Published in final edited form as:

IEEE Int Conf Healthc Inform. 2017 August ; 2017: 374–379. doi:10.1109/ICHI.2017.41.

Estimating Disease Onset Time by Modeling Lab Result Trajectories via Bayes Networks

Wonsuk Oh^{*}, Pranjul Yadav[†], Vipin Kumar[†], Pedro J. Caraballo[‡], M. Regina Castro[§], Michael S. Steinbach[†], and Gyorgy J. Simon^{*,†}

^{*}Institute for Health Informatics, University of Minnesota

[†]Department of Computer Science and Engineering, University of Minnesota

[‡]Division of General Internal Medicine, Mayo Clinic

[§]Division of Endocrinology, Diabetes, Metabolism & Nutrition, Mayo Clinic

^{††}Department of Medicine, University of Minnesota

Abstract

The true onset time of a disease, particularly slow-onset diseases like Type 2 diabetes mellitus (T2DM), is rarely observable in electronic health records (EHRs). However, it is critical for analysis of time to events and for studying sequences of diseases. The aim of this study is to demonstrate a method for estimating the onset time of such diseases from intermittently observable laboratory results in the specific context of T2DM. A retrospective observational study design is used. A cohort of 5,874 non-diabetic patients from a large healthcare system in the Upper Midwest United States was constructed with a three-year follow-up period. The HbA1c level of each patient was collected from earliest and the latest follow-up. We modeled the patients' HbA1c trajectories through Bayesian networks to estimate the onset time of diabetes. Due to non-random censoring and interventions unobservable from EHR data (such as lifestyle changes), naïve modeling of HbA1c through linear regression or modeling time-to-event through proportional hazard model leads to a clinically infeasible model with no or limited ability to predict the onset time of diabetes. Our model is consistent with clinical knowledge and estimated the onset of diabetes with less than a six-month error for almost half the patients for whom the onset time could be clinically ascertained. To our knowledge, this is the first study of modeling long-term HbA1c progression in non-diabetic patients and estimating the onset time of diabetes.

I. Introduction

The recent adoption of the electronic health record (EHR) [1], [2] provides us with an opportunity to use it for advanced analytics and thus improving patient health outcomes [3]–[6]. Many of these analyses rely on the onset time of diseases. For example, accurate onset time is required for time-to-event outcome analyses, and also for analyses that are concerned with sequences in which diseases develop. Especially for slow-onset diseases like hyperlipidemia (HLD), hypertension (HTN), and type 2 diabetes mellitus (T2DM), onset time is not directly observable from the EHRs [7]. EHRs store the time when a problem was discovered or recorded, which can be significantly (years) different from the onset time, when the disease actually started. In this study, we construct a model that can reliably

estimate the onset time for slow onset diseases from intermittently observable EHR data elements, most notably from laboratory results. We demonstrate this method through estimating the onset time of T2DM from HbA1c.

T2DM is a progressive metabolic disease, defined by chronically elevated blood sugar levels. The American Diabetes Association (ADA) [8] defines diabetes as a condition in which glycated hemoglobin A1c (HbA1c) levels exceed 6.5%. T2DM is a fast growing public health concern in the United States [9]. Approximately, 29.1 million Americans (9.3 % of the total population) are suffering from diabetes in 2014 [10]. Diabetes has severe consequences, with a significant impact on the quality of life [11], [12]. It is known to be the leading cause of kidney failure and blindness and the seventh-leading cause of death in the United States [9]. Since diabetes is a non-reversible progressive chronic disease [13], [14], prevention and timely diagnosis are of key importance [15]. For improve preventative care for T2DM, many risk prediction models have been developed [7], [16]–[18] and did successful demonstrating high prediction accuracy. However, these models are mainly focused on risks of developing T2DM for a specific time period but do not directly address the exact onset time of T2DM. Arguably, estimating onset time is the most proactive approach to prevent or delay the progression to T2DM. In this paper, we aim to estimate the onset time of diabetes, the earliest time when a non-diabetic patient has HbA1c in excess of 6.5%.

