-years all-cause mortality in outpatients with type 2 diabetes mellitus: a longitudinal observational study. *Cardiovasc Diabetol*. 2014;13:135. doi:10.1186/s12933-014-0135-6

75. Franklin S, Larson M, Khan S, et al. Does the Relation of Blood Pressure to Coronary Heart Disease Risk Change With Aging?: The Framingham Heart Study. *Circulation*. 2001;103(9):1245-1249.

76. Vaccarino V, Holford TR, Krumholz HM. Pulse Pressure and Risk for Myocardial Infarction and Heart Failure in the Elderly. *J Am Coll Cardiol*. 2000;36(1):130-138. doi:10.1016/S0735-1097(00)00687-2

77. Franklin SS, Khan SA, Wong ND, Larson MG, Levy D. Is pulse pressure useful in predicting risk for coronary heart Disease? The Framingham heart study. *Circulation*. 1999;100(4):354-360. doi:10.1161/01.cir.100.4.354

78. Said MA, Eppinga RN, Lipsic E, Verweij N, van der Harst P. Relationship of Arterial Stiffness Index and Pulse Pressure With Cardiovascular Disease and Mortality. *J Am Heart Assoc*. 2018;7(2):e007621. doi:10.1161/JAHA.117.007621

79. Amada IA, Tratton IM, H.A.W. N, et al. Association of systolic blood pressure with macrovascular and microvascular complications of type 2 diabetes (UKPDS 36): Prospective observational study. *Br Med J*. 2000;321(7258):412-419. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed5&NEWS=N&AN=2000275675.

80. Zhao W, Katzmarzyk PT, Horswell R, et al. Aggressive blood pressure control increases coronary heart disease risk among diabetic patients. *Diabetes Care*. 2013;36(10):3287-3296. doi:10.2337/dc13-0189

81. Zhao W, Katzmarzyk PT, Horswell R, et al. Blood pressure and stroke risk among diabetic patients. *J Clin Endocrinol Metab*. 2013;98(9):3653-3662. doi:10.1210/jc.2013-1757

82. Zhao W, Katzmarzyk PT, Horswell R, et al. Blood pressure and heart failure risk among diabetic patients. *Int J Cardiol*. 2014;176(1):125-132. doi:10.1016/j.ijcard.2014.06.051

83. Al-Delaimy WK, Manson JE, Solomon CG, et al. Smoking and Risk of Coronary Heart Disease Among Women with Type 2 Diabetes Mellitus. *Arch Intern Med*. 2002;162(3):273-279. doi:10.1001/archinte.162.3.273

84. Price JF, Mowbray PI, Lee AJ, Rumley A, Lowe GDO, Fowkes FGR. Relationship between smoking and cardiovascular risk factors in the development of peripheral arterial disease and coronary artery disease. Edinburgh Artery Study. *Eur Heart J*. 1999;20(5):344-353. doi:10.1053/euhj.1998.1194

85. Fowkes FGR, Housley E, Riemersma RA, et al. Smoking, Lipids, Glucose Intolerance, and Blood Pressure as Risk Factors for Peripheral Atherosclerosis Compared with Ischemic Heart Disease in the Edinburgh Artery Study. *Am J Epidemiol*. 1993;103(5):1771-1774. doi:10.1016/j.amjmed.2014.11.036.

86. Lagani V, Chiarugi F, Manousos D, et al. Realization of a service for the long-term risk assessment of diabetes-related complications. *J Diabetes Complications*. 2015;29(5):691-698. doi:10.1016/j.jdiacomp.2015.03.011

87. Sacchi L, Dagliati A, Segagni D, Leporati P, Chiovato L, Bellazzi R. Improving risk-stratification of Diabetes complications using temporal data mining. *2015 37th Annu Int Conf IEEE Eng Med Biol Soc*. 2015:2131-2134. doi:10.1109/EMBC.2015.7318810

88. Marini S, Trifoglio E, Barbarini N, et al. A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes. *J Biomed Inform*. 2015;57:369-376. doi:10.1016/j.jbi.2015.08.021

