Modeling Trajectories for Diabetes Complications

Pranjul Yadav [*] Lisiane Pruinelli [†] Andrew Hangsleben [*] Sanjoy Dey [*]
Katherine Hauwiller [*] Bonnie L Westra [†‡] Connie W Delaney [†‡] Vipin Kumar [*]
Michael Steinbach [*] Gyorgy J Simon [‡]

Abstract

Diabetes mellitus (DM) is a prevalent and costly disease and if not managed effectively, it leads to complications in almost every body system. Evidence-based guidelines for prevention and management of DM exist, but they ignore the trajectory along which the disease developed. With the implementation of electronic health records (EHRs), sufficiently detailed data are available to elucidate diabetes trajectories and sequences of diabetes-related comorbidities across subpopulations. As a first step, we developed a Diabetes Mellitus Complication Index (DMCI) using EHR data to summarize the patients' condition pertinent to diabetes into a single score. Next, we modeled trajectories of developing diabetic complications based on groups of patients with varying diabetic complications at baseline. Such knowledge can form the basis for future trajectory-centered guidelines.

1 Introduction and Background

Diabetes mellitus (DM) affects 11.3% (25.6 million) of Americans age 20 or older and is the seventh leading cause of death in the United States[1]. There is considerable research on risk factors to predict and manage diabetic outcomes[1]. Without appropriate management of diabetes, patients are at risk for secondary diseases in almost every body system at later time points. Evidence based practice (EBP) guidelines for management and prevention of diabetic complications synthesize the latest scientific evidence. While EBP guidelines have been shown to improve care, they neither consider the patient's trajectory nor the sequence of events that lead up to the patient's current conditions. In this work, we show that such information is invaluable; a patient's risk of developing further complications depends on their trajectory thus far.

Simple disease models describing a single typical diabetes trajectory as a sequence of successively worsening conditions exist[2]. However, these models were aimed more at patient education than at a physiologically accurate description of the evolution of the underlying disease pathology. Such simple models obviously cannot form the basis of evidence based guidelines.

In heterogeneous diseases, analyzing the data on a per-subpopulation basis has been shown to elucidate more interesting patterns than analyzing the entire population[3, 4]. In this work, we hypothesize, with abundant supporting evidence[5], that diabetes and the underlying metabolic syndrome follows multiple trajectories. We aim to develop a methodology that is capable of elucidating scientifically accurate diabetes trajectories retrospectively from the extensive clinical data repository of a large Midwestern health system. Specifically, we study a diabetic population and track changes to their health over time in terms of diabetes-related comorbidities as documented in the electronic health record (EHR).

Diabetes, its severity and the ensuing complications can be described most accurately through a large number of correlated EHR data elements, including associated diagnoses, laboratory results and vitals. The relationships among these data elements, known as multicollinearity, render efforts to track patients' conditions across time fraught with data overfitting issues. To contain the collinearity problem we summarize the patients' condition into a single dimension (a single score), which we term the Diabetes Mellitus Complication Index (DMCI).

The development of severity indices from EHRs looks back on a rich history. Even in the context of DM, several risk scores for diabetes from EHRs have been developed[6]. Most risk score models focus on predict-

[*]Department of Computer Science And Engineering, University of Minnesota
[†]School of Nursing, University of Minnesota
[‡]Institute for Health Informatics, University of Minnesota

ing the risk of diabetes rather than the risk of the associated complications. Two risk scores have specifically focused on diabetes complications[7, 8] to predict outcomes; however their diabetes complication indices were limited to the use of complications based on International Classification of Diseases (ICD) codes alone[8] or asking patients if they were ever informed that they had DM complications[5]. In a diabetic population like ours, good predictors of the complications do not necessarily coincide with good predictors of diabetes given that the metabolic syndromes in our patients have already evolved past diabetes. The inclusion of additional variables, such as lab results and vital signs may provide useful information for early prediction of complications. This necessitates the development of a new diabetes complication index to be used in our effort to study patient trajectories.

Our work makes the following novel contributions. First, we develop DMCI which summarizes a patient's health in terms of post-diabetic complications into a single score. Second, through the use of this score, we track a patient's health and show that distinct trajectories in diabetes can be identified, demonstrating the need and laying the foundation for future clinical EBP guidelines that take trajectories into account.