Modeling the trajectory of HbA1c from EHR data poses several key challenges. First, patient visits are intermittent, so the actual progression of HbA1c is unobservable: we may not be able to directly observe the HbA1c level of a patient at a particular time t . Even though we may not be able to observe the HbA1c level directly, the patient can still contribute partial information to the model: if the patient was out of control (HbA1c $\geq 6.5\%$) at an earlier time T ($T < t$), then we expect the HbA1c to be out of control at t ; or if the patient was under control (HbA1c $< 6.5\%$) at a later time T ($T > t$), then the patient should be under control at t . Not all models can make use of this partial information. Second, since HbA1c tends to increase over time, patients will either become diabetic or receive preventive diabetes drugs and hence get censored. This virtually guarantees that censoring in our study is not random. Third, some essential data elements, most notably patient education and change in lifestyle interventions, are missing from the structured EHR. Lifestyle interventions are the first line of defense in both the prevention and the management of diabetes, yet related information is unavailable in structured format in the EHR.

The natural choice for estimating onset time is by modeling HbA1c trajectory using a linear regression model or by modeling time-to-event using a proportional hazards model. Unfortunately, both models constructed from EHR data lack the capabilities to address many of the above challenges. In fact, if we built a linear model to predict HbA1c, we would erroneously find that HbA1c actually decreases over time. HbA1c would not typically decrease without intervention; it appears to decrease because of non-random censoring and patients undertaking lifestyle interventions. A proportional hazards model we constructed to directly estimate the onset time of diabetes also provides poor estimates for the same reasons: the non-informative censoring assumption is violated and the intervention is not observable and hence is unaccounted for.

In this manuscript, we propose a novel approach that uses a Bayes network [19]–[21] to model HbA1c level progression in non-diabetic patients and to subsequently estimate the onset time. Our model addresses all four of the above challenges. It models the (unobservable) actual and observed HbA1c level separately. While we can only observe a patient's HbA1c intermittently, we can estimate the (hidden) actual HbA1c at any time. Our model also takes partial information into account. If we observe a patient to have HbA1c in excess of 6.5% at time T , then his latent HbA1c level is expected to be above 6.5% any time afterwards until the patient receives intervention; and similarly, if we observe a patient to have HbA1c level less than 6.5% at time T , we expect his latent HbA1c to be less than 6.5% at any time before T . This separation between the latent and observed HbA1c is the key to adjusting for non-random censoring. In addition, we include a latent variable for the lifestyle intervention, allowing us to further reduce bias.

We evaluated the HbA1c progression model on a cohort from a large healthcare system in the Upper Midwest United States. We demonstrated that the resultant model reflects the actual changes in HbA1c level well, and we also showed that the model has the ability to accurately estimate the time to the onset of diabetes.

II. Materials and Methods

A. Data

A retrospective observational study design is used to construct predictive models for glycated hemoglobin (HbA1c) level using patient's baseline characteristics. We collected clinical data from a large healthcare system in the Upper Midwest region of the United States. Vital signs, diagnoses (ICD-9-CM) and laboratory results are available after 2006, while medications are available after 2010. The baseline for each patient is established on a patient's first observation date on or after January 1, 2011. We use a retrospective period (2006-baseline) to establish the patient's baseline characteristics and we track HbA1c measurement during the follow-up period (baseline-December 31, 2013) until loss to follow-up or until the patient shows indication of T2DM including T2DM medication prescription, out-of-control HbA1c level (≥ 6.5) or a related diagnosis code. Our cohort has 5,874 adult non-diabetic patients with at least two HbA1c measurements during follow-up. Table 1 presents the baseline characteristics of our cohort.

B. Bayesian network

In this manuscript, we construct a progression model for glycated hemoglobin A1c (HbA1c) level using a Bayesian network. Bayes networks explicitly describe dependence relationships among variables, allowing us to incorporate our prior beliefs. Our central prior belief is that HbA1c does not decrease without lifestyle change, therapy, or some other kind of intervention and that prediabetic patients receive lifestyle intervention, which is not recorded in their EHR.

1) HbA1c progression model—We present the graph for our HbA1c progression model in Figure 1. Clear circle nodes are observable and shaded circle nodes are latent (not directly observable). The formulation in Figure 1 assumes that $Z(t)$, which we will describe later this

section, is independent of $Z(\tau)$, $t - \tau$, given time t and the observed variables. This is a reasonable assumption, because $Z(t)$ directly incorporates t as a predictor. Time is always measured relative to baseline in years.

Nodes X and A_0 describe the patient's baseline characteristics: $A_{0,j}$ denotes the observed baseline HbA1c level and X_j denotes the baseline comorbidities for patient j .

