ORIGINAL ARTICLE

Type 2 Diabetes Mellitus Trajectories and Associated Risks

Wonsuk Oh,[1] Era Kim,[1] M. Regina Castro,[2] Pedro J. Caraballo,[3] Vipin Kumar,[4] Michael S. Steinbach,[4] and Gyorgy J. Simon[1,5,*]

Abstract

Disease progression models, statistical models that assess a patient's risk of diabetes progression, are popular tools in clinical practice for prevention and management of chronic conditions. Most, if not all, models currently in use are based on gold standard clinical trial data. The relatively small sample size available from clinical trial limits these models only considering the patient's state at the time of the assessment and ignoring the trajectory, the sequence of events, that led up to the state. Recent advances in the adoption of electronic health record (EHR) systems and the large sample size they contain have paved the way to build disease progression models that can take trajectories into account, leading to increasingly accurate and personalized assessment. To address these problems, we present a novel method to observe trajectories directly. We demonstrate the effectiveness of the proposed method by studying type 2 diabetes mellitus (T2DM) trajectories. Specifically, using EHR data for a large population-based cohort, we identified a typical trajectory that most people follow, which is a sequence of diseases from hyperlipidemia (HLD) to hypertension (HTN), impaired fasting glucose (IFG), and T2DM. In addition, we also show that patients who follow different trajectories can face significantly increased or decreased risk.

Key words: big data analytics; data mining

Introduction

The advent of electronic health record (EHR) systems has paved the way to perform large-scale data analytics to uncover new medical knowledge that was previously inaccessible. EHR systems store information about entire populations and offer long follow-up times. In this study, we work with data from a premier healthcare provider in the Midwestern United States, which pioneered the adoption of EHR systems in the region, allowing us access to nearly 13 years of follow-up time for a relatively large number of patients. Such long follow-up, in turn, allows us to study disease trajectories that lead to type 2 diabetes mellitus.

Type 2 diabetes mellitus (T2DM) is one of the fastest growing public health concerns in the United States.[1] There are 29.1 million patients (9.3% of the US populations) suffering from diabetes in 2014.[2] Diabetes, which is the seventh leading cause of death in the United States, is known to be a nonreversible (incurable) chronic disease,[3,4] leading to severe complications,[1,5] including chronic kidney disease, amputation, blindness, and various cardiac and vascular diseases. Early identification of patients at high risk is regarded as the most effective clinical tool to prevent or delay the development of T2DM, allowing patients to change their lifestyle or receive medication earlier. In turn, these interventions can help decrease the risk of diabetes by 30%–60%.[6,7] Many risk models[8–10] aiming at early identification of patients at high risk are widely used in the clinical settings. These models typically only consider the patient's current state at the time of the assessment and ignore the trajectory, the sequence of events, that led up to the state.

The motivating hypothesis for our work is to study whether the trajectory influences the risk of diabetes. Diabetes is a heterogeneous disorder involving complex biological mechanisms. In our study, we discovered multiple trajectories to diabetes that can help some of the underlying mechanisms and their associated risk of developing diabetes.

Access to 13 years of follow-up allows us to make inferences about trajectories leading up to T2DM. Since many diseases are progressive (worsen over time),

[1]Institute for Health Informatics; [4]Department of Computer Science and Engineering; University of Minnesota, Minneapolis, Minnesota.
[2]Endocrinology, Diabetes, Metabolism and Nutrition; [3]General Internal Medicine; [5]Department of Health Sciences Research; Mayo Clinic, Rochester, Minnesota.

*Address correspondence to: Gyorgy J. Simon, Health Sciences Research, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, E-mail: gsimon@gmail.com

EHR data with its large sample size and long follow-up time allow us the opportunity to study the progression of these diseases. However, due to the nature of EHR data, unlocking this potential is challenging. The challenge stems from two compounding factors. First, EHR data were not designed to be a research platform; thus, some critical data elements are not directly observable and need to be inferred. Second, chronic conditions have slow onset, and as a result, the onset time is not only unobservable but also difficult to estimate accurately. The purpose of this article was twofold: (1) we first describe the challenges we faced in using EHR data and the methods we developed to overcome those challenges, and (2) we then describe the interesting findings we uncovered.

