USING MACHINE LEARNING TO PREDICT MICROVASCULAR COMPLICATIONS IN PATIENTS WITH TYPE 1 DIABETES

By

QINGQING XU

A dissertation submitted to the
Department of Pharmaceutical Health Outcomes and Policy, College of Pharmacy
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN

PHARMACEUTICAL HEALTH OUTCOMES AND POLICY

Chair of Committee: Sujit S. Sansgiry, MS, PhD

Committee Member: Susan Abughosh, PhD

Committee Member: E. James Essien, MD, DrPH

Committee Member: Vijay N. Joish, PhD

Committee Member: Shubhada Sansgiry, PhD

University of Houston

June 2020

DEDICATED TO

Kaigui Xu, Hongfen Wang, Xianyun Li, Yuzhi Yu (grandparents)

Dongyuan Xu and Hanping Li (parents), Jie Wang and Xuanxian Xu (in-laws)

Yifan Wang (husband)

Jocelyn Anne Wang (daughter) and Andrew Xu Wang (son)

ACKNOWLEDGMENTS

My path towards earning this doctorate degree and finishing my dissertation research would not be possible without the guidance, supports, challenge, encouragement and help of a lot of people. First and foremost, I sincerely thank Dr. Sujit S. Sansgiry for being my advisor and mentor over the past five years. He has patiently and inspiringly guided me through all stages of my research, and continuously helped me build confidence and encouraged me to chase any opportunities. Without him, I would never have explored the breadth and depth in health outcomes research. Because of him, I have enjoyed every single moment of learning. I also thank him for rejoicing me when I had my son in my second year of graduate study, and for relieving stress and planning things out for me when my husband had an accident that I had to postpone me proposal defense. It would be impossible to count all the ways that he has influenced my life and shaped me into the person that I am today. A PhD advisor like him only comes once in my lifetime.

I would like to thank my committee, Dr. Susan Abughosh for always being encouraging and asking probing questions, Dr. E. James Essien for teaching me to systematically and calmly handle any problem, Dr. Shubhada Sansgiry for providing insightful comments, discussing with me back and forth on research details and pointing me to the right direction to look for answers, and last but as importantly, Dr. Vijay N. Joish for his critical inputs in guiding the direction of my dissertation.

I want to acknowledge the following clinicians, Dr. Archana Sadhu at Houston Methodist Hospital, Dr. Aziz Shaibani at Baylor College of Medicine and Nerve and Muscle Center of Texas, Dr. Bernadette Asias-Dinh at University of Houston College of Pharmacy, and Dr. Carolyn R Carman, Dr. Jennifer Tasca and Dr. Joe L Wheat from University of Houston College of Optometry, who took their time and advised me on the operational definitions of diabetic nephropathy, retinopathy and neuropathy. I want to thank Dr. Yifan Wang, Dr. Sifei Liu, Dr. Shuyan Huang, Dr. Ying Lin and Dr. Yibin Liao for their help in model training using machine learning.

A special thanks to Benjamin Lewing and Ravi Goyal, for making a supportive team, encouraging me when I’m worried and helping me preparing for my defense and job search. I thank Rutu Paranjpe, Zahra Majd and Ning Lyu for standing by me and cheering me up, and all students and colleagues for making the graduate school so delightful. I thank all professors at the College who has taught and helped me during my graduate years. I would also like to thank Danielle Armstrong, Melissa Nieto and Christen Gould for their kind assistance that makes any administrative process painless.

Lastly, I would like to dedicate my dissertation work to my beloved family who always believe in me and support me. To my grandparents, Kaigui Xu, Hongfen Wang, Xianyun Li and Yuzhi Yu, who teach me unconditional love and perseverance, always put my needs first, encourage me to face the challenges in life and rejoice in my smallest achievements. To my father, Dongyuan Xu and my mother, Hanping Li, who have taken turns traveling between China and the US over the past 5 years and helping me take care of the kids and the family. My pursuit of the PhD degree would be impossible without them. To my in-laws who have supported us financially and shed stress from us. To my dearest husband, Yifan Wang, who has been a dependable and strong partner, sharing with me the ups and downs in life, inspiring me in research and trusting me with the kids. Last but not the least, to my sweetest daughter, Jocelyn Anne Wang, who is so thoughtful, patient and independent at such a young age, who appreciates the limited time that I could spend with her, helps me with house chores and takes good care of her brother when I cannot. And to my lovely son, Andrew Xu Wang, who is so easy-going and knows how to entertain himself. You both have always been my source of joy and taught me what life is all about.

ABSTRACT

Background: Diabetic microvascular complications can lead to long-term morbidity and mortality, significantly drive healthcare costs, and impair quality of life of patients with type 1 diabetes (T1D). Early prediction and prevention of microvascular complications, including nephropathy, retinopathy, and neuropathy in T1D patients can support informed clinical decision making and potentially delay the progression to long-term adverse outcomes. Although machine learning (ML) methods have been applied for disease prediction in healthcare, there is very limited research using advanced ML methods (e.g., neural networks) for the prediction of microvascular complications in T1D patients. Moreover, there is no study that has explicitly compared the performance of different predictive models. In addition, none of the predictive models in previous studies incorporated A1C variability as a predictor, specifically in ML models.

Objectives: The first objective of this study was to develop and compare predictive models, namely, ML and conventional statistical models for 3 microvascular complications (diabetic nephropathy, retinopathy, and neuropathy) in T1D patients. The second objective of this study was to develop and compare predictive models, namely, ML and conventional statistical models and evaluate whether A1C variability can help better predict each of the 3 microvascular complications (diabetic nephropathy, retinopathy and neuropathy) in T1D patients.

Methods: This was a factorial experimental study using retrospective real-world registry data. Adult T1D patients participating in the T1D Exchange Clinic Registry and met the eligibility criteria were included for the analysis. Baseline characteristics of eligible T1D patients that were measured between 2010 and 2012 were used to predict three microvascular complications that were measured till 2017. Two ML methods, i.e., support vector machine (SVM) and neural network (NN) and one conventional statistical method, i.e., logistic regression (LR) were used to develop predictive models. The three microvascular complications, i.e., diabetic nephropathy, retinopathy and neuropathy were operationalized as binary variables (yes/no). Predictors for each microvascular complication were selected. Specifically, A1C variability was manipulated into the following 5 levels: a) single A1C, b) mean A1C, c) combination single, d) combination mean, and e) multiple. Models were first developed through 10-fold cross-validation on the train set. Then the model was fit on the entire train set and evaluated on the test set. Hence, for each microvascular complication, 11 (10+1) predictive models were developed using each modeling method with each predictor set. A total of 495 models (11 x 5 predictor sets x 3 modeling method x 3 microvascular complications) were developed, 165 models for each microvascular complication. Performance measure was operationalized as F1 score. Factorial analysis of variance (ANOVA) was used to test research hypotheses. Post hoc Tukey-Kramer test was performed to evaluate which levels within a factor were significantly different. An alpha level of <0.05 was used to determine statistical significance of an association. Data preparation process, summary statistics, correlation analysis and LR were performed using SAS 9.4 (SAS Institute, Inc. Cary, NC). Predictive modelling by SVM and ANN were performed through Scikit-learn 0.22.1 and the Keras application programming interface (API) of TensorFlow™ online version 1.0.0.

Results: A total of 4476, 3595, and 4072 patients met the eligibility criteria and included in the cohort of nephropathy, retinopathy, and retinopathy, respectively. Within each cohort, 510 (11%), 659 (18%) and 579 (14%) developed nephropathy, retinopathy, and neuropathy, respectively during the follow-up period. Patients of the three cohorts were on average 38-40 (±14.5-15.4) years and had been diagnosed with T1D for an average (±SD) of 19-21 (±11.3-12.5) years. Slightly more than half (53-55%) of patients were women. For the first objective, the mean (±SD) F1 score of 33 LR models were 0.19±0.10, lower than that of 33 SVM models (0.38 ± 0.03) and 33 NN models (0.38 ± 0.03). Two-way ANOVA indicated a significant interaction between the effects of modeling method and microvascular complication on performance measure (F1 scores, $p<.0001$). ML models performed significantly better than LR models within each study cohort. Post hoc Tukey-Cramer test indicated there was no statistical difference between F1 scores of SVM and NN models. For objective 2, three-way ANOVA indicated significant interactions between modeling method, microvascular complication and A1C variability. Hence, two-way ANOVA was performed within each cohort. F test indicates that A1C variability had significant effect on F1 score of the nephropathy cohort when the modeling method was NN ($F=6.78$, $p<.0001$). Post hoc Tukey-Kramer test indicates that mean F1 scores of the nephropathy cohort from NN models using d) combination mean or e) multiple were significantly higher than using b) mean A1C or c) combination single. In the cohort of retinopathy, there is no effect of A1C variability on performance measure. Lastly, in the cohort of neuropathy, F test indicates the A1C variability had significant effect on performance measure when the modeling method was LR ($F=8.19$, $p<.0001$). Post hoc Tukey-Kramer test indicates that mean F1 score of the neuropathy cohort from LR models using e) multiple was significantly lower than using other A1C variability measures. Across all three cohorts, ML models performed significantly better than LR models.

Conclusion: The study indicates that ML models compared to LR models produced significantly higher F1 scores for predicting all three types of microvascular complications irrespective of which A1C variability measure was used. The study indicates that it is better to use A1C variability combination mean or multiple for evaluating A1C variability when predicting diabetic nephropathy in T1D patients using NN machine learning models. Future research is needed to develop decision support systems that can advise clinicians based on the results from predictive models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Artificial Intelligence & Machine Learning

Artificial intelligence (AI) is a wide-ranging area in computer science. There is no unanimous definition of AI. Russell and Norvig (2009) defined AI in terms of its goals: “AI is the field that aims at building systems that think/act rationally (like humans)” (Russel & Norvig, 2009; Bringsjord & Govindarajulu, 2018). AI techniques have been widely applied across industries, including manufacturing, retail, travel and hospitality, financial services, energy, feedstock, utilities, and healthcare and life sciences (Tripathi, 2016). Movie recommendations, speech recognition, Google's customization of individual searches based on previous web data, and driving a car using GPS navigation are some of the examples of AI applications that have already remarkably changed and improved our lives (Tripathi, 2016).

Machine learning (ML) is a sub-domain of AI. ML refers to the process that allows computers to learn automatically without human assistance to achieve the aim of learning from data. It stems from statistics and computer science and is the way to realize AI (Geron, 2017).

Big Data Has Facilitated Application of ML in Health Care

The term “big data” vividly describes the complex, diverse, and massive amount of data that is available nowadays (Murdoch & Detsky, 2013). It not only refers to the data per se, but also the science of managing, integrating, analysing, and sharing data (Manyika et al., 2011). In health care, “big data” pools include claims and cost data (owned by payers and providers), clinical data (owned by providers), pharmaceutical research and development (R&D) data (owned by pharmaceutical companies and academia), and patient behaviour data (owned by consumers and stakeholders outside health care such as retail and apparel) (Groves, Kayyali, Knott, & Van Kuiken, 2013). These databases have been utilized to answer research questions in health outcomes research for a long time. Typically, researchers try to learn from the data in order to either predict future events/health outcomes or understand relationships between variables (Breiman, 2001).

Because of the remarkable capability, efficiency, and flexibility of ML algorithms to handle data and achieve a solution, there has been a rapid expansion of ML application to the health care sector (Murdoch & Detsky, 2013). In fact, over a hundred start-up companies have emerged and applied ML to specialties of patient data and risk analytics, medical research, imaging and diagnostics, lifestyle management and marketing, mental health, emergency room and surgery, inpatient care and hospital management, drug discovery, virtual assistants, wearables, and clinical decision support software.(Mazzanti, Shirka, Gjergo, & Hasimi, 2018)

Terminology and Classification of ML

As developed by computer scientists, in ML terminology, “variables” are called “attributes”. Attributes in combination with their values are termed “features”, although in many cases, features and attributes are used interchangeably. “Outcomes” or “dependent variables” in health outcomes research are referred to as “labels” or the “solution”. There are various criteria for classifying ML systems and very often these criteria can be used in combination for classification purposes. Based on the extent and type of supervision an ML system receives during the data learning process, it can be broadly categorized as supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In supervised learning, the data used for learning contains information of the desired solution (i.e., label or the dependent variable). Typical tasks of supervised ML are classification, in which the dependent variable or label is categorical variable and regression, in which the dependent variable is continuous (Geron, 2017).

Overview of Predictive Modeling in Healthcare

Predictive modelling refers to the process of developing a mathematical tool or model to predict the probability of an outcome (Geisser 1993; Kuhn & Johnson 2013). A predictive model for a health outcome such as a disease is a model that outputs the likelihood or risk of a disease based on the input information from a patient (Steyerberg, 2019). In health care, the outcome can be, but not limited to a clinical/disease status, hospitalization, health resource utilization and expenditure, medication adherence, and patient satisfaction. Input information can be patient demographics, clinical characteristics, and lifestyle factors that are available from electronic medical records, patient claims, or survey data (Steyerberg, 2019). Once a model is developed and validated, it can be applied to predict future events in patients. Healthcare stakeholders including payers and providers can use predictive models for decision support such as risk stratification and targeting patients for interventions (Steyerberg, 2019).

Two Approaches to Achieve Prediction

Prediction can be achieved through two approaches: conventional statistical methods and advanced ML models (Shalev-Shwartz & Ben-David, 2014).

Conventional statistical predictive model is a formalization of relationship between variables in the form of mathematical equations (Shalev-Shwartz & Ben-David, 2014). Conventional statistical methods assume a stochastic data model. In other words, they assume observed data are from a random probability distribution. The outcome to be predicted can be represented as a function of independent input variables plus random noise (Shalev-Shwartz

& Ben-David, 2014). Commonly used statistical methods include regression, logistic regression (LR) and time-to-event or survival analysis.

On the other hand, ML predictive model is an algorithm that operates on input variables to predict the outcome variable(s) (Shalev-Shwartz & Ben-David, 2014). ML methods usually do not assume a parametric model between independent and dependent variables and are more liberal in techniques and approaches to achieve prediction (Contreras & Vehi, 2018). Commonly used advanced ML methods include linear support vector machines (SVMs), artificial neural networks (NNs), classification and regression trees (CARTs) & random forests (RFs) and k-nearest neighbors (Geron, 2017). The comparison of conventional statistical modeling versus ML modeling is summarized in **Table 1**.

Table 1. Comparison of conventional statistical and ML modeling methods

	Statistical Modeling	**ML Modeling**
Definition	"Parametric formalizations of relationships between independent and dependent variables in the form of mathematical equations"	"Algorithms that operate on independent variables to predict the dependent variable(s) without clear formalization of the relationship"
Commonly used methods	Linear regressions, logistic regressions, Cox models	SVMs, NNs, CARTs & RFs, k-nearest neighbors
Assumptions	Rigid assumptions about the relationship and data distributions	No rigid assumptions about the problem and data distributions in general
Training	No 'training' process	'Training' is needed to tune the model
Techniques used for modeling	Conservative in techniques and approaches	More liberal in techniques and approaches
Predictors	Often require independent predictor variables and less number of predictors	Can handle multicollinearity, redundancy in data and 'wide' data

Performance Measures of Predictive Models

The performance of predictive models can be evaluated mainly by two types of measures: basic single-threshold measures and threshold-free measures (He & Garcia, 2009). Commonly used single-threshold measures include accuracy, sensitivity, specificity, precision and F1 score; and commonly used threshold-free measure includes area under receiver-operating characteristics curve (AUC) (Jiao & Du, 2016). The confusion matrix and the calculation of single-threshold measures are listed in **Tables 2** and **3.** As ML models can be 'trained', a single performance metric can be chosen as the target for improvement. Hence, it is critical to choose the appropriate performance metric in order to serve the researchers' specific prediction goals.

Table 2. The confusion matrix

	Predicted Positive	**Predicted Negative**	**Total**
Actual Positive	True Positive (TP)	False Negative (FN)	TP + FN
Actual Negative	False Positive (FP)	True Negative (TN)	FP + TN
Total	TP + FP	FN + TN	TP + TN + FP + FN

Table 3. The formula of single-threshold performance metrics

Performance Metric	**Formula**
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Precision	$\frac{TP}{TP + FP}$
F1 score	$2 \times \frac{Sensitivity \times Precision}{(Sensitivity + Precision)}$

Choosing the Right Performance Measure for Classification of Imbalanced Data

In healthcare, we face classification problems a lot, e.g., to categorize patients into diseased/non-diseased, high-risk/low-risk or case/control groups. In most cases, the data contains unequal number of cases and controls and specifically, the number of cases is less than the number of controls. This is the simplest manifestation of imbalanced data. More generally, imbalanced data refers to the unequal representation of different levels of the class (Li & Mao, 2014). The imbalance nature of the data not only makes correct prediction of the less represented class more difficult, but also results in misleading perceptions of model performance based on commonly used performance metrics, such as accuracy and AUC (He & Garcia, 2009; Valverde-Albacete & Peláez-Moreno, 2014; Akosa, 2017). For example, among a total of 1000 individuals, 10 are ‘diseased’ and 990 are not. In the most extreme case, a model correctly predicted the 990 non-diseased individuals while misclassifying those 10 patients as non-diseased, the accuracy of the model is as high as 99%. However, the model fails to identify any diseased patients. This exemplifies the so-called ‘accuracy paradox’ where a high accuracy does not indicate a ‘good’ model performance (Valverde-Albacete & Peláez-Moreno, 2014; Akosa, 2017). This causes a problem especially when researchers aim to correctly identify the ‘diseased’ cases. Accuracy and AUC are calculated based on the predictive model’s capability of identifying both ‘cases’ and ‘controls’. If we focus more on the correct identification of the ‘cases’ or the less prevalent class from imbalanced data when developing predictive models, the F1 score (the harmonic mean of precision and sensitivity) is a better indicator for model performance (He & Garcia, 2009; Jiao & Du, 2016).

ML Predictive Models in Diabetes Management

The number of published articles in Google Scholar that involve both diabetes and ML have increased remarkably, from around 500 in the year of 2000 to over 10,000 in 2017 (Contreras & Vehi, 2018). A review of literature from PubMed on ML in diabetes management published between 2010 and 2018 found a total of 141 English articles, majority of which were published between 2015 and 2018. These literatures cover diverse aspects of diabetes management, the top three being closed-loop systems ("artificial pancreas", 22%), daily-life support in diabetes management (e.g., a decision support system or DSS that monitors a patient's diet, physical activity, medication use, and glucose measurements and applies ML algorithms to learn from recorded data in order to assist patients and clinicians with informed decision making, 21%), and real-time blood glucose prediction (e.g., prediction of blood glucose excursion using data captured by continuous glucose monitor or CGM, 19%). Other areas include risk and patient stratification (13%), detection of adverse glycemic events (10%), insulin bolus calculators and advisory systems (9%), and detection of meals, exercise and faults (6%) (Contreras & Vehi, 2018).

As ML is a very powerful tool in prediction, this study tries to apply ML algorithms to research in type 1 diabetes (T1D). Following is an overview of T1D, its associated complications, and consequential clinical, economic, and social impacts.

Overview of Type 1 Diabetes and its Management

T1D is a chronic progressive disease characterized by elevated blood glucose level, abnormalities of carbohydrate, fat, and protein metabolism (Bluestone, Herold, & Eisenbarth, 2010; Chiang, Kirkman, Laffel, & Peters, 2014; Todd, 2010). Common symptoms of T1D include frequent urination, excessive thirst, extreme hunger, unusual weight loss, increased fatigue and irritability, and blurry vision (Atkinson, Eisenbarth, & Michels, 2014). It's usually diagnosed at a younger age (in children and adolescents) (Chiang et al., 2014) and slightly more common in boys and men (*Global report on diabetes*, 2016; Ostman et al., 2008). Seasonal variations also exist, in which more T1D cases are diagnosed in autumn and winter (Moltchanova, Schreier, Lammi, & Karvonen, 2009) and birth in the spring is associated with a higher chance of having T1D (Kahn et al., 2009). Worldwide, there are around 23 million individuals affected by the disease (*Global report on diabetes*, 2016; Cho et al., 2018). In the United States (U.S.), over 1.5 million people have T1D with 40,000 new cases diagnosed every year (*Type 1 Diabetes*, 2019). Treating T1D and its complications is expensive: the total cost is approximately $15 billion every year in the US (Tao, Pietropaolo, Atkinson, Schatz, & Taylor, 2010). A recent study found that the per patient per year (PPPY) cost for T1D was over $18,817, which was significantly higher than the costs for treating type 2 diabetes (T2D) (Joish et al., 2020).

Glycemic Control in T1D Management

Glycemic control is critical in preventing and slowing the progression of diabetic microvascular complications (Association, 2019d). Glycated hemoglobin (A1C) level is a useful indicator of blood glucose control. It estimates a patient’s blood glucose level over a period of three months (Ontario, 2018). Excellent glycemic control can substantially reduce the incidence of ESRD, retinopathy, neuropathy, myocardial Infarction, stroke, and all-cause mortality. It can also improve patients’ QoL and reduce healthcare costs due to avoidable complications (Herman et al., 2018). Thus, treatment guidelines usually recommend a certain A1C level as a goal to assist clinicians and patients in judging whether their diabetes are well managed or not. The American Diabetes Association (ADA) 2019 guidelines sets a glycemic target of A1C < 7.0% for many nonpregnant adult patients (Association, 2018a). However, normoglycemia is not achieved by around 80% of adult T1D patients, even with the many advances in treatment modalities (Juarez, Ma, Kumasaka, Shimada, & Davis, 2014; Foster et al., 2019)

Treatment for T1D

Insulin therapies are essential in helping patients achieve glycemic control and are the current standard of care for T1D patients (Association, 2019b). They are categorized as rapid-acting (aspart, lispro, glulisine, and insulin human), short-acting (regular R), intermediate-acting (NPH or isophane insulin), and long-acting (glargine, detemir, albulin, and degludec) based on the drug's time of onset and duration of action. Short-acting and rapid-acting insulins are used at meal times (bolus) and are often used together with an intermediate-acting or long-acting insulin, which keeps consistent blood glucose levels during periods of fasting (basal) (Association, 2019c). Since 2000, newer generations of insulin and its analogues as well as their modes have been developed (**Appendix 1**). Other advances in diabetes management include devices for glucose monitoring such as blood glucose meters and continuous glucose monitors (CGM), closed loop systems, and transplantation (Aathira & Jain, 2014). Lifestyle management including diabetes self-management education and support, nutrition therapy (weight management and carbohydrates), and physical activity is also important (Association, 2019a).

Three Types of T1D Related Microvascular Complications

T1D is associated with chronic complications. Elevated glucose level can promote pathological change of the blood vessels (such as sclerosis and abnormal proliferation of vascular endothelial cells inside the capillary), which can affect the kidneys, eyes, and nerves and lead to diabetic nephropathy, retinopathy, and neuropathy (Fowler, 2008).

The prevalence of diabetic nephropathy or kidney disease among T1D patients is around 15%-40% (Viswanathan, 2015). Microalbuminuria is the earliest phenotype of diabetic nephropathy and has an annual incidence rate of 2-3% (Marshall, 2012). Certain race/ethnicity groups including South Asians, American Hispanics, and African-Americans are at higher risk of developing diabetic nephropathy (Ameh, Okpechi, Agyemang, & Kengne, 2019). Diabetic nephropathy is associated with long term macrovascular complications such as end-stage renal disease (ESRD)/renal failure and cardiovascular diseases (Fowler, 2008; Viswanathan, 2015) and it can significantly drive health care cost: Patients with diabetic nephropathy have between $3,580 - $12,830 more costs than patients without (Zhou et al., 2017).

Diabetic retinopathy is the most common microvascular complication among the three types (Fowler, 2008). It is associated with other two types of microvascular complications, macrovascular complications and blindness (Fong, Aiello, Ferris, & Klein, 2004; Pearce, Simó, Lövestam-Adrian, Wong, & Evans, 2018), adversely impacts health-related quality of life (HRQoL) (Chen et al., 2010), and drives healthcare resource utilization(Candrilli, Davis, Kan, Lucero, & Rousculp, 2007).

Diabetic neuropathy is a group of complications that is composed of both diabetic peripheral neuropathy (DPN) and diabetic autonomic neuropathy (DAN) (Association, 2018b). DPN is symptomized by numbness, burning, and tingling pain in extremities, although up to 50% of patients can be symptomless (Association, 2018b). The prevalence of DPN based on European data ranges from 6% to 34% in diabetic patients (Alleman et al., 2015). DPN increases the risk for foot ulceration and amputation, which further associates with mortality and worse HRQoL in diabetic patients (Alleman et al., 2015; Pop-Busui et al., 2017). The PPPY medical costs for diabetic patients with DPN ranged between $12,492 and $30,755 in 2015, which were significantly higher than those patients with diabetes only ($6,632) (Sadosky et al., 2015). On the other hand, DAN is less prevalent than DPN. DAN is a group of disorders including gastroparesis, constipation or diarrhea, bladder dysfunction, erectile impotence, and cardiovascular autonomic dysfunction (CAN) (Boulton et al., 2005).

These three types of microvascular complications are often synergic and if not well managed, can lead to poor prognosis, adversely impact HRQoL, and drive healthcare costs (Atkinson et al., 2014; Kähm, Laxy, Schneider, & Holle, 2019). The costs for treating diabetes-related complications in T1D patients was estimated to be $7,816 PPPY (Joish et al., 2020). Hence, early screening and prevention of these complications is critical in T1D management (Association, 2019d).

Current Screening Approach for the Three Types of Complications

The ADA treatment guidelines recommend annual screening for nephropathy, retinopathy, and neuropathy starting at five years after diagnosis of T1D (Association, 2019d). For subgroups of patients with or without specific conditions, timing and frequency of examinations can be changed. For example, all T1D patients with comorbid hypertension should have nephropathy assessment at least once a year. More frequent eye examination is recommended for patients with existing evidence of retinopathy (Association, 2019d).

Predictive Models may Enhance the Screening and Prevention of the Three Types of Microvascular Complications

However, there is still space for improvement in screening and prevention of microvascular complications.

For the diagnosis of diabetic nephropathy, which also applies to all clinical tests, positive results will need to be confirmed by a second or repeated tests due to differences in laboratory methods, urine samples, and definition of nephropathy (de Jong & Curhan, 2006). A 2017 study suggested that utilization of kidney disease risk scores may be helpful and cost-effective in identifying at-risk patients (Yarnoff et al., 2017).