We model the possible intervention a patient may have received as a latent variable I_j , which depends on the baseline HbA1c level. Patients who become prediabetic (HbA1c ≥ 5.7) or suffer from multiple comorbidities are likely to receive advice to change their lifestyle (exercise, eating habits) in order to prevent or delay progression to overt diabetes. The fact that the patient received such advice is not recorded in our data and whether the patient complies with such advice is also unobservable. We adopt a three-level lifestyle intervention since those who have HbA1c over 5.7 [8] need to consider lifestyle intervention and those with HbA1c over 6.0 [22] receive more aggressive lifestyle intervention with cautious follow-up schedule. Accordingly, I_j can take three values; i_{no} if the patient received no intervention; $i_{5.7}$ if the patient received intervention consistent in intensity with intervention typically given to patients with HbA1c between 5.7 and 6.0 mg/dL; and $i_{6.0}$ if the patient received intervention typical of patients with HbA1c between 6.0 and 6.5 mg/dL. Thus, we model the latent intervention variable I as

$$p(I_j = i \mid A_{0,j}) = \begin{cases} 1, & \text{if } i = i_{no} \text{ and } A_{0,j} < 5.7 \\ 1, & \text{if } i = i_{5.7} \text{ and } 5.7 \leq A_{0,j} \text{ and } A_{0,j} < 6.0 \\ 1, & \text{if } i = i_{6.0} \text{ and } 6.0 \leq A_{0,j} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Intervention as defined above is deterministic; it is formulated as a probabilistic model to fit into the Bayes network framework.

In order to address the problem of non-random censoring, the cornerstone of our methodology is the separation between the unobservable actual HbA1c level at time t , $Z_j(t)$, and the observed HbA1c level at time t , $A_j(t)$. We can compute the hidden HbA1c level at any time t , while the observed HbA1c $A_j(t)$ level is only available at one time point T_j for each patient; it is unknown at any other time point.

We assume $Z_j(t)$ is identical to a linear combination of time t , the baseline HbA1c level $A_{0,j}$, the patient's baseline comorbidities X_j , and the latent intervention I_j .

$$p(Z_j(t) = z \mid A_{0,j}, X_j, I_j = i) = \begin{cases} 1, & \text{if } z = A_{0,j} + X_j\xi + t\beta + (I_j = i_{5.7})\delta_1 + (I_j = i_{6.0})\delta_2 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $Z(t)$ is forced to take the value predicted by the linear model; ξ , β , and δ are coefficients of X , t , and the interventions, respectively.

The observed HbA1C level $A_j(t)$ of patient j at time t is identical to $Z_j(t)$ with Gaussian noise when it is observed (i.e. $O_j(t) = 1$ and is unknown otherwise.

$$p(A_j(T) = a \mid O_j(t) = o, Z_j(t) = z) = \begin{cases} \Phi(z - a, \sigma^2), & \text{if } a \text{ is known and } o = 1 \\ 1, & \text{if } a \text{ is unknown and } o = 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where ϕ and Φ denote the probability density function (PDF) and the cumulative distribution function (CDF) of the standard normal distribution. $A_j(t)$ is unknown at all time points $t \neq T_j$ and needs to coincide with $Z_j(t)$ at $t = T_j$. We would like to point out that $A_j(t)$ does not need to exactly coincide with the model prediction from Eq (2); there is a Gaussian error term in Eq (3) that allows for differences between the model prediction and $Z_j(t)$ and thus $A_j(t)$.

Finally, $E_j(t) = 1$ denotes that patient j has had an event at time t or before, namely there exists a time $T_j \leq t$, such that $Z_j(T_j) \geq 6.5$.

$$p(E_j(t) = e \mid Z_j(t) = z) = \begin{cases} 1, & \text{if } e = 1 \text{ and } 6.5 \leq z \\ 1, & \text{if } e = 0 \text{ and } z < 6.5 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Similarly to the intervention, $E_j(t)$ is also deterministic; we simply use a probabilistic notation to cast it into the Bayes network formalism.

The separation of the actual and observed HbA1c level is the key concept of our Bayesian network model. If the patient j already had an event, then $Z_j(t)$ must be ≥ 6.5 ; otherwise $p(E_j(t))$ becomes 0. Similarly, $Z_j(t)$ must be < 6.5 for patients under observation who will not suffer an event at T_j , i.e. $A_j(T_j) < 6.5$. The likelihood, which we will discuss in the following section, becomes 0 if this assertion does not hold.

