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

Era Kim, MS[1,2], David S. Pieczkiewicz, PhD[1], M. Regina Castro, MD[3], Pedro J. Caraballo, MD[3], Gyorgy J. Simon, PhD[1]
[1]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA;
[2]Medica Research Institute, Minnetonka, MN, USA and OptumLabs Visiting Fellow;
[3]Mayo Clinic, Rochester, MN, USA

Abstract

Because deterioration in overall metabolic health underlies multiple complications of Type 2 Diabetes Mellitus, a substantial overlap among risk factors for the complications exists, and this makes the outcomes difficult to distinguish. 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 aimed to examine feasibility of our proposed methodology that separates these two roles, thereby, improving interpretation of predictions and helping prioritize which complication to target first. To separate these two roles, we built models for six complications utilizing Multi-Task Learning—a machine learning technique for modeling multiple related outcomes by exploiting their commonality—in 80% of EHR data (N=9,793) from a university hospital and validated them in remaining 20% of the data. Additionally, we externally validated the models in claims and EHR data from the OptumLabs™ Data Warehouse (N=72,720). Our methodology successfully separated the two roles, revealing distinguishing outcome-specific risk factors without compromising predictive performance. We believe that our methodology has a great potential to generate more understandable thus actionable clinical information to make a more accurate and timely prognosis for the patients.

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)[1]. 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[2–4]. 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[5–7], 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)[8] 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[9,10]. 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[11] 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.

Materials and Methods

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.

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 1).

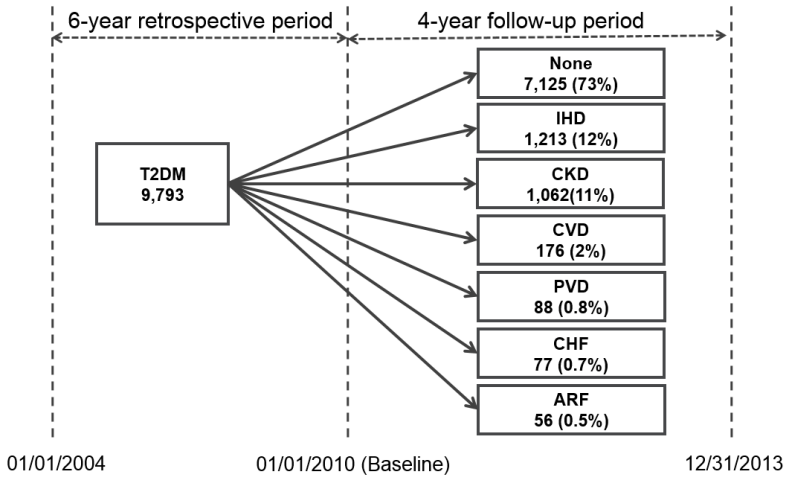

Figure 1. Study design

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

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.

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[8]. 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 1) and selection procedure.

Baseline Patient Characteristics in UMMC and OLDW Datasets

Table 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.)[12], and they had signs of established CKD based on GFR[13]. Compared to UMMC patients, OLDW patients were older and had better HbA1c, better lipids, better kidney function, but higher SBP and DBP.

Table 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

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

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

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 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[14] 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.

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[15]. 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.

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.

Coefficients from Multi-Task Learning Methodology

Figure 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.

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 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 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 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[16]. 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[17,18].

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 3) with α coefficients from General Progression Model (Figure 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

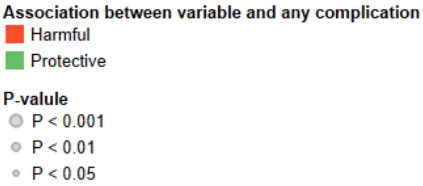

Figure 3. Coefficients for Risk of Developing *Any* Complication

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)[19].

General Progression showed that a unit increase in HbA1c conferred a HR of 1.039 (exp(.0385)) on all complications uniformly (Figure 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 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 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.

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[20] in Figure 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.

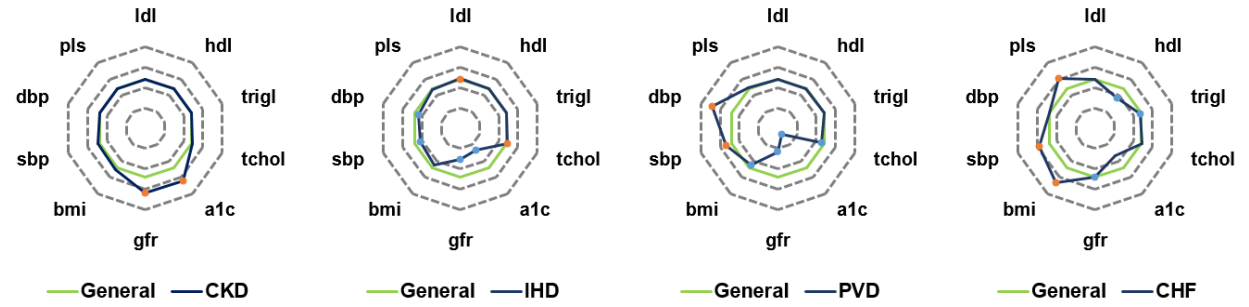

Figure 4. Characteristic Shapes of Differential Markers for CKD, IHD, PVD, and CHF

IHD, PVD, and CHF are well-known concomitant macrovascular complications. They share similar pathophysiology and are believed to have similar risk factors[21]. Given these facts, distinguishing among them without a methodology like ours is more difficult. In Figure 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[16]. 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).