Many studies tried to establish a screening schedule of eye examination that would be more efficient in managing T1D. The Diabetes Control and Complications Trial (DCCT) and Epidemiology of Diabetes Interventions and Complications (EDIC) group recommended an individualized eye screening approach based on patient's state of retinopathy in 2017 (Nathan et al., 2017). Status of retinopathy is categorized into no retinopathy, mild, moderate, or severe non-proliferative retinopathy, and advanced retinopathy (including proliferative diabetic retinopathy, clinically significant macular edema, or previous self-reported treatment with panretinal or focal photocoagulation, intraocular glucocorticoids, or anti-VEGF agents). They reported that patients with lower risk of retinopathy progression (such as those with no retinopathy) can receive less frequent screening (i.e., at 4-year or 3-year intervals) whereas those at higher risk need to receive more frequent eye exams (i.e., at 6-month or 3-month intervals). Personalized screening schedules would result in 58% reduction (10.7 fewer) of retinal examinations and cost savings of approximately $1 billion (43% decrease) over a 20-year period compared to annual screening after 5 years (Nathan et al., 2017). A 2016 systematic review compared cost-effectiveness of eye exam by clinic camera and telemedicine and concluded that telemedicine screening can save cost and improve access, especially in low- and middle-income countries, where nearly 80% of all diabetic patients live (Pasquel et al., 2015). Researchers also revealed lack of compliance in receiving eye examinations in low-socioeconomic-status patients (Margaret M. Byrne et al., 2014; Pasquel et al., 2015). One study found community-based retinal screening can be cost-effective (slightly over $100 per person screened) (M. M. Byrne et al., 2014).

Lastly, for the screening of diabetic neuropathy, a study compared different screening tests for DPN, including Michigan Neuropathy Screening Instrument (MNSI), Semmes-Weinstein Monofilament (SWM), vibration sensation and ankle reflex, in terms of simplicity, reliability, and accuracy (Al-Geffari, 2012). The author indicated that even though methods correlated with each other, they often came to very different conclusions. A combination of screening methods for diabetic neuropathy would increase sensitivity and specificity (Al-Geffari, 2012). A more recent study examined effectiveness of different screening approaches (Brown, Pribesh, Baskette, Vinik, & Colberg, 2017). It echoed previous findings that different tools can be used in combination and suggested that future study is needed to refine and develop new screening methods. Significant increase in cost occurred during the diagnostic period compared to the baseline period. Similarly, a retrospective study using Health and Retirement Study (HRS) - Medicare Claims linked database indicated that research is needed to improve efficiency in DPN evaluation (Callaghan et al., 2012).

Inefficiency of healthcare resource and expenditure use is one aspect of concern. Moreover, patients may not benefit the most following current screening approaches. This is because with current screening guidelines, patients would probably assume or have the misconception that their risks of developing certain microvascular complications are equal after 5 years, which is not true. This would especially pose a problem for patients at higher risk of disease progression. This may partly explain the low compliance of patients to attend annual screening (Molitch et al., 2004).

A economic study in 2003 suggested that a predictive risk model was the most efficient tool for screening patients compared to lab tests, although it may not be the most accurate method, probably due to lack of accuracy in models developed by conventional statistical methods (Zhang et al., 2003). A predictive model to differentiate patients who are at risk for each of the three microvascular complications can be useful in informing healthcare providers and patients and may change patients' perceptions and potentially change their health behavior, including but not limited to attending screening appointments, becoming more watchful for signs of disease progression, better compliance to insulin therapies and glucose monitoring, healthier diet and more exercise. Predictive models for microvascular complications in T1D patients may also facilitate intervention in at-risk patients and result in long-term cost savings.

Research Objective

Hence, this study intended to develop and compare predictive models for diabetic nephropathy, retinopathy, and neuropathy in T1D patients using both conventional statistical and ML methods. It also aimed to incorporate predictors that were not included in previous studies in prediction and assess whether inclusion of the predictor would impact the prediction of each of the three microvascular complications. This study directly compared the performance of conventional statistical methods and ML methods in prediction by using the same predictors for each microvascular complication. It supplements current knowledge in understanding relationship between patient, clinical and contextual characteristics and each complication. It may serve as a preliminary screening tool to identify at-risk patients for further confirmatory laboratory tests and help patients and their health care providers (HCPs) for better informed T1D management.

In CHAPTER 2, microvascular complication risk factors and previous predictive models for each of the three microvascular complication in T1D patients that employed ML methods will be reviewed. The knowledge gap and the research questions will be discussed.

CHAPTER 2

Literature Review

Risk Factors for Microvascular Complications

With major improvement in diabetes management, progression to long term macrovascular morbidity and mortality is delayed (Association, 2019d). Because microvascular complications can put patients at risk of developing major morbidity and mortality, the ADA guideline emphasizes the importance of screening for, preventing, and delaying the progression of diabetic nephropathy, retinopathy, and neuropathy (Association, 2019d). Extensive researches have been conducted to assess risk factors (besides hyperglycemia) for diabetic complications. Common ones include older age, certain races, longer duration of T1D, dyslipidemia, hypertension, overweight and obesity, smoking, and inactive lifestyle (*Risk factors for complications*, 2018; Association, 2019d). Retinopathy itself is a risk factor for the other two types of microvascular complications (Association, 2019d). Ulceration is a specific risk parameter for neuropathy (Donnelly, Emslie-Smith, Gardner, & Morris, 2000). On the other hand, use of (angiotensin-converting enzyme) ACE inhibitors may reduce the risk of progressing to microvascular complications in T1D patients (Donnelly et al., 2000).

Impact of Glycemic Variability on Microvascular Complications

A level of A1C <= 7% was established as the gold standard of glycaemic control from the DCCT, the largest clinical trial in T1D patients in the U.S. However, patients with similar mean A1C levels had quite differential risk of developing retinopathy (Group, 1995). Thus, researchers have been looking for other parameters to account for diabetes progression. There is on-going debate on the association between glycemic variability (both short-term and long-term) and diabetes complications. A SLR implied that within-day glucose variability (or short-term variability) could predict complications in type 2 diabetes (T2D) patients independent of A1C levels. However, the evidence for T1D patients is inconclusive (Nalysnyk, Hernandez-Medina, & Krishnarajah, 2010). A SLR and meta-analysis in 2015 indicates that in both T1D and T2D patients, A1C variability (or long-term variability) was adversely associated with both micro- and macro- vascular complications and mortality independently of the mean A1C value (Gorst et al., 2015). Nevertheless, many factors can contribute to the variation rather than the true biological variability (Sacks, 2011). Future research is needed to better understand the role of glycemic variability in the progression of diabetic complications and apply it in clinical risk assessment (Nalysnyk et al., 2010; Gorst et al., 2015).

Comparison of A1C and Glucose Variability

A1C is formed by the attachment of glucose to haemoglobin and it is contained by red blood cells (erythrocyte). Because the lifespan of erythrocytes is around 120 days, an A1C usually indicates the glucose level over a period of three months (Nathan et al., 2008). Commonly used measures of A1C variability include standard deviation (SD: measures how much values differs from the group mean), adjusted SD (accounting for the number of measures) and coefficient of variation (CV: = SD/mean). Biological variation of A1C within a non-diabetic individual over time is usually minimal (Kilpatrick, Maylor, & Keevil, 1998), whereas variation between individuals is greater (Sacks, 2011). Unlike blood glucose level, which can be affected by numerous pre-analysis factors such as food ingestion, prolonged fasting, exercise, medications, venous stasis, posture, sample handling, the source of blood, acute disease that can alter glucose concentration, and even acute stress, A1C is mainly influenced by an individual’s erythrocyte life span, race, and presence of iron-deficiency anemia (Sacks, 2011). Hence, A1C variability provides a more stable estimate for glucose variation of an individual. Although inconclusive, greater extent of glycaemic variability, especially long-term A1C variability can put T1D patients at higher risk of diabetes complications independent of mean A1C (Nalysnyk et al., 2010; Gorst et al., 2015). A 2018 study evaluated different ways of measuring A1C variability and found that adjusted standard deviation (adj-SD) of A1C was the best predictor of all-cause mortality among T2D patients in terms of statistical significance and odds ratio plus its 95% confidence interval (Orsi et al., 2018). Hence, SD of multiple A1C values were used as one operationalization of A1C variability in this study.

Predictive Models for Microvascular Complications in T1D Patients Using ML

Based on previous knowledge, predictive models for diabetes complications has ensued to assist informed clinical decision making (Lagani, Koumakis, Chiarugi, Lakasing, & Tsamardinos, 2013; Cichosz, Johansen, & Hejlesen, 2015; Lagani et al., 2015; Kavakiotis et al., 2017; Dagliati et al., 2018). Two published SLRs revealed that most existing prediction models in diabetes research were about long-term macrovascular outcomes such as cardiovascular diseases or mortality and were based on data from patients with T2D alone or a mixture of T2D (majority) and T1D patients (Lagani et al., 2013; Cichosz et al., 2015). How much are those findings applicable to T1D patients is unknown. We have also witnessed an emerging trend in the methodology used in the models: although conventional statistical methods (e.g. LR, Cox model) were adopted quite often, newer machine learning (ML) algorithms have been applied to the field (Kavakiotis et al., 2017; Contreras & Vehi, 2018; Dagliati et al., 2018).

A SLR was conducted to identify predictive models for microvascular complications in T1D patients using ML algorithms and published in the Journal of Medical Artificial Intelligence (Xu, Wang, & Sansgiry, 2019). A total of six studies were found, among which, four studies used data obtained from T1D patients alone and two used data from both T1D and T2D patients (Skevofilakas, Zarkogianni, Karamanos, & Nikita, 2010; Vergouwe et al., 2010;

Aspelund et al., 2011; Lagani et al., 2015; Kazemi, Moghimbeigi, Kiani, Mahjub, & Faradmal, 2016; Ravizza et al., 2019). To briefly summarize the findings, only one study developed predictive models for all three types of microvascular complications whereas the other five focused on the prediction of either diabetic nephropathy, retinopathy or neuropathy. The outcomes of diabetic nephropathy and retinopathy were predicted 3 times, respectively and diabetic nephropathy predicted twice. There is considerable variation in the definition of each microvascular complication, due to which it is hard to directly compare the performance of predictive models for the same microvascular complication from different studies. There is a paucity of large contemporary longitudinal real-world data to evaluate disease progression in T1D patients, especially in the United States (Xu, Wang, & Sansgiry, 2019).

Common predictors used across studies and across three types of microvascular complications included age, gender, diabetes duration, body mass index (BMI), blood pressure, lipid level, and mean or a single HbA1C value. The study using the DCCT/EDIC data is most robust in terms of comprehensiveness of predictors – in addition to previous mentioned factors, they also included measures of insulin use (insulin regimen, total insulin daily dosage), additional patient demographics (marital status and occupation), post pubescent diabetes duration, presence of neuropathy, past history of severe hypoglycemia (SH) and hospitalization(s) due to diabetic ketoacidosis (DKA), family history of T1D and other types of diabetes, and even measures on patient attempted suicide and specific ideal body weight (Lagani et al., 2015). A SLR in 2017 summarized common clinical, environmental, and genetic risk factors for DPN, and indicated that future research is needed to confirm the relationship between psychological factors and progression of DPN (Hébert, Veluchamy, Torrance, & Smith, 2017). We did not find any study that incorporated A1C variability as a predictor.

Attempted ML algorithms included classification and regression tree (CART) and random forest (RF) (CART/RF, n=3), support vector machines (SVMs, n=2), logistic regression (LR, n=2) and neural networks (NNs, n=1) (Xu, Wang, & Sansgiry, 2019). Within ML models, SVMs and NNs were reported to perform better than other models in these studies. Hence, these two methods were chosen for our research.

Model performance was evaluated using either AUC (n=4) or accuracy (n=2) (Xu, Wang, & Sansgiry, 2019). Moreover, none of these models targeted to improve the F1 score. How well these models can identify patients at risk is questionable, especially considering the imbalanced nature of the data.

Research Objectives

The first objective of this study was to develop and compare predictive models, namely, ML and conventional statistical models for 3 microvascular complications (diabetic nephropathy, retinopathy, and neuropathy) in T1D patients.

The second objective of this study was to develop and compare predictive models, namely, ML and conventional statistical models and evaluate whether A1C variability can help better predict each of the 3 microvascular complications (diabetic nephropathy, retinopathy and neuropathy) in T1D patients.

Significance

Predictive models can serve as a convenient and less expensive way for patient risk identification (Zhang et al., 2003). While there are six studies that have developed predictive models for microvascular complications in T1D patients, none of them focused on enhancing the F1 score of the models, which is a better indicator for a model's capability of identifying patients at risk with imbalanced data. Moreover, none of previous studies explicitly compared the performance of different modeling methods. This study adds to current knowledge by explicitly comparing the performance of two ML methods and conventional logistic regression using the same predictor sets. The development of these predictive models for diabetic microvascular complications has the following clinical implications: first, the focus on improving F1 score can better help identify those patients who are at higher risk for each microvascular complication (Harrell, Lee, & Mark, 1996), which can bring these high-risk patients to the attention of their HCPs; HCPs can use the estimates to make informed decisions. For example, they can order confirmatory lab tests earlier and provide more appropriate treatment and education for high-risk patients. Second, the study provides a better understanding of relative importance of risk factors for each microvascular complication among T1D patients. Specifically, the effect of A1C variability on T1D prognosis was evaluated. This can supplement current knowledge in terms of how multiple A1C measures of patients can be utilized in clinical settings for decision support. Specifically, HCPs can record multiple A1C values and calculate their standard deviations to represent A1C variability. Algorithms can also be developed to evaluate the variability in A1C and the information can be used in predictive models for identifying high risk patients for microvascular complications. The knowledge of risk factors can be used in designing future clinical trials involving T1D patients such as patient stratification based on important risk factors. Third, it can assess the relative therapeutic benefit of different types of contemporary insulin therapies as well as diabetes management modalities. Last but not the least, it may help in efficiently allocating health care resources based on patients' needs and thus, potentially save health care cost. Predictive models can be used as a preliminary screening tool in hospitals and other primary care settings to improve efficiency as well as test accuracy. For example, patients at lower risk can be ordered less frequent lab test for certain complications.

Innovation

This was among the first studies that utilized experimental design to explicitly compare the performance of different predictive modelling methods for each of the three microvascular complications. This was also among the first to develop predictive models for diabetic nephropathy, retinopathy and neuropathy in T1D patients with a specific focus on enhancing the F1 score, which is a better indicator of a model's capability of identifying 'cases', whereas previous studies focused on other metrics such as accuracy and AUC, which may not be an appropriate indicator of model performance, especially when the data is imbalanced. Recent scientific findings point to the delay of progression to long-term macrovascular complications and emphasize the importance of early screening and prevention of micro-vascular complications. Although annual physical examination of feet, eyes, and urine lab works are recommended for patients who have been diagnosed with T1D for at least 5 years, the screening approach can be individualized provided a dependable predictive model that identifies individual's risk. This study is innovative in that it is among the first to utilize advanced ML methods, including SVMs and NNs for the prediction of microvascular complications among T1D patients. Only a few studies predicted microvascular complications in T1D patients in the United States. And even fewer studies predicted diabetic neuropathy among T1D patients. This study is among the first to comprehensively assess patient risk of developing both peripheral and autonomic neuropathy in T1D patients. Furthermore, A1C variability was incorporated into the predictive models for the first time.

In CHAPTER 3, the theories that guided this research will be discussed and followed by specific aims and hypotheses.

CHAPTER 3

Theoretical Framework

The conceptualization of this study was based on the Statistical Learning Theory and Andersen Behavioral Model. Statistical Learning Theory was used to guide the model development and Andersen Behavioral Model was used to guide the predictor selection.

Statistical Learning Theory

The statistical learning theory was used to guide model development, validation and comparison (Shalev-Shwartz & Ben-David, 2014). Specifically, we focused on prediction for binary classification. According to the statistical learning framework, a formal model contains the following:

1. **Input**: includes a domain set, X, a label set y, and training data $S = (X, y)$. The domain set, X refers to all the objects that we want to classify. In this study, the domain set is all T1D patients. One domain point is an individual patient and is referred to as an instance. The instance can be represented by a vector of features, or characteristics of T1D patients (e.g., age, gender, etc.). X is also referred to as the instance space. The label set y refers to the classes we want to predict. In this study, y is a two-element set $\{0,1\}$ where 0 denotes non-diseased (e.g., not having diabetic nephropathy) and 1 denotes diseased (e.g., having diabetic nephropathy). The training data $S = (X, y)$ are the data we have access to.
2. **Output**: a classifier/predictive rule $h: X \rightarrow y$ that can be used to predict future domain points. In this study, the classifier can be conventional logistic regression (LR) models and advanced ML models (SVMs or NNs).
3. **A data-generation model**: we assume X are generated by an unknown probability distribution D. There is a correct labeling/classifying function $f: X \rightarrow y$ that applies

to all instances that we want to learn. This condition can be relaxed as not all label y can be fully determined by the unknown features of X.

4. **Measure of success**: The error of a predictive model, $h: X \rightarrow y$ is defined as the probability that it does not predict the correct label on a random data point generated by the underlining distribution D. It is denoted as L, or loss of a predictive model, when $h(x) \neq f(x)$.

The adapted model from the statistical learning theory is depicted in **Figure 1**.

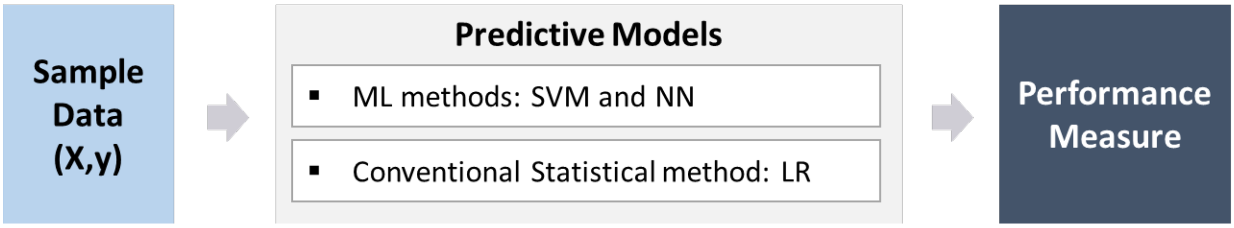


Figure 1. Adapted model using the statistical learning theory

Empirical Risk Minimization

Conventional statistical models such as LR do not normally have a training process. They rely heavily on predictor selection and use the whole training data for modeling and use the test set for evaluation. ML, on the other hand, can train the models to minimize the loss/error L. As the training data is the only information that is known to us, ML methods try to minimize the error based on the training data. This error is referred to as the empirical error and the process of its minimization is called empirical risk minimization (ERM) (Shalev-Shwartz & Ben-David, 2014).

Overfitting

As a ML model trains and learns, it may predict on the training set excellently, yet poorly on the test/new data. This is called overfitting. To prevent overfitting, different ML methods take different approaches to prevent overfitting so that the model can perform well on the training set and potentially as well on the test/new data.

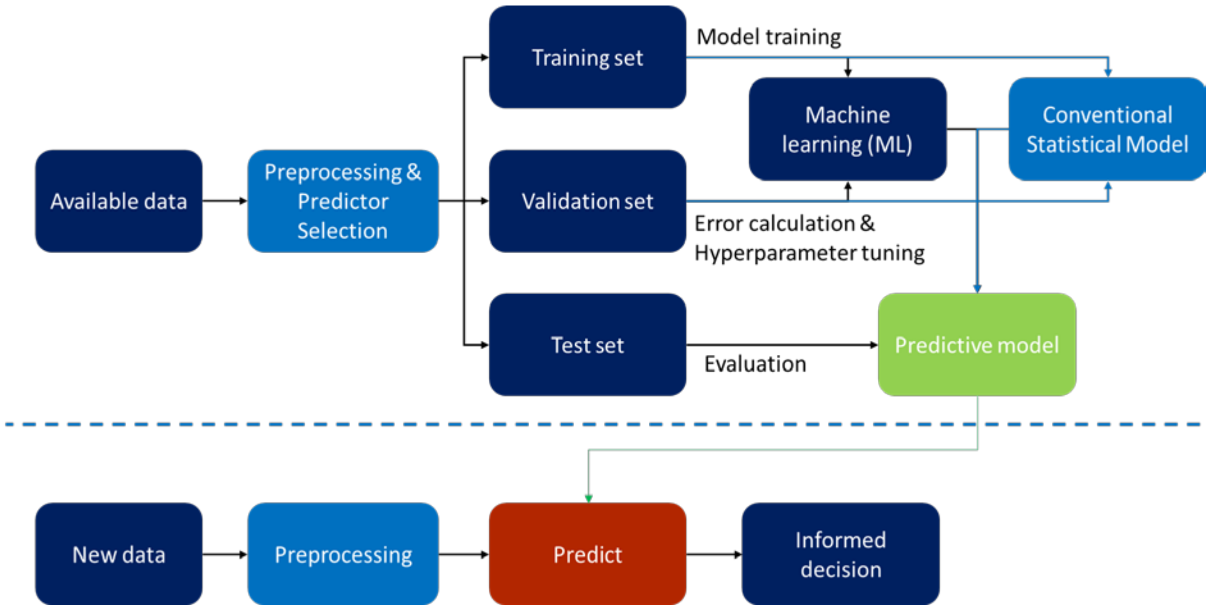


Figure 2. Illustration of general process of predictive modeling

The general process of the predictive modeling process is illustrated in **Figure 2**. The explanation of the three modeling methods is provided below.

Logistic Regression

A LR uses the maximum likelihood estimation and makes the following four assumptions of the label y (Vittinghoff, Glidden, Shiboski, & McCulloch, 2012):

1) y follows a binomial distribution.
2) The expected mean of y is given by the logit function:

$$E[y|X] = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

3) Values of y are statistically independent.
4) The residuals after fitting the model should be normally distributed (for this assumption however, LR is very robust to violation of normality, especially when sample size is large enough).

The sigmoid (S-shaped) logistic function outputs a number between 0 and 1, which is the probability of the outcome belongs to a class (e.g., diseased or not diseased). The most commonly used cutoff point for labeling is 0.5., i.e., if the probability is below 0.5, it predicts y to be 0 or 'not diseased', and 1 or 'diseased' otherwise. The advantages of LR include few assumptions made for predictors (such as their distributions), easy interpretation of parameter estimates, and known statistical significance of each predictor. For these advantages, LR has been widely used in health care research and accepted by clinicians (Vittinghoff et al., 2012). However, the implicit assumption of linear relationship of risk with respect to the log-odds parametric transformation may not hold (Westreich, Lessler, & Funk, 2010). Also, LR requires "long" data, in which the observations is more than the predictors used in modeling. Violation of either assumption or a small sample size will yield poor estimates (Geron, 2017).

One note for LR is that, as with other conventional statistical methods, it cannot handle correlated predictors. For the measure of A1C variability, it usually has to adopt a summary measure such as the mean-A1C or SD-A1C to represent the variability of A1C measured at different time points.

Support Vector Machines (SVMs)

An SVM is an algorithm that can utilize a variety of parametric and nonparametric models for classification. An SVM attempts to find the best dividing hyperplane that maximizes the margin between classes (called "large margin classification") (Geron, 2017). The instances that locate at the edge of the separating hyperplane will determine the margin between classes and the decision of the best hyperplane, and hence, they are called 'support vectors'. SVMs can easily handle high-dimensional data, and they do not assume a parametric relationship between the model predictors and outcome.

Figure 3 illustrates how an SVM works on a two-dimensional plane to classify patients as diseased or not diseased (with two features or predictors of mean A1C and BMI).

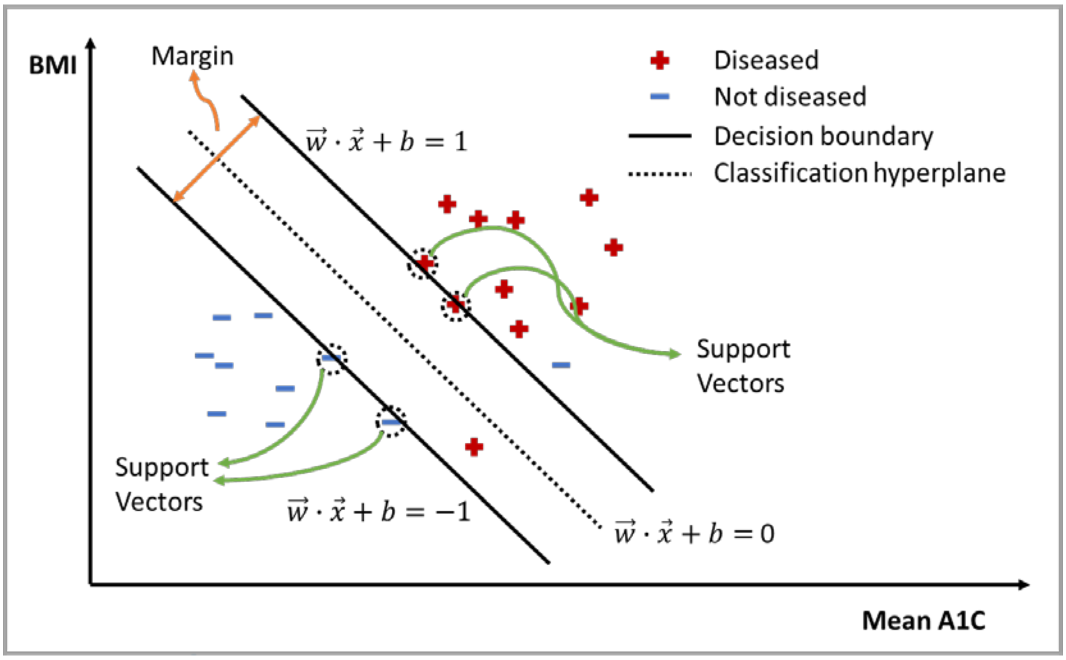


Figure 3. Illustration of an SVM (made-up example, not based on actual data)

Assuming the labels of y are +1 (diseased) or −1 (not diseased). A linear SVM classifier is based on a linear discriminant function of the form $f(X) = \vec{w} \cdot \vec{x} + b$. The vector $\vec{w}$ is the weight vector, and b is called the bias. The classification hyperplane is defined by $\vec{w} \cdot \vec{x} + b = 0$, as illustrated, a line (with 2 predictors). The line is perpendicular to vector $\vec{w}$ and go through the origin. When more predictors are incorporated to inform the classification, the classification hyperplane will become a plane in three dimensions, and more generally, a hyperplane in higher dimensions. When $\vec{w} \cdot \vec{x} + b \geq 1$, an instance is categorized as diseased, and when $\vec{w} \cdot \vec{x} + b \leq -1$, an instance is categorized as not diseased. The SVM optimizes the weight vector by minimizing the 'hinge loss':

$$l = \max{(0, 1 - t \cdot y)}$$

Where $y = f(X)$ and $t = \pm 1$ $(the\ intended\ output\ of\ class)$. When y is predicted correctly, hinge loss l is 0; when predicted y is far from t, l gets larger. The weight vector can be optimized to minimize the loss l. SVMs are also capable of non-linear classification.

SVMs are sensitive to the scales (or data distribution) of predictors. Hence, predictors will need to be pre-processed such as standardized before the step of modelling (Geron, 2017). However, in practice, we usually don't standardize predictors in a logistic model because of easy interpretability of parameter estimates.

Hyperparameters for SVMs

SVMs can be tuned to improve prediction via certain hyperparameters. These hyperparameters are not directly estimated from the data but specified *a priori* by the researcher. The hyperparameters for an SVM include 1) the soft-margin constant C and 2) parameters of the kernel function (Geron, 2017).