2) Parameters learning—In the previous section, we described the structure of our network model; this section is concerned with estimating the values of the parameters ξ, β, δ . The maximum likelihood estimation (MLE) is widely used for estimating the parameters of a network model. Specifically, the data likelihood is

$$\mathcal{L}(\xi, \beta, \delta \mid A_0, X, O(t), A(t), E(t)) = \prod_t \prod_j p(A_{0,j}, X_j, O(t)_j, A(t)_j, E_j(t); \xi, \beta, \delta) \quad (5)$$

where the probability density function (PDF) of the data, $p(A_{0,j}, X_j, O(t)_j, A(t)_j, E_j(t); \xi, \beta, \delta)$, is the marginal distribution over the latent variables $Z(t)_j$ and I_j as following:

$$p(A_{0,j}, X_j, O(t)_j, A(t)_j, E_j(t); \xi, \beta, \delta) = \int_z \sum_i p(A_{0,j}, X_j, O(t)_j, A(t)_j, E_j(t), Z_j(t) = z, I_j = i; \xi, \beta, \delta) dz$$

(6)

We seek the parameter values ξ, β, θ that maximize the likelihood in (5) using the PDF (6). Any off-the-shelf algorithm can be used to compute the parameters [23].

3) Prediction—Our goal is to estimate the predicted HbA1c level at time t . Since our model has hidden nodes that we cannot observe directly, we need to find $Z(t) = z$ that maximizes $p(Z(t) / A_0, X, O(t), E(t))$ for given time t . Recall that the latent intervention variables are required for estimating $Z(t)$, but we cannot directly observe them, thus we seek

$$\begin{aligned} & \arg \max_z p(Z_t = z \mid A_0, X, O(t), E(t)) \\ &= \sum_i [p(Z(t) = z \mid A_0, X, O(t), E(t), I = i) \times p(I = i \mid A_0, X)] \end{aligned} \quad (7)$$

III. Results

The goal of this work is to estimate the onset time of diseases, diabetes in our concrete application, through modeling patients' HbA1c trajectory via Bayesian network. Onset time is the earliest time when a patient's HbA1c exceeds 6.5. Unfortunately, due to the intermittent nature of patient visits, the actual onset time is observable only for relatively few patients (492 out of 5,874). Thus we are unable to evaluate the performance of the proposed model on the entire cohort. Instead, we evaluate our model in two parts. In the first part, we evaluate the performance of the proposed model on the entire population for estimating HbA1c. Our method was not designed for this task, the purpose of this evaluation is merely to demonstrate that our method can achieve reasonable performance on the entire cohort even when we compare it to a standard technique, i.e. linear regression, optimized for that task. In the second part, we evaluate our model for the intended purpose of the algorithm, namely for predicting the onset time of diabetes. Naturally, we can only use the 492 patients, for whom the onset time is known. We compare our model with the Cox proportional hazards model, using median survival time as the estimate for onset time. The 10-folds cross-validation is used to evaluate model accuracy. All implementation and analysis is conducted with the use of R version 3.3.1.

In total, we built three models: the proposed model, a multivariate linear regression model with the observed HbA1c as the dependent variable and baseline HbA1c, baseline co-morbidities, follow-up time and the latent interventions and a Cox proportional hazards

model with the same variables except for follow-up time. Analogously to the proposed model, we considered obesity (Obese), high cholesterol (hyperlipidemia; HLD) and high blood pressure (hypertension; HTN) as comorbidities.

Tables 2, 3 and 4 show the predictors, the coefficients, and the empirical p-values estimated through bootstrap estimation with 1000 replications, for the three models. We interpret the coefficients of the multivariate linear regression model as follows. A unit increase in baseline HbA1c (A_0) increases the HbA1c at time t by .729; being obese increases the HbA1c at time t by .037 and receiving the latent intervention does not change the HbA1c significantly. The other variables can be interpreted similarly. Of particular interest is the coefficient of time: during each additional year of follow-up time, the HbA1c level increases by -0.008 , on average (i.e. it decreases by .008). This negative coefficient is problematic for two reasons. First, our understanding is that HbA1c increases unless the patient receives interventions. Second, this model cannot be used to estimate the time to onset of diabetes, because the estimated HbA1c level is predicted to decrease over time, suggesting that the patient will never develop diabetes. This is a clear example demonstrating the challenges of modeling the HbA1c trajectory from EHR data. Additionally, the model obviously contradicts clinical knowledge as well.