Specifically, we define trajectories as sequences in which patients develop comorbidities as they progress toward T2DM. Besides T2DM, we consider three important comorbidities: hyperlipidemia (HLD, high cholesterol or unbalance of the various lipids), hypertension (HTN, high blood pressure), and impaired fasting glucose (IFG, elevated fasting plasma glucose). We infer the typical (most frequent) trajectory and enumerate the atypical trajectories that our data support. We build predictive models to determine whether following an atypical trajectory is associated with different risks of diabetes.

We perform our analysis on a large community-based cohort derived from EHR system in the Rochester Epidemiology Project[11] consisting of patients who received their primary care at the Mayo Clinic. The data have nearly 13 years of unfragmented follow-up, making it the largest and cleanest EHR-derived data set of its kind. In this article, we show that a single typical trajectory exists, and it is consistent with the trajectory that is commonly used for diabetes patient education. We enumerate several atypical trajectories that cover ∼27% of the diabetes cases observed in our data set and assess the excess risk (if any) they confer on patients following them.

Data and Challenges

Data

The study cohort consists of Mayo Clinic primary care patients residing in Olmsted County, MN. During the study period from 1999 to 2013, when complete EHR data were available, we have 70 k patients with research consent. Informed consent was obtained from patients during each visit, and consent information was stored in the EHR. Demographic information, diagnosis codes

encoded as ICD-9-CT, laboratory results, vital signs, and medication data were collected for this period.

Challenges

To establish trajectories, sequences in which the disease develops, we should only consider new (*incident*) diagnoses (as opposed to preexisting conditions) along with their onset dates. Surprisingly, this information is difficult to infer from the EHR system for the following reasons.

Secondary use of EHR data. EHR systems were originally developed for documenting patients' state for reimbursement purposes. The presence of diagnosis codes in the EHR is driven by billing rules. They may be present because the corresponding condition was tested, possibly newly discovered, or was complicating the treatment of other conditions. There is no designation in the EHR whether a diagnosis is incident or preexisting. Moreover, diagnoses may be missing (no reimbursement was requested for the condition) and can be false positive (the patient was merely tested for a condition).

Slow-onset conditions. The second issue concerns the onset date. The development of T2DM as well as the comorbidities that commonly precede it can take decades. The signs for these conditions are subtle and can remain undetected for years. Establishing the onset time for these conditions is challenging. Instead of trying to estimate the onset date, we only assume that it happened before the earliest recording date. Another issue regarding the slow progression is that even with 13 years of follow-up, we can only observe partial trajectories, that is, the development of only a few new conditions. Therefore, if we tried to observe, rather than infer, the sequences, we would focus on patients with the fastest progression, possibly biasing the results.

Study Design

A retrospective observational study design is adapted. We use January 1, 2005, as the baseline for our study. The period before the baseline, that is, 1999–2004, is called a prebaseline period. We use the prebaseline period to determine patients' baseline diabetes status and comorbidities by retrospectively examining their medical history through laboratory measurements, vitals, and diagnoses. Of particular interest is the presence of T2DM-related comorbidities HLD, HTN, and IFG at the baseline. We set a follow-up period of 2005–2013 to follow the patients and record whether they developed diabetes. The incidence of T2DM and its date

during the follow-up period were determined by a chart review.

The construction of the study cohort is described in Table 1. We included all adult patients with research consent and no diabetes diagnosis code at baseline. There are 69,747 such patients. From this cohort, we excluded patients with a high suspicion of diabetes (389 patients with fasting plasma glucose >125 mg/dL or those taking diabetes medications), unknown glucose value (14,559 patients), undetermined lipid status (1023 patients), and unknown blood pressure (498 patients). Our final study cohort consists of 43,509 patients, and 4795 of the 43,509 patients (11%) developed diabetes during the follow-up period.