89. Kazemi M, Moghimbeigi A, Kiani J, Mahjub H, Faradmal J. Diabetic peripheral neuropathy class prediction by multicategory support vector machine model : a cross-sectional study. *Epidemiol Health*. 2016;38. doi:10.4178/epih.e2016011

90. DuBrava S, Mardekian J, Sadosky A, et al. Using Random Forest Models to Identify Correlates of a Diabetic Peripheral Neuropathy Diagnosis from Electronic Health Record Data. *Pain Med*. 2016:107-115. doi:10.1093/pm/pnw096

91. Huang GM, Huang KY, Lee TY, Weng J. An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. *BMC Bioinformatics*. 2015;16(Suppl 1):S5. doi:10.1186/1471-2105-16-S1-S5

92. R.K.K. L, Y. W, R.C.W. M, et al. Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: A prospective case-control cohort analysis. *BMC Nephrol*. 2013;14(1). http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L52698127%5Cnhttp://dx.doi.org/10.1186/1471-2369-14-162.

93. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama*. 2016;304(6):649-656. doi:10.1001/jama.2016.17216

94. Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *J Diabetes Sci Technol*. 2015;9(1):86-90. doi:10.1177/1932296814554260

95. Association AD. Economic costs of diabetes in the U.S. in 2012. *Diabetes Care*. 2013;36(4):1033-1046. doi:10.2337/dc12-2625

96. Narayan KMV, Boyle JP, Thompson TJ, Sorensen SW, Williamson DF. Lifetime risk for diabetes mellitus in the United States. *JAMA*. 2003;290(14):1884-1890. doi:10.1001/jama.290.14.1884

97. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*. 2008;31 Suppl 1:S55-60. doi:10.2337/dc08-S055

98. Gerich JE. The genetic basis of type 2 diabetes mellitus: impaired insulin secretion versus impaired insulin sensitivity. *Endocr Rev*. 1998;19(4):491-503. doi:10.1210/edrv.19.4.0338

99. Doria A, Patti M-E, Kahn CR. The emerging genetic architecture of type 2 diabetes. *Cell Metab*. 2008;8(3):186-200. doi:10.1016/j.cmet.2008.08.006

100. Froguel P, Velho G, Vionnet N. Genetics and diabetes. *Sang Thromb Vaiss*. 1994;6(5 SUPPL):39-46.

101. Fowler MJ. Microvascular and Macrovascular Complications of Diabetes. *Clin diabetes*. 2008;26(2):77-82. doi:10.2337/diaclin.26.2.77

102. Long AN, Dagogo-Jack S. Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection. *J Clin Hypertens (Greenwich)*. 2011;13(4):244-251. doi:10.1111/j.1751-7176.2011.00434.x

103. Zhang X, Saaddine JB, Chou C-F, et al. Prevalence of diabetic retinopathy in the United States, 2005-2008. *JAMA*. 2010;304(6):649-656. doi:10.1001/jama.2010.1111

104. Varma R. From a population to patients: the Wisconsin epidemiologic study of diabetic retinopathy. *Ophthalmology*. 2008;115(11):1857-1858. doi:10.1016/j.ophtha.2008.09.023

105. Centers for Disease Control and Prevention. National Diabetes Fact Sheet, 2011.

106. Sowers JR, Epstein M, Frohlich ED. Diabetes, Hypertension, and Cardiovascular Disease : An Update. *Hypertension*. 2001;37(4):1053-1059. doi:10.1161/01.HYP.37.4.1053

107. Beltrán-Sánchez H, Harhay MO, Harhay MM, McElligott S. Prevalence and trends of metabolic syndrome in the adult U.S. population, 1999-2010. *J Am Coll Cardiol*. 2013;62(8):697-703. doi:10.1016/j.jacc.2013.05.064

108. Isomaa B, Almgren P, Tuomi T, et al. Cardiovascular Morbidity and Mortality Associated With the Metabolic Syndrome. *Diabetes ...*. 2001;24(4). http://care.diabetesjournals.org/content/24/4/683.short. Accessed March 3, 2015.