Coefficients from Reference Methodology

Figure 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.

Association between variable and complication
Harmful Protective

Statistical significance
P < 0.001 P < 0.01 P < 0.05

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 5. Coefficients from Baseline Methodology

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 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 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.

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 1). However, except CKD, C-Indices were still within 95CIs.

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

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

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[22]. 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.

Acknowledgements

This work was partially supported by the NIH award LM011972 and the NSF award IIS 1602198. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. Alberti KGMM, Eckel RH, Grundy SM, et al. Harmonizing the Metabolic Syndrome. *Circulation*. 2009;120(16):1640-1645. doi:10.1161/CIRCULATIONAHA.109.192644.
2. Saltiel AR, Kahn CR. Insulin signaling and the regulation of glucose and lipid metabolism. *Nature*. 2001;414(6865):799-806.
3. Kim E, Oh W, Pieczkiewicz DS, Castro MR, Caraballo PJ, Simon GJ. Divisive Hierarchical Clustering towards Identifying Clinically Significant. *AMIA 2014 Symp Proc*. 2014.
4. 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.
5. 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.
6. 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.
7. 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.
8. Cambridge M n. p. OptumLabs. OptumLabs and OptumLabs Data Warehouse (OLDW) Pre-Approved Language. PDF. Repro.
9. 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.
10. 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.
11. 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.
12. CDC - Risk Factors for Complications. https://www.cdc.gov/diabetes/statistics/risk_factors_national.htm. Accessed April 10, 2017.
13. What are the Stages of Chronic Kidney Disease (CKD)? National Kidney Foundation. https://www.kidney.org/atoz/content/gfr. Accessed June 27, 2016.
14. Cox DR, Society S, Methodological SB. Regression Models and Life-Tables. *J R Stat Soc*. 1972;34(2):187-220. doi:10.2307/2985181.
15. Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. *J Mach Learn Res*. 2010;11:1833-1863. doi:10.1109/ICDM.2009.108.
16. 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.
17. 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.
18. Isomaa B, Almgren P, Tuomi T, et al. Cardiovascular Morbidity and Mortality Associated With the Metabolic Syndrome. *Diabetes Care*. 2001;24(4).

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

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

21. 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.

22. 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.

Appendices

Variable	P-value						
	General	CKD	ARF	IHD	PVD	CHF	CVD
a1c	0.008	0.049	0.340	0.027	0.005	0.058	0.327
ldl	0.329	0.117	0.289	0.034	0.265	0.259	0.256
hdl	0.058	0.281	0.343	0.123	0.145	< 0.001	0.283
trigl	0.014	0.297	0.307	0.059	0.079	< 0.001	0.278
tchol	0.021	0.057	0.25	0.039	0.023	0.059	0.234
gfr	< 0.001	< 0.001	0.265	< 0.001	< 0.001	0.251	0.255
gfr_norm	< 0.001	< 0.001	0.261	< 0.001	< 0.001	0.025	0.237
bmi	0.042	0.067	0.333	0.051	0.034	< 0.001	0.272
pls	0.188	0.290	0.320	0.105	0.076	< 0.001	0.294
sbp	0.01	0.076	0.31	0.002	0.003	< 0.001	0.272
dbp	0.006	0.261	0.307	0.029	< 0.001	0.289	0.266
never_smoker	0.002	0.089	0.017	0.098	< 0.001	< 0.001	0.317
age	< 0.001	0.289	0.311	0.045	< 0.001	< 0.001	< 0.001
male	0.107	0.094	0.321	0.019	< 0.001	0.008	0.312

Table A-1. P-values of Coefficients from Multi-Task Learning Methodology

Variable	P-value					
	CKD	ARF	IHD	PVD	CHF	CVD
a1c	0.033	< 0.001	0.132	0.023	0.164	0.007
ldl	0.052	< 0.001	0.043	0.101	0.289	0.045
hdl	0.088	0.014	0.101	0.333	< 0.001	0.024
trigl	0.333	0.010	0.023	0.025	0.003	0.162
tchol	0.269	< 0.001	0.029	0.045	0.239	0.271
gfr	< 0.001	< 0.001	0.006	< 0.001	0.021	0.019
gfr_norm	< 0.001	< 0.001	0.007	< 0.001	0.258	0.033
bmi	0.085	0.102	0.101	0.118	< 0.001	0.184
pls	0.348	< 0.001	0.121	0.081	< 0.001	0.088
sbp	0.055	0.005	0.024	< 0.001	< 0.001	0.007
dbp	0.154	0.004	0.117	< 0.001	0.138	0.159
never_smoker	0.344	< 0.001	0.016	< 0.001	< 0.001	0.128
age	< 0.001	< 0.001	0.021	< 0.001	< 0.001	< 0.001
male	0.347	< 0.001	0.059	< 0.001	0.002	0.041

Table A-2. P-values of Coefficients from Baseline Methodology