Table 4 shows that the coefficients for Eq (2) in the Bayesian network model. The interpretation of the coefficient in the Bayesian network model is analogous to that of the multivariate linear model: a year increase in follow-up time t increases the HbA1c level by 0.320, on average. Assuming this rate of increase in HbA1c, a patient would progress from almost prediabetic levels (say, HbA1c of 5.5%) to overt diabetes (HbA1c 6.5%) in 3 years, which is reasonably consistent with screening interval in current guidelines [8], [24].

A. Estimating HbA1c level

We compare the proposed model with a multivariate linear regression model on predicting HbA1c levels. Both models offer reasonable performance; although the linear regression model has the slightly higher correlation (.569) with the actual HbA1c values than the proposed model (.501). We wish to emphasize that the proposed model was not designed to provide as accurate HbA1c prediction as possible; but rather our goal was to construct a model that is physiologically feasible and offers good prediction of diabetes onset time. We argue that despite the small deficit in predictive performance, the proposed model is actually better. The linear model cannot predict HbA1c levels in excess of 6.5; thus it is totally inadequate for predicting the onset time of diabetes: its prediction is that no patient will ever develop diabetes.

Since the multivariate linear regression model has no ability to predict diabetic (> 6.5) HbA1c levels, we cannot use it to estimate diabetes onset time, and thus we exclude it from further evaluation.

B. Estimating T2DM Onset Time

We now return to our original problem of estimating the onset time of diabetes, the time when the HbA1c level exceeds 6.5. Unfortunately, for the vast majority of patients, the onset time is not observable. We identified 492 patients who have an observed HbA1c level

between 6.3 and 6.7 and we simply call the corresponding observation time as the observed onset time, although the actual onset time can be slightly different. We estimate onset time as the earliest time when a patient's estimated HbA1c exceeds 6.5. We can evaluate our method by comparing the estimated and the observed onset time.

We cannot expect the estimated onset time to match the observed onset time exactly. Changes in HbA1c are not instantaneous; researchers tend to think of HbA1c as a 3-month running average of the blood glucose level. Accordingly, our observed onset time could differ from the actual onset time by almost 2 months (the time it takes for the HbA1c to progress from 6.3 to 6.5 according to our model). Therefore, we believe that the best possible estimate is within 2 months of the actual onset time. Thus these observed onset times are not exact, but are appropriate as a "silver standard".

1) Evaluating the proposed method against the "silver standard"—Figure 2 shows the cumulative density distribution of the prediction error. We computed the prediction error as the absolute difference between the predicted onset time and the onset time observed from the EHR data. The x-axis represents the prediction error (measured in years), and the y-axis represents the cumulative probability of that prediction error denotes the percentage of patients who have a prediction error smaller than the corresponding value on the x-axis. For example, 43.5% of patients (y-axis) have a prediction error less than half year (0.5 on the x-axis). The solid black line represents the cumulative density curve of the prediction error from the Bayesian network. The figure shows that a quarter of the patients (22.0 %) have a prediction error less than three months, which is our theoretical lowest error; and almost half of the patients (43.5 %) have an error less than half year.

2) Comparing the proposed method to the Cox proportional hazards model—The gray dashed line in Figure 2 represents the cumulative density of the prediction error from the Cox model. We use median survival time, which is the earliest time when the survival probability is equal or less than 0.5, as the estimated onset time.

Not only does the Cox model show significantly lower prediction ability as compared with the Bayesian network model, but it failed to make a prediction for 261 out of the 492 patients, because these patients had a survival probability higher than 0.5 at the end of follow-up. The cumulative density of the prediction error from the Cox model reveals that only 1.2% of the patients have a prediction error equal or less than three months and 4.5% of the patients have a prediction error equal or less than half year. Since the 3-year to 5-year interval is regarded as cost-effective screening interval [24], the Cox model with 53.0% of the patients having a prediction error of more than 3-years, has limited utility in this application.

IV. Summary and Discussion

Electronic health records (EHRs) contain rich, longitudinal, and large volumes of EHR data, which offers us a new way of conducting clinical research. Many research questions rely on the accurate estimation of onset time for diseases. Because of the inherent limitations of EHR data, we cannot always directly extract this information from EHRs as they contain the

recorded time for diseases rather than the onset time. For slow-onset diseases, the onset time and the recorded time can be substantially different; they can differ by years. In this paper, we propose a novel model to estimate the onset time of chronic diseases primarily based on the defining laboratory results.