109. Alberti KGMM, Eckel RH, Grundy SM, et al. Harmonizing the Metabolic Syndrome. *Circulation*. 2009;120(16):1640-1645. doi:10.1161/CIRCULATIONAHA.109.192644

110. Alexander CM, Landsman PB, Teutsch SM, Haffner SM. NCEP-Defined Metabolic Syndrome, Diabetes, and Prevalence of Coronary Heart Disease Among NHANES III Participants Age 50 Years and Older. *Diabetes*. 2003;52(5):1210-1214. doi:10.2337/diabetes.52.5.1210

111. Beckman JA, Creager MA, Libby P. Diabetes and atherosclerosis: epidemiology, pathophysiology, and management. *JAMA J Am Med Assoc*. 2002;287(19):2570-2581.

112. Wilson PWF, D'Agostino RB, Parise H, Sullivan L, Meigs JB. Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus. *Circulation*. 2005;112(20):3066-3072. doi:10.1161/CIRCULATIONAHA.105.539528

113. R.A. C, C.I. A, E.J. D, et al. A review of treatment response in type 2 diabetes: Assessing the role of patient heterogeneity. *Diabetes, Obes Metab*. 2010;12(10):845-857. doi:10.1111/j.1463-1326.2010.01248.x

114. Zagar A, Chen L, Boye KS, et al. Understanding Heterogeneity in Response to Antidiabetes Treatment:A Post Hoc Analysis Using SIDES, a Subgroup Identification Algorithm. 2013;7(2):420-430.

115. Alexander CM, Landsman PB, Teutsch SM, Haffner SM. Prevalence of Coronary Heart Disease Among NHANES III Participants Age 50 Years and Older. 2003;52(May).

116. Kim E, Oh W, Pieczkiewicz DS, Castro MR, Caraballo PJ, Simon GJ. Divisive Hierarchical Clustering towards Identifying Clinically Significant. *AMIA 2014 Symp Proc*. 2014.

117. Oh W, Kim E, Castro MR, et al. Type 2 Diabetes Mellitus Trajectories and Associated Risks. *Big Data*. 2016;4(1):25-30. doi:10.1089/big.2015.0029

118. Croft P, Altman DG, Deeks JJ, et al. The science of clinical practice: Disease diagnosis or patient prognosis? Evidence about "what is likely to happen" should shape clinical practice. *BMC Med*. 2015;13(1):1-8. doi:10.1186/s12916-014-0265-4

119. Oh W, Yadav P, Kumar V, et al. Estimating Disease Onset Time by Modeling Lab Result Trajectories via Bayes Networks. In: *Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI)*. ; 2017:374-379. doi:10.1109/ICHI.2017.41

120. Gleissner C a, Galkina E, Nadler JL, Ley K. Mechanisms by which diabetes increases cardiovascular disease. *Drug Discov Today Dis Mech*. 2007;4(3):131-140. doi:10.1016/j.ddmec.2007.12.005

121. Sobel BE. Optimizing cardiovascular outcomes in diabetes mellitus. *Am J Med*. 2007;120(9 Suppl 2):S3-11. doi:10.1016/j.amjmed.2007.07.002

122. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *ACP J Club*. 2002;136(2):A11-A11. doi:10.1136/ebm.7.2.36

123. Duckworth W, Abraira C, Moritz T, et al. Glucose control and vascular complications in veterans with type 2 diabetes. *N Engl J Med*. 2009;360(2):129-139. doi:10.1056/NEJMoa0808431

124. Doran KM, Raven MC, Rosenheck RA. What drives frequent emergency department use in an integrated health system? National data from the Veterans Health Administration. *Ann Emerg Med*. 2013;62(2):151-159. doi:10.1016/j.annemergmed.2013.02.016

125. Gaede P, Lund-Andersen H, Parving H-H, Pedersen O. Effect of a Multifactorial Intervention on Mortality in Type 2 Diabetes. *N Engl J Med*. 2008;358(6):580-591. doi:10.1056/NEJMoa0706245

126. Zoungas S, De Galan BE, Ninomiya T, Grobbee D, Hamet P, Heller S et al. Combined Effects of Routine Blood Pressure Lowering and Intensive Glucose Control on Macrovascular and Microvascular Outcomes in Patients With Type 2 Diabetes. *Diabetes Care*. 2009;32(11):2068-2074.