To determine whether a patient has a particular comorbidity at the baseline, we use phenotyping algorithms. Phenotyping algorithms[12,13] are simple classifiers that infer the presence of a disease based on diagnoses, laboratory results, vitals, and medications. Specifically, in this study, we constructed three ordinal variables for IFG, HTN, and HLD as combinations of diagnoses, abnormal laboratory results (or vitals), and medications. The American Diabetes Association guidelines[14] were followed to determine whether a laboratory result (or vital sign) is normal. Table 2 lists these variables and the number of patients. Except for HLD, the majority of patients do not have a comorbidity at baseline in our cohort. Table 3 presents the baseline characteristics of the remaining variables.

Methods

Extracting the typical and atypical trajectories

We define a diabetes *trajectory* as a sequence of comorbidities (i.e., HDL, HTN, and IFG) potentially leading up to diabetes. The ordering of these comorbidities is denoted by an arrow ($\rightarrow$). For example, suppose we have three comorbidities A, B, and C, and the trajectory A $\rightarrow$ B $\rightarrow$ C indicates that A is followed by B and B is followed by C. These conditions are generally assumed

Table 1. Study population

Description	Count
Inclusion	
Patients age (18 at January 1, 2005	69,747
Exclusion	
Diabetic patients	−389 = 69,358
Patients with unknown glucose	−14,559 = 54,717
Patients with unknown lipid	−1023 = 53,862
Patients with unknown BP	−498 = 53,598
Nondiabetic patients who did not survive 5 years	−10,089 = 43,509

Table 2. Study variables for IFG, HTN, and HLD accounting for severity

Risk factor	Description	Count
IFG		
ifg.no	FPG ≤100	35,110
ifg.pre1	100 < FPG ≤110	6797
ifg.pre2	110 < FPG ≤125	1602
HTN		
htn.no	No indication of HTN	29,603
htn.untx	No drug is needed, and only one blood pressure result is elevated	5355
htn.tx	Treatment needed	8551
HLD		
hld.no	No indication of HLD	12,092
hld.untx	No therapeutic need, but some indication of hyperlipidemia exists (laboratory or diagnosis)	25,439
hld.tx	Treatment needed	5978
Obesity		
obese.no	BMI <25	13,061
obese.overweight	25 ≤ BMI <30	10,642
obese.obese	diagnosis or BMI ≥30	12,188

BMI, body mass index; FPG, fasting plasma glucose; HLD, hyperlipidemia; IFG, impaired fasting glucose.

to follow many different sequences (trajectories). We call the trajectory followed by most patients *typical* and label all other trajectories *atypical*.

We only know that at baseline, a patient has already developed a set of comorbidities (say) A, B, and C, but we could not directly observe in which order these comorbidities were developed. We can, however, estimate it. Suppose B follows A, A $\rightarrow$ B. If B indeed follows A, every time we encounter B, we should also encounter A. Therefore, the probability Pr(A|B) should be high. Accordingly, we define the probability of A $\rightarrow$ B as

$$Pr(A \rightarrow B) = Pr(A \mid B). \quad (1)$$

Let us extend this to calculate the probability of an entire trajectory. We define the probability of a trajectory as the likelihood of observing the data set under

Table 3. Baseline characteristics for variables not in Table 2

Risk factor	Description	Finding
Demographic		
Age	Age (mean ± SD)	46 ± 16
Gender	Gender (% male)	42.16
Tobacco	Smoking status (past or current smoker %)	14.92
Diagnoses		
Renal	Renal disease (prevalence %)	1.40
IHD	Ischemic heart disease (prevalence %)	6.31
CVD	Cardiovascular disease (prevalence %)	2.02
PVD	Peripheral vascular disease (prevalence %)	1.10
CHF	Congestive heart failure (prevalence %)	0.92
Carotid	Carotid artery disease (prevalence %)	0.86

SD, standard deviation.