1) Soft-margin constant C (also called the C hyperparameter) is used to balance the trade-off between margin maximization and violations of the margin (errors on the training set: observations that fall within the margin or are even misclassified). A smaller C value will lead to a larger margin but more margin violations. In practice, C is varied through a wide range of values and the optimal value is assessed through cross-validation using the training set.
2) Kernel parameters are used to affect the decision boundary. The degree of the polynomial kernel and the width parameter of the Gaussian kernel can be specified to

make an SVM model more flexible. The lowest degree polynomial is the linear kernel (or no kernel at all). A linear SVM usually works well in many cases (Geron, 2017). Radial Basis Function (RBF) is another commonly used kernel function. It is expressed as

$$K(x_i, x_j) = \exp\left(-\gamma||x_i - x_j||\right)^2$$

SVM tries to find the minimum of the following problem:

$$\min_{\alpha_i} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \exp\left(-\gamma||x_i - x_j||\right)^2 - \sum_{i=1}^{n} \alpha_i$$

In which γ is a number. The default value in Sci-Kit Learn SVC classifier is 'scale', which equals to 1 / (n_predictors * X.var()), where X represents the matrix of predictors and var() calculate the variance matrix of X. $\sum_{i=1}^{n} y_i \alpha_i = 0, 0 \leq \alpha_i \leq C$. The minimum value of the above problem depends on both hyperparameters C and γ.

Neural Networks (NNs)

NNs were originally designed to mimic the behavior of biological neurons. Each individual biological neuron can receive and transmit signals to thousands of other neurons, and it seems that they are organized in consecutive layers. Together they constitute a complex biological neural system (Geron, 2017). A NN was first invented in 1943 by the neurophysiologist Warren McCulloch and the mathematician Walter Pitts (McCulloch & Pitts, 1943). NNs have evolved over the years and became one of the most powerful ML algorithms in handling large and complex problems (Geron, 2017).

NNs are usually composed of an input layer, one or more 'hidden' layers, and an output layer. Each layer can have multiple neurons (nodes). For each training instance, the NN feeds the predictor values to the neurons in the input layer, randomly assigns weights to multiply the value, computes the weighted sum plus a bias term (usually 1) to feed to the neurons in the consecutive hidden layer. This process is repeated until reaching the output layer, yielding the probability of the outcome. Thus, the algorithm makes a prediction each time (forward pass). This probability is compared to the observed value (yes-1 vs no-0) to calculate the error. Then the model goes through each layer in reverse to measure the error contribution from each connection and adjust the weights at each connection to reduce the error (reverse pass). This type of NNs is called feed forward NNs where the connection between neurons do not form any cycles (Geron, 2017).

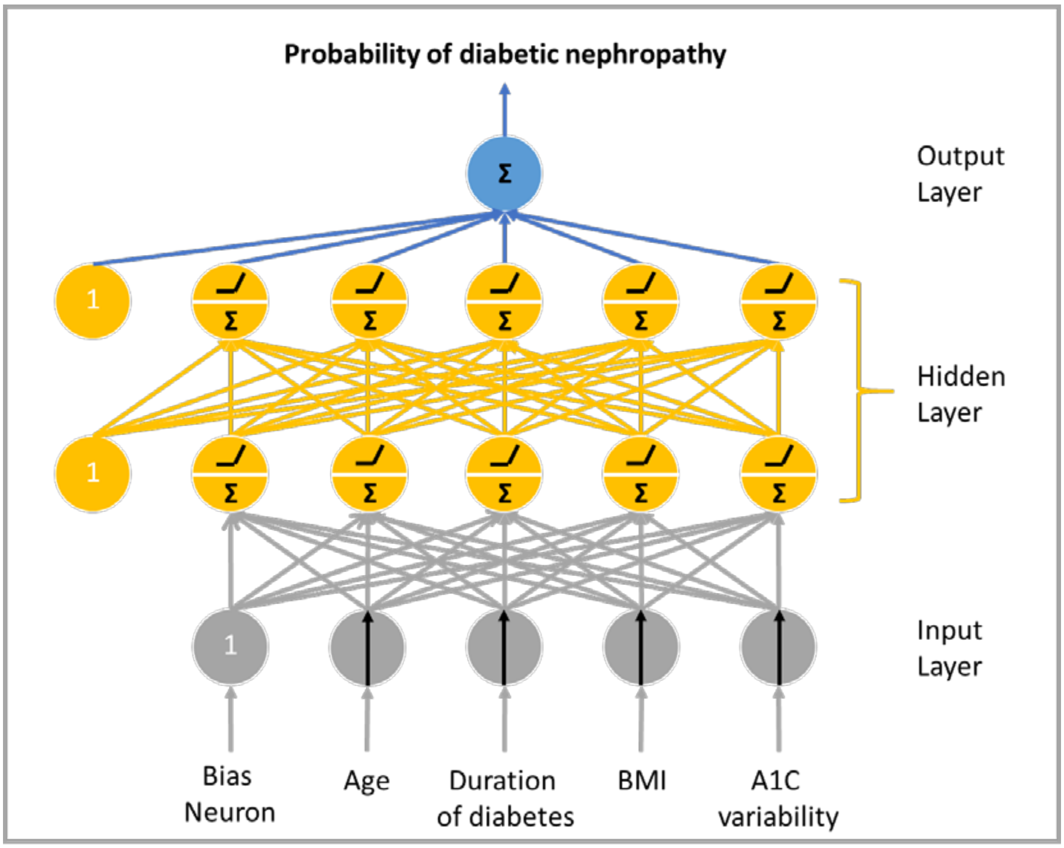


Figure 4. Illustration of an ANN with two hidden layers

Figure 4 illustrates a feedforward NN with the input layer of 4 neurons (age, T1D duration, BMI and A1C variability), 2 hidden layers and an output layer for predicting diabetic nephropathy. Note that this is an example of fully connected NNs, i.e., every neuron in the previous layer is connected to each and every neuron in the consecutive layer. The connections can be randomly dropped, making NN more flexible in modeling.

The prediction error or loss l is calculated as the binary cross-entropy loss or log loss, which is often used for binary classification problems:

$$l = -\,(y\log(p) + (1 - y)\log(1 - p))$$

Where y is the label and p is the predicted probability. The loss l is minimized through the process of gradient descent, in which the gradient is the slope of the loss function. The amount that the weight is adjusted is called the "learning rate".

Hyperparameters for NNs

There are many hyperparameters for NNs, including 1) the number of hidden layers, 2) the number of neurons per layer, 3) percentage of randomly dropped connections at each layer, 4) the type of activation function in each layer, 5) the weight initializing logic, 6) the learning rate, 7) the number of iterations/epochs for training, and 8) the l_2 penalty.

1) Number of hidden layers: For many cases, a single hidden layer would work well provided it has enough neurons (Geron, 2017). But a NN with more hidden layers (also called a deep NN or deep learning) can model complex functions using much fewer neurons than a shallow NN and thus can be trained faster (Geron, 2017).
2) Number of neurons per layer: This is defined by the researcher and it usually depends on the number of layers as well. Cross-validation is often used to find the optimal number. A simple approach to determine the number of hidden layers and number of neurons is to start from a model with more layers and neurons than we actually needed, then use early stop to prevent it from overfitting (Geron, 2017).

3) Percentage of randomly dropped connections at each layer: Similar to the number of layers and neurons, this is defined by the researcher and can be tuned through validation.
4) Type of activation function in each layer: Different activation function can be defined in each layer. Commonly used activation functions include step (Heaviside step or sign), logistic (sigmoid), hyperbolic tangent, and RuLU (rectified linear activation unit, y = max(0, x)) functions (Geron, 2017). RuLU is commonly used in hidden layers.
5) Weight initializing logic: Weights in NNs are generated by random number generators and are usually initiated with small values close to zero. After each round of learning, the weights increase to achieve lower loss (Hastie, Robert., & Friedman, 2009). A random number generator is a mathematical function that produces random sequences of numbers (Shalev-Shwartz & Ben-David, 2014). By default, the random number generator uses a seed to initiate the number generation process. The seed is usually the current time in milliseconds in most implementations to ensure different sequences of numbers being generated every time. By specifying the seed with a number (such as 42), the random number generator will produce the same sequences of numbers every time it runs.
6) Learning rate: The amount that each time the weight is adjusted is called the learning rate. As the loss for NNs is nonconvex, meaning there may be many local minima or lowest loss. If a learning rate is too large, the function may jump over the local minima and fail to find the optimal solution. When the learning rate is too small, the model may take too long to learn and be not efficient (Hastie et al., 2009).
7) The number of iterations/epochs for training: One training epoch refers to the one time that a NN learns from the entire training set. Number of training epochs

determines how many time the NN will learn from the entire training data (Hastie et al., 2009). The optimal number of epochs depends on the loss of both training and evaluation data sets. Because if we keep training the model on the train set, it will reach the global minimum of loss for the training set, but a model fitting the training set too well may not predict the validation set well. Hence, the loss for both train and validation set need to be monitored and the training epochs can be stopped when the validation loss does not decrease any further.

8) Lastly, a regularization term l_2 can be added to the model to minimize the value of weights and prevent overfitting the training data (and hence the model can fit the test data better):

$$l_2\ penalty = l_2 \sum_{i=0}^{n} x_i^2$$

One last note about NNs is that as there are many hyperparameters to be specified, the source of randomness sometimes cannot be controlled. In other words, the results may change after each time of running the model, even though we tried to control for randomness, e.g., by setting a seed for weight initialization. The common approach to tackle this is to repeat running the same model for 10 or 100 times and use the mean performance for its evaluation (Ripley, 1996).

Both SVMs and NNs share one main drawback that make the application of these methods in health care research challenging – the resulting predictive model is unintuitive for interpretation, especially to clinicians, who are always looking for causality. So, the point here is, no one method is inherently better than another. The choice of the algorithms depends on the research purpose, and the performance of the model depends largely on the data itself (Steyerberg, Eijkemans, Harrell, & Habbema, 2001; Geron, 2017).

Predictor Selection – Andersen Behaviour Model

The Andersen Behavioral Model (ABM) was used to guide the predictor selection of the research (Andersen, 2008). According to ABM, complex contextual, individual, and health behavioral factors can influence health outcomes directly and/or through other characteristics (**Figure 5**). Contextual and individual characteristics are categorized into predisposing, enabling, and need factors, and health behaviors are divided into personal health practices, process of medical care, and use of personal health services. Outcome measures can be perceived health, evaluated health, as well as satisfaction.

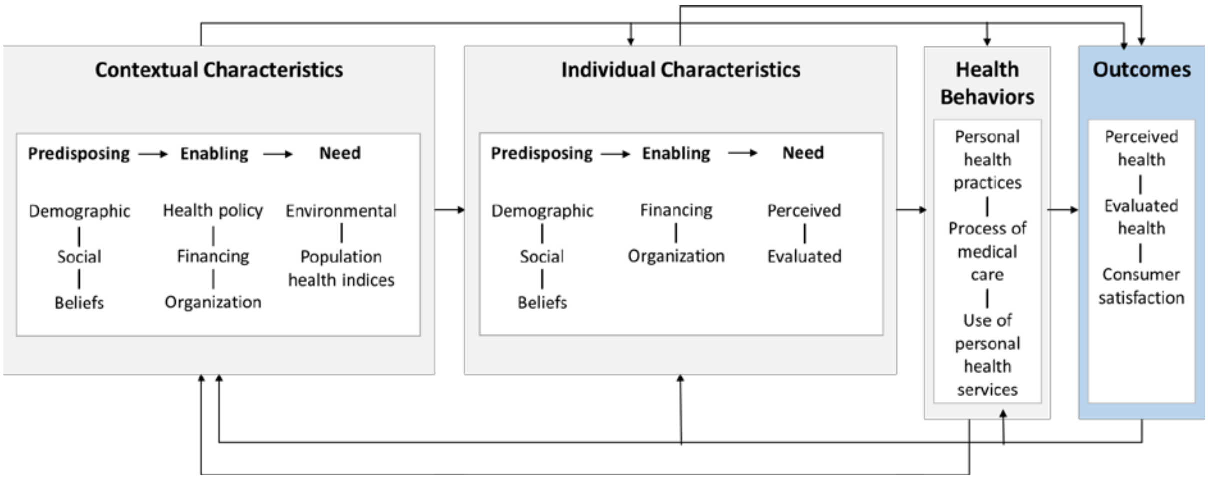


Figure 5. Andersen Behavioral Model

In the present study, ABM was adapted to evaluate individual and health behavioral determinants of diabetic microvascular complications. The outcomes were evaluated health – medical chart indication of diabetic nephropathy, retinopathy, and neuropathy (yes/no). Predicting factors were grouped into individual predisposing (e.g., socio-demographics), enabling (e.g., insurance) and need (e.g., comorbidities) factors, and health behavioral factors.

Individual predisposing factors considered include patient's age, gender, race, T1D duration, education level, and marital status. Enabling factors include patient employment status, household income (per capita), and insurance type. Need factors include A1C variability, BMI, blood pressure, cholesterol level and history of past medical conditions (comorbidities). Health behavioral factors include type of insulin used and insulin delivery method, use of CGM (yes/no), use of other medications (including other antidiabetics, ACE inhibitors or angiotensin II receptor blockers (ARBs)) and smoking status.

Specifically, in this study, the second main objective was to understand how A1C variability affects the prediction of each microvascular complication, especially within ML models. A1C variability was manipulated into the following 5 levels:

a) Single A1C: 1 single A1C value, which serves as the reference (no variability)
b) Mean A1C: the mean of multiple A1C values
c) Combination single: 2 variables, single A1C and the standard deviation (SD) of multiple A1C values (SD A1C)
d) Combination mean: 2 variables, the mean of multiple A1C values (mean A1C) and their SD (SD A1C)
e) Multiple: the multiple individual A1C values and their SD (SD A1C)

These 5 levels are nominal and not arranged in any order. By adding one of each of the five levels to the other selected covariates, a total of 5 predictor sets were developed and used for prediction of each microvascular complication. The adapted model from ABM is illustrated in **Figure 6**.

Figure 6. Model conceptualization using Andersen Behavioral Model

The final proposed model using constructs from both statistical learning theory and ABM is illustrated in **Figure 7**.

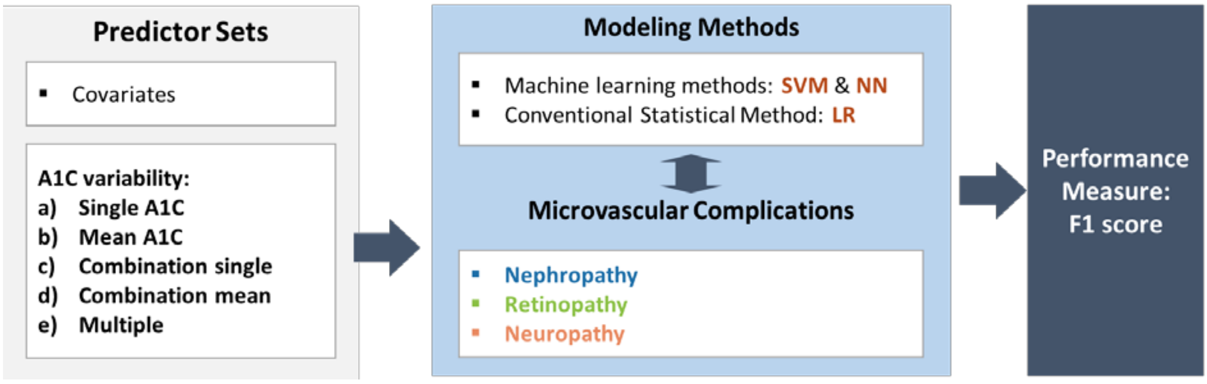


Figure 7. Proposed model

Research Hypotheses

H1: There is no significant difference in performance measures (F1 score) between ML (SVM and NN) and conventional statistical (LR) methods for predicting three microvascular complications (diabetic nephropathy, retinopathy and neuropathy) in T1D patients.

H2: There is no significant difference in performance measure (F1 score) for ML (SVM and NN) and conventional statistical (LR) methods using A1C variability to predict each microvascular complication (diabetic nephropathy, retinopathy and neuropathy) in T1D patients.

In CHAPTER 4, the methodology for this research will be explained in details.

CHAPTER 4

Methods

Due to the applied nature of this research, in order to compare the performance of different modeling methods on predicting the three different outcomes as well as evaluate the effect of modeling methods on utilizing different predictor sets, the following steps were conducted:

1) Determination of study design
2) Formation of three cohorts: diabetic nephropathy, retinopathy, and neuropathy
3) Operational definition of study measures
4) Splitting of train and test sets
5) Selection of predictors
6) Model development in each cohort using LR, SVM and NN with different predictor sets
7) Performance evaluation and comparison using F1 score

The overview of the study design is illustrated in **Figure 8**. Following that, detailed methods for each step will be explained.

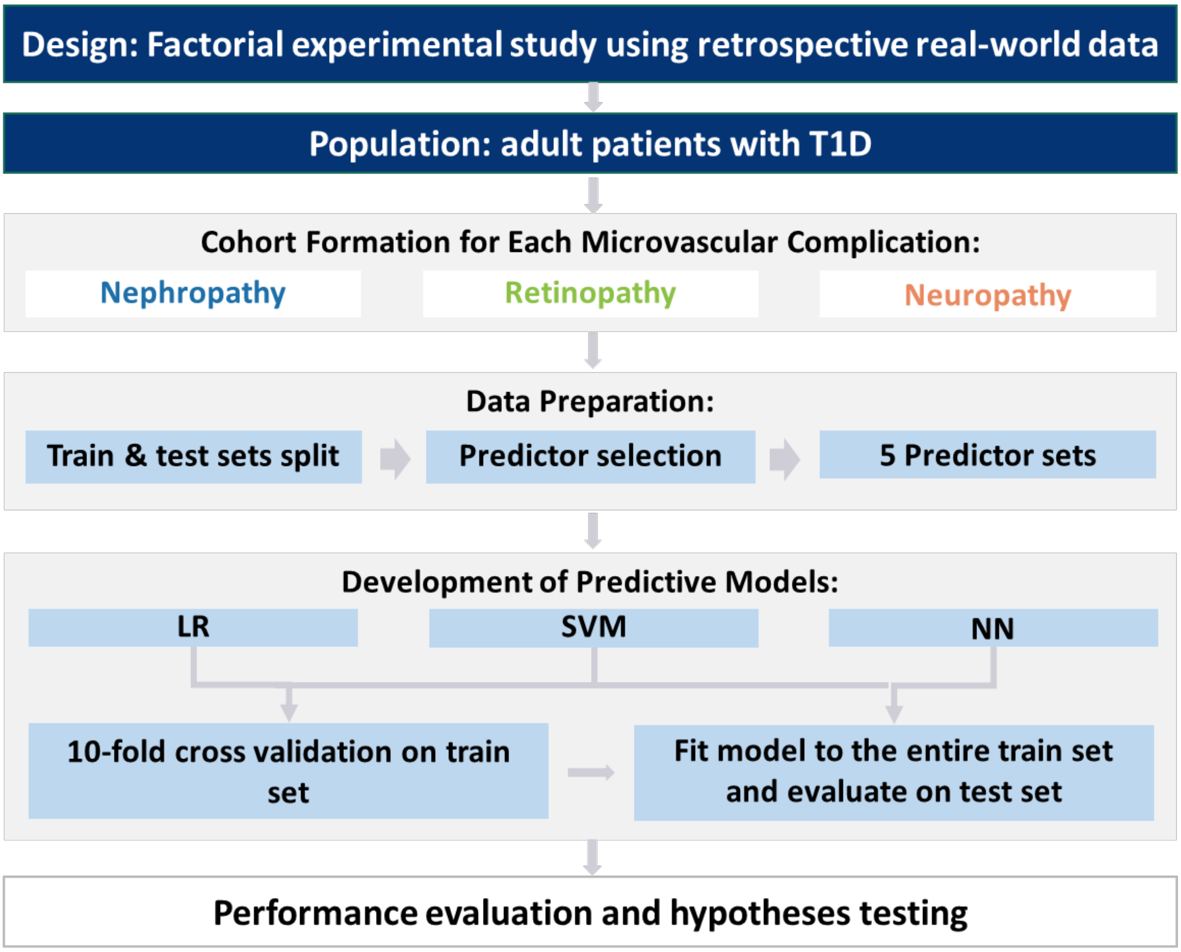


Figure 8. Overview of the study design

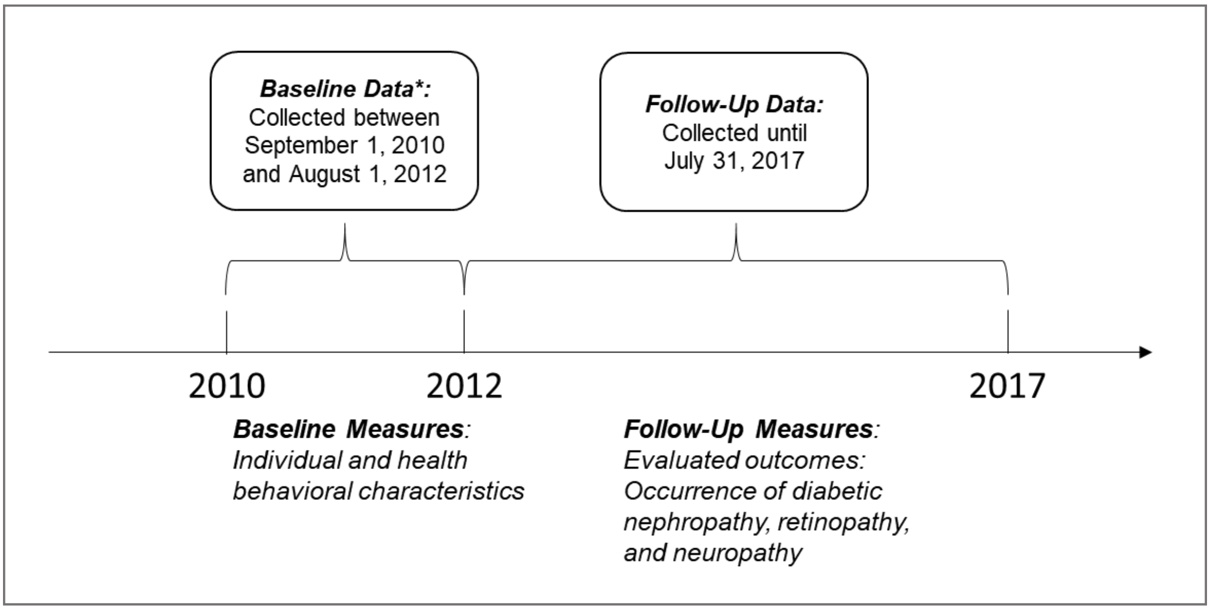
Study Design

This study adopted a factorial experimental design to evaluate model performance by three types of modeling methods (i.e., SVM, NN and LR), three types of outcomes (i.e., diabetic nephropathy, retinopathy and neuropathy) and different predictor sets. The statistical learning theory was used to guide model development and evaluation. Three cohorts of patients based on the three outcomes were formed and within each cohort, data were split into train and test sets. The ABM was used to guide predictor selection. Individual and health behavioral factors were considered for predicting evaluated outcomes, i.e., diabetic nephropathy, retinopathy, and neuropathy. LR, SVM and NN were applied to develop the predictive models. Factorial analysis of variance (ANOVA) was used to evaluate the effect of modeling method, study outcomes, and predicting sets, specifically, measures of A1C variability on model performance (F1 score).

Data Source & Patient Population

The T1D Exchange Clinic Registry was used for the study. The registry was established by T1D Exchange, a nonprofit research organization that dedicates to drive research and improve care for T1D patients (https://t1dexchange.org/about/). A detailed description of the database was published previously (Beck et al., 2012). Briefly, the registry enrolled participants from clinical centers of T1D patients that represents most locations throughout the U.S. It extracts patient information from clinic chart including diagnoses, procedures, pharmacy, demographics, and lab test results. It also administers a participant questionnaire at enrollment and at annual follow-ups that assesses a participant’s health behavior and distress measures. The contents of the questionnaire have changed slightly over the years.

Three sets of data were used for this study. The dataset collected between September 1, 2010 and August 1, 2012 was used to assess baseline characteristics of adult T1D patients. These baseline measures were used to predict outcomes of microvascular complications (nephropathy, retinopathy, and neuropathy) recorded in the dataset collected between April 30, 2015 and July 31, 2016 or between May 1, 2016 and July 31, 2017 (**Figure 9**). Baseline dataset included information from 25,762 subjects at 68 sites and constituted of five files: A1C, labs, medical conditions, medications, and subject. The follow-up datasets contain 20,842 patients from 73 clinics (2015-2016) and 18,743 participants from 79 clinic sites (2016-2017), respectively, and each was composed of four files: A1C, medical conditions, medications, and subject. The datasets were anonymized, but the same patients can be identified and linked by the same patient ID number. Although the registry data is cross-sectional in nature, it contains multiple A1C values for the same patient, which made it possible to assess variability of A1C or long-term glycemic variability.



*The registry data is cross-sectional in nature, although multiple A1C values can be available up to 10 years from the exam date.

Figure 9. Study timeline

Operational Definition of Study Measures

1. Outcome measures

Study outcomes are the three types of microvascular complications, i.e. diabetic nephropathy (kidney disease), retinopathy, and neuropathy assessed in the follow-up datasets. Each outcome is defined as a binary variable: "yes" as having the outcome and "no" as not having the outcome. Presence of each outcome is captured by measures from participant questionnaire as well as medical conditions recorded using Medical Dictionary for Regulatory Activities Terminology (MedDRA) terms in clinic chart. The operational definitions of nephropathy, retinopathy, and neuropathy are provided in **Appendix 3**.

2. Baseline measures

Baseline measures include individual characteristics and health behavioral factors. These measures were considered as predictors to be incorporated into the predictive models. Predictors considered in this study were chosen based on previous literature as well as considering attainability from patients clinic records (Lagani et al., 2015).