127. Blumenthal D. Launching HITECH. *N Engl J Med*. 2010;362(5):382-385. doi:10.1056/NEJMp0912825

128. Terry K. EHR Adoption: U.S. Remains The Slow Poke. Information Week. http://www.informationweek.com/healthcare/electronic-health-records/ehr-adoption-us-remains-the-slow-poke/d/d-id/1107410. Published 2012.

129. Frankovich J, Longhurst CA, Sutherland SM. Evidence-Based Medicine in the EMR Era. *N Engl J Med*. 2011;365(19):1758-1759. doi:10.1056/NEJMp1108726

130. Herrin J, Graca B, Nicewander D, et al. The effectiveness of implementing an electronic health record on diabetes care and outcomes. *Health Serv Res*. 2012;47(4):1522-1540.

doi:10.1111/j.1475-6773.2011.01370.x

131. Cebul RD, Love TE, Jain AK, Hebert CJ. Electronic health records and quality of diabetes care. *N Engl J Med*. 2011;365(9):825-833. doi:10.1056/NEJMsa1102519

132. Ciemins EL, Coon PJ, Fowles JB, Min S. Beyond health information technology: critical factors necessary for effective diabetes disease management. *J Diabetes Sci Technol*. 2009;3(3):452-460.

133. Kern LM, Barrón Y, Dhopeshwarkar R V, Edwards A, Kaushal R. Electronic health records and ambulatory quality of care. *J Gen Intern Med*. 2013;28(4):496-503. doi:10.1007/s11606-012-2237-8

134. Dorr D, Bonner L, Cohen A, et al. Informatics systems to promote improved care for chronic illness: a literature review. *J Am Med Informatics Assoc*. 2007;14(2):156-163. doi:10.1197/jamia.M2255.Introduction

135. Weber V, Bloom F, Pierdon S, Wood C. Employing the electronic health record to improve diabetes care: a multifaceted intervention in an integrated delivery system. *J Gen Intern Med*. 2008;23(4):379-382. doi:10.1007/s11606-007-0439-2

136. O'Connor P, Sperl-Hillen J, Rush W, et al. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. *Ann Fam Med*. 2011;9(1):12-21. doi:10.1370/afm.1196.INTRODUCTION

137. Crosson J, Ohman-Strickland P, Hahn K, Shaw E, Orzano A, Crabtree B. Electronic medical records and diabetes quality of care: results from a sample of family medicine practices. *Ann Fam Med*. 2007;5(3):209-215. doi:10.1370/afm.696

138. Crosson J, Ohman-Strickland P, Cohen D, Clark E, Crabtree B. Typical electronic health record use in primary care practices and the quality of diabetes care. *Ann Fam Med*. 2012;10(3):221-227. doi:10.1370/afm.1370.INTRODUCTION

139. Chawla N, Davis D. Bringing big data to personalized healthcare: a patient-centered framework. *J Gen Intern Med*. 2013;28(3):660-665. doi:10.1007/s11606-013-2455-8

140. Sperl-HillenPalattao K, Amundson J, Ekstrom H, Rush B, O'Connor P. Outpatient EHR-based diabetes clinical decision support that works: Lessons learned from implementing diabetes wizard. *Diabetes Spectr*. 2010;23(3):150-154. doi:10.2337/diaspect.23.3.150

141. Ely J, Osheroff J, Chambliss M, Ebell M, Rosenbaum M. Answering physicians' clinical questions: obstacles and potential solutions. *J Am Med Inform Assoc*. 2005;12(2):217-224. doi:10.1197/jamia.M1608

142. Yoo I, Alafaireet P, Marinov M, et al. Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst*. 2012;36(4):2431-2448. doi:10.1007/s10916-011-9710-5

143. Hand DJ, Mannila H, Smyth P. *Principles of Data Mining*. MIT press; 2001.

144. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*. 2006;2:59-77. doi:10.1177/117693510600200030

145. Lisboa PJ, Taktak AFG. The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks*. 2006;19(4):408-415. doi:10.1016/j.neunet.2005.10.007

146. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17. doi:10.1016/j.csbj.2014.11.005

147. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. *Mol Pharm*. 2016;13(5):1445-1454. doi:10.1021/acs.molpharmaceut.5b00982

148. Obenshain M. Application of data mining techniques to healthcare data. *Infect Control*.

2004;25(8):690-695. doi:10.1086/502460

149. Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. 2011.