2 Data Preparation

After Institutional Review Board (IRB) approval, a de-identified data set was obtained from a Midwest University's clinical data repository (CDR). The CDR contains over 2 million patients from a single Midwest health system that has 8 hospitals and 40 clinics. Data elements included various EHRs attributes, such as demographic information (age, gender), vital signs: systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse, and body mass index (BMI); and laboratory test results: glomerular filtration rate (GFR), hemoglobin A1c, low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL), triglycerides and total cholesterol. Further ICD-9 codes related to both Type 1 and Type 2 DM, and their accompanied complications such as ischemic heart disease (IHD), cerebrovascular disease (CVD), chronic kidney disease (CKD), congestive heart failure (CHF), peripheral vascular disease (PVD), Diabetic Foot, and Opthalmic complications were used in this study.

3 Study Design and Cohort Selection

For our study, we used Jan. 1, 2009 as a baseline. The study cohort consists of patients with type 1 or type 2 DM at baseline, identified in billing transactions. Patient were included if they had at least two A1c results at least 6 months apart after baseline. Patients with no laboratory results or vitals before 2009 were excluded on the basis that they show no indication of receiving primary care at the health system. The final cohort consists of 13,360 patients. Patients' initial DMCI was determined at baseline, and their health (in terms of the DMCI score) was followed until last the follow-up. The mean time for follow-up was 1568 with a standard deviation of 263 days.

4 Diabetes Risk Score Development

The novel DMCI was developed using Cox proportional hazards survival modeling techniques. Each of the 6 complications (CKD, CVD, CHF, PVD, IHD, Diabetic Foot) were modeled through a separate Cox regression model using patients who did not already present with the complication at baseline. Cox Proportional Hazard Models are survival models which estimate the hazard $\lambda_j(t)$ for patient j at time t based on covariates Z_j and a baseline hazard $\lambda_o(t)$. The hazard function has the form

$$\lambda_j(t/Z_j) = \lambda_o(t)exp(Z_j\beta)$$

The co-efficient vector β is estimated through maximizing the partial likelihood. The partial likelihood can be maximized using the Newton-Raphson algorithm. The partial likelihood has the form

$$L(\beta) = \pi_{i:C_i=1} \frac{\theta_i}{\sum_{j:Y_j \geq Y_i} \theta_j}$$

θ_j has the form $exp(Z\beta)$ and let C_i be the indication function. C_i is 1 if the event occurred and $C_i = 0$ to represent censoring. The baseline hazard is common to all patients. Besides the complications (except for the one we are modeling), age, gender, obesity, hypertension and hyperlipidemia diagnosis, laboratory test results and vitals, were included as covariates. Backwards elimination was employed for variable selection. Each of the 7 regression models (one for each complication) provided an estimate of the coefficients, which can be interpreted as the relative risk of developing the complication in question.

The DMCI score is the weighted sum of the linear prediction from the seven regression models. The seven weights are determined by the performance of the corresponding regression model on a leave-out validation set. Table 1 represents weights assigned to each model, with the respective complication as the outcome.

Table 1. Weights For Individual Regression Model

Complication	Model Weight
CHF	0.787
IHD	0.569
CVD	0.694
PVD	0.688
CKD	0.758
FOOT	0.712

Therefore, the DMCI score can be thought of as approximately 7 times the relative risk a patient faces in developing a complication (any diabetic complication). Patient's risk from individual regression model was computed using the equation below,

$r_i = Z_i * \beta$, where in r_i denotes the patients risk, Z_i represents the covariates and β are the coefficients learned.

5 Subpopulation Trajectory Extraction

Using the DMCI score, the health status trajectory of every patient from 2009 onwards was calculated. Patients and their trajectories (time stamped sequence of DMCI scores) were grouped by complications. First, we considered a single complication at a time, creating seven categories: patients presenting with CKD, CVD, etc. at baseline. A patient presenting with multiple complications falls into all applicable categories. Next, we considered pairs of complications: e.g. a possible category consists of patients with IHD and diabetic foot problems.

For every category (sub-population of patients), the shape of the DMCI score trajectory was determined through segmented linear regression with 3 knots. One can think about these regression models as a straight line with one elbow (at $\hat{x}$). These trajectories can be expressed in the form

$$\text{y} = \begin{cases} a * x + b, & \text{if } x < \hat{x} \\ c * x + d, & \text{if } x \geq \hat{x} \end{cases}$$

where in a,b,c,d ϵ R.