A1C variability:

The main predictor evaluated in this study is A1C variability, which refers to the change or fluctuation in glycosylated hemoglobin A1C level (%) over long term (from one visit to the next). In this study, it was operationalized using the last 3 A1C values (%) that were measured at least 3 months apart at baseline. Measures of A1C variability was manipulated as 5 levels:

a) Single A1C: defined as the last A1C measured at baseline. It does not reflect any A1C variability. It serves as the reference level. Single A1C:

b) Mean A1C: defined as the mean of last 3 A1C values at baseline

c) Combination single: 2 variables; defined as the last A1C (single A1C) and the standard deviation of the last 3 A1C values (SD A1C)

d) Combination mean: 2 variables; defined as the mean of the last 3 A1C values (mean A1C) and their standard deviation (SD A1C)

e) Multiple: 4 variables; defined as the individual values of the last 3 A1C at baseline and their standard deviation (SD A1C)

As the most recent A1C (%) is commonly used in clinical settings as an indicator for glucose control, a single most recent A1C value was used as reference level for A1C variability. Mean-A1C and SD-A1C are used in previous literature as operationalization for A1C variability and hence, are used in this study.(Orsi et al., 2018)

Covariates:

Other covariates considered as predictors include demographics (age, gender, race, marital status, education level, income, employment status, insurance coverage), T1D duration, blood pressure (mmHg), BMI, cholesterol level (LDL, HDL, triglyceride levels, lipid fasting status), microalbuminuria status (yes/no), baseline comorbidities including diabetic retinopathy, diabetic nephropathy, diabetic neuropathy, cardiovascular conditions (hypertension, dyslipidemia, CAD, PVD, cardiac arrythmia, cerebrovascular accident), endocrine diseases (hypothyroidism or Hashimoto disease, hyperthyroidism or Grave's disease and others), gastrointestinal diseases (Celiac disease, vitamin B12 deficiency/pernicious anemia, IBD), musculoskeletal /connective tissue conditions (rheumatoid arthritis, osteoporosis, Lupus, Sjogrens, dermatomyositis), psychiatric conditions (depression, anxiety, ADHD, psychosis, eating disorders), and skin conditions (vitiligo, psoriasis, necrobiosis lipoidica diabeticorum, alopecia areata), and health behavioral factors include insulin used at baseline (insulin delivery method and name/type of insulin), use of other antidiabetics (DPP4 inhibitors, GLP1 agonists, metformin, pramlintide & others), use of ACE inhibitors or ARBs, use of continuous glucose monitor (CGM), and smoking status (Ever smoked and smoking status at baseline). They are defined using measures from the baseline dataset. The operational definitions of all variables are provided in **Appendix 3**.

Cohort Formation

1. Eligibility Criteria Across Cohorts:

<u>Inclusion criteria</u>: Across cohorts, participants need to meet all of the following inclusion criteria to be eligible for the study:

- Patients who had a definite T1D (see **Appendix 2** for definition of definite T1D defined by the registry) and had records in the T1D Exchange Registry data collected during both baseline and one of the follow-up period.
- Age >=18 years at baseline
- Had a non-missing value for age of the diagnosis of T1D
- Had a non-missing value of A1C at exam
- Had >= 2 additional A1C measures that were assessed at least 3 months apart from each other to evaluate A1C variability

<u>Exclusion criteria</u>: Across cohorts, participants who meet any of the following criteria will be excluded from the analyses:

- Participants with a history of cancer, including abdominal tumor, acoustic neuroma, basal cell carcinoma, bladder cancer, bone cancer, bone giant cell tumor, bone marrow transplant, brain tumor, breast cancer, breast ductal carcinoma, cancer (NOS), cancer of skin (excl melanoma), cervical cancer, chronic lymphocytic leukemia, colon adenoma, colon cancer, colorectal cancer, endometrial cancer, esophageal cancer, Hodgkin's lymphoma, kidney cancer, leukemia, liver cancer, lung cancer, lung cancer metastatic, lymphoma, malignant breast neoplasm, malignant melanoma, malignant melanoma of eyelid, meningioma, multiple myeloma, neoplasm (NOS), ovarian cancer, ovarian neoplasia, pituitary adenoma, prostate cancer, renal cell carcinoma, skin carcinoma, thyroid cancer, uterine cancer, and vulvar cancer.
- Participants with a history of kidney, pancreas or islet cell transplantation, end-stage renal disease (ESRD) or kidney/renal failure (including receiving dialysis or kidney transplant) any time during the study period
- Participants who were pregnant at the time of exam
- Participants who were transgender

2. Eligibility Criteria for the Cohort of Diabetic Nephropathy:

Inclusion criteria: For the cohort evaluating development of diabetic nephropathy (kidney disease), in addition to the above eligibility criteria across cohorts (described in **Eligibility Criteria Across Cohorts**), patients need to meet all of the following criteria:

- Had no missing information for the measure of diabetic nephropathy in the follow-up data

Exclusion criteria: participants who meet any of the following criteria were excluded:

- Had a clinic chart indication of diabetic nephropathy/kidney disease in the baseline data

- Had a clinic chart indication of history of renal failure in the baseline data
- Had a clinic chart indication of receiving dialysis for renal failure in the baseline data
- Had a clinic chart indication of kidney cancer in the baseline data
- Was taking angiotensin-converting-enzyme (ACE) inhibitors or angiotensin II receptor blockers (ARBs) for diabetic nephropathy (microalbuminuria) in the baseline data

3. Eligibility Criteria for the Cohort of Diabetic Retinopathy:

Inclusion criteria: For the cohort evaluating development of diabetic retinopathy, in addition to the above eligibility criteria across cohorts (described in **Eligibility Criteria Across Cohorts**), patients need to meet all of the following criteria:

- Had no missing information for the outcome of diabetic retinopathy in the follow-up data

Exclusion criteria: participants who meet any of the following criteria were excluded:

- Had a clinic chart indication of retinopathy in the baseline data
- Had a clinic chart indication of blindness in the baseline data
- Had a clinic chart indication or patient-report of having been treated for diabetic retinopathy in either eye (including laser, injections to the eye, vitrectomy) in the baseline data
- Had received cataract surgery or treatment for glaucoma reported by the patient in the baseline data
- Was taking angiotensin-converting-enzyme (ACE) inhibitors or angiotensin II receptor blockers (ARBs) for diabetic retinopathy in the baseline data

4. Eligibility Criteria for the Cohort of Diabetic Neuropathy:

Inclusion criteria: For the cohort evaluating development of diabetic Neuropathy, in addition to the above eligibility criteria across cohorts (described in **Eligibility Criteria Across Cohorts**), patients need to meet all of the following criteria:

- Had no missing information for the outcome of diabetic neuropathy in the follow-up data

Exclusion criteria: participants who meet any of the following criteria were excluded:

- Had a clinic chart indication of diabetic peripheral neuropathy in the baseline data
- Had a clinic chart indication of presence of foot ulcer in the baseline data
- Had a clinic chart indication of erectile or sexual dysfunction in the baseline data
- Had a clinic chart indication of Charcot joint in the baseline data
- Had a clinic chart indication of orthostatic hypotension with fixed heart rate in the baseline data
- Had a clinic chart indication of tachycardia with fixed heart rate in the baseline data
- Had a clinic chart indication of gastroparesis in the baseline data
- Had a clinic chart indication of medical history of amputation of toe or amputation below/above knee in the baseline data

5. Criteria for Handling Missing Data:

As with any real-world data, the registry database is prone to missing values. Although imputing missing values using certain algorithms such as regression or random forest can potentially reduce bias and increase sample size for prediction modelling, we want to focus our efforts to use available information that's already in the database and assess and interpret associations between complete baseline measures and the outcomes during follow-up. Thus, we are going to delete observations that are missing values on any predictor variable of interest for the respective cohort. For example, as diabetes duration is associated with prognosis of all three microvascular complications, observations that have missing information on T1D duration will be removed from our analysis. Candidate predictors are identified from previous literature as discussed previously.

Train Set and Test Set

Once applying the inclusion and exclusion criteria across board and for each microvascular complication, three cohorts of data were obtained: one for diabetic nephropathy, one for diabetic retinopathy, and the last for diabetic neuropathy. Before taking a closer look into the data, a test set was separated from each cohort and kept intact, which was used for model performance evaluation.

A stratified random sampling approach based on the outcome variable (i.e., whether or not the patient progressed to diabetic nephropathy, retinopathy, and neuropathy) was used to select 20% data as the test set for each cohort (Geron, 2017). This ensured that in both training and test sets, similar proportions of patients were affected by each microvascular complication.

Predictor Selection

Predictors were selected based on previous literature, univariate analyses and correlation analyses of the train set for each cohort. Descriptive statistics (mean, SD, median, range, count, percentage) of diabetes related microvascular complications, treatment patterns, and patient demographics were calculated for each cohort during baseline period. Nominal variables that had more than two levels were transformed to binary dummies for each level. This was because ML algorithms usually assume that two nearby values are more similar than two distant values, which is not true for nominal variables such as race, insurance type or income categories (Geron, 2017).

Univariate comparisons between patients who progressed to each microvascular complication versus those not in each cohort were made using t tests for normally distributed continuous variables, Wilcoxon rank sum tests for non-normally distributed continuous variables, and chi-square and Fisher's exact tests for categorical variables. Pearson's correlation analyses were performed to evaluate correlations between potential predictors and each outcome.

Once covariates for each cohort were determined, the 5 different A1C measures were added to the model, making the following 5 predictor sets (PSs) for each cohort:

Feature Manipulation for ML Models

Feature manipulation refers to the process of transforming input values (Geron, 2017). This is one of the key steps in data preparation, because some ML algorithms don't perform well when there is huge difference in feature scales (e.g., systolic blood pressure may range from 90 to 140 whereas mean A1C ranges between 4%-14%) (Geron, 2017). Three methods were tried in this study: "min-max scaling" (also called "normalization"), standardization, and robust scaler.

Min-max scaling works by subtracting the min value and dividing by the max minus the min, which yields values ranging from 0 to 1 (Geron, 2017). Standardization works by first subtracting the mean from the input value and dividing it by the standard deviation (SD) so that the transformed (or standardized) values have a zero mean and unit variance. Standardization does not restrict the values to a specific range (Geron, 2017). Robust scaler, on the other hand, subtracts the median from the input value and further divides the value by the interquartile range (IQR) (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html).

Over-Sampling

As only a relatively small proportion of patients developed the outcomes of interest during the follow-up period, Synthetic Minority Oversampling TEchnology (SMOTE) was used to over-sample the minority/disease-positive group to balance the data (Chawla, Bowyer, Hall , & Kegelmeyer, 2002). SMOTE created synthetic samples by interpolation/perturbation. It bootstrapped over the cases, found the k-nearest neighbours (default k is 5) of each case, and calculated the difference between the sample's characteristics vector and their neighbors'. The difference was weighted by a random number between 0 and 1, added to the original sample and hence, created 'synthetic' samples. SMOTE has been widely applied in predictive modelling in health care research such as in the study that predicted breast cancer in diabetic patients (Hsieh et al., 2019) and the study that predicted diabetes mellitus (Alghamdi et al., 2017). In this study, synthetic samples were created so that in each training set, cases (disease-positive) and controls (disease-negative) had equal number.

Determination of Sample Size

There were two aspects of sample size consideration for this study: 1) sample size consideration for predictive modeling; and 2) sample size consideration for statistical hypothesis testing.

1. Sample Size Consideration for Predictive Modeling

One crucial factor affecting the performance of prediction models is the sample size of the data. The number of data observations/instances used for developing predictive models relative to the candidate predictors used in the model should be large enough for robust prediction. As more advanced ML algorithms are very robust in prediction using "wide" data (i.e. less observations relative to more predicting variables), this sample size estimation was based on the smallest sample needed for LR. A general rule of thumb is events per variable (EPV) >=10. That is, the ratio of the number of events, i.e. number of observations in the smaller of the two outcome groups, divided by the number of degrees of freedom (parameters excluding the intercept term) should be at least 10 (Harrell et al., 1996). Assuming 15 predictors in the LR model, multiply it by 10 yields a number of 150, which implies that at least 150 observations with the outcome of interest (in our case, occurrence of diabetic nephropathy, retinopathy, and neuropathy) should be enough for modeling using LR. With more advanced ML algorithms, this number would be even less.

A feasibility test on the data revealed that a total of 5,010 adult patients whose information were collected in both baseline and 2016-2017 follow-up (which contains a relatively smaller sample compared to the 2015-2016 sample) period. Assuming the occurrence of each microvascular complication is 10%, we would have about 500 patients with each of the outcomes. Thus, we would have more than enough patients in each cohort for prediction modeling.

2. Sample Size Consideration for Statistical Hypothesis Testing

For the testing of hypothesis 1, power analysis for a two-way 3 by 3 factorial ANOVA was conducted in G*Power 3.0.10 to determine a sufficient sample size using an alpha (α) of 0.05, a power (1-β) of 0.80, numerator degree of freedom (df) of 2 (based on the df of the main effect 'modeling method') and number of groups of 9 (3x3) (**Table 4**) (Cohen, 1992).

Table 4. Sample size estimates based on different effect sizes for hypothesis 1 (keep constant of α= 0.05, power = 0.80, numerator df = 2 and number of groups = 9).

Effect size	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**
Sample size	244	158	111	82	64

As we applied 3 modelling methods to 3 microvascular complications with 11 F1 scores obtained for each model (10 F1 scores from 10-fold cross validation and 1 F1 score from the test set), a total of 99 F1 scores would be obtained. Hence, the study was powered to test an effect size of 0.35.

For the testing of hypothesis 2, power analysis for a two-way 3 by 5 factorial ANOVA was conducted in G*Power 3.0.10 to determine a sufficient sample size using an alpha (α) of 0.05, a power (1-β) of 0.80, numerator degree of freedom of 8 (df = (3-1) x (5-1)), based on the df of the interaction effect 'modeling method' and 'predictor sets') and number of groups of 15 (3x5) (**Table 5**) (Cohen, 1992).

Table 5. Sample size estimates based on different effect sizes for hypothesis 2 (keep constant of α= 0.05, power = 0.80, numerator df = 8 and number of groups = 15).

Effect size	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**
Sample size	284	249	176	131	103

For each cohort, we applied 3 modelling methods to 5 different levels of A1C variability with 11 F1 scores obtained for each model. That resulted in a total of 165 F1 scores. Hence, the study was powered enough to test an effect size of 0.35.

Data Analysis

Estimation of Model Performance

In order to estimating prediction error, 10-fold stratified cross validation was applied to the train set of each cohort for each modeling method. The 10-fold cross validation approach is the most widely used method for error estimation (Hastie et al., 2009). More generally, a k-fold cross validation divides the data into k equal-sized folds. Each time, the model leaves 1 fold out for validation and uses the rest k-1 folds for model fitting. Hence, each time the model can be validated on a different validation dataset. Considering the imbalance nature of the data, stratified sampling approach based on the study outcome in each cohort was taken for the fold generation (Geron, 2017).

In addition, the whole train set was fit to the predictive models and tested on the test set. Hence, a total of 11 performance measures were obtained for each modeling method with each predictor set in each cohort.

Prediction Using LR

In the cohort of diabetic nephropathy, prediction via LR was conducted using the afore-defined predictor sets a) through d). 10-fold cross validation was conducted on the train set. Then the entire train set was fit on LR and evaluated on the test set.

Specifically, for the last predictor set e), which included 3 A1C values as well as the SD of A1C, a generalized estimating equation (GEE) with logit link was employed to accommodate multicollinearity between A1C values measured at different time points (Hardin, 2005). The GEE model takes the form of

$$g(\mu_{ij}) = log(\mu_{ij}/(1 - \mu_{ij})) = X'_{ij}\beta$$

Where $j = 1, 2, 3, \ldots n_i$, and

$i = 1, 2, 3$, representing the jth measurement on the ith patient.

The distribution was binomial (proportion):

$$V(\mu_{ij}) = \mu_{ij}/(1 - \mu_{ij})$$

The regression parameter vector β was estimated taking into account of the covariance structure of correlated measures. In this study, the correlated measures were the 3 A1C values. Age and T1D duration at the time when the 3 A1C were measured were recalculated. All other variables, including the study outcomes and other predictor covariates were assumed to be the same at the 3 different time points. An exchangeable working correlation matrix was assumed and used for the models:

$$\text{Exchangeable Corr}(y_{ij}, y_{ik}) = \begin{cases} 1 & j = k \\ \alpha & j \neq k \end{cases}$$

The fitted models from GEE was applied to the evaluation sets in the 10-fold cross validation as well as test sets to evaluate model performance.

This same process was repeated for the cohorts of diabetic retinopathy and neuropathy.

Prediction Using SVM

In the cohort of diabetic nephropathy, prediction via SVM was conducted using afore-defined 5 predictor sets a) to e). In order to train the SVM models, different values of soft-margin constant C, kernel functions and feature scaling methods were tried on the train set. Through 10-fold cross-validation, model performance in terms of F1 score was calculated on the 10 validation sets and the final model hyperparameters were chosen based on the highest mean F1 score of the 10 validation sets. These hyperparameters and feature scaling method were then applied to the entire train set and evaluated on the test set to obtain the 11^{th} F1 score (Geron, 2017).

This same process was repeated for the cohorts of diabetic retinopathy and neuropathy.

Prediction Using NN

In the cohort of diabetic nephropathy, prediction via NN was conducted using aforementioned 5 predictor sets a) to e). A seed of number 42 was set to the weight initializing logic to reduce source of randomness. In order to train the NN models, different values of the following 7 hyperparameters were tried on the train set: 1) the number of hidden layers, 2) the number of neurons per layer, 3) percentage of randomly dropped connections at each layer, 4) the type of activation function in each layer, 5) the learning rate, 6) the number of iterations/epochs for training, and 7) the $l_2\ penalty$. Each time, a set of 7 hyperparameters were trialed on the train set through 10-fold cross validation. The loss and F1 score of both train and validation folds were plotted against each epoch of training and the mean F1 score was calculated. Then one of the hyperparameters was changed to see how that would affect the loss and F1 score curve. The number of epochs was determined by the point where the loss curve of the validation set stopped decreasing. The final model hyperparameters were chosen based on the highest mean F1 score of the 10 validation folds (Geron, 2017).

This same process was repeated for the cohorts of diabetic retinopathy and neuropathy.

Performance Measure

Model's performance was evaluated in terms of F1 score. F1 score ranges between 0 and 1, with higher F1 score indicating better performance of the model. A probability cut point of 0.5 was used to classify observations as events or nonevents. In each cohort, 11 F1 scores were obtained for each modeling method with each predictor set. In addition, in order to explain a model's capability in identifying patients who were at risk and interpret F1 score, models' sensitivity and precision were provided.

Statistical Hypotheses

<u>Statistical Hypothesis 1</u>:

$$\mu_{LR} = \mu_{SVM} = \mu_{NN}$$

Where μ_{LR} = F1 scores for models using logistic regression method

μ_{SVM} = F1 scores for models using support vector machine method

μ_{NN} = F1 scores for models using neural network method

<u>Statistical Hypothesis 2</u>: There is no significant difference in the performance of LR, SVM and NN models in utilizing A1C variability for the prediction of diabetic nephropathy, retinopathy and neuropathy, respectively.

Hypothesis 2a: $\mu_{Nep\_LR\_PS_i} = \mu_{Nep\_SVM\_PS_i} = \mu_{Nep\_NN\_PS_i}$

Where Nep represents cohort of nephropathy and PS represents predictor sets,

$\mu_{Nep\_LR\_PS_i}$ = F1 scores of logistic regression models with different predictor sets

$\mu_{Nep\_SVM\_PS_i}$ = F1 scores of support vector machine models with different predictor sets

$\mu_{Nep\_NN\_PS_i}$ = F1 scores of neural networks models with different predictor sets

Hypothesis 2b: $\mu_{Ret\_LR\_PS_i} = \mu_{Ret\_SVM\_PS_i} = \mu_{Ret\_NN\_PS_i}$

Where Ret represents cohort of retinopathy and PS represents predictor sets,

$\mu_{Ret\_LR\_PS_i}$ = F1 scores of logistic regression models with different predictor sets

$\mu_{Ret\_SVM\_PS_i}$ = F1 scores of support vector machine models with different predictor sets

$\mu_{Ret\_NN\_PS_i}$ = F1 scores of neural networks models with different predictor sets

Hypothesis 2c: $\mu_{Neu\_LR\_PS_i} = \mu_{Neu\_SVM\_PS_i} = \mu_{Neu\_NN\_PS_i}$

Where Neu represents cohort of neuropathy and PS represents predictor sets,

$\mu_{Neu\_LR\_PS_i}$ = F1 scores of logistic regression models with different predictor sets

$\mu_{Neu\_SVM\_PS_i}$ = F1 scores of support vector machine models with different predictor sets

$\mu_{Neu\_NN\_PS_i}$ = F1 scores of neural networks models with different predictor sets

Statistical Analysis

Descriptive statistics (mean, SD, median, range, count, %) were generated for all study measures. Univariate analyses (t tests for normally distributed variables, Wilcoxon rank sum test for non-normally distributed continuous variables, chi-square and Fisher's exact tests for categorical variables) were conducted to evaluate unadjusted association between baseline characteristics and each outcome. Pearson's correlation analyses were conducted to assess the correlation between baseline characteristics and each outcome. Factorial analysis of variance (ANOVA) was used to test research hypotheses. Post hoc Tukey-Kramer test was performed to evaluate which levels within a factor were significantly different.

An alpha level of less than 0.05 was used to determine statistical significance of an association. SAS 9.4 (SAS Institute, Inc. Cary, NC) was used to perform data preparation, descriptive, univariate and correlation analyses, multiple LR and GEE; the application programming interface (API) of SciKit-Learn version 0.22.1 (*http://scikit-learn.org/*) (Pedregosa et al., 2011) and TensorFlow.keras (*http://tensorflow.org/*) (Abadi, 2015) were used to implement SVM and NN models.

Protection of Human Subjects

This study was retrospective in nature. We analyzed observational data of T1D patients collected by the T1D Exchange Clinic Registry. The data contains only HIPPA-compliant de-identified patient information. No recruitment of patients or intervention was involved or imposed on the subjects or their healthcare providers in the study. There is no potential risk to patients or their health care providers. This study approved as an exempt category by the Institutional Review Board (IRB) on ethics of human research at University of Houston before study initiation. The study was also conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki, the International Conference on Harmonization Good Clinical Practice (ICH GCP) and Good Epidemiology Practices (GEP), and other applicable regulatory requirements.

In CHAPTER 5, the results of the research will be presented.

CHAPTER 5

Results

This chapter presents results from the analysis. First, the formation of three patient cohorts, i.e., diabetic nephropathy, retinopathy and neuropathy is outlined in the attrition table. Then the results from each cohort will be presented, including a summary of the characteristics of the entire cohort (patient demographics, clinical characteristics, treatments and A1C measures), results from correlation analyses of the train set, parameters of the LR models, hyperparameters of the final SVM and NN models, and the performance metrics from the LR, SVM and NN models. Lastly, the results of statistical hypothesis testing are presented.

Patient Attrition

A total of 4476, 3595, and 4072 patients met the eligibility criteria for the cohort of nephropathy, retinopathy, and retinopathy, respectively. **Figure 10** provides the patient attrition chart. For each cohort, 80% of data were used for model training, leaving 20% for model testing. The training sets of nephropathy, retinopathy, and neuropathy were composed of 3580, 2875, and 3257 patients, respectively.

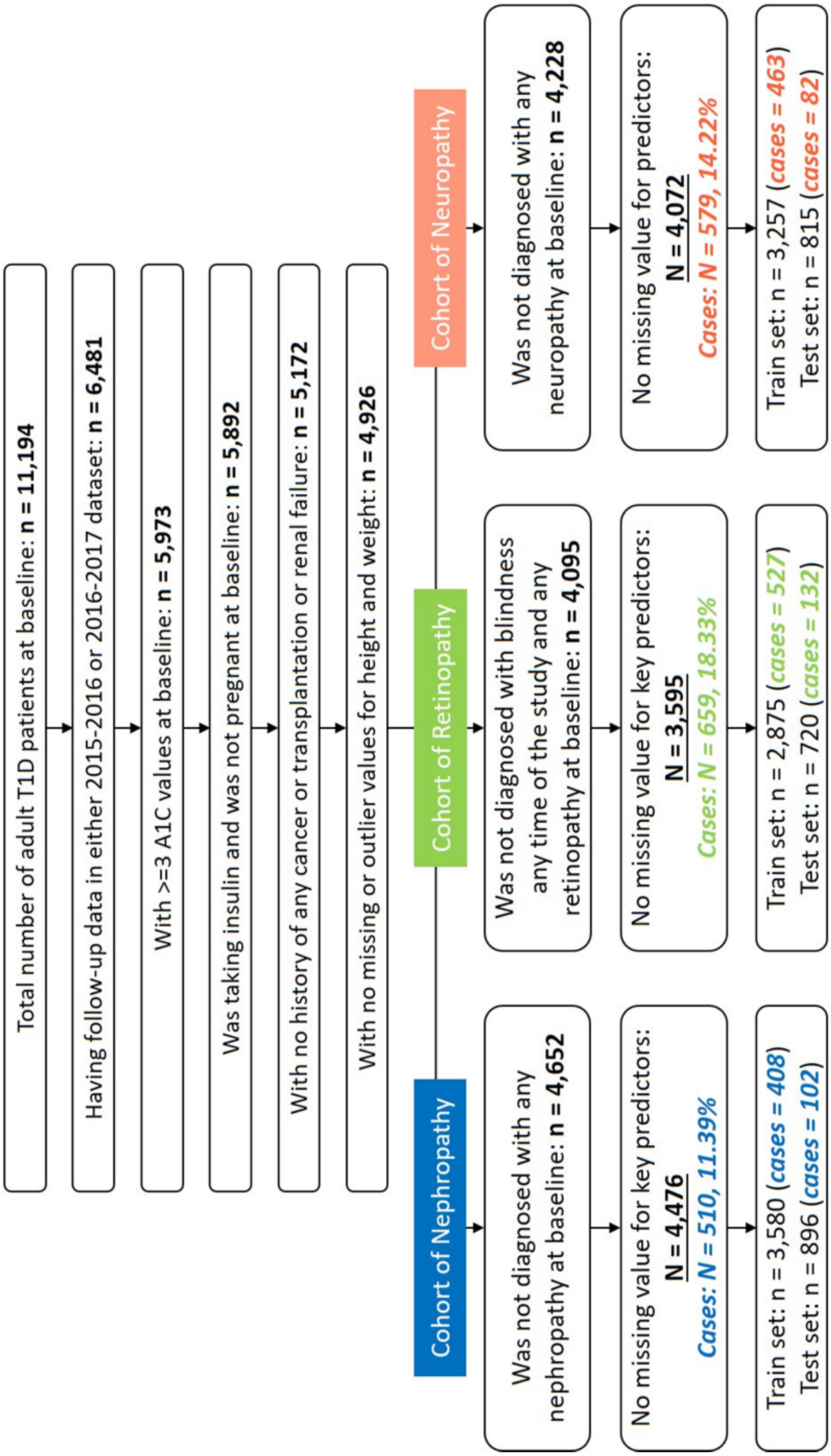


Figure 10. Patient attrition chart

Cohort of Nephropathy

Baseline Characteristics

Among the 4,476 patients in the nephropathy cohort, 510 (11%) developed diabetic nephropathy (cases) during the follow-up period. Slightly more than half (53%) were women. The mean (±SD) age of patients in the case group was 48 (±16.7) years, significantly older than those who did not develop nephropathy during follow-up (controls) (38±14.8, $p<0.0001$). The baseline demographic, clinical, and treatment characteristics between patients in the case and control groups were summarized in **Table 6**.

Demographics: Univariate analyses indicated that compared to the control group (20±12.1 years), cases had had T1D for a longer period at baseline (26±14.1 years, $p<0.0001$). Less proportion of the case group than the control group had bachelor's or above education (46.0% vs 51.8%, $p<0.05$), had commercial health insurance (77.5% vs 86.3%, $p<0.0001$), and worked full-time or part-time (53.9% vs 64.6%, $p<0.0001$). A greater proportion of the case group than the control group were married or living together (65.3% vs 57.8%, $p=0.001$), had below $50k household income (35.1% vs 24.6%, $p<0.0001$), and had ever smoked (36.1% vs 27.8%, $p<0.0001$). The two groups were similar in other demographics.