150. Centers for Disease Control and Prevention. diabetes report card 2012: national and state profile of diabetes and its complications. 2012.

151. Lindström J, Peltonen M, Eriksson JG, et al. Improved lifestyle and decreased diabetes risk over 13 years: long-term follow-up of the randomised Finnish Diabetes Prevention Study (DPS). *Diabetologia*. 2013;56(2):284-293. doi:10.1007/s00125-012-2752-5

152. Knowler WC, Barrett-Connor E, Fowler SE, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med*. 2002;346(6):393-403. doi:10.1056/NEJMoa012512

153. Simon GJ, Caraballo PJ, Therneau TM, Cha SS, Castro MR, Li PW. Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus. *Knowl Data Eng IEEE Trans*. 2013;PP(99):1-13. doi:10.1109/TKDE.2013.76

154. Simon GJ, Schrom J, Castro MR, Li PW, Caraballo PJ. Survival association rule mining towards type 2 diabetes risk assessment. *AMIA Annu Symp Proc*. 2013;2013:1293.

155. Schrom JR, Caraballo PJ, Castro MR, Simon GJ. Quantifying the Effect of Statin Use in Pre-Diabetic Phenotypes Discovered Through Association Rule Mining.

156. Simon GJ, Kumar V, Li PW. A simple statistical model and association rule filtering for classification. *Proc 17th ACM SIGKDD Int Conf Knowl Discov data Min - KDD '11*. 2011:823. doi:10.1145/2020408.2020550

157. Tan P-N, Steinbach M, Kumar V. Cluster Analysis: Basic Concepts and Algorithms. In: *Introduction to Data Mining*. Addison-Wesley; 2005.

158. Rocca W a, Yawn BP, St Sauver JL, Grossardt BR, Melton LJ. History of the Rochester Epidemiology Project: half a century of medical records linkage in a US population. *Mayo Clin Proc*. 2012;87(12):1202-1213. doi:10.1016/j.mayocp.2012.08.012

159. American Diabetes Association. Executive summary: standards of medical care in diabetes—2014. *Diabetes Care*. 2014;37 Suppl 1(January):S5-13. doi:10.2337/dc14-S005

160. Davis RB, Anderson JR. Exponential survival trees. *Stat Med*. 1989;8(8):947-961.

161. Olshen LBJHFRA, Stone CJ. Classification and regression trees. *Wadsworth Int Gr*. 1984.

162. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines. 1997.

163. Srikant R, Agrawal R. Mining generalized association rules. *Futur Gener Comput Syst*. 1997;13(2-3):161-180. doi:10.1016/S0167-739X(97)00019-8

164. Turner RC, Millns H, Neil HA., et al. Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: United Kingdom prospective diabetes study (UKPDS: 23). *Bmj*. 1998;316(7134):823-828. doi:10.1136/bmj.316.7134.805

165. Nichols GA, Gullion CM, Koro CE, Ephross SA, Brown JB. The Incidence of Congestive Heart Failure in Type 2 Diabetes. *Diabetes Care*. 2004;27(8):1879-1884. doi:10.2337/diacare.27.8.1879

166. Cambridge M n. p. OptumLabs. OptumLabs and OptumLabs Data Warehouse (OLDW) Pre-Approved Language. PDF. Repro.

167. Zhang P, Sun Z, Wang F, Hu J. Towards Computational Drug Repositioning : A Comparative Study of Single- task and Multi-task Learning. In: *AMIA 2009 Symposium ProceedingsAnnual Symposium Proceedings Vol. 2015*. Vol 401. ; 2015:169-170.

168. Bickel S, Bogojeska J, Lengauer T, Scheffer T. Multi-Task Learning for HIV Therapy Screening. In: *Proceedings of the 25th International Conference on Machine Learning*.