The performance of our model is particularly reassuring when compared to multivariate linear regression, the standard method for modeling continuous measures, such as HbA1c. Linear regression, with no ability to compensate for informative censoring (patients who are removed from the study for taking antidiabetic drugs are very likely to develop diabetes) constructed a model that predicted decreasing HbA1c levels. Not only is this model not supported by our knowledge of the pathophysiology of metabolic degeneration, but it also cannot be used for predicting onset times, because the decreasing HbA1c levels suggest that the patient will never develop diabetes.

We further evaluated our method in terms of its ability to predict the onset time for diabetes on a smaller cohort in which diabetes onset times are known. We found that we could estimate the onset time within 6 months error for almost half of the patients; and within 3 months accuracy for almost quarter of the patients. Our prediction error was lowest for patients who progressed to diabetes in a bit over a year. Our prediction error increased sharply for patients with an onset time 2+ years into the future.

An alternative to the proposed model could be a Cox proportional hazards model using one of the standard methods for estimating time to onset of diabetes. Using median survival as an estimate for time to onset of diabetes, the Cox model offered poor performance with a prediction error in excess of 3 years for almost half of the population.

We demonstrated this method through the concrete application of estimating the onset time of T2DM. To the best of our knowledge [25], [26], this is the first study of modeling the trajectory of HbA1c level and estimating the onset time of diabetes. The methodology is not specific to T2DM; we believe that it can be generalized to other slow onset diseases in a straightforward manner.

Acknowledgments

The work described in this manuscript was supported by NIH grant LM011972, NSF grants IIS-1602394 and IIS-1602198. The views expressed in this manuscript are those of the authors and do not necessarily reflect the views of NIH and NSF.

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*. 2012; 13(6):395–405.
2. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*. 2013; 20(e2):e206–e211. [PubMed: 24302669]
3. Yadav, P., Steinbach, MS., Pruinelli, L., Westra, B., Delaney, C., Kumar, V., Simon, GJ. Forensic Style Analysis with Survival Trajectories. *Proceedings of the 2015 IEEE International Conference on Data Mining*; 2015; IEEE; p. 1069-1074.