Table 4. The five most likely trajectories

No.	Trajectory	Likelihood
1	HLD → HTN → IFG → DM	0.100
2	HLD → HTN → DM → IFG	0.067
3	HLD → IFG → HTN → DM	0.058
4	HTN → HLD → IFG → DM	0.044
5	HLD → IFG → DM → HTN	0.040

the assumption that it was generated by the trajectory in question. Suppose there are four comorbidities A, B, C, and D, and the trajectory is A → B → C → D. Patients following this trajectory may have progressed to different stages: Some patients may have progressed all the way to D, others to C, some to only A or B, and yet others may not present with any symptoms yet but will follow the trajectory once they start progressing. In patients who have already progressed to D along this trajectory, we should see A, B, and C with very high probability, that is, $\Pr(A,B,C|D)$ should be high. In other patients following the same trajectory, who have only progressed to C, we should see A and B with high probability, that is, $\Pr(A,B|C)$ should be high. We define the probabilities for patients who have only progressed to A or B analogously, giving us the probability of the trajectory as

$$\Pr(A \to B \to C \to D) = \Pr(A,\ B,\ C \mid D) \\ \times \Pr(A,\ B \mid C) \times \Pr(A \mid B). \quad (2)$$

Note that the same patient can be counted multiple times. For example, a patient presenting with A and B at baseline is counted not only for the sequence A → B → C → D but also for A → B → D → C, as well as for B → A → C → D among others. Therefore, the likelihood does not coincide with the percentage of patients following this trajectory.

Type 2 diabetes risk modeling with trajectories

To address the association between the different trajectories and the risk of developing diabetes, we constructed a multivariate logistic regression model for diabetes outcome using demographics (Table 2), glucose level, staged comorbidities (Table 3), and three trajectories (Table 5). Data analysis was conducted in R version 3.2.3.

Results

In this section, we show the typical trajectory extracted from our data and subsequently enumerate the atypical trajectories. We then investigate whether the atypical trajectories are associated with increased risk of developing T2DM.

Typical trajectory

In Table 4, we present the five most likely trajectories selected based on the likelihood. The likelihoods are generally low because the probability of progression to T2DM itself is low. The most likely trajectory in our study cohort is HLD → HTN → IFG → DM, which coincides with the trajectory that is commonly used for patient education. We also observe that the most likely trajectory is far more likely than the others. Counting the number of patients (Table 5) who show no evidence of following a different trajectory confirms that the vast majority of the patients follow this trajectory. This satisfies our definition of typical trajectory.

Atypical trajectories

There is evidence in our data that patients can follow trajectories different from the typical one (i.e., HLD → HTN → IFG → DM). In our definition, a patient is said to *follow* a trajectory HLD → HTN → IFG → DM if and only if his sequence of comorbidities is constant with that trajectory. Formally, if the patient has k comorbidities, these have to coincide with the first k diseases along the trajectory. A patient follows the *typical* trajectory if his comorbidities are consistent with the typical trajectory and an *atypical* trajectory if he shows evidence to contrary. For example, a patient with comorbidities {HLD, HTN} follows the typical trajectory we identified, while a patient with comorbidities

Table 5. Typical and atypical trajectories

No.	Trajectory	HLD	HTN	IFG	Count	DM	Pr(DM)
1	Typical	N	N	N	8795	235	0.037
2	Typical	Y	N	N	16,307	948	0.058
3	Typical	Y	Y	N	8297	950	0.115
4	Typical	Y	Y	Y	3485	1362	0.391
5	Atypical with skipping HLD	N	Y	N	1711	92	0.054
6	Atypical with skipping HLD	N	Y	Y	413	84	0.203
7	Atypical with skipping HTN	Y	N	Y	3328	992	0.298
8	Atypical with skipping HLD and HTN	N	N	Y	1173	132	0.113

{HLD, IFG} shows evidence to contrary, that is, skipped HTN, because HTN preceded IFG along the trajectory. A patient who has no comorbidity is assumed to follow the typical trajectory. We enumerate the atypical trajectories based on which conditions of the typical trajectory are skipped.

Table 5 shows the typical and atypical trajectories with detailed information. Each row in the table corresponds to a patient group, presenting with a set of comorbidities at baseline. Patients with T2DM at baseline are excluded, so we omit T2DM from the table. The column "No." is simply an arbitrary identifier assigned to the group. We also show the total number of patients in this group and the number of cases, patients who developed T2DM during the follow-up period. We assign these patient groups to trajectories, which we also show in the table. For instance, we assign No. 7 group to the atypical trajectory. Patients in group No. 7 present with HLD and IFG but not with HTN. We do not know whether they developed IFG or HLD first, but it is inconsequential: The fact that they have IFG and no HTN offers evidence that they did not follow the typical trajectory. In the typical trajectory, patients develop HTN before they develop IFG; thus, every patient with IFG should also present with HTN.