Vol 1. ; 2008:56-63. doi:10.1145/1390156.1390164

169. CDC - Risk Factors for Complications. https://www.cdc.gov/diabetes/statistics/risk_factors_national.htm. Accessed April 10, 2017.

170. What are the Stages of Chronic Kidney Disease (CKD)? National Kidney Foundation. https://www.kidney.org/atoz/content/gfr. Accessed June 27, 2016.

171. Cox DR, Society S, Methodological SB. Regression Models and Life-Tables. *J R Stat Soc*. 1972;34(2):187-220. doi:10.2307/2985181

172. Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. *J Mach Learn Res*. 2010;11:1833-1863. doi:10.1109/ICDM.2009.108

173. Malik S, Wong ND, Franklin SS, et al. Impact of the metabolic syndrome on mortality from coronary heart disease, cardiovascular disease, and all causes in United States adults. *Circulation*. 2004;110(10):1245-1250. doi:10.1161/01.CIR.0000140677.20606.0E

174. Chambers JM, Cleveland WS, Kleiner B TP. *Graphical Methods for Data Analysis*. Taylor & Francis Ltd; 1983.

175. Osborn CY, Groot M, Wagner JA. Racial and Ethnic Disparities in Diabetes Complications in the Northeastern United States: the Role of Socioeconomic Status. *J Natl Med Assoc*. 2013;105(1):51-58.

176. Florez JC. Precision medicine in diabetes: Is it time? In: *Diabetes Care*. Vol 39. ; 2016:1085-1088. doi:10.2337/dc16-0586

177. Perveen S, Shahbaz M, Keshavjee K, et al. A Systematic Machine Learning Based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression. *Comput Struct Biotechnol J*. 2017;13(December 2016):1445-1454. doi:10.1177/117693510600200030

178. Bengio Y, Delalleau O, Simard C. Decision Trees Do Not Generaliza To New Variations. *Comput Intell*. 2010;26(4):449-467.

179. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Crown WH. Optum labs: Building a novel node in the learning health care system. *Health Aff*. 2014;33(7):1187-1194. doi:10.1377/hlthaff.2014.0038

180. OptumLabs. OptumLabs and OptumLabs Data Warehouse (OLDW) Descriptions and Citation. *Cambridge, MA n.p,*. 2018;PDF(Reproduced with permission from OptumLabs).

181. American Diabetes Association (ADA). Standards of Medical Care in Diabetes - 2017. *Diabetes Care*. 2017;40 (sup 1)(January):s4-s128. doi:10.2337/dc17-S001

182. Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B. and Wei, L.J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. 2011;30(10):1105-1117. doi:10.1109/TMI.2012.2196707.Separate

183. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. Translating the A1C assay into estimated average glucose values. *Diabetes Care*. 2008;31(8):1473-1478. doi:10.2337/dc08-0545

184. Kim E, Pieczkiewicz DS, Castro MR, Caraballo PJ, Simon GJ. Multi-Task Learning to Identify Outcome-Specific Risk Factors that Distinguish Individual Micro and Macrovascular Complications of Type 2 Diabetes. *AMIA 2018 Informatics Summit Proc*. 2018.

185. Evans GW, Byington RP, David C, et al. Effects of Intensive Blood-Pressure Control in Type 2 Diabetes Mellitus. *N Engl J Med*. 2010;362(17):1575-1585. doi:10.1056/NEJMoa1001286

186. Schulz KF, Altman DG, Moher D, et al. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Med*. 2010;8(1):18. doi:10.1186/1741-

7015-8-18

187. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Int J Surg*. 2014;12(12):1495-1499. doi:10.1016/j.ijsu.2014.07.013

188. Bossuyt PM, Reitsma JB, Bruns DE, et al. RESEARCH METHODS & REPORTING STARD 2015 : an updated list of essential items for. *Radiographies*. 2015;277(3):1-9. doi:10.1136/bmj.h5527

189. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol*. 2015;67(6):1142-1151. doi:10.1016/j.eururo.2014.11.025