Residual sum of squares (RSS) was used as the objective function to obtain the coefficients of the segmented linear regression and the location of the elbow point. RSS has the form

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

In the equation above, y_i is the risk at i_{th} time stamp and $\hat{y_i}$ is the corresponding risk computed using segmented linear regression.

6 Results

Table 2 provides the count of patients in various cohorts. Ophthalmic conditions are no longer considered as they have insufficient patient coverage (less than 100 patient). Table 3 provides the count for various populations with comorbidities. We modeled ophthalmic comorbidities, but there were insufficient numbers, so we excluded this group from further analysis.

Table 2. Patient Counts for Single DM Comorbidity

Comorbidity	Count	Comorbidity	Count
IHD	4398	CHF	741
CVD	986	PVD	662
CKD	742	Foot	267

Table 3. Patient Counts for DM Comorbidities

Comorbidity	Count	Comorbidity	Count
IHD, CVD	457	IHD, PVD	379
IHD, CKD	361	IHD, Foot	662
IHD, CHF	478		

Figure 1 presents the DMCI trajectory for varying subpopulations. The horizontal axis denotes time since baseline in days and the vertical axis corresponds to the DMCI score. Each curve in the graph represents a subpopulation defined by a single complication. For example, the bottommost curve corresponds to patients presenting with CHF at baseline. Their average risk of developing a complication (other than CHF, which they already have) is 4.4 at baseline, It increases steadily for approximately 550 days, at which point it reaches 4.7 and then it becomes flat (stops increasing materially going forward). As observed from the graph, the average risk associated with patients diagnosed with CKD is comparatively higher than that of patients diagnosed with CHF.

Figure 1 shows that (i) subpopulations defined by various complications at baseline have a different average risk at baseline. This information is readily incorporated into existing indices and guidelines. The figure also shows that (ii) these patients have different patterns of risk moving forward. For example, the risk of developing a complication increases sharply for CHF patients for 550 days and then becomes flat. In contrast, the risk of IHD increases steadily (but at a lower rate) throughout the observation period; and CKD (topmost curve) increases at a much lower rate.

Table 4. Distribution of Scores for Different Subgroups

Complication	Min-Risk	Risk-25	Risk-50	Risk-75	Max-Risk
IHD	-6.02	1.86	3.92	5.98	22.68
CHF	-5.17	1.98	3.86	5.91	12.55
PVD	-5.23	2.07	3.90	6.20	14.37
CKD	-5.62	1.77	3.94	5.94	14.71
CVD	-6.72	2.16	4.00	6.04	14.37
Diabetic Foot	-4.64	2.08	3.95	6.08	14.37