The number of patients who followed atypical trajectories is substantial. From the table, we can see that 6626 of the 43,509 patients (15%) followed atypical trajectories, and more importantly, 1300 of 4795 (27%) *cases* (patients who developed T2DM) followed atypical trajectories.

Atypical trajectories and the risk of developing type 2 diabetes

To study whether the trajectory influences the patients' risk of progression to diabetes, we have built a regression model for incident diabetes, which, besides the usual comorbidities, also includes the trajectory as an independent variable. Table 6 shows the predictors and their coefficient estimates. The predictors describing HLD, HTN, IFG, and obesity are ordinal; their levels are ordered: "no" (no sign of disease) is less severe than "untx" (no treatment needed) and "untx" is less severe than "tx" (treated). The effect of each level is measured relative to the next lower level. For instance, the effect of hld.tx is measured relative to hld.untx: Requiring treatment for hyperlipidemia (hld.tx) increases the log odds of progression to diabetes by 0.29 relative to patients who do not require treatment for hyperlipidemia (hld.untx).

Table 6. Predictors and coefficient estimates from the type 2 diabetes predictive model

Variable	Coefficient estimate (SE)	p
Intercept	−13.55 (0.44)	<0.001
Age	0.01 (0.00)	<0.001
Male	−0.16 (0.39)	<0.001
Gluc	0.11 (0.00)	<0.001
hld.untx	0.38 (0.08)	<0.001
hld.tx	0.29 (0.09)	<0.001
htn.untx	0.19 (0.07)	0.005
htn.tx	0.26 (0.06)	<0.001
ifg.pre1	0.20 (0.07)	0.005
ifg.pre2	0.00 (0.12)	0.978
obese.ovrwght	0.10 (0.05)	0.048
obese.obese	0.45 (0.05)	<0.001
trajskip.htn	0.24 (0.08)	0.002
trajskip.hl	−0.06 (0.13)	0.650
trajskip.both	−0.54 (0.16)	<0.001

SE, standard error.

Two of the atypical trajectories are significant. The atypical trajectory where patients skip HTN (patients with HLD and IFG but without HTN) increased the log odds of developing T2DM by 0.24 compared to the typical trajectory. At first, this appears as if the *lack* of HTN *increased* the risk. The risk of T2DM depends on the deterioration of the underlying metabolic health, and the comorbidities, including HTN, are imperfect indicators of the deterioration of the metabolic health. One probable explanation is that the metabolic health of the patients with HLD and IFG (but without HTN) has deteriorated just as far as that of the patient with HTN, but their blood pressure has not yet increased sufficiently to meet the HTN diagnosis criteria. In such patients, the deterioration of the underlying condition, which typically manifests itself in the HTN disease, cannot contribute to the diabetes risk *through the HTN variable*, but its detrimental effect is captured through the trajectory variable.

The atypical trajectory where patients skip both HLD and HTN altered the log odds of developing T2DM by −0.54 (i.e., decrease it by 0.54) compared to the typical trajectory. In patients following the typical trajectory, IFG increases the (log odds of the) risk of diabetes by 0.20 or $0.20 + 0.00 = 0.20$ depending how far the patient has progressed. However, in the absence of both HLD and HTN, these 0.20 overestimate the patients' actual risk; thus, the trajectory adjusts the risk (downward). In other words, for patients who present neither with HLD nor with HTN, elevated fasting glucose is not as damaging (with regard to diabetes) as we would expect assuming that IFG is independent of these conditions.

Discussion

In this work, we studied a novel approach to infer disease progression from EHRs. EHRs with their large sample size and long follow-up time are becoming increasingly popular for population-based disease progression studies. However, unreliable diagnostic codes in the EHR data combined with the slow onset of many of the chronic diseases make it virtually impossible for us to directly observe trajectories, sequences in which the diseases develop. In this work, we described methods to sidestep or overcome these issues and discover interesting previously unknown knowledge.