Clinical characteristics: Patients in the case group on average had higher BMI (28±5.9 vs 27±4.8, $p<0.0001$), SBP (124±15.0 vs 120±13.1, $p<0.0001$), and triglyceride level (101±76.5 vs 89±78.4, $p<0.0001$) than the control group. As of notice, more than a third (36.1%) of patients had their lipid fasting status unknow. Among the 2,741 patients whose fasting status were indicated, 69.4% were fasting and 30.6% were not fasting. Hence, it was less meaningful to compare patients' lipid levels directly, especially when some of them were fasting and others were not.

Medical history: The case group had a greater percentage of patients with a history of microalbuminuria (18.0% vs 2.4%, p<0.0001), diabetic retinopathy (33.1% vs 12.7%, p<0.0001) and neuropathy (28.2% vs 10.7%, p<0.0001), cardiovascular conditions including hypertension (50.8% vs 26.1%, p<0.0001), dyslipidemia (52.0% vs 35.4%, p<0.0001), coronary artery disease (CAD) (10.4% vs 2.8%, p<0.0001) and peripheral vascular disease (PVD) (2.3% vs 0.5%, p<0.001), cardiac arrythmia (2.3% vs 0.8%, p<0.001), hypothyroidism or Hashimoto disease (28.8% vs 21.6%, p=0.0002), gastrointestinal diseases (6.7% vs 4.1%, p<0.01), rheumatoid arthritis (RA) or osteoporosis (12.2% vs 4.3%, p<0.0001), depression (22.0% vs 12.1%, p<0.0001), and anxiety (6.1% vs 4.2%, p<0.05).

Treatment: The two groups did not differ much in their treatment at baseline except that a greater proportion of the case group had used ACE inhibitors or ARBs (51.0% vs 27.5%, p<0.0001). More patients in the control group were using insulin aspart injection (Novolog®) (46.3% vs 41.0%, p<0.05) compared to the case group.

A1C Measures: Measures of A1C were summarized in Table 7. Compared to the control group, the case group was on average higher in their most recent A1C level (8.1±1.6 vs 7.7±1.3, p<0.0001), mean-A1C (8.1±1.5 vs 7.7±1.2, p<0.0001), SD-A1C (0.5±0.4 vs 0.4±0.4, p<0.001) and CV-A1C (0.06±0.04 vs 0.05±0.04, p<0.05). In order to understand the frequency of A1C measure for each patient, the gap/time difference between the last 2 A1C values were evaluated. The gap between the last two A1C measures was on average 6.0 months (range 3.0-130.0 months) with a median gap of 4.0 months. Among the 3580 patients in the train set, there were only 38 (1%) patients whose gap between the last two A1C values were over 24 months.

Table 6. Baseline characteristics of patients in the nephropathy cohort

Characteristics	Cohort of Nephropathy						
	Total		Nephropathy: No		Nephropathy: Yes		
	N	%	N	%	N	%	P-value
	4,476	**100.00%**	**3,966**	**88.61%**	**510**	**11.39%**	
Age at baseline							**<0.0001** †
Mean (SD)	40	15.39	38	14.85	48	16.74	
Median (range)	38	18.0-86.8	37	18.0-85.8	50	18.3-86.8	
Age group							**<0.0001**
18-27 years	886	19.79%	813	20.50%	73	14.31%	
28-37 years	1,317	29.42%	1,235	31.14%	82	16.08%	
38-47 years	885	19.77%	801	20.20%	84	16.47%	
48-64 years	1,131	25.27%	936	23.60%	195	38.24%	
≥65 years	257	5.74%	181	4.56%	76	14.90%	
Age at T1D Diagnosis							**<0.0001** †
Mean (SD)	19	13.36	19	13.00	22	15.54	
Median (range)	15	0.0-76.0	15	0.0-76.0	17	0.0-76.0	
T1D Duration							**<0.0001** †
Mean (SD)	21	12.49	20	12.09	26	14.15	
Median (range)	18	0.6-66.0	18	0.6-66.0	25	1.2-63.4	
Gender							**0.032**
Female	2,381	53.19%	2,087	52.62%	294	57.65%	
Male	2,095	46.81%	1,879	47.38%	216	42.35%	
Race							0.276
White	4,084	91.24%	3,623	91.35%	461	90.39%	
Black/African American	116	2.59%	98	2.47%	18	3.53%	
Hispanic or Latino	165	3.69%	150	3.78%	15	2.94%	
Others	111	2.48%	95	2.40%	16	3.14%	
Education Level	n = 4,350		n = 3,861		n = 489		**0.010**
Less than bachelor's degree	2,123	48.80%	1,859	48.15%	264	53.99%	
Bachelor's degree	1,387	31.89%	1,260	32.63%	127	25.97%	
Master's, professional, or doctorate	840	19.31%	742	19.22%	98	20.04%	
Insurance Coverage	n = 4,126		n = 3,660		n = 466		**<0.0001**
Commercial health insurance	3,520	85.31%	3,159	86.31%	361	77.47%	

Table 6. *Continued*

Characteristics	Cohort of Nephropathy						
	Total		Nephropathy: No		Nephropathy: Yes		
	N	%	N	%	N	%	P-value
	4,476	**100.00%**	**3,966**	**88.61%**	**510**	**11.39%**	
Government-sponsored insurance	491	11.90%	402	10.98%	89	19.10%	
Not specified	115	2.79%	99	2.70%	16	3.43%	
Marital Status	n = 4,432		n = 3,925		n = 507		**0.001**
Married or living together	2,599	58.64%	2,268	57.78%	331	65.29%	
Divorced, separated, single, or widowed	1,833	41.36%	1,657	42.22%	176	34.71%	
Annual household income	n = 3,492		n = 3,105		n = 387		**<0.0001**
<$50,000	899	25.74%	763	24.57%	136	35.14%	
$50,000 to < $100,000	1,316	37.69%	1,171	37.71%	145	37.47%	
>=$100,000	1,277	36.57%	1,171	37.71%	106	27.39%	
Employment Status							**<0.0001**
Working full time or part-time at baseline	2,838	63.40%	2,563	64.62%	275	53.92%	
Student or homemaker	904	20.20%	842	21.23%	62	12.16%	
Unemployed, retired, disabled or other	734	16.40%	561	14.15%	173	33.92%	
Smoking Status							
Yes, smoking at baseline	406	9.07%	352	8.88%	54	10.59%	0.205
Not smoking at baseline, but smoked before	1,285	28.71%	1,101	27.76%	184	36.08%	**<0.0001**
BMI (kg/m^2)	27.10	4.97	26.95	4.82	28.22	5.89	**<0.0001** †
Mean (SD)	27.10	4.97	26.95	4.82	28.22	5.89	
Median (range)	26.37	12.47-65.57	26.26	12.47-56.05	27.49	15.71-65.57	
BMI category							**<0.0001**
Under or normal weight	1,693	37.82%	1,533	38.65%	160	31.37%	
overweight	1,742	38.92%	1,554	39.18%	188	36.86%	

Table 6. *Continued*

Characteristics	Cohort of Nephropathy						
	Total		Total		Total		Total
	N	%	N	%	N	%	P-value
	4,476	**100.00%**	**3,966**	**88.61%**	**510**	**11.39%**	
obese	1,041	23.26%	879	22.16%	162	31.76%	
Blood Pressure (mmHg)	n = 4,365		n = 3,864		n = 501		
Diastolic blood pressure							**0.005**
Mean (SD)	71.98	8.47	72.11	8.46	70.98	8.43	
Median (range)	71	40-111	71	40-111	70	42-100	
Systolic blood pressure							**<0.0001** †
Mean (SD)	120.74	13.40	120.33	13.12	123.85	15.02	
Median (range)	120	60-198	120	82-195	124	60-198	
Cholesterol Levels							
HDL value	n = 3,978		n = 3,521		n = 457		0.842†
Mean (SD)	61.12	17.93	61.14	17.69	60.95	19.69	
Median (range)	59	14-162	59	14-162	57	23-155	
LDL value	n = 4,201		n = 3,716		n = 485		0.334†
Mean (SD)	92.08	27.61	91.91	27.10	93.35	31.21	
Median (range)	90	3-281	90	3-266	89	16-281	
Triglycerides value	n = 3,896		n = 3,439		n = 457		**<0.0001** ‡
Mean (SD)	90.79	78.24	89.41	78.38	101.24	76.47	
Median (range)	73	0-3000	72.00	0-3000	81.00	26-1058	
Lipids Fasting Status	n = 4,287		n = 3,797		n = 490		**<0.0001**
Fasting	1,901	44.34%	1,728	45.51%	173	35.31%	
Not Fasting	840	19.59%	739	19.46%	101	20.61%	
Unknown	1,546	36.06%	1,330	35.03%	216	44.08%	
Microalbuminuria at baseline (Yes)	186	4.16%	94	2.37%	92	18.04%	**<0.0001**
Comorbidities at Baseline							
Diabetic retinopathy	672	15.01%	503	12.68%	169	33.14%	**<0.0001**
Diabetic neuropathy	570	12.73%	426	10.74%	144	28.24%	**<0.0001**
Cardiovascular conditions							
Hypertension	1,293	28.89%	1,034	26.07%	259	50.78%	**<0.0001**
Dyslipidemia	1,670	37.31%	1,405	35.43%	265	51.96%	**<0.0001**
CAD	166	3.71%	113	2.85%	53	10.39%	**<0.0001**
PVD	32	0.71%	20	0.50%	12	2.35%	**<0.0001** §

Table 6. *Continued*

Characteristics	Cohort of Nephropathy						
	Total		Total		Total		Total
	N	%	N	%	N	%	P-value
	4,476	**100.00%**	**3,966**	**88.61%**	**510**	**11.39%**	
PVD or amputation (knee or toe)	39	0.87%	26	0.66%	13	2.55%	**0.0002**§
Cardiac arrythmia	44	0.98%	32	0.81%	12	2.35%	**0.001**
Cerebrovascular accident	25	0.56%	21	0.53%	4	0.78%	0.520§
Endocrine diseases							
Hypothyroidism or Hashimoto disease	1,002	22.39%	855	21.56%	147	28.82%	**0.0002**
Hyperthyroidism or Grave's disease	92	2.06%	80	2.02%	12	2.35%	0.615
Other endocrine diseases	37	0.83%	33	0.83%	4	0.78%	1.000§
Gastrointestinal diseases	197	4.40%	163	4.11%	34	6.67%	**0.008**
Musculoskeletal/Connective Tissue conditions							
RA or osteoporosis	233	5.21%	171	4.31%	62	12.16%	**<0.0001**
Psychiatric conditions							
Depression	593	13.25%	481	12.13%	112	21.96%	**<0.0001**
Anxiety	197	4.40%	166	4.19%	31	6.08%	0.050
ADHD	87	1.94%	75	1.89%	12	2.35%	0.477
Psychosis	17	0.38%	13	0.33%	4	0.78%	0.120§
Eating disorders	28	0.63%	24	0.61%	4	0.78%	0.552§
Skin conditions	101	2.26%	89	2.24%	12	2.35%	0.876
CGM use							0.091
Yes	982	21.94%	885	22.31%	97	19.02%	
No	3,494	78.06%	3,081	77.69%	413	80.98%	
Insulin use							
Type of insulin analog							
Insulin lispro (Humalog)	2,299	51.36%	2,019	50.91%	280	54.90%	0.089
Insulin aspart (Novolog)	2,044	45.67%	1,835	46.27%	209	40.98%	**0.024**
Insulin detemir (Levemir)	162	3.62%	146	3.68%	16	3.14%	0.536
Insulin glargine (Lantus)	1,454	32.48%	1,291	32.55%	163	31.96%	0.789

Table 6. *Continued*

Characteristics	Cohort of Nephropathy						
	Total		Total		Total		Total
	N	%	N	%	N	%	P-value
	4,476	**100.00%**	**3,966**	**88.61%**	**510**	**11.39%**	
Insulin delivery method at baseline							0.601
Pump only	2,677	59.81%	2,365	59.63%	312	61.18%	
Injections/pens only	1,716	38.34%	1,524	38.43%	192	37.65%	
Both pump and injections/pens	83	1.86%	77	1.95%	6	1.18%	
Use of Other Medications for Blood Glucose Control	358	8.00%	314	7.92%	44	8.63%	0.578
Use of ACE inhibitors or ARBs	1,352	30.21%	1,092	27.53%	260	50.98%	**<0.0001**

† Indicates p value was based on t test with unequal variance; ‡ Indicates p value was based on Wilcoxon rank sum test because the variable was not normally distributed; § Indicates p value was based on Fisher's exact test.

Abbreviations: SD: standard deviation; BMI: Body mass index, calculated as the body mass in kilograms divided by the square of the body height in meters (kg/m2); SD: standard deviation; HDL: high-density lipoprotein; LDL: low-density lipoprotein; CAD: coronary artery disease; ADHD: Attention-deficit/hyperactivity disorder; RA: rheumatoid arthritis; IBD: Inflammatory bowel disease; PVD: peripheral vascular disease; CHF: congestive heart failure; CVA: cerebral vascular accident; TIA: transient ischemic attack. See “**Appendix 3”** for operational definitions of all variables.

Table 7. Baseline A1C measures of patients in the nephropathy cohort

Characteristics	Cohort of Nephropathy						
	Total		Nephropathy: No		Nephropathy: Yes		
	N	%	N	%	N	%	P-value
	4,476	**100.00%**	**3,966**	**88.61%**	**510**	**11.39%**	
Single A1C							**<0.0001†**
Mean (SD)	7.75	1.31	7.70	1.26	8.15	1.61	
Median (range)	7.50	4.00-15.60	7.50	4.00-15.60	7.90	5.10-15.00	
Mean A1C							**<0.0001†**
Mean (SD)	7.75	1.25	7.70	1.20	8.13	1.49	
Median (range)	7.57	4.07-14.00	7.53	4.07-14.00	7.77	5.40-13.97	
Quartiles of mean A1C							**<0.0001**
Quartile I	1,078	24.08%	991	24.99%	87	17.06%	
Quartile II	1,195	26.70%	1,070	26.98%	125	24.51%	
Quartile III	1,081	24.15%	955	24.08%	126	24.71%	
Quartile IV	1,122	25.07%	950	23.95%	172	33.73%	
SD A1C							**0.0007‡**
Mean (SD)	0.43	0.39	0.42	0.38	0.50	0.44	
Median (range)	0.34	0.00-5.15	0.32	0.00-5.15	0.36	0.00-3.76	
Quartiles of SDA1C							**<0.0001**
Quartile I	1,291	28.84%	1,155	29.12%	136	26.67%	
Quartile II	948	21.18%	854	21.53%	94	18.43%	
Quartile III	1,108	24.75%	1,002	25.26%	106	20.78%	
Quartile IV	1,129	25.22%	955	24.08%	174	34.12%	
CV A1C							**0.017‡**
Mean (SD)	0.05	0.04	0.05	0.04	0.06	0.04	
Median (range)	0.04	0.00-0.52	0.04	0.00-0.52	0.05	0.00-0.31	
Quartiles of CV A1C							**0.002**
Quartile I	1,115	24.91%	986	24.86%	129	25.29%	
Quartile II	1,123	25.09%	1,019	25.69%	104	20.39%	
Quartile III	1,119	25.00%	1,002	25.26%	117	22.94%	
Quartile IV	1,119	25.00%	959	24.18%	160	31.37%	

† Indicates p value was based on t test with unequal variance; ‡ Indicates p value was based on Wilcoxon rank sum test because the variable was not normally distributed; § Indicates p value was based on Fisher's exact test. See "**Appendix 3**" for operational definitions of all variables.

Predictor Selection

Predictors were selected based on univariate and correlation analyses of the train set as well as previous literature. Significant characteristics from univariate analyses of the train set were similar as significant factors from univariate analyses of the entire cohort. Pearson's correlation analyses were conducted on the train set to evaluate correlation between predictors and the outcome variable as well as test for multi-collinearity of predictor variables. Although most predictors were significantly correlated with the outcome variable (diabetic nephropathy), the absolute values of correlation coefficient were between 0.03-0.25: the top three correlated predictors were history of microalbuminuria (ρ=0.254), age (ρ=0.203), and history of diabetic retinopathy (ρ=0.201).

Among predictors, most recent A1C level was strongly ($|\rho|>0.7$) correlated with mean-A1C (ρ=0.926) but weakly correlated with SD-A1C (ρ=0.365) or CV-A1C (ρ=0.203); history of hypertension was strongly correlated with use of ACE inhibitors and ARBs (ρ=0.710); use of Humalog was strongly and negatively correlated with Novolog (ρ=-0.885); age, marital status, and working status were moderately correlated with each other ($0.4<|\rho|<0.5$); age, duration of T1D, history of hypertension, history of dyslipidemia, and use of ACE inhibitors or ARBs were also moderately correlated with each other ($0.4<|\rho|<0.5$).

Considering previous literature, results from univariate analysis and correlation analysis of the train set, the following 21 variables were selected: A1C variability, age, duration of T1D, BMI, household income (>=100k vs <100k), insurance type, marital status (married vs others), smoking status (ever smoked vs never), comorbidities including microalbuminuria, diabetic retinopathy, diabetic neuropathy, hypertension, dyslipidemia, CAD, PVD, hypothyroidism or Hashimoto disease, gastrointestinal diseases, RA or osteoporosis, depression and anxiety, and use of Novolog vs other insulins. When incorporating into machine learning models, multi-level categorical variables were dummy coded (0/1).

Predictive Models by LR

With each predictor set, a total of 11 LR models were developed: 10 from ten-fold cross-validation of the train set and 1 final model based on the entire train set and evaluated on the test set. The odds ratios (ORs) and their 95% confidence intervals (CIs) of the final model with each predictor set were reported in the following **Tables 8a** through **8e**.

Final model LR-Nep-A: While controlling for other covariates, unit increase in A1C would increase a patient's odds of developing diabetic nephropathy by 0.33 (OR 1.33, 95%CI 1.22-1.46, $p<0.0001$); one year older in age would raise the odds by 0.03 (OR 1.03, 95%CI 1.02-1.04, $p<0.0001$). The odds of developing diabetic nephropathy in patients with history of microalbuminuria were on average 6.67 (95%CI 4.62-9.63, $p<0.0001$) times that of patients without microalbuminuria. Having a medical history of diabetic retinopathy, diabetic neuropathy, hypertension, or musculoskeletal/connective tissue conditions also increase a patient's odds of developing diabetic nephropathy, while having commercial insurance decreases the odds by about a fourth (OR 0.75, 95%CI 0.57-0.97, $p<0.05$) (**Table 8a**).

Final model LR-Nep-B: This model indicates similar associations between predictors and diabetic nephropathy. Unit increase in mean-A1C would increase a patient's odds of developing diabetic nephropathy by 0.33 (OR 1.33, 95%CI 1.21-1.46, $p<0.0001$) while controlling for other covariates (**Table 8b**).

Final model LR-Nep-C & LR-Nep-D: Both models indicate that in addition to A1C or mean-A1C, SD-A1C is a significant predictor for the outcome of diabetic nephropathy. The odds of developing diabetic nephropathy increased by an average of 0.34 – 0.40 with unit increase in SD-A1C (LR-Nep-C: OR 1.40, 95%CI 1.05-1.85, p<0.05; LR-Nep-D: OR 1.33, 95%CI 1.00-1.76, p<0.05) (**Tables 8c** & **8d**).

Final model GEE-Nep-E: The GEE model indicates that while controlling for other covariates, both A1C values and SD-A1C over time were significantly associated with diabetic nephropathy. Unit increase in A1C would increase a patient's odds of developing diabetic nephropathy by 0.001(OR 1.00, 95%CI 1.000-1.001, p<0.01) whereas unit increase in SD-A1C would increase the odds by 0.61 (OR 1.61, 95%CI 1.25-2.07, p<0.001) (**Table 8e**).

Table 8a. Final LR model for prediction of development of diabetic nephropathy using predictor set with single A1C

LR-Nep-A	OR	95% CI	*P value*
Single A1C	1.334	1.222 - 1.457	**<0.0001**
Age at Exam (years)	1.032	1.022 - 1.043	**<0.0001**
Duration of T1D (years)	0.999	0.988 - 1.009	0.809
BMI (kg/m^2)	1.019	0.997 - 1.042	0.091
Annual household income: >=100K *vs* <100K	0.781	0.589 - 1.034	0.084
Commercial insurance *vs* Others	0.745	0.570 - 0.974	**0.031**
Married *vs* divorced, separated, single (never married), or widowed	0.988	0.752 - 1.297	0.931
Smoking status: ever smoked *vs* never	0.932	0.727 - 1.194	0.577
Comorbidities at baseline			
Microalbuminuria	6.670	4.621 - 9.628	**<0.0001**
Diabetic retinopathy	1.769	1.332 - 2.351	**<0.0001**
Diabetic neuropathy	1.451	1.087 - 1.936	**0.012**
Hypertension	1.369	1.042 - 1.799	**0.024**
Dyslipidemia	0.949	0.737 - 1.222	0.686
CAD	1.229	0.785 - 1.926	0.367
PVD	1.198	0.488 - 2.941	0.693
Hypothyroidism or Hashimoto disease	1.087	0.836 - 1.414	0.534
Gastrointestinal diseases	1.515	0.947 - 2.424	0.083
Musculoskeletal/connective tissue conditions	1.662	1.124 - 2.456	**0.011**
Depression	1.311	0.967 - 1.777	0.081
Anxiety	1.271	0.773 - 2.090	0.344
Use of Novolog *vs* Other insulins	0.866	0.688 - 1.091	0.222

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; See "**Appendix 3"** for operational definitions of all variables.

Table 8b. Final LR model for prediction of development of diabetic nephropathy using predictor set with mean A1C

LR-Nep-B	OR	95% CI	*P value*
Mean A1C	1.326	1.209 - 1.455	**<0.0001**
Age at Exam (years)	1.032	1.022 - 1.044	**<0.0001**
Duration of T1D (years)	0.999	0.988 - 1.010	0.847
BMI (kg/m^2)	1.019	0.997 - 1.042	0.085
Annual household income: >=100K *vs* <100K	0.785	0.593 - 1.041	0.092
Commercial insurance *vs* Others	0.740	0.566 - 0.967	**0.028**
Married *vs* divorced, separated, single (never married), or widowed	0.985	0.750 - 1.293	0.912
Smoking status: ever smoked *vs* never	0.929	0.725 - 1.191	0.562
Comorbidities at baseline			
Microalbuminuria	6.491	4.497 - 9.369	**<0.0001**
Diabetic retinopathy	1.748	1.315 - 2.323	**0.0001**
Diabetic neuropathy	1.453	1.089 - 1.937	**0.011**
Hypertension	1.347	1.025 - 1.771	**0.033**
Dyslipidemia	0.939	0.729 - 1.209	0.624
CAD	1.235	0.789 - 1.935	0.356
PVD	1.209	0.495 - 2.952	0.677
Hypothyroidism or Hashimoto disease	1.076	0.828 - 1.399	0.583
Gastrointestinal diseases	1.520	0.951 - 2.428	0.080
Musculoskeletal/connective tissue conditions	1.655	1.119 - 2.447	**0.012**
Depression	1.311	0.969 - 1.776	0.079
Anxiety	1.252	0.762 - 2.057	0.375
Use of Novolog *vs* Other insulins	0.872	0.692 - 1.098	0.243

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; See "**Appendix 3**" for operational definitions of all variables.

Table 8c. Final LR model for prediction of development of diabetic nephropathy using predictor set with combination single

LR-Nep-C	OR	95% CI	*P value*
Single A1C	1.286	1.173 - 1.411	**<0.0001**
SD A1C	1.396	1.054 - 1.848	**0.020**
Age at baseline (years)	1.033	1.022 - 1.044	**<0.0001**
Duration of T1D (years)	1.000	0.989 - 1.010	0.946
BMI (kg/m^2)	1.021	0.998 - 1.043	0.068
Annual household income: >=100K *vs* <100K	0.781	0.589 - 1.035	0.085
Commercial insurance *vs* Others	0.766	0.585 - 1.003	0.053
Married *vs* divorced, separated, single (never married), or widowed	0.987	0.751 - 1.296	0.923
Smoking status: ever smoked *vs* never	0.929	0.725 - 1.191	0.560
Comorbidities at baseline			
Microalbuminuria	6.613	4.575 - 9.559	**<0.0001**
Diabetic retinopathy	1.774	1.335 - 1.359	**<0.0001**
Diabetic neuropathy	1.431	1.072 - 1.910	**0.015**
Hypertension	1.359	1.034 - 1.786	**0.028**
Dyslipidemia	0.951	0.739 - 1.226	0.700
CAD	1.229	0.784 - 1.925	0.369
PVD	1.202	0.492 - 2.940	0.687
Hypothyroidism or Hashimoto disease	1.095	0.842 - 1.424	0.497
Gastrointestinal diseases	1.514	0.945 - 2.424	0.084
Musculoskeletal/connective tissue conditions	1.641	1.111 - 2.425	**0.013**
Depression	1.292	0.952 - 1.752	0.100
Anxiety	1.279	0.777 - 2.105	0.334
Use of Novolog *vs* Other insulins	0.869	0.690 - 1.095	0.235

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; See "**Appendix 3**" for operational definitions of all variables.

Table 8d. Final LR model for prediction of development of diabetic nephropathy using predictor set with combination mean

LR-Nep-D	OR	95% CI	*P value*
Mean A1C	1.275	1.152 - 1.411	**<0.0001**
SD A1C	1.328	1.003 - 1.759	**0.047**
Age at baseline (years)	1.033	1.022 - 1.044	**<0.0001**
Duration of T1D (years)	1.000	0.989 - 1.011	0.974
BMI (kg/m^2)	1.021	0.999 - 1.044	0.064
Annual household income: >=100K *vs* <100K	0.784	0.592 - 1.040	0.091
Commercial insurance *vs* Others	0.755	0.577 - 0.988	**0.041**
Married *vs* divorced, separated, single (never married), or widowed	0.980	0.746 - 1.288	0.887
Smoking status: ever smoked *vs* never	0.928	0.724 - 1.189	0.556
Comorbidities at baseline			
Microalbuminuria	6.466	4.473 - 9.346	**<0.0001**
Diabetic retinopathy	1.756	1.321 - 1.335	**0.0001**
Diabetic neuropathy	1.434	1.074 - 1.914	**0.014**
Hypertension	1.338	1.018 - 1.760	**0.037**
Dyslipidemia	0.943	0.732 - 1.215	0.651
CAD	1.237	0.790 - 1.938	0.353
PVD	1.214	0.499 - 2.954	0.669
Hypothyroidism or Hashimoto disease	1.083	0.833 - 1.409	0.549
Gastrointestinal diseases	1.519	0.950 - 2.429	0.081
Musculoskeletal/connective tissue conditions	1.638	1.108 - 2.421	**0.013**
Depression	1.297	0.958 - 1.758	0.093
Anxiety	1.262	0.767 - 2.076	0.360
Use of Novolog *vs* Other insulins	0.874	0.694 - 1.101	0.254

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; See "**Appendix 3**" for operational definitions of all variables.