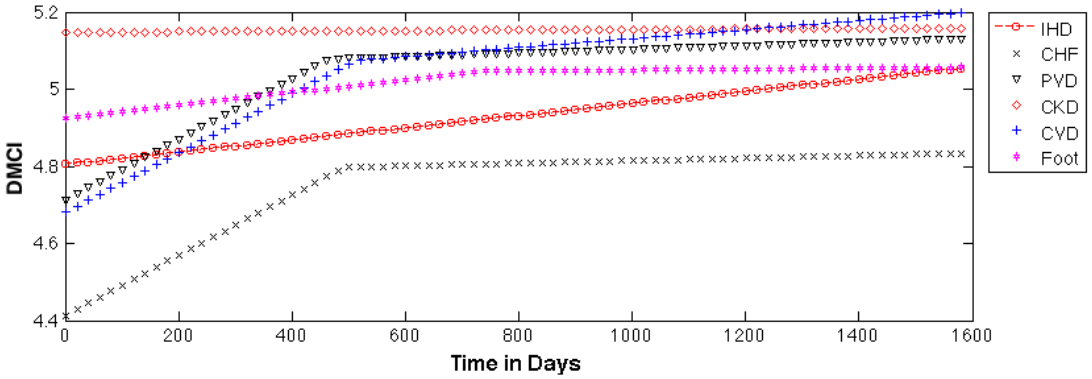

Figure 1: Health status trajectory for varying subpopulations

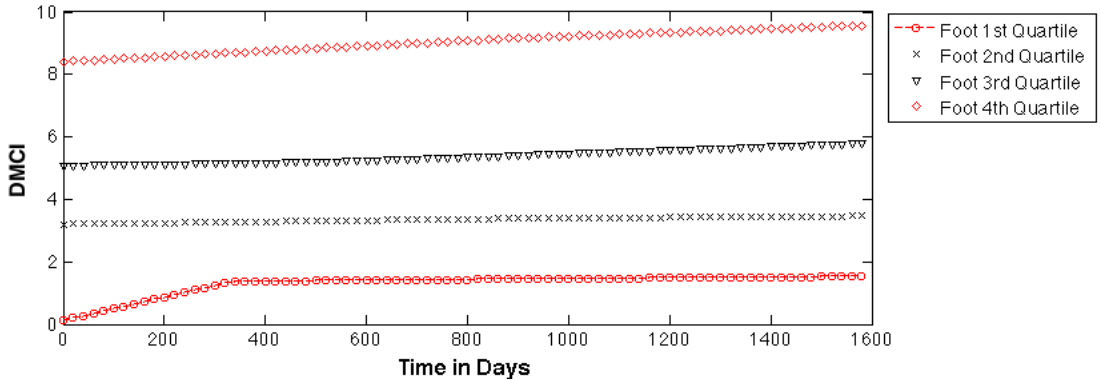

Figure 2: Shape of the individual quartiles for patients diagnosed with diabetic foot

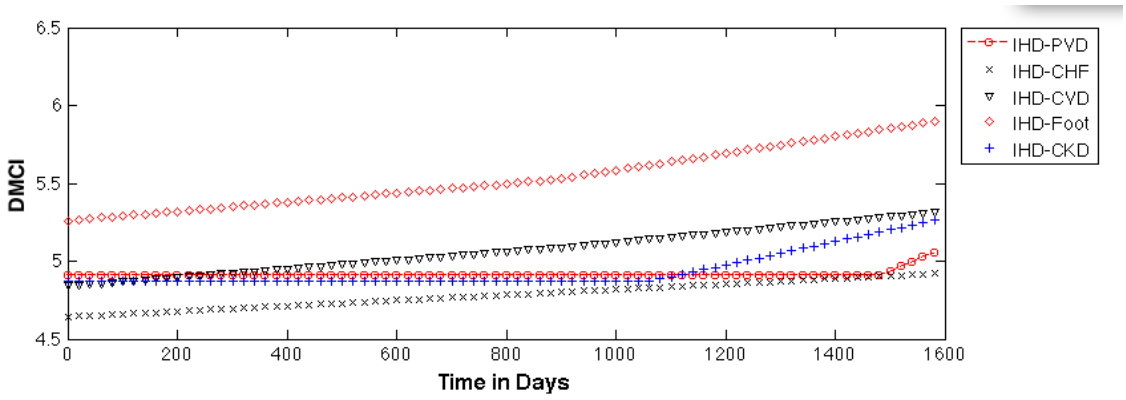

Figure 3: Shape of the individual quartiles for patients diagnosed with diabetic foot

Figure 1 presents the average risk for each population. To illustrate the distribution of the risk, in Table 3, we provide the interquartile range of the DMCI score in each subpopulation. Using the information from table 1, the risk trajectories of patients belonging to the top 25 in their respective subgroups were analyzed. Figure 2 presents the average behavior for the highest-risk quartile.

In order to investigate whether the shape of the health-risk trajectory for each quartile within a subgroup is similar, the patterns for each quartile for multiple subpopulations were explored. In Figure 2, the shape of the individual quartiles for patients diagnosed with diabetic foot is depicted. The figure shows that having different risk at baseline only tells a part of the story. These patients not only have different risks, but they also exhibit different progression patterns: their DMCI curves have different shapes.

Figure 3, depicts the trajectories of patients with IHD and an additional complication. The results suggest that even in a subpopulation defined by a single complication, significant heterogeneity exists, as evidenced by differing shapes of the trajectory curves.