Specifically, we overcame the problem of unreliable diagnostic codes through phenotyping. Phenotyping refers to the combined use of diagnosis codes, laboratory results, and medications to determine whether a patient presents with a condition at a given time. As phenotypes, we created an ordinal variable for each condition of interest, which, besides indicating the presence of a condition, also encoded its severity.

Solving the issue of onset dates is more challenging, and we sidestepped it by simply assuming that the onset date occurred before the earliest recording date. Even if we managed to estimate the onset dates accurately, the prebaseline period (and the 13 years of follow-up in general) was insufficient to observe entire trajectories. Instead of directly observing, we inferred the trajectories from snapshots. We used likelihood estimation to find a typical trajectory, which coincided with the trajectory that is commonly used for diabetes patient education.

We found that in the context of diabetes, some atypical trajectories had a significant effect on the risk of progression to T2DM. We observed that skipping HTN increased the risk of T2DM by approximately the same amount as HTN itself, and we also observed that having high blood glucose without HTN or HLD is not as damaging as one would expect under the assumption that these conditions affect the risk of T2DM independently. These are novel findings that were previously not known and not even studied.

Given the popularity of EHR data as a research platform, we expect larger sample sizes and longer follow-up times in the future. With the explosive growth of wearable health devices providing real-time physiological measurements, we may be able to infer the onset dates with better accuracy. Unfortunately, these improvements will not be able to completely eliminate the issues addressed in this work. The need for using historic data will remain, and along with it, the uncertainty in the historic data will remain, as well. Methods, such as the ones proposed in this article, will still be required to help unlock the full potential of historic data.

Author Disclosure Statement

No competing financial interests exist.

References

1. Centers for Disease Control and Prevention. Diabetes Report Card 2012: National and State Profile of Diabetes and Its Complications. Atlanta, GA: U.S. Department of Health and Human Services, 2012.
2. Centers for Disease Control and Prevention. National diabetes statistics report: Estimates of diabetes and its burden in the United States, 2014. National Diabetes Statistics Report. Atlanta, GA: U.S. Department of Health and Human Services, 2014.
3. Saudek CD. Can diabetes be cured? JAMA 2009; 301:1588–1590.
4. Buse JB, Caprio S, Cefalu WT, et al. How do we define cure of diabetes? Diabetes Care 2009; 32:2133–2135.
5. Forbes JM, Cooper ME. Mechanisms of diabetic complications. Physiol Rev 2013; 93:137–188.
6. Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002; 346:393–403.
7. Tuomilehto J, Lindström J, Eriksson JG, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med 2001;344:1343–1350.
8. Wilson PWF, Meigs JB, Sullivan L, et al. Prediction of incident diabetes mellitus in middle-aged adults: The Framingham Offspring Study. Arch Intern Med 2007; 167:1068–1074.
9. Eddy DM, Schlessinger L. Archimedes: A trial-validated model of diabetes. Diabetes Care 2003; 26:3093–3101.
10. Clarke PM, Gray AM, Briggs A, et al. A model to estimate the lifetime health outcomes of patients with Type 2 diabetes: The United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68). Diabetologia 2004; 47:1747–1759.
11. Rocca WA, Yawn BP, St Sauver JL, et al. History of the Rochester Epidemiology Project: Half a century of medical records linkage in a US population. Mayo Clin Proc 2012; 87:1202–1213.
12. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: Results of the eMERGE consortium. Sci Transl Med 2011; 3:79re1.
13. Kim E, Oh W, Pieczkiewicz DS, et al. Divisive hierarchical clustering towards identifying clinically significant pre-diabetes subpopulations. AMIA Annu Symp Proc 2014; 2014:1815–1824.
14. American Diabetes Association. Executive summary: Standards of medical care in diabetes—2014. Diabetes Care 2014; 37:S5–S13.

Abbreviations Used

EHR = electronic health record
HLD = hyperlipidemia
HTN = hypertension
IFG = impaired fasting glucose
T2DM = type 2 diabetes mellitus