Table 8e. Final GEE model for prediction of development of diabetic nephropathy using predictor set with multiple

GEE-Nep-E	OR	95% CI	*P value*
Individual A1C	1	1.000 - 1.001	**0.005**
SD A1C	1.607	1.247 - 2.071	**0.0002**
Age at baseline (years)	1.0003	1.0002 - 1.0005	**0.0002**
Duration of T1D (years)	1.008	0.997 - 1.018	0.164
BMI (kg/m^2)	1.018	0.996 - 1.040	0.106
Annual household income: >=100K *vs* <100K	0.774	0.587 - 1.019	0.068
Commercial insurance *vs* Others	0.694	0.529 - 0.909	**0.008**
Married *vs* divorced, separated, single (never married), or widowed	1.153	0.895 - 1.485	0.271
Smoking status: ever smoked *vs* never	1.034	0.803 - 1.332	0.793
Comorbidities at baseline			
Microalbuminuria	6.313	4.337 - 9.188	**<0.0001**
Diabetic retinopathy	1.931	1.442 - 2.586	**<0.0001**
Diabetic neuropathy	1.608	1.203 - 2.149	**0.001**
Hypertension	1.639	1.268 - 2.119	**0.0002**
Dyslipidemia	1.127	0.876 - 1.449	0.351
CAD	1.604	0.966 - 2.339	0.070
PVD	1.385	0.564 - 3.402	0.477
Hypothyroidism or Hashimoto disease	1.163	0.900 - 1.504	0.248
Gastrointestinal diseases	1.490	0.926 - 2.397	0.100
Musculoskeletal/connective tissue conditions	2.022	1.376 - 2.972	**0.0003**
Depression	1.344	0.993 - 1.821	0.056
Anxiety	1.261	0.761 - 2.089	0.367
Use of Novolog *vs* Other insulins	0.872	0.693 - 1.099	0.246

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; See "**Appendix 3"** for operational definitions of all variables.

Predictive Models by SVM

Using each predictor set, 11 SVM models were developed by Sci-Kit Learn SVC classifier: 10 from ten-fold cross-validation of the train set and 1 final model based on the entire train set and evaluated on the test set. Predictors were pre-processed using RobustScaler but without scaling (i.e., removing the median only). SMOTE was used to oversample the minority group (cases) so that there were equal numbers of cases and controls for modeling. Random state was set to be 42 to ensure repeatable weight initiation. The kernel function was set to be 'rbf' and γ as 'scale' for all models. The hyperparameter Cs used for the final trained models with the 5 predictor sets are as follows: a) SVM-Nep-A: C= 10.8; b) SVM-Nep-B: C=5.5; c) SVM-Nep-C: C=10.5; d) SVM-Nep-D: C=15.5; and e) SVM-Nep-E: C=4.6.

Predictive Models by NN

Using each predictor set, 11 NN models were developed using the TensorFlow.keras package: 10 from ten-fold cross-validation of the train set and 1 final model based on the entire train set and evaluated on the test set.

The final hyperparameters were selected based on the loss and accuracy curves of the train and validation set through the process of ten-fold cross validation. Each time, one hyperparameter was tuned to see how it impacted the loss curve and accuracy. The loss curve of the validation set was bumpy but gradually declining until flatten off. The plateau of the loss curve of the validation set indicated that the training can be stopped, even though the loss curve of the train set was still declining. With larger learning rate, fewer number of epochs was needed for reaching the plateau, but the learning curve can be bumpier. However, after we tried both ways – smaller learning rate with more epochs of training and larger learning rate with fewer epochs of learning – the highest F1 score can be achieved were similar, at

around the value of 0.6. Examples of the accuracy and loss curves of the train and validation set of NN models using the 5 predictor sets are provided in **Appendix 5**.

The final NN models were trained without scaling of the predictors. SMOTE was used to oversample the minority group (cases) so that there were equal numbers of cases and controls for modeling. All the final NN models comprised 1 input layer, 1 output layer with the 'sigmoid' activation function, and 2 hidden layers with the 'ReLU' activation function and a l_2 penalty of 0.005. The first hidden layer comprised 128 nodes and the second 64 nodes. The connections between the hidden layers and the consecutive layers can be randomly dropped by 50%. The Adam optimization algorithm was used for training. The learning rate and epochs used for the final NN models with the 5 predictor sets are as follows:

a) NN-Nep-A: learning rate = 0.01 and epochs = 50;

b) NN-Nep-B: learning rate = 0.01 and epochs = 50;

c) NN-Nep-C: learning rate = 0.01 and epochs = 50;

d) NN-Nep-D: learning rate = 0.00005 and epochs = 200; and

e) NN-Nep-E: learning rate = 0.00001 and epochs = 300.

As there are multiple sources of randomness, each model was repeated 10 times and the average performance metrics were calculated and reported.

Model Performance

The F1 scores of LR, SVM and NN models in the cohort of nephropathy by A1C variability are plotted in **Figure 11**. The performance measures of all models were provided in **Appendix 5**.

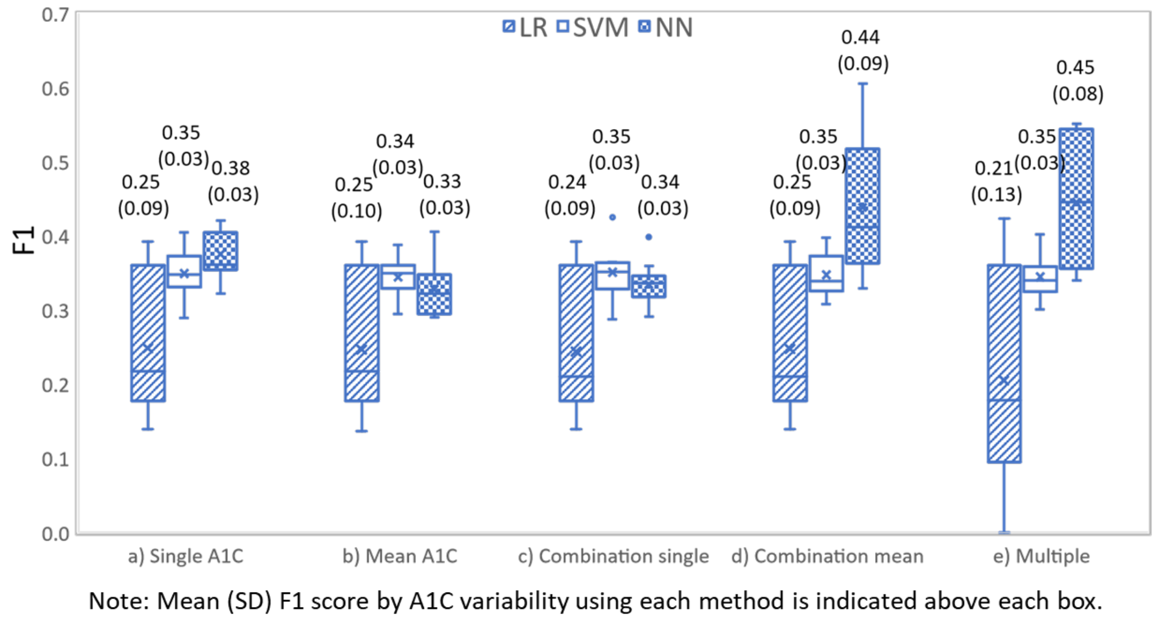


Figure 11. Box plot of F1 scores of nephropathy cohort by modeling method and A1C variability

Cohort of Retinopathy

Baseline Characteristics

Among the 3,595 patients in the retinopathy cohort, 659 (18%) developed diabetic retinopathy (cases) during the follow-up period. Slightly more than half (53%) were women. The mean (±SD) age of patients in the case group was 41 (±14.1) years, significantly older than those who did not develop nephropathy during follow-up (controls) (38±14.5, p<0.0001). The baseline demographic, clinical, and treatment characteristics between patients in the case and control groups were summarized in **Table 9**.

Demographics: Univariate analyses indicated that compared to the control group (18±11.1 years), cases had had T1D for a longer period at baseline (24±10.9 years, p<0.0001). Less proportion of the case group than the control group had commercial health insurance (82.4% vs 86.8%, p<0.05). A greater proportion of the case group than the control group had below $50k household income (26.5% vs 20.0%, p<0.001), worked either full-time or part-time

(68.3% vs 66.8%, p<0.0001), and were smoking at baseline (12.4% vs 8.3%, p<0.001) or ever smoked (31.7% vs 27.4%, p<0.05). The two groups were similar in other demographics.

Clinical characteristics: Patients in the case group on average had higher BMI (28±4.9 vs 27±4.8, p<0.0001), SBP (121±13.7 vs 120±12.8, p<0.05), and triglyceride level (92±66.5 vs 89±84.3, p<0.05) than the control group. Similar to the nephropathy cohort, more than a third (35.2%) of patients had their lipid fasting status unknow.

Medical history: The case group had a greater percentage of patients with a history of microalbuminuria (7.9% vs 3.8%, p<0.0001), diabetic nephropathy (6.5% vs 2.7%, p<0.0001) and neuropathy (17.7% vs 8.2%, p<0.0001), cardiovascular conditions including hypertension (34.0% vs 24.3%, p<0.0001), dyslipidemia (40.7% vs 24.3%, p<0.01) and CAD (4.7% vs 2.0%, p<0.0001), and depression (17.0% vs 11.6%, p<0.001).

Treatment: The two groups did not differ much in their treatment at baseline except that a greater proportion of the case group had used CGM (25.9% vs 22.2%, p<0.05) and ACE inhibitors or ARBs (37.0% vs 26.5%, p<0.0001).

A1C Measures: Measures of A1C were summarized in **Table 10**. Compared to the control group, the case group was on average higher in their most recent A1C level (7.9±1.4 vs 7.6±1.3, p<0.0001) and mean-A1C (7.9±1.3 vs 7.6±1.2, p<0.0001). But the SD-A1C and CV-A1C between cases and controls did not differ significantly. The gap between the last two A1C measures was on average 6.1 months (range 3.0-130.0 months) with a median gap of 4.0 months. Among the 2875 patients in the train set, there were only 34 (1%) patients whose gap between the last two A1C values were over 24 months.

Table 9. Baseline characteristics of patients in the retinopathy cohort

Characteristics	Cohort of Retinopathy						
	Total		Retinopathy: No		Retinopathy: Yes		
	N	%	N	%	N	%	
	3,595	100.00%	2,936	81.67%	659	18.33%	P-value
Age at baseline							**<0.0001**
Mean (SD)	39	14.49	38	14.53	41	14.07	
Median (range)	37	18.0 - 85.8	37	18.0 - 82.2	39	18.1 - 85.8	
Age group							**<0.0001**
18-27 years	778	21.64%	616	20.98%	162	24.58%	
28-37 years	1,069	29.74%	930	31.68%	139	21.09%	
38-47 years	753	20.95%	610	20.78%	143	21.70%	
48-64 years	852	23.70%	667	22.72%	185	28.07%	
≥65 years	143	3.98%	113	3.85%	30	4.55%	
Age at T1D							**<0.0001**
Mean (SD)	20	13.38	20	13.55	17	12.26	
Median (range)	16	0.0 - 76.0	16	0.0 - 76.0	13	0.0 - 65.0	
T1D Duration							**<0.0001**
Mean (SD)	19	11.34	18	11.11	24	10.87	
Median (range)	17	0.6 - 60.0	16	0.6 - 60.0	24	1.2 - 58.5	
Gender							0.992
Female	1,910	53.13%	1,560	53.13%	350	53.11%	
Male	1,685	46.87%	1,376	46.87%	309	46.89%	
Race							0.222
White	3,294	91.63%	2,691	91.66%	603	91.50%	
Black/African American	87	2.42%	67	2.28%	20	3.03%	
Hispanic or Latino	123	3.42%	107	3.64%	16	2.43%	
Others	91	2.53%	71	2.42%	20	3.03%	
Education Level	n = 3,517		n = 2,873		n = 644		0.723
Less than bachelor's degree	1,613	45.86%	1,324	46.08%	289	44.88%	
Bachelor's degree	1,192	33.89%	965	33.59%	227	35.25%	
Master's, professional, or doctorate	712	20.24%	584	20.33%	128	19.88%	
Insurance Coverage							**0.013**
Commercial insurance	3,092	86.01%	2,549	86.82%	543	82.40%	

Table 9. *Continued*

Characteristics	Cohort of Retinopathy						
	Total		Retinopathy: No		Retinopathy: Yes		P-value
	N	%	N	%	N	%	
	3,595	100.00%	2,936	81.67%	659	18.33%	
Government-sponsored insurance	395	10.99%	304	10.35%	91	13.81%	
Not specified	108	3.00%	83	2.83%	25	3.79%	
Marital Status	n = 3,568		n = 2,913		n = 655		0.333
Married or living together	2,119	59.39%	1,719	59.01%	400	61.07%	
Divorced, separated, single, or widowed	1,449	40.61%	1,194	40.99%	255	38.93%	
Annual household income	n = 2,913		n = 2,374		n = 539		**0.0007**
<$50,000	690	23.69%	547	23.04%	143	26.53%	
$50,000 to < $100,000	1,100	37.76%	873	36.77%	227	42.12%	
>=$100,000	1,123	38.55%	954	40.19%	169	31.35%	
Employment Status							**<0.0001**
Working full time or part-time at baseline	2,412	67.09%	1,962	66.83%	450	68.29%	
Student or homemaker	724	20.14%	647	22.04%	77	11.68%	
Unemployed, retired, disabled or other	459	12.77%	327	11.14%	132	20.03%	
Smoking Status							
Yes, smoking at baseline	327	9.10%	245	8.34%	82	12.44%	**0.0009**
Not smoking at baseline, but smoked before	1,014	28.21%	805	27.42%	209	31.71%	**0.027**
BMI (kg/m^2)							**<0.0001**
Mean (SD)	26.94	4.81	26.78	4.77	27.64	4.93	
Median (range)	26.27	11.16 - 65.57	26.15	11.16 - 65.57	26.80	15.99 - 51.16	
BMI category							**0.0003**
Under or normal weight	1,376	38.28%	1,156	39.37%	220	33.38%	
overweight	1,432	39.83%	1,173	39.95%	259	39.30%	
obese	787	21.89%	607	20.67%	180	27.31%	
Blood Pressure (mmHg)	n = 3,502		n = 2,857		n = 645		

Table 9. *Continued*

Characteristics	Cohort of Retinopathy						
	Total		Retinopathy: No		Retinopathy: Yes		
	N	%	N	%	N	%	
	3,595	100.00%	2,936	81.67%	659	18.33%	P-value
Diastolic blood pressure							0.588
Mean (SD)	72.29	8.37	72.26	8.31	72.45	8.62	
Median (range)	71	40 - 111	71	40 - 111	72	50 - 100	
Systolic blood pressure							**0.013**
Mean (SD)	120.29	13.01	120.00	12.84	121.49	13.68	
Median (range)	120	60 - 195	120	84 - 195	120	60 - 174	
Cholesterol Levels							
HDL value	n = 3,183		n = 2,597		n = 586		0.074
Mean (SD)	61.54	17.96	61.81	18.05	60.34	17.55	
Median (range)	59.00	14.00 - 162.00	59.00	14.00 - 155.00	57.00	17.00 - 162.00	
LDL value	n = 3,373		n = 2,748		n = 625		0.999
Mean (SD)	91.99	27.51	91.99	27.61	92.00	27.09	
Median (range)	90.00	3.00 - 266.00	90.00	3.00 - 266.00	88.00	22.00 - 192.00	
Triglycerides value	n = 3,118		n = 2,643		n = 575		**0.020‡**
Mean (SD)	89.60	81.31	89.04	84.29	92.11	66.52	
Median (range)	71.00	0.00 - 3000.00	71.00	0.00 - 3000.00	77.00	17.00 - 941.00	
Lipids Fasting Status							**0.0007**
Fasting	1538	44.63%	1,279	45.52%	259	40.72%	
Not Fasting	675	19.59%	567	20.18%	108	16.98%	
Unknown	1233	35.78%	964	34.31%	269	42.30%	
Microalbuminuria at baseline (Yes)	163	4.63%	111	3.78%	52	7.89%	**<0.0001**
Comorbidities at Baseline							
Diabetic nephropathy	122	3.39%	79	2.69%	43	6.53%	**<0.0001**
Diabetic neuropathy	359	9.99%	242	8.24%	117	17.75%	**<0.0001**
Cardiovascular conditions							
Hypertension	939	26.12%	715	24.35%	224	33.99%	**<0.0001**
Dyslipidemia	1,274	35.44%	1,006	24.26%	268	40.67%	**0.002**
CAD	91	2.53%	60	2.04%	31	4.70%	**<0.0001**
PVD	13	0.36%	9	0.31%	4	0.61%	0.274§
Cardiac arrythmia	28	0.78%	25	0.85%	3	0.46%	0.296§
Cerebrovascular accident	9	0.25%	7	0.19%	2	0.30%	0.673§

Table 9. *Continued*

Characteristics	Cohort of Retinopathy						
	Total		Retinopathy: No		Retinopathy: Yes		
	N	%	N	%	N	%	
	3,595	100.00%	2,936	81.67%	659	18.33%	P-value
Endocrine diseases							
Hypothyroidism or Hashimoto disease	795	22.11%	646	22.00%	149	22.61%	0.734
Hyperthyroidism or Grave's disease	72	2.00%	59	2.01%	13	1.97%	0.951
Other endocrine diseases	21	0.58%	15	0.51%	6	0.91%	0.253§
Gastrointestinal diseases	160	4.45%	131	4.46%	29	4.40%	0.945
Musculoskeletal/Connective Tissue conditions							
RA or osteoporosis	156	4.34%	124	4.22%	32	4.86%	0.472
Psychiatric conditions							
Depression	452	12.57%	340	11.58%	112	17.00%	**0.0002**
Anxiety	155	4.31%	129	4.39%	26	3.95%	0.609
ADHD	67	1.86%	53	1.81%	14	2.12%	0.584
Psychosis	10	0.28%	6	0.20%	4	0.61%	0.093
Eating disorders	21	0.58%	15	0.51%	6	0.91%	0.253
Skin conditions	79	2.20%	64	2.18%	15	2.28%	0.879
CGM use							**0.037**
Yes	822	22.87%	651	22.17%	171	25.95%	
No	2,773	77.13%	2,285	77.83%	488	74.05%	
Insulin use							
Type of insulin analog							
Insulin lispro (Humalog)	1,834	51.02%	1,488	50.68%	346	52.50%	0.398
Insulin aspart (Novolog)	1,658	46.12%	1,370	46.66%	288	43.70%	0.168
Insulin detemir (Levemir)	123	3.42%	99	3.37%	24	3.64%	0.731
Insulin glargine (Lantus)	1,170	32.55%	950	32.36%	220	33.38%	0.611
Participant insulin delivery method at time of most recent exam							0.537
Pump only	2,169	60.33%	1,779	60.59%	390	59.18%	
Injections/pens only	1,356	37.72%	1,101	37.50%	255	38.69%	
Both pump and injections/pens	70	1.95%	56	1.90%	14	2.12%	

Table 9. *Continued*

Characteristics	Cohort of Retinopathy						
	Total		Retinopathy: No		Retinopathy: Yes		
	N	%	N	%	N	%	
	3,595	**100.00%**	**2,936**	**81.67%**	**659**	**18.33%**	**P-value**
Use of Other Medications for Blood Glucose Control	288	8.01%	226	7.70%	62	9.41%	0.144
Use of ACE inhibitors or ARBs	1,022	28.43%	778	26.50%	244	37.03%	**<0.0001**

† Indicates p value was based on t test with unequal variance; ‡ Indicates p value was based on Wilcoxon rank sum test because the variable was not normally distributed; § Indicates p value was based on Fisher's exact test.

Abbreviations: SD: standard deviation; BMI: Body mass index, calculated as the body mass in kilograms divided by the square of the body height in meters (kg/m2); SD: standard deviation; HDL: high-density lipoprotein; LDL: low-density lipoprotein; CAD: coronary artery disease; ADHD: Attention-deficit/hyperactivity disorder; RA: rheumatoid arthritis; IBD: Inflammatory bowel disease; PVD: peripheral vascular disease; CHF: congestive heart failure; CVA: cerebral vascular accident; TIA: transient ischemic attack. See “**Appendix 3”** for operational definitions of comorbidities and treatments.

Table 10. Baseline A1C measures of patients in the retinopathy cohort

Characteristics	Cohort of Retinopathy						
	Total		Retinopathy: No		Retinopathy: Yes		
	N	%	N	%	N	%	P-value
	3,595	**100.00%**	**2,936**	**81.67%**	**659**	**18.33%**	
Single A1C							**<0.0001†**
Mean (SD)	7.69	1.29	7.63	1.26	7.95	1.40	
Median (range)	7.50	4.00 - 15.00	7.50	4.80 - 15.00	7.70	4.00 - 14.40	
Mean A1C							**<0.0001†**
Mean (SD)	7.69	1.23	7.63	1.20	7.95	1.34	
Median (range)	7.50	4.07 - 14.00	7.47	4.90 - 14.00	7.67	4.07 - 13.73	
Quartiles of mean A1C							**<0.0001**
Quartile I	918	25.54%	795	27.08%	123	18.66%	
Quartile II	907	25.23%	745	25.37%	162	24.58%	
Quartile III	871	24.23%	718	24.46%	153	23.22%	
Quartile IV	899	25.01%	678	23.09%	221	33.54%	
SD A1C							0.131‡
Mean (SD)	0.42	0.39	0.42	0.38	0.45	0.43	
Median (range)	0.32	0.00 - 5.15	0.32	0.00 - 5.15	0.35	0.00 - 3.76	
Quartiles of SD A1C							**0.030**
Quartile I	1,043	29.01%	862	29.36%	181	27.47%	
Quartile II	687	19.11%	565	19.24%	122	18.51%	
Quartile III	942	26.20%	785	26.74%	157	23.82%	
Quartile IV	923	25.67%	724	24.66%	199	30.20%	
CV A1C							0.575‡
Mean (SD)	0.05	0.04	0.05	0.04	0.05	0.04	
Median (range)	0.04	0.00 - 0.52	0.04	0.00 - 0.52	0.04	0.00 - 0.34	
Quartiles of CV A1C							0.199
Quartile I	896	24.92%	734	25.00%	162	24.58%	
Quartile II	901	25.06%	733	24.97%	168	25.49%	
Quartile III	899	25.01%	752	25.61%	147	22.31%	
Quartile IV	899	25.01%	717	24.42%	182	27.62%	

† Indicates p value was based on t test with unequal variance; ‡ Indicates p value was based on Wilcoxon rank sum test because the variable was not normally distributed; § Indicates p value was based on Fisher's exact test. See "**Appendix 3**" for operational definitions of all variables.

Predictor Selection

Predictors were selected based on univariate and correlation analyses of the train set as well as previous literature. Significant characteristics from univariate analyses of the train set were similar to those significant factors from univariate of the entire cohort. Pearson's correlation analyses were conducted on the train set to evaluate correlation between predictors and the outcome variable as well as test for multi-collinearity of predictor variables. Although most predictors were significantly correlated with the diabetic retinopathy, the absolute values of correlation coefficient were between 0.02-0.21: the top three correlated predictors were history of duration of T1D (ρ=0.207), mean A1C (ρ=0.105), and history of diabetic neuropathy (ρ=0.104).

Among predictors, most recent A1C level was strongly ($|\rho|>0.7$) correlated with mean A1C (ρ=0.925) but weakly correlated with SD-A1C (ρ=0.357); history of hypertension was strongly correlated with use of ACE inhibitors and ARBs (ρ=0.717); age and working status were moderately correlated with each other ($0.4<|\rho|<0.5$); age, duration of T1D, history of hypertension, history of dyslipidemia, and use of ACE inhibitors or ARBs were also moderately correlated with each other ($0.4<|\rho|<0.5$).

Considering previous literature, results from univariate analysis and correlation analysis, the following 15 variables were selected: A1C variability, age, duration of T1D, BMI, household income (>=100k vs <100k), insurance type, smoking status at baseline (yes vs no), comorbidities including microalbuminuria or diabetic retinopathy, diabetic neuropathy, dyslipidemia, CAD, depression or psychosis, use of CGM, and use of ACE inhibitors or ARBs at baseline. When incorporating into machine learning models, multi-level categorical variables were dummy coded (0/1).

Predictive Models by LR

With each of the 5 predictor sets, a total of 11 LR models were developed: 10 from ten-fold cross-validation of the train set and 1 final model based on the entire train set and evaluated on the test set. The ORs and their 95% CIs of the final model with each predictor set were reported in the following **Tables 11a** through **11e**.

Final model LR-Ret-A: While controlling for other covariates, unit increase in most recent A1C would increase a patient's odds of developing diabetic retinopathy by 0.24 (OR 1.24, 95%CI 1.15-1.34, $p<0.0001$); one year older in age reduced the odds by 0.01 (OR 0.99, 95%CI 0.98-0.99, $p<0.05$), whereas one more year having T1D increased the odds of diabetic retinopathy by 0.05 (OR 1.05, 95%CI 1.04-1.06, $p<0.0001$); unit increase in BMI would raise the odds by 0.03 (OR 1.03, 95%CI 1.01-1.05, $p<0.0001$). The odds of developing diabetic retinopathy in patients with history of microalbuminuria or diabetic nephropathy were on average 1.50 (95%CI 1.05-2.16, $p<0.05$) times that of patients without these conditions. Similarly, a medical history of diabetic neuropathy would put a patient at higher risk of developing diabetic retinopathy (1.5 times of the odds) than a patient without diabetic neuropathy (**Table 11a**).

Final model LR-Ret-B: This model indicates similar associations between predictors and diabetic retinopathy. Unit increase in mean A1C would increase a patient's odds of developing diabetic nephropathy by 0.27 (OR 1.27, 95%CI 1.17-1.38, $p<0.0001$) while controlling for other covariates (**Table 11b**).

Final model LR-Ret-C, LR-Ret-D & GEE-Ret-E: None of the three models indicate a significant association between SD A1C and the outcome of diabetic retinopathy (**Tables 11c – 11e**).