7 Discussion And Conclusion

The purpose of this study was to model patients' progression towards diabetes complications through the use of a novel DMCI derived from EHR data. The DMCI was used to stage the patients' health in terms of diabetic complications. Results clearly demonstrated the existence of multiple trajectories in diabetes thereby alluding to the complex heterogeneity of the disease. Specifically, we divided patients into multiple (potentially overlapping) subpopulations based on their baseline complications and confirmed that patients with different baseline complications have different risks of developing additional complications. Secondly, we have also shown that these patient subpopulations differ not only in their risk but also in the temporal behavior of their risk: patients in certain subpopulations 'accrue 'risk at a higher rate initially and at a slower rate later, while the DMCI score in patients in other subpopulations increases at a steady rate throughout the follow-up period. Thirdly, we have also demonstrated that the trajectories differ even within the same patient subpopulation. Patients presenting with additional complications (e.g. a second complication on top of IHD) have different risks and different trajectories. Finally, we have also shown that when we stratify patients within the same subpopulation by their baseline risk, they exhibit different trajectories. This can naturally be a consequence of these patients suffering from additional complications explaining their increased relative baseline risk.

These findings support a conclusion in a previous study that patient subgroups vary by level of severity. Dey et al. [4] used a national convenience sample of 581 Medicare-certified HHC agencies' EHRs for 270,634 patients to understand which patients are likely to improve in their mobility and found that mobility status at admission was the single strongest predictor of mobility improvement[4]. However, very different patterns were apparent when conducting the analysis within the level of severity for mobility at admission.

An interesting finding in our study is that patients with diabetic foot problems have the highest severity at base line, and more so when combined with IHD. This finding may be associated with the strict relationship between glycemic control and microvascular complications. Foot problems are associated both with nerve and vascular damage, creating a risk for infections. Uncontrolled glucose further exacerbates the potential for severe infec-

tions and potential amputations. Patients with diabetic foot complications are likely to continue having an increasing risk for additional problems, as foot problems are a leading cause of hospital admission, amputation, and mortality in diabetes patients.[9]
Through our previous work[3] in investigating diabetic subpopulations and their risk of mortality, we have already gained an appreciation of the immense heterogeneity of diabetes and the metabolic syndrome. Studying trajectories expands this heterogeneity along a new dimension. While the preliminary work presented in this study merely offers a glimpse at the complexity of diabetes and its complications, it unquestionably demonstrates the value of trajectories in understanding patient progression and possibly prognosis. Further research in this direction will undoubtedly lead to improvements in EBP guidelines by taking trajectories into account.
Limitations of this study include the secondary use of EHR data and its associated challenges. The data in this study represent care provided in a single health system; the study needs replication in additional health settings and under different clinical conditions. The DMCI score was developed from EHR data retrospectively and independent validation would be beneficial.

8 Acknowledgement

This study is supported by National Science Foundation (NSF) grant: IIS-1344135. Contents of this document are the sole responsibility of the authors and do not necessarily represent official views of the NSF.

This was partially supported by Grant Number 1UL1RR033183 from the National Center for Research Resources (NCRR) of the National Institutes of Health (NIH) to the University of Minnesota Clinical and Translational Science Institute (CTSI).

References

[1] *Centers for Disease Control and Prevention (CDC). National diabetes fact sheet. Available from: http://www.cdc.gov/diabetes/pubs/pdf/ndfs2011.pdf.*

[2] B. Ramlo-Halsted, S. Edelman *The natural history of type 2 diabetes: practical points to consider in developing prevention and treatment strategies*, Clinical diabete, 2000.

[3] V. Kumar and P. Yadav, *Progression and risk assessment of comorbid conditions in type 2 Diabetes Mellitus*, Biomedical Informatics and Computational Biology (BICB) Symposium, Rochester, MN, 2014.

[4] S. Dey, *Data mining to predict mobility outcomes in home health care*

[5] T. Tuomi, N. Santoro, S. Caprio, M. Cai, J. Weng, L. Groop. *The many faces of diabetes: a disease with increasing heterogeneity.* Lancet, 2014; 383: pp. 1084-94.

[6] G. Collins, S. Mallett, O. Omar, L. Yu. *Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting.* BMC Med. 2011;9:103-7015-9-103.

[7] B. Fincke, J. Clark , M. Linzer , et al. *Assessment of long-term complications due to type 2 diabetes using patient self-report: the diabetes complications index.* J Ambul Care Manage. 2005;28(3): pp. 262-273.

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

[9] N. Avitabile, A. Banka, V. Fonseca. *Glucose control and cardiovascular outcomes in individuals with diabetes mellitus: lessons learned from the megatrials.* Heart Failure Clinics. 2012;8(4):pp. 513-522.