190. Ahmed I, Debray TPA, Moons KGM, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Med Res Methodol*. 2014;14(1):3. doi:10.1186/1471-2288-14-3

191. Abo-Zaid G, Sauerbrei W, Riley RD. Individual participant data meta-analysis of prognostic factor studies: state of the art? *BMC Med Res Methodol*. 2012;12:56. doi:10.1186/1471-2288-12-56

192. Dekkers OM, von Elm E, Algra A, Romijn JA, Vandenbroucke JP. How to assess the external validity of therapeutic trials: A conceptual approach. *Int J Epidemiol*. 2010;39(1):89-94. doi:10.1093/ije/dyp174

193. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng*. 2005;17(3):299-310. doi:10.1109/TKDE.2005.50

194. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;45(4):427-437. doi:10.1016/j.ipm.2009.03.002

195. Bibal A. Interpretability of Machine Learning Models and Representations : an Introduction. *ESANN 2016 proceedings, Eur Symp Artif Neural Networks, Comput Intell Mach Learn*. 2016;(April):77-82.

196. Lipton ZC. The Mythos of Model Interpretability. *ICML Work Hum Interpret Mach Learn*. 2016;(Whi).

197. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. 2017;(February 2017):1-11. doi:10.1093/bib/bbx044

198. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ; 2016:1135-1144. doi:10.1145/1235

199. Maier W, Holle R, Hunger M, et al. The impact of regional deprivation and individual socio-economic status on the prevalence of Type 2 diabetes in Germany. A pooled analysis of five population-based studies. *Diabet Med*. 2013;30(3):e78-86. doi:10.1111/dme.12062

200. Hu R, Shi L, Rane S, Zhu J, Chen CC. Insurance, Racial/Ethnic, SES-related Disparities in Quality of Care Among US Adults with Diabetes. *J Immigr Minor Heal*. 2014;16(4):565-575. doi:10.1007/s10903-013-9966-6

201. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: A clinical example. *J Clin Epidemiol*. 2003;56(9):826-832. doi:10.1016/S0895-4356(03)00207-5

202. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model transparency and validation: A report of the ISPOR-SMDM modeling good research practices task force-7. *Med Decis Mak*. 2012;32(5):733-743.

doi:10.1177/0272989X12454579

203. Chawla, A., Chawla, R. and Jaggi S. Microvasular and macrovascular complications in diabetes mellitus: Distinct or continuum? *Indian J Endocrinol Metab*. 2016;20(4):546.

204. Chrysant SG, Chrysant GS. The Age-Related Hemodynamic Changes of Blood Pressure and Their Impact on the Incidence of Cardiovascular Disease and Stroke: New Evidence. *J Clin Hypertens*. 2014;16(2):87-90. doi:10.1111/jch.12253

205. Safar ME, Balkau B, Lange C, et al. Hypertension and vascular dynamics in men and women with metabolic syndrome. *J Am Coll Cardiol*. 2013;61(1):12-19. doi:10.1016/j.jacc.2012.01.088

206. Safar ME. Pulse pressure, arterial stiffness, and cardiovascular risk. *Curr Opin Cardiol*. 2000;15:258-263.

207. Chrysant SG, Chrysant GS. Effectiveness of lowering blood pressure to prevent stroke versus to prevent coronary events. *Am J Cardiol*. 2010;106(6):825-829. doi:10.1016/j.amjcard.2010.05.006

208. Messerli FH, Mancia G, Conti CR, et al. Dogma disputed: can aggressively lowering blood pressure in hypertensive patients with coronary artery disease be dangerous? *Ann Intern Med*. 2006;144(12):884-893. doi:10.1016/j.jemermed.2006.11.010

209. Madhavan S, Ooi WL, Cohen H, Alderman MH. Relation of pulse pressure and blood pressure reduction to the incidence of myocardial infarction. *Hypertension*. 1994;23(3):395-401. doi:10.1161/01.HYP.23.3.395

210. Freitas AA. Comprehensible Classification Models – a position paper. *ACM SIGKDD Explor Newsl*. 2013;15(1):1-10. doi:10.1145/2594473.2594475