4. Westra BL, Landman S, Yadav P, Steinbach MS. Secondary Analysis of an Electronic Surveillance System Combined with Multi-focal Interventions for Early Detection of Sepsis. *Applied Clinical Informatics*. 2017; 8(1):47–66. [PubMed: 28097288]
5. Yadav, P., Pruinelli, L., Hangsleben, A., Dey, S., Hauwiller, K., Westra, BL., Delaney, CW., Kumar, V., Steinbach, MS., Simon, GJ. Modeling Trajectories for Diabetes Complications. *Proceedings of the 4th Workshop on Data Mining for Medicine and Healthcare*. 2015 SIAM International Conference on Data Mining; 2015;
6. Yadav, P., Prunelli, L., Hoff, A., Steinbach, MS., Westra, BL., Kumar, V., Simon, GJ. Causal Inference in Observational Data. 2016. [Online]. Available: <http://arxiv.org/abs/1611.04660>
7. Oh W, Kim E, Castro MR, Caraballo PJ, Kumar V, Steinbach MS, Simon GJ. Type 2 Diabetes Mellitus Trajectories and Associated Risks. *Big Data*. 2016; 4(1):25–30. [PubMed: 27158565]
8. American Diabetes Association. Standards of Medical Care in Diabetes-2016. *Diabetes Care*. 2016; 39(Supplement 1)
9. Centers for Disease Control and Prevention. Tech Rep. U.S. Department of Health and Human Services; Atlanta, GA: 2012. Diabetes Report Card 2012: National and State Profile of Diabetes and Its Complications.
10. Centers for Disease Control and Prevention. Tech Rep. U.S. Department of Health and Human Services; Atlanta, GA: 2014. National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014.
11. Huang ES, Brown SES, Ewigman BG, Foley EC, Meltzer DO. Patient Perceptions of Quality of Life With Diabetes-Related Complications and Treatments. *Diabetes Care*. 2007; 30(10):2478–2483. [PubMed: 17623824]
12. Bradley C, Speight J. Patient perceptions of diabetes and diabetes therapy: assessing quality of life. *Diabetes/Metabolism Research and Reviews*. 2002; 18(S3):S64–S69. [PubMed: 12324988]
13. Saudek CD. Can Diabetes Be Cured? *JAMA*. 2009; 301(15):1588. [PubMed: 19366779]
14. Buse JB, Caprio S, Cefalu WT, Ceriello A, Del Prato S, Inzucchi SE, McLaughlin S, Phillips GL, Robertson RP, Rubino F, Kahn R, Kirkman MS. How Do We Define Cure of Diabetes? *Diabetes Care*. 2009; 32(11):2133–2135. [PubMed: 19875608]
15. Diabetes Prevention Program Research Group. Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. *New England Journal of Medicine*. 2002; 346(6):393–403. [PubMed: 11832527]
16. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. *BMJ*. Nov 28; 2011 343(1):d7163–d7163. [Online]. Available: <http://www.bmj.com/cgi/doi/10.1136/bmj.d7163>. [PubMed: 22123912]
17. Kim, E., Oh, W., Pieczkiewicz, DS., Castro, MR., Caraballo, PJ., Simon, GJ. Divisive Hierarchical Clustering towards Identifying Clinically Significant Pre-Diabetes Subpopulations. *Proceedings of the 2014 AMIA Annual Symposium*; 2014;
18. Simon GJ, Caraballo PJ, Therneau TM, Cha SS, Castro MR, Li PW. Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus. *IEEE Transactions on Knowledge and Data Engineering*. 2015; 27(1):130–141.
19. Heckerman, D. *Innovations in Bayesian Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. A Tutorial on Learning with Bayesian Networks; p. 33-82.
20. Lucas PJF, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*. 2004; 30(3):201–214. [PubMed: 15081072]
21. Orphanou K, Stassopoulou A, Keravnou E. Temporal abstraction and temporal Bayesian networks in clinical domains: A survey. *Artificial Intelligence in Medicine*. 2014; 60(3):133–149. [PubMed: 24529699]
22. World Health Organization. Tech Rep. 2011. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation.
23. Boyd, S., Vandenberghe, L. *Convex Optimization*. Cambridge University Press; 2004.
24. Kahn R, Alperin P, Eddy DM, Borch-Johnsen K, Buse JB, Feigelman J, Gregg EW, Holman RR, Kirkman MS, Stern MP, Tuomilehto J, Wareham NJ. Age at initiation and frequency of screening to detect type 2 diabetes: a cost-effectiveness analysis. *The Lancet*. 2010; 375(9723):1365–1374.

25. Yadav, P., Steinbach, M., Kumar, V., Simon, G. ACM Computing Surveys. 2017. Mining Electronic Health Records: A Survey.
26. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal. 2017; 15:104–116. [PubMed: 28138367]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1.
Glycated hemoglobin (HbA1c) level progression model

Fig. 2.
The cumulative density of the prediction error in years.

TABLE I

Demographics, comorbidity, and HbA1c level data at baseline

Risk Factor	No. (%) of sample
Demographic	
Age, mean $\pm$ SD	57 $\pm$ 14
Sex: male	47.2 %
Comorbidity	
Hyperlipidemia (HLD)	82.3 %
Hypertension (HTN)	70.8 %
Obese (BMI $\geq$ 30)	55.5 %
Glycated hemoglobin (HbA1c)	
Normal (<5.7 percent)	40.8 %
Prediabetes (5.7–6.4 percent)	59.2 %

TABLE II

Predictors and coefficient estimates from the multivariate linear regression model

	Coefficient	p-value
(Intercept)	1.556	< .001
A_0	0.729	< .001
time t	-0.008	0.346
$i_{5,7}$	0.001	0.967
$i_{6,0}$	0.040	0.027
Obese	0.037	< .001
HLD	-0.025	0.109
HTN	0.013	0.265

TABLE III

Predictors and coefficient estimates from the Cox proportional hazards model

	Coefficient	p-value
A_0	3.622	< .001
$i_{5,7}$	0.109	.732
$i_{6,0}$	0.225	.323
Obese	0.444	.002
HLD	-0.814	< .001
HTN	0.121	.382

TABLE IV

Predictors and coefficient estimates from the multivariate linear regression model

	Coefficient	p-value
time t	0.320	< .001
$i_{5,7}$	-0.089	.008
$i_{6,0}$	0.298	< .001
Obese	-0.048	< .001
HLD	-0.098	< .001
HTN	-0.074	< .001