Table 11a. Final LR model for prediction of development of diabetic retinopathy using predictor set with single A1C

LR-Ret-A	OR	95% CI	*P value*
Single A1C	1.242	1.149 - 1.342	**<0.0001**
Age at baseline (years)	0.988	0.978 - 0.998	**0.016**
T1D duration (years)	1.050	1.039 - 1.061	**<0.0001**
BMI (kg/m^2)	1.028	1.007 - 1.049	**0.009**
Annual household income: >=100K *vs* <100K	0.848	0.673 - 1.069	0.163
Commercial insurance *vs* Others	0.872	0.651 - 1.168	0.359
Employment Status			**0.0004**
Working full time or part time (reference)	-	-	-
Student or homemaker	0.573	0.416 - 0.790	**0.0007**
Unemployed, retired, disabled or other	1.323	0.980 - 1.786	0.068
Smoking status at baseline: yes *vs* no	1.160	0.836 - 1.610	0.374
Comorbidities at baseline			
Microalbuminuria or Diabetic nephropathy	1.502	1.046 - 2.157	**0.028**
Diabetic neuropathy	1.496	1.096 - 2.042	**0.011**
Dyslipidemia	0.831	0.661 - 1.046	0.115
CAD	1.126	0.646 - 1.964	0.675
Depression or psychosis	1.138	0.858 - 1.509	0.370
Use of CGM at baseline: yes *vs* no	1.260	0.999 - 1.590	0.051
Use of ACE inhibitors or ARBs: yes *vs* no	1.011	0.792 - 1.290	0.930

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; See "**Appendix 3**" for operational definitions of all variables.

Table 11b. Final LR model for prediction of development of diabetic retinopathy using predictor set with mean A1C

LR-Ret-B	OR	95% CI	*P value*
Mean A1C	1.275	1.174 - 1.385	**<0.0001**
Age at baseline (years)	0.988	0.978 - 0.998	**0.024**
T1D duration (years)	1.051	1.040 - 1.062	**<0.0001**
BMI (kg/m^2)	1.028	1.007 - 1.049	**0.009**
Annual household income: >=100K *vs* <100K	0.862	0.684 - 1.087	0.209
Commercial insurance *vs* Others	0.876	0.654 - 1.174	0.375
Employment Status	1.124	0.809 - 1.561	0.486
Working full time or part time (reference)			**0.0003**
Student or homemaker	-	-	-
Unemployed, retired, disabled or other	0.548	0.412 - 0.784	**0.0006**
Smoking status at baseline: yes *vs* no	1.317	0.976 - 1.778	0.072
Comorbidities at baseline			
Microalbuminuria or Diabetic nephropathy	1.478	1.029 - 2.122	**0.035**
Diabetic neuropathy	1.474	1.080 - 2.012	**0.015**
Dyslipidemia	0.825	0.655 - 1.037	0.099
CAD	1.140	0.654 - 1.985	0.644
Depression or psychosis	1.118	0.843 - 1.484	0.439
Use of CGM at baseline: yes *vs* no	1.245	0.987 - 1.571	0.064
Use of ACE inhibitors or ARBs: yes *vs* no	1.009	0.790 - 1.288	0.944

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; See "**Appendix 3**" for operational definitions of all variables.

Table 11c. Final LR model for prediction of development of diabetic retinopathy using predictor set with combination single

LR-Ret-C	OR	95% CI	*P value*
Single A1C	1.224	1.126 - 1.329	**<0.0001**
SD A1C	1.144	0.876 - 1.494	0.324
Age at baseline (years)	0.988	0.978 - 0.998	**0.019**
T1D duration (years)	1.051	1.040 - 1.062	**<0.0001**
BMI (kg/m^2)	1.028	1.007 - 1.050	**0.008**
Annual household income: >=100K *vs* <100K	0.849	0.674 - 1.070	0.166
Commercial insurance *vs* Others	0.882	0.658 - 1.184	0.404
Employment Status			**0.0004**
Working full time or part time (reference)	-	-	-
Student or homemaker	0.576	0.418 - 0.795	**0.0008**
Unemployed, retired, disabled or other	1.316	0.975 - 1.777	0.073
Smoking status at baseline: yes *vs* no	1.157	0.834 - 1.606	0.384
Comorbidities at baseline			
Microalbuminuria or Diabetic nephropathy	1.494	1.040 - 2.146	**0.030**
Diabetic neuropathy	1.484	1.087 - 2.027	**0.013**
Dyslipidemia	0.831	0.661 - 1.046	0.115
CAD	1.128	0.647 - 1.967	0.670
Depression or psychosis	1.134	0.855 - 1.505	0.382
Use of CGM at baseline: yes *vs* no	1.255	0.994 - 1.583	0.056
Use of ACE inhibitors or ARBs: yes *vs* no	1.013	0.794 - 1.293	0.917

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; See "**Appendix 3**" for operational definitions of all variables.

Table 11d. Final LR model for prediction of development of diabetic retinopathy using predictor set with combination mean

LR-Ret-D	OR	95% CI	*P value*
Mean A1C	1.266	1.157 - 1.386	**<0.0001**
SD A1C	1.054	0.804 - 1.382	0.704
Age at baseline (years)	0.988	0.978 - 0.999	0.025
T1D duration (years)	1.051	1.040 - 1.062	<0.0001
BMI (kg/m^2)	1.028	1.007 - 1.050	0.008
Annual household income: >=100K *vs* <100K	0.862	0.684 - 1.087	0.209
Commercial insurance *vs* Others	0.879	0.656 - 1.179	0.389
Employment Status			**0.0004**
Working full time or part time (reference)	-	-	-
Student or homemaker	0.570	0.413 - 0.787	**0.0006**
Unemployed, retired, disabled or other	1.314	0.973 - 1.774	0.075
Smoking status at baseline: yes *vs* no	1.123	0.809 - 1.561	0.488
Comorbidities at baseline			
Microalbuminuria or Diabetic nephropathy	1.475	1.027 - 2.120	**0.035**
Diabetic neuropathy	1.469	1.076 - 2.006	**0.016**
Dyslipidemia	0.825	0.656 - 1.038	0.101
CAD	1.140	0.655 - 1.986	0.642
Depression or psychosis	1.118	0.842 - 1.484	0.440
Use of CGM at baseline: yes *vs* no	1.243	0.985 - 1.568	0.066
Use of ACE inhibitors or ARBs: yes *vs* no	1.010	0.791 - 1.289	0.939

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; See "**Appendix 3**" for operational definitions of all variables.

Table 11e. Final GEE model for prediction of development of diabetic retinopathy using predictor set with multiple

GEE-Ret-E	OR	95% CI	*P value*
Individual A1C	1	1.000 - 1.0001	**0.0005**
SD A1C	1.194	0.949 - 1.500	0.13
Age at baseline (years)	0.969	0.961 - 0.978	**<0.0001**
T1D Duration (years)	1.032	1.023 - 1.041	**<0.0001**
BMI (kg/m^2)	1.026	1.005 - 1.047	**0.013**
Annual household income: >=100K *vs* <100K	0.865	0.688 - 1.087	0.213
Commercial insurance *vs* Others	0.826	0.622 - 1.096	0.185
Employment Status			
Working full time or part time (reference)	ref	-	-
Student or homemaker	3.259	2.197 - 4.834	**<0.0001**
Unemployed, retired, disabled or other	2.143	1.577 - 2.912	**<0.0001**
Smoking status at baseline: yes *vs* no	1.139	0.820 - 1.582	0.437
Comorbidities at baseline			
Microalbuminuria or Diabetic nephropathy	1.523	1.065 - 2.175	**0.021**
Diabetic neuropathy	1.764	1.294 - 2.405	**0.0003**
Dyslipidemia	1.027	0.822 - 1.281	0.816
CAD	1.610	0.924 - 2.903	0.092
Depression or psychosis	1.213	0.908 - 1.620	0.191
Use of CGM at baseline: yes *vs* no	1.254	1.0004 - 1.572	0.049
Use of ACE inhibitors or ARBs: yes *vs* no	1.252	0.992 - 1.580	0.058

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; See "**Appendix 3**" for operational definitions of all variables.

Predictive Models by SVM

Using each of the 5 predictor sets, 11 SVM models were developed by Sci-Kit Learn SVC classifier: 10 from ten-fold cross-validation of the train set and 1 final model based on the entire train set and evaluated on the test set. Predictors were pre-processed using RobustScaler without scaling (i.e., removing the median only). SMOTE was used to oversample the cases so that there were equal numbers of cases and controls for modeling. Random state was set to be 42 to ensure repeatable weight initiation. The kernel function was set to be 'rbf' and γ as 'scale' for all models. The hyperparameter Cs used for the final trained models with the 5 predictor sets are as follows: a) SVM-Ret-A: C= 2.7; b) SVM-Ret-B: C=6.5; c) SVM-Ret-C: C=1.6; d) SVM-Ret-D: C=5.6; and e) SVM-Ret-E: C=0.2.

Predictive Models by NN

Using each of the 5 predictor sets, 11 NN models were developed using the TensorFlow.keras package: 10 from ten-fold cross-validation of the train set and 1 final model based on the entire train set and evaluated on the test set.

The final hyperparameters were selected based on the loss and accuracy curves of the train and validation set through the process of ten-fold cross validation. Each time, one hyperparameter was tuned to see how it impacted the loss curve and accuracy. The loss curve of the validation set was bumpy but gradually declining until flatten off. The plateau of the loss curve of the validation set indicated that the training can be stopped, even though the loss curve of the train set was still declining. With larger learning rate, fewer number of epochs was needed for reaching the plateau, but the learning curve can be bumpier. However, after we tried both ways – smaller learning rate with more epochs of training and larger learning rate with fewer epochs of learning – the highest F1 score can be achieved were similar, at

around the value of 0.65. Examples of the accuracy and loss curves of the train and validation set of NN models using the 5 predictor sets are provided in **Appendix 5**.

The final NN models were trained without scaling of the predictors. SMOTE was used to oversample the cases so that there were equal numbers of cases and controls for modeling. All final NN models comprised 1 input layer, 1 output layer with the 'sigmoid' activation function, and 3 hidden layers with the 'ReLU' activation function and a l_2 penalty of 0.005. The Adam optimization algorithm was used for training. The learning rate, epochs, nodes in the 1st, 2nd and 3rd hidden layers, and percentage of randomly dropped connections between consecutive layers (indicated as percentages in parentheses after number of nodes) used for the final NN models with the 5 predictor sets were as follows:

a) NN-Ret-A: learning rate = 0.01; epochs = 50; 1st hidden layer: 50 nodes (random drop = 50%); 2nd hidden layer: 50 nodes (random drop = 50%); 3rd hidden layer: 50 nodes (random drop = 50%).

b) NN-Ret-B: the same as NN-Ret-A.

c) NN-Ret-C: learning rate = 0.01; epochs = 50; 1st hidden layer: 50 nodes (random drop = 30%); 2nd hidden layer: 50 nodes (random drop = 30%); 3rd hidden layer: 50 nodes (random drop = 30%).

d) NN-Ret-D: learning rate = 0.01; epochs = 100; 1st hidden layer: 128 nodes (random drop = 50%); 2nd hidden layer: 64 nodes (random drop = 50%); 3rd hidden layer: 64 nodes (random drop = 50%).

e) NN-Ret-E: learning rate = 0.01; epochs = 40; 1st hidden layer: 128 nodes (random drop = 50%); 2nd hidden layer: 64 nodes (random drop = 50%); 3rd hidden layer: 64 nodes (random drop = 50%).

As there are multiple sources of randomness, each model was repeated 10 times and the average performance metrics were calculated and reported.

Model Performance

The F1 scores of LR, SVM and NN models in the retinopathy cohort by A1C variability are plotted in **Figure 12**. The performance measures of all models are provided in **Appendix 6**.

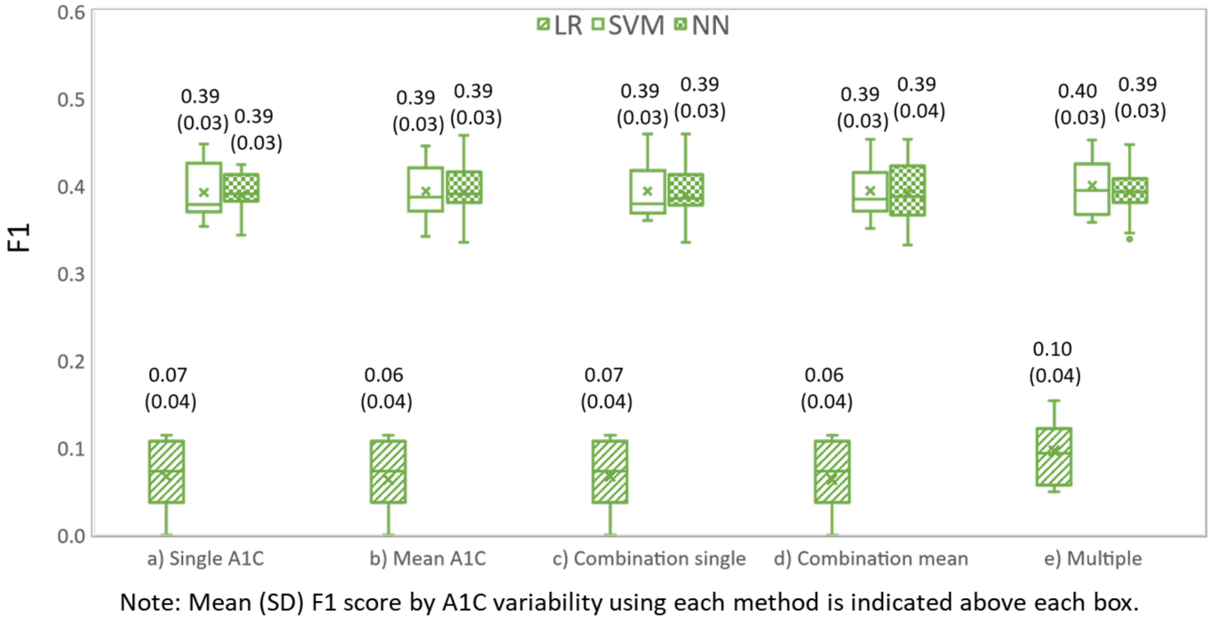


Figure 12. Box plot of F1 scores of retinopathy cohort by modeling method and A1C variability

Cohort of Neuropathy

Baseline Characteristics

Among the 4,072 patients in the neuropathy cohort, 579 (14%) developed diabetic neuropathy (cases) during the follow-up period. Slightly more than half (55%) were women. The mean (±SD) age of patients in the case group was 49 (±14.5) years, significantly older than those who did not develop nephropathy during follow-up (controls) (36±14.2, p<0.0001). The baseline demographic, clinical, and treatment characteristics between patients in the case and control groups were summarized in **Table 12**.

<u>Demographics</u>: Univariate analyses indicated that compared to the control group (18±12.8 years), cases had had T1D for a longer period at baseline (27±13.5 years, p<0.0001). Less proportion of the case group than the control group had commercial health insurance (78.9% vs 87.3%, p<0.0001) and worked full-time or part-time (63.6% vs 65.0%, p<0.0001). A greater proportion of the case group than the control group were married or living together (71.7% vs 54.6%, p<0.0001), had below $50k household income (33.5% vs 23.9%, p<0.0001), and had ever smoked (37.8% vs 25.4%, p<0.0001) or were smoking at baseline (12.4% vs 8.6%, p<0.01). The two groups were similar in other demographics.

<u>Clinical characteristics</u>: Patients in the case group on average had higher BMI (28±4.9 vs 27±5.5, p<0.01) and SBP (123±15.1 vs 120±12.8, p<0.0001) than the control group. Similar to the nephropathy and retinopathy cohorts, more than a third (35.3%) of patients had their lipid fasting status unknow.

<u>Medical history</u>: The case group had a greater percentage of patients with a history of microalbuminuria (11.4% vs 4.1%, p<0.0001), diabetic retinopathy (31.6% vs 9.6%, p<0.0001) and nephropathy (9.1% vs 3.2%, p<0.0001), cardiovascular conditions including hypertension (49.0% vs 23.4%, p<0.0001), dyslipidemia (50.4% vs 32.2%, p<0.0001), CAD

(8.1% vs 2.0%, p<0.0001), hypothyroidism or Hashimoto disease (25.6% vs 21.3%, p<0.05), RA or osteoporosis (10.4% vs 3.6%, p<0.0001), and depression (16.2% vs 10.8%, p<0.0001).

Treatment: The two groups did not differ much in their treatment at baseline except that a greater proportion of the case group had used ACE inhibitors or ARBs (50.8% vs 26.2%, p<0.0001) and other medications for blood glucose control (including DPP4 Inhibitors, GLP1 agonists, metformin, pramlintide or other medications reported by participant or indicated in medication records) (10.2% vs 7.6%, p<0.05).

A1C Measures: Measures of A1C were summarized in **Table 13**. Univariate analyses indicated that compared to the control group, the case group was on average higher in their most recent A1C level (7.8±1.4 vs 7.7±1.3, p<0.05) and mean A1C (7.9±1.3 vs 7.7±1.2, p<0.05). But the SD A1C and CV A1C between cases and controls did not differ significantly. The gap between the last two A1C measures was on average 6.1 months (range 3.0-130.0 months) with a median gap of 4.0 months. Among the 3257 patients in the train set, there were only 41 (1%) patients whose gap between the last two A1C values were over 24 months.

Table 12. Baseline characteristics of patients in the neuropathy cohort

Characteristics	**Cohort of Neuropathy**						
	Total		**Neuropathy: No**		**Neuropathy: Yes**		**P-value**
	N	**%**	**N**	**%**	**N**	**%**	
	4,072	**100.00%**	**3,493**	**85.78%**	**579**	**14.22%**	
Age at baseline							**<0.0001**
Mean (SD)	38	14.98	36	14.21	49	14.53	
Median (range)	36	18.0 - 86.8	34	18.0 - 86.8	49	18.5 - 85.8	
Age group							**<0.0001**
18-27 years	862	21.17%	770	22.04%	92	15.89%	
28-37 years	1,317	32.34%	1,269	32.33%	48	8.29%	
38-47 years	797	19.57%	673	19.27%	124	21.42%	
48-64 years	899	22.08%	673	19.27%	226	39.03%	
≥65 years	197	4.84%	108	3.09%	89	15.37%	
Age at T1D Diagnosis							**<0.0001**†
Mean (SD)	18	13.22	18	12.82	22	14.82	
Median (range)	14	0.0 - 72.0	14	0.0 - 72.0	18	0.0 - 66.0	
T1D Duration							**<0.0001**†
Mean (SD)	20	11.95	18	11.23	27	13.50	
Median (range)	17	0.6 - 62.6	16	0.6 - 62.6	26	1.2 - 62.4	
Gender							0.385
Female	2,239	54.99%	1,911	54.71%	328	56.65%	
Male	1,833	45.01%	1,582	45.29%	251	43.35%	
Race/Ethnicity							0.526
White Non-Hispanic	3,693	90.69%	3,167	90.67%	526	90.85%	
Black/African American	112	2.75%	93	2.66%	19	3.28%	
Hispanic or Latino	158	3.88%	135	3.86%	23	3.97%	
Others	109	2.68%	98	2.81%	11	1.90%	
Education Level	n = 3,971		n = 3,411		n = 560		0.269
Less than bachelor's degree	1,930	48.60%	1,675	49.11%	255	45.54%	
Bachelor's degree	1,284	32.33%	1,089	31.93%	195	34.82%	
Master's, professional, or doctorate	757	19.06%	647	18.97%	110	19.64%	
Insurance Coverage	n = 3,741		n = 3,196		n = 545		**<0.0001**
Commercial health insurance	3,220	86.07%	2,790	87.30%	430	78.90%	

Table 12. *Continued*

Characteristics	Cohort of Neuropathy						
	Total		Neuropathy: No		Neuropathy: Yes		P-value
	N	%	N	%	N	%	
	4,072	100.00%	3,493	85.78%	579	14.22%	
Government-sponsored insurance	415	11.09%	313	9.79%	102	18.72%	
Not specified	106	2.83%	93	2.91%	13	2.39%	
Marital Status	n = 4,037		n = 3,460		n = 477		**<0.0001**
Married or living together	2,304	57.07%	1,890	54.62%	414	71.75%	
Divorced, separated, single, or widowed	1,733	42.93%	1,570	45.38%	163	28.25%	
Annual household income (self-reported)	n = 3,182		n = 2,725		n = 457		**<0.0001**
<$50,000	805	25.30%	652	23.93%	153	33.48%	
$50,000 to < $100,000	1,199	37.68%	1,041	38.20%	158	34.57%	
>=$100,000	1,178	37.02%	1,032	37.87%	146	31.95%	
Employment Status							**<0.0001**
Working full time or part-time at baseline	2,638	64.78%	2,270	64.99%	368	63.56%	
Student or homemaker	880	21.61%	840	24.05%	40	6.91%	
Unemployed, retired, disabled or other	554	13.61%	383	10.96%	171	29.53%	
Smoking Status							
Yes, smoking at baseline	371	9.11%	299	8.56%	72	12.44%	**0.003**
Not smoking at baseline, but smoked before	1,108	27.21%	889	25.45%	219	37.82%	**<0.0001**
BMI (kg/m^2)							**0.008†**
Mean (SD)	27.02	4.99	26.93	4.90	27.57	5.47	
Median (range)	26.26	11.16 - 65.57	26.19	11.16 - 56.05	26.63	17.87 - 65.57	
BMI category							**0.028**
Under or normal weight	1,589	39.02%	1,379	39.48%	210	36.27%	

Table 12. *Continued*

Characteristics	Cohort of Neuropathy						
	Total		Neuropathy: No		Neuropathy: Yes		P-value
	N	%	N	%	N	%	
	4,072	**100.00%**	**3,493**	**85.78%**	**579**	**14.22%**	
overweight	1,553	38.14%	1,341	38.39%	212	36.61%	
obese	930	22.84%	773	22.13%	157	27.12%	
Blood Pressure (mmHg)	n = 3,965		n = 3,400		n = 565		
Diastolic blood pressure							**<0.0001†**
Mean (SD)	72.13	8.40	72.39	8.27	70.54	9.02	
Median (range)	71.00	42.00 - 111.00	72.00	46.00 - 111.00	70.00	42.00 - 102.00	
Systolic blood pressure							**<0.0001†**
Mean (SD)	120.44	13.19	119.99	12.79	123.15	15.11	
Median (range)	120.00	82.00 - 198.00	120.00	82.00 - 198.00	122.00	86.00 - 178.00	
Cholesterol Levels							
HDL value	n = 3,612		n = 3,105		n = 507		0.852
Mean (SD)	61.35	17.84	61.33	17.71	61.49	18.62	
Median (range)	59.00	14.00 - 162.00	59.00	14.00 - 162.00	59.00	23.00 - 140.00	
LDL value	n = 3,814		n = 3,263		n = 551		0.100
Mean (SD)	92.60	28.00	92.91	27.93	90.78	28.39	
Median (range)	90.00	3.00 - 281.00	91.00	3.00 - 281.00	88.00	26.00 - 205.00	
Triglycerides value	n = 3,547		n = 3,047		n = 500		0.969‡
Mean (SD)	90.85	81.88	90.77	83.25	91.32	73.08	
Median (range)	72.00	0.00 - 3000.00	72.00	0.00 - 3000.00	72.00	13.00 - 941.00	
Lipids Fasting Status	n = 3,901		n = 3,345		n = 556		0.908
Fasting	1,745	44.73%	1,492	44.60%	253	45.50%	
Not Fasting	779	19.97%	671	20.06%	108	19.42%	
Unknown	1,377	35.30%	1,182	35.34%	195	35.07%	

Table 12. *Continued*

Characteristics	Cohort of Neuropathy						
	Total		Neuropathy: No		Neuropathy: Yes		P-value
	N	%	N	%	N	%	
	4,072	100.00%	3,493	85.78%	579	14.22%	
Microalbuminuria at baseline (Yes)	208	5.11%	142	4.07%	66	11.40%	**<0.0001**
Comorbidities at Baseline							
Diabetic retinopathy	520	12.77%	337	9.65%	183	31.61%	**<0.0001**
Diabetic nephropathy	166	4.08%	113	3.24%	53	9.15%	**<0.0001**
Cardiovascular conditions							
Hypertension	1,100	27.01%	816	23.36%	284	49.05%	**<0.0001**
Dyslipidemia	1,417	34.80%	1,125	32.21%	292	50.43%	**<0.0001**
CAD	117	2.87%	70	2.00%	47	8.12%	**<0.0001**
PVD	18	0.44%	12	0.34%	6	1.04%	**0.033**§
Cardiac arrythmia	31	0.76%	20	0.57%	11	1.90%	**0.003**§
Cerebrovascular accident	17	0.42%	9	0.26%	8	1.38%	**0.001**§
Endocrine diseases							
Hypothyroidism or Hashimoto disease	893	21.93%	745	21.33%	148	25.56%	**0.023**
Hyperthyroidism or Grave's disease	81	1.99%	66	1.89%	15	2.59%	0.263
Other endocrine diseases	30	0.74%	26	0.74%	4	0.69%	1.000§
Gastrointestinal diseases	172	4.22%	142	4.07%	30	5.18%	0.216
Musculoskeletal/ Connective Tissue conditions							
RA or osteoporosis	185	4.54%	125	3.58%	60	10.36%	**<0.0001**
Psychiatric conditions							
Depression	473	11.62%	379	10.85%	94	16.23%	**0.0002**
Anxiety	170	4.17%	141	4.04%	29	5.01%	0.279

Table 12. *Continued*

Characteristics	Cohort of Neuropathy						
	Total		Neuropathy: No		Neuropathy: Yes		P-value
	N	%	N	%	N	%	
	4,072	100.00%	3,493	85.78%	579	14.22%	
ADHD	78	1.92%	71	2.03%	7	1.21%	0.181
Psychosis	15	0.37%	12	0.34%	3	0.52%	0.461§
Eating disorders	22	0.54%	19	0.54%	3	0.52%	1.000§
Skin conditions	79	1.94%	63	1.80%	16	2.76%	0.121
CGM use							0.059
Yes	897	22.03%	752	21.53%	145	25.04%	
No	3,175	77.97%	2,741	78.47%	434	74.96%	
Insulin use							
Type of insulin analog							
Insulin lispro (Humalog)	2,103	51.65%	1,804	51.65%	299	51.64%	0.998
Insulin aspart (Novolog)	1,856	45.58%	1,607	46.01%	249	43.01%	0.179
Insulin detemir (Levemir)	142	3.49%	120	3.44%	22	3.80%	0.658
Insulin glargine (Lantus)	1,320	32.42%	1,131	32.38%	189	32.64%	0.900
Participant insulin delivery method at time of most recent exam							0.519
Pump only	2,446	60.07%	2,103	60.21%	343	59.24%	
Injections/pens only	1,545	37.84%	1,317	37.30%	228	39.38%	
Both pump and injections/pens	81	1.99%	73	2.09%	8	1.39%	
Use of Other Medications for Blood Glucose Control	324	7.96%	265	7.59%	59	10.19%	**0.032**
Use of ACE inhibitors or ARBs	1,211	29.74%	917	26.25%	294	50.78%	**<0.0001**

† Indicates p value was based on t test with unequal variance; ‡ Indicates p value was based on Wilcoxon rank sum test because the variable was not normally distributed; § Indicates p value was based on Fisher's exact test.

Abbreviations: SD: standard deviation; BMI: Body mass index, calculated as the body mass in kilograms divided by the square of the body height in meters (kg/m2); SD: standard deviation; HDL: high-density lipoprotein; LDL: low-density lipoprotein; CAD: coronary artery disease; ADHD: Attention-deficit/hyperactivity disorder; RA: rheumatoid arthritis; IBD: Inflammatory bowel disease; PVD: peripheral vascular disease; CHF: congestive heart failure; CVA: cerebral vascular accident; TIA: transient ischemic attack. See "**Appendix 3"** for operational definitions of predictor variables.

Table 13. Baseline A1C measures of patients in the neuropathy cohort

Characteristics	Cohort of Neuropathy						
	Total		Neuropathy: No		Neuropathy: Yes		
	N	%	N	%	N	%	
	4,072	100.00%	3,493	85.78%	579	14.22%	P-value
Single A1C							**0.041†**
Mean (SD)	7.74	1.30	7.72	1.29	7.85	1.37	
Median (range)	7.50	4.00 - 15.60	7.50	4.00 - 15.60	7.60	5.10 - 15.60	
Mean A1C							**0.011†**
Mean (SD)	7.74	1.25	7.72	1.23	7.87	1.34	
Median (range)	7.53	4.07 - 14.00	7.53	4.07 - 14.00	7.63	5.13 - 13.88	
Quartiles of mean A1C							0.159
Quartile I	1,000	24.56%	865	24.76%	135	23.32%	
Quartile II	994	24.41%	864	24.74%	130	22.45%	
Quartile III	1,060	26.03%	912	26.11%	148	25.56%	
Quartile IV	1,018	25.00%	852	24.39%	166	28.67%	
SD A1C							0.552‡
Mean (SD)	0.43	0.37	0.42	0.36	0.45	0.43	
Median (range)	0.35	0.00 - 4.60	0.35	0.00 - 4.42	0.35	0.00 - 4.60	
Quartiles of SD A1C							0.453
Quartile I	1,160	28.49%	1,005	28.77%	155	26.77%	
Quartile II	861	21.14%	731	20.93%	130	22.45%	
Quartile III	1,027	25.22%	889	25.45%	138	23.83%	
Quartile IV	1,024	25.15%	868	24.85%	156	26.94%	
CV A1C							0.874‡
Mean (SD)	0.05	0.04	0.05	0.04	0.05	0.04	
Median (range)	0.04	0.00 - 0.52	0.04	0.00 - 0.47	0.05	0.00 - 0.52	
Quartiles of CV A1C							0.655
Quartile I	1,018	25.00%	868	24.85%	150	25.91%	
Quartile II	1,021	25.07%	888	25.42%	133	22.97%	
Quartile III	1,017	24.98%	868	24.85%	149	25.73%	
Quartile IV	1,016	24.95%	869	24.88%	147	25.39%	

† Indicates p value was based on t test with unequal variance; ‡ Indicates p value was based on Wilcoxon rank sum test because the variable was not normally distributed; § Indicates p value was based on Fisher's exact test. See "**Appendix 3**" for operational definitions of all variables.

Predictor Selection

Predictors were selected based on univariate and correlation analyses of the train set as well as previous literature. Significant characteristics from univariate analyses of the train set were similar to those significant factors from univariate analyses of the entire cohort. Pearson's correlation analyses were conducted on the train set to evaluate correlation between predictors and the outcome variable as well as test for multi-collinearity of predictor variables. Although most predictors were significantly correlated with the diabetic retinopathy, the absolute values of correlation coefficient were between 0.03-0.30: the top three correlated predictors were age ($\rho=0.303$), history of diabetic retinopathy ($\rho=0.228$), and duration of T1D ($\rho=0.242$).

Among predictors, most recent A1C level was weakly correlated with SD A1C ($\rho=0.371$); history of hypertension was strongly correlated with use of ACE inhibitors and ARBs ($\rho=0.720$); age, duration of T1D, history of hypertension, history of dyslipidemia, and use of ACE inhibitors or ARBs were moderately correlated with each other ($0.4<|\rho|<0.5$); age, marital status, and working status were also moderately correlated with each other ($0.4<|\rho|<0.5$).

Considering previous literature, results from univariate analyses and correlation analyses, the following 21 variables were selected: A1C variability, age, duration of T1D, BMI, household income (>=100k vs <100k), insurance type, employment status, smoking status (ever smoked vs never), comorbidities including microalbuminuria or diabetic nephropathy, diabetic retinopathy, hypertension, dyslipidemia, CAD, PVD, cardiac arrythmia, CVA, hypothyroidism or Hashimoto disease, gastrointestinal diseases, musculoskeletal/connective tissue conditions (RA or osteoporosis) and depression, and use of other medications for blood glucose control (yes vs no). When incorporating into machine learning models, multi-level categorical variables were dummy coded (0/1).

Predictive Models by LR

With each predictor set, a total of 11 LR models were developed: 10 from ten-fold cross-validation of the train set and 1 final model based on the entire train set and evaluated on the test set. The ORs and their 95% CIs of the final model with each predictor set were reported in the following **Tables 14a** through **14e**.

Final model LR-Neu-A: While controlling for other covariates, unit increase in A1C would increase a patient's odds of developing diabetic neuropathy by 0.23 (OR 1.23, 95%CI 1.13-1.34, p<0.0001); one year older in age would raise the odds by 0.04 (OR 1.04, 95%CI 1.03-1.05, p<0.0001); one year longer in having T1D would raise the odds of diabetic neuropathy by 0.02 (OR 1.02, 95%CI 1.01-1.03, p<0.0001). The odds of developing diabetic neuropathy in patients with history of microalbuminuria or diabetic nephropathy were on average twice (95%CI 1.42-2.80, p<0.0001) that of patients without. Having a medical history of diabetic retinopathy or depression also increase a patient's odds of developing diabetic neuropathy. Interestingly, history of PVD decreases the odds by 0.72 (OR 0.28, 95%CI 0.08-0.98, p<0.05), whereas use of other blood glucose control medication versus not increased the odds of diabetic neuropathy (OR 1.51, 95%CI 1.05-2.17, p<0.05) (**Table 14a**).

Final model LR-Neu-B: This model indicates similar associations between predictors and diabetic neuropathy. Unit increase in mean A1C would increase a patient's odds of developing diabetic nephropathy by 0.28 (OR 1.28, 95%CI 1.17-1.40, p<0.0001) while controlling for other covariates (**Table 14b**).

Final model LR-Neu-C: This model indicates that in addition to a single A1C, SD A1C is a significant predictor for the outcome of diabetic neuropathy. The odds of developing diabetic neuropathy increased by 0.45 with unit increase in SD-A1C (OR 1.45, 95%CI 1.07-1.96, $p<0.05$) (**Tables 14c**).

Final model LR-Neu-D: This model did not indicate a significant association between SD A1C and the outcome of diabetic neuropathy (**Tables 14d**).

Final model GEE-Nep-E: The GEE model indicates that while controlling for other covariates, A1C values over time were not significantly associated with diabetic neuropathy. Unit increase in SD A1C would increase the odds by 0.36 (OR 1.36, 95%CI 1.03-1.80, $p<0.05$) (**Table 14e**).

Table 14a. Final LR model for prediction of development of diabetic neuropathy using predictor set with single A1C

LR-Neu-A	OR	95% CI	*P value*
Single A1C	1.230	1.127 - 1.342	**<0.0001**
Age at baseline (years)	1.042	1.031 - 1.053	**<0.0001**
T1D Duration (years)	1.021	1.011 - 1.032	**<0.0001**
BMI (kg/m^2)	0.985	0.963 - 1.008	0.189
Annual household income: >=100K *vs* <100K	0.803	0.623 - 1.035	0.090
Commercial insurance *vs* Others	0.832	0.631 - 1.098	0.193
Employment Status			**0.002**
Working full time or part time (reference)	-	-	-
Student or homemaker	0.514	0.335 - 0.789	**0.002**
Unemployed, retired, disabled or other	1.263	0.942 - 1.694	0.118
Smoking status: ever smoked *vs* never	1.355	1.076 - 1.706	**0.010**
Comorbidities at baseline			
Microalbuminuria or diabetic nephropathy	1.994	1.422 - 2.796	**<0.0001**
Diabetic retinopathy	1.711	1.295 - 2.262	**0.0002**
Hypertension	1.091	0.846 - 1.407	0.502
Dyslipidemia	0.938	0.740 - 1.188	0.594
CAD	0.831	0.510 - 1.355	0.458
PVD	0.277	0.078 - 0.978	**0.046**
Cardiac arrythmia	1.857	0.780 - 4.422	0.162
CVA	1.519	0.473 - 4.880	0.482
Hypothyroidism or Hashimoto disease	0.863	0.669 - 1.115	0.260
Gastrointestinal diseases	1.262	0.763 - 2.087	0.364
Musculoskeletal/connective tissue conditions	0.997	0.658 - 1.509	0.988
Depression	1.383	1.027 - 1.862	**0.033**
Use of other medications for blood glucose control: yes *vs* no	1.509	1.049 - 2.170	**0.027**

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; CVA: Cerebrovascular accident See "**Appendix 3"** for operational definitions of all variables.

Table 14b. Final LR model for prediction of development of diabetic neuropathy using predictor set with mean A1C

LR-Neu-B	OR	95% CI	*P value*
Mean-A1C	1.279	1.167 - 1.402	**<0.0001**
Age at baseline (years)	1.043	1.032 - 1.054	**<0.0001**
T1D Duration (years)	1.022	1.012 - 1.032	**<0.0001**
BMI (kg/m^2)	0.984	0.962 - 1.007	0.176
Annual household income: >=100K *vs* <100K	0.819	0.635 - 1.056	0.124
Commercial insurance *vs* Others	0.835	0.633 - 1.101	0.202
Employment Status			**0.002**
Working full time or part time (reference)	-	-	-
Student or homemaker	0.508	0.331 - 0.781	**0.002**
Unemployed, retired, disabled or other	1.261	0.940 - 1.691	0.122
Smoking status: ever smoked *vs* never	1.340	1.064 - 1.688	**0.013**
Comorbidities at baseline			
Microalbuminuria or diabetic nephropathy	1.957	1.394 - 2.746	**0.0001**
Diabetic retinopathy	1.685	1.274 - 2.229	**0.0003**
Hypertension	1.074	0.832 - 1.387	0.581
Dyslipidemia	0.927	0.731 - 1.176	0.534
CAD	0.832	0.510 - 1.357	0.462
PVD	0.267	0.075 - 0.949	**0.041**
Cardiac arrythmia	1.852	0.775 - 4.442	0.165
CVA	1.564	0.481 - 5.079	0.457
Hypothyroidism or Hashimoto disease	0.855	0.661 - 1.104	0.229
Gastrointestinal diseases	1.264	0.765 - 2.091	0.361
Musculoskeletal/connective tissue conditions	1.000	0.660 - 1.516	0.998
Depression	1.367	1.015 - 1.842	**0.040**
Use of other medications for blood glucose control: yes *vs* no	1.521	1.057 - 2.189	**0.024**

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; CVA: Cerebrovascular accident See "**Appendix 3"** for operational definitions of all variables.

Table 14c. Final LR model for prediction of development of diabetic neuropathy using predictor set with combination single

LR-Neu-C	OR	95% CI	*P value*
Most recent A1C	1.182	1.078 - 1.297	**0.0004**
SD-A1C	1.447	1.069 - 1.959	**0.017**
Age at baseline (years)	1.043	1.032 - 1.054	**<0.0001**
T1D Duration (years)	1.022	1.012 - 1.033	**<0.0001**
BMI (kg/m^2)	0.986	0.964 - 1.009	0.232
Annual household income: >=100K *vs* <100K	0.800	0.621 - 1.032	0.086
Commercial insurance *vs* Others	0.855	0.647 - 1.130	0.270
Employment Status			**0.002**
Working full time or part time	-	-	-
Student or homemaker	0.519	0.338 - 0.797	**0.003**
Unemployed, retired, disabled or other	1.262	0.941 - 1.692	0.121
Smoking status: ever smoked *vs* never	1.363	1.082 - 1.716	**0.009**
Comorbidities at baseline			
Microalbuminuria or diabetic nephropathy	1.962	1.398 - 2.755	**<0.0001**
Diabetic retinopathy	1.706	1.290 - 2.256	**0.0002**
Hypertension	1.091	0.845 - 1.408	0.504
Dyslipidemia	0.942	0.743 - 1.195	0.625
CAD	0.829	0.509 - 1.352	0.453
PVD	0.283	0.081 - 0.993	**0.049**
Cardiac arrythmia	1.739	0.726 - 4.166	0.215
CVA	1.468	0.456 - 4.762	0.520
Hypothyroidism or Hashimoto disease	0.865	0.669 - 1.117	0.266
Gastrointestinal diseases	1.270	0.768 - 2.103	0.352
Musculoskeletal/connective tissue conditions	0.984	0.650 - 1.489	0.938
Depression	1.266	1.014 - 1.841	**0.040**
Use of other medications for blood glucose control: yes *vs* no	1.504	1.146 - 2.165	**0.028**

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; CVA: Cerebrovascular accident See “**Appendix 3”** for operational definitions of all variables.

Table 14d. Final LR model for prediction of development of diabetic neuropathy using predictor set with combination mean

LR-Neu-D	OR	95% CI	*P value*
Mean-A1C	1.232	1.115 - 1.362	**<0.0001**
SD-A1C	1.331	0.979 - 1.809	0.068
Age at baseline (years)	1.044	1.033 - 1.055	**<0.0001**
T1D Duration (years)	1.022	1.012 - 1.033	**<0.0001**
BMI (kg/m^2)	0.986	0.963 - 1.008	0.215
Annual household income: >=100K *vs* <100K	0.814	0.631 - 1.051	0.114
Commercial insurance *vs* Others	0.851	0.645 - 1.124	0.257
Employment Status			**0.002**
Working full time or part time (reference)	-	-	-
Student or homemaker	0.513	0.334 - 0.789	**0.002**
Unemployed, retired, disabled or other	1.260	0.939 - 1.690	0.123
Smoking status: ever smoked *vs* never	1.349	1.070 - 1.700	**0.011**
Comorbidities at baseline			
Microalbuminuria or diabetic nephropathy	1.937	1.379 - 2.721	**0.0001**
Diabetic retinopathy	1.684	1.272 - 2.228	**0.0003**
Hypertension	1.076	0.833 - 1.390	0.573
Dyslipidemia	0.933	0.735 - 1.184	0.569
CAD	0.830	0.509 - 1.354	0.456
PVD	0.274	0.078 - 0.966	**0.044**
Cardiac arrythmia	1.765	0.735 - 4.237	0.204
CVA	1.518	0.467 - 4.933	0.488
Hypothyroidism or Hashimoto disease	0.857	0.663 - 1.107	0.238
Gastrointestinal diseases	1.270	0.767 - 2.102	0.352
Musculoskeletal/connective tissue conditions	0.990	0.654 - 1.500	0.963
Depression	1.358	1.007 - 1.830	**0.045**
Use of other medications for blood glucose control: yes *vs* no	1.515	1.053 - 2.180	**0.025**

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; CVA: Cerebrovascular accident See “**Appendix 3”** for operational definitions of all variables.

Table 14e. Final GEE model for prediction of development of diabetic neuropathy using predictor set with multiple

LR-Neu-E	OR	95% CI	*P value*
Individual A1C	1	1.000 - 1.0001	0.052
SD A1C	1.362	1.029 - 1.803	**0.031**
Age at baseline (years)	1.009	1.001 - 1.017	**0.024**
T1D Duration (years)	0.991	0.983 - 0.999	**0.034**
BMI (kg/m^2)	0.98	0.958 - 1.003	0.089
Annual household income: >=100K *vs* <100K	0.847	0.656 - 1.092	0.199
Commercial insurance *vs* Others	0.867	0.668 - 1.125	0.283
Employment Status			
Working full time or part time (reference)	ref	-	-
Student or homemaker	4.935	3.082 - 7.901	**<0.0001**
Unemployed, retired, disabled or other	2.942	1.954 - 4.429	**<0.0001**
Smoking status: ever smoked *vs* never	1.401	1.114 - 1.761	**0.003**
Comorbidities at baseline			
Microalbuminuria or diabetic nephropathy	1.871	1.329 - 2.635	**0.0003**
Diabetic retinopathy	2.659	2.018 - 3.505	**<0.0001**
Hypertension	1.539	1.199 - 1.976	**0.001**
Dyslipidemia	1.191	0.940 - 1.508	0.147
CAD	1.201	0.733 - 1.970	0.467
PVD	0.410	0.123 - 1.360	0.145
Cardiac arrythmia	1.881	0.733 - 4.829	0.189
CVA	1.873	0.550 - 6.382	0.316
Hypothyroidism or Hashimoto disease	1.095	0.855 - 1.403	0.47
Gastrointestinal diseases	1.357	0.848 - 2.169	0.203
Musculoskeletal/connective tissue conditions	1.277	0.821 - 1.986	0.278
Depression	1.345	0.987 - 1.831	0.060
Use of other medications for blood glucose control: yes *vs* no	1.359	0.937 - 1.970	0.106

Abbreviations: OR: odds ratio; CI: confidence interval; CAD: coronary artery disease; PVD: peripheral vascular disease; CVA: Cerebrovascular accident See "**Appendix 3"** for operational definitions of all variables.

Predictive Models by SVM

Using each predictor set, 11 SVM models were developed by Sci-Kit Learn SVC classifier: 10 from ten-fold cross-validation of the train set and 1 final model based on the entire train set and evaluated on the test set. Predictors were pre-processed using RobustScaler without scaling (i.e., removing the median only). SMOTE was used to oversample the cases so that there were equal numbers of cases and controls for modeling. Random state was set to be 42 to ensure repeatable weight initiation. The kernel function was set to be 'rbf' and γ as 'scale' for all models. The hyperparameter Cs used for the final trained models with the 5 predictor sets are as follows: a) SVM-Neu-A: C= 4.3; b) SVM-Neu-B: C=5.8; c) SVM-Neu-C: C=8.4; d) SVM-Neu-D: C=6; and e) SVM-Neu-E: C=3.8.

Predictive Models by NN

Using each predictor set, 11 NN models were developed using the TensorFlow.keras package: 10 from ten-fold cross-validation of the train set and 1 final model based on the entire train set and evaluated on the test set.

The final hyperparameters were selected based on the loss and accuracy curves of the train and validation set through the process of ten-fold cross validation. Each time, one hyperparameter was tuned to see how it impacted the loss curve and accuracy. The loss curve of the validation set was bumpy but gradually declining until flatten off. The plateau of the loss curve of the validation set indicated that the training can be stopped, even though the loss curve of the train set was still declining. With larger learning rate, fewer number of epochs was needed for reaching the plateau, but the learning curve can be bumpier. However, after we tried both ways – smaller learning rate with more epochs of training and larger learning rate with fewer epochs of learning – the highest F1 score can be achieved were similar, at

around the value of 0.6. Examples of the accuracy and loss curves of the train and validation set of NN models using the 5 predictor sets are provided in **Appendix 4**.

The final NN models were trained without scaling of the predictors. SMOTE was used to oversample the cases so that there were equal numbers of cases and controls for modeling. All final NN models comprised 1 input layer, 1 output layer with the 'sigmoid' activation function, and 3 hidden layers with the 'ReLU' activation function and a l_2 penalty of 0.005. The 1st hidden layer comprised 128 nodes, the 2nd 64 nodes and the 3rd 64 nodes. The connections between each hidden layer and the consecutive layers can be randomly dropped by 50%. The Adam optimization algorithm was used for training with a learning rate of 0.01. The epochs used for the final NN models with the 5 predictor sets are as follows:

a) NN-Neu-A: epochs = 50;

b) NN-Neu-B: epochs = 50;

c) NN-Neu-C: epochs = 50;

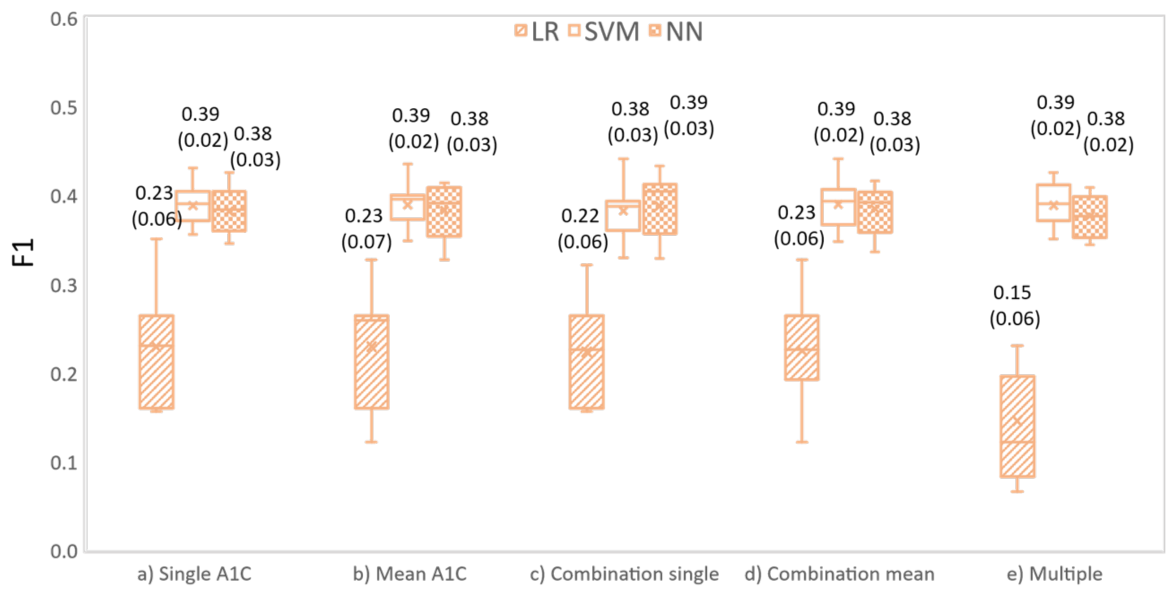
d) NN-Neu-D: epochs = 40; and

e) NN-Neu-E: epochs = 40.

As there are multiple sources of randomness, each model was repeated 10 times and the average performance metrics were calculated and reported.

Model Performance

The F1 scores of LR, SVM and NN models in the neuropathy cohort by A1C variability are plotted in **Figure 13**. The performance measures of all models are provided in **Appendix 7**.

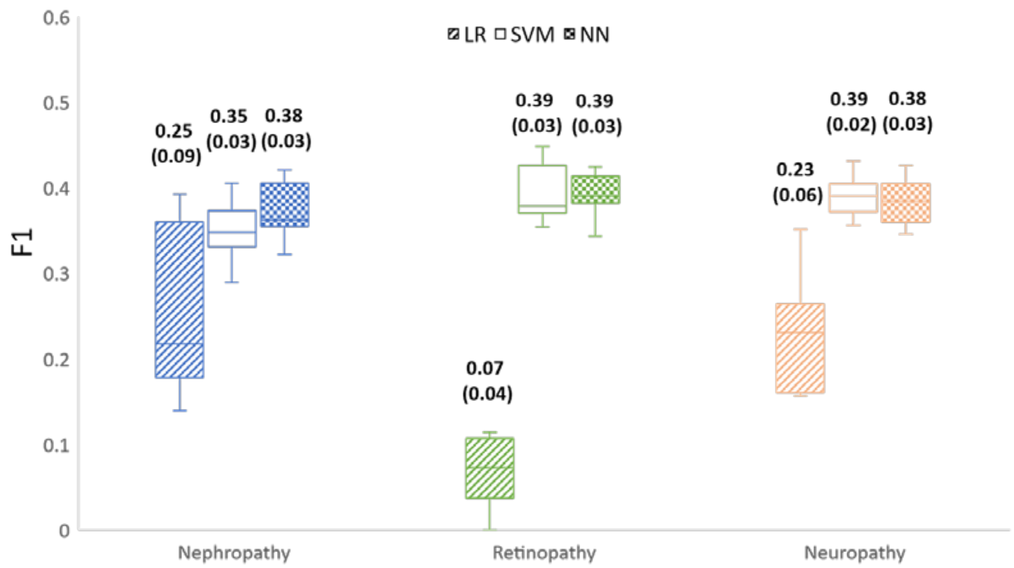


Note: Mean (SD) F1 score by A1C variability using each method is indicated above each box.

Figure 13. Box plot of F1 scores of retinopathy cohort by modeling method and A1C variability

Testing of Statistical Hypotheses

The F1 scores using predictor set with single A1C are plotted in **Figure 14**.



Note: Mean (SD) F1 score of each cohort using each method is indicated above each box.

Figure 14. Box plot of F1 scores using predictor sets with single A1C by modeling method and microvascular complication

For statistical hypothesis 1, two-way ANOVA indicated a significant interaction between the effects of modeling method and microvascular complication on F1 scores, $F(4, 90) = 21.75$, $p<.0001$ (**Table 15**). There was statistically significant difference in F1 scores between ML and LR models ($F=403.92$, $p<0.0001$).

Table 15. Two-way ANOVA testing the effect of modeling method and microvascular complication on F1 scores (n=99)

Source	DF	Sum of Squares	Mean Squares	F Value	P
Modeling method	2	0.861	0.431	202.08	<.0001
Microvascular complication	2	0.048	0.024	11.29	<.0001
Interaction term	4	0.185	0.046	21.75	<.0001
Model	8	1.095	0.137	64.22	<.0001
Error	90	0.192	0.002		
Corrected Total	98	1.287			
Contrast: ML vs LR	1	0.861	0.861	403.92	<.0001

An examination of the interaction indicates that the interaction was mainly between the modeling methods of SVM and NN (**Figure 15**).

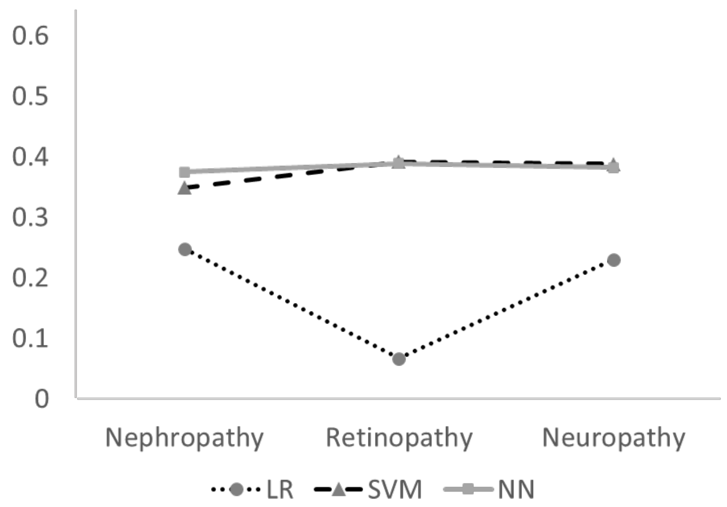


Figure 15. Interaction plot for F1 scores by microvascular complication and modeling method

Post hoc Tukey-Kramer test was further conducted to determine adjusted pairwise differences between modeling methods. There was significant difference between LR and SVM ($p<0.0001$) and between LR and NN ($p<0.0001$). There was no significant difference between the method of SVM and NN ($p>0.05$) (**Figure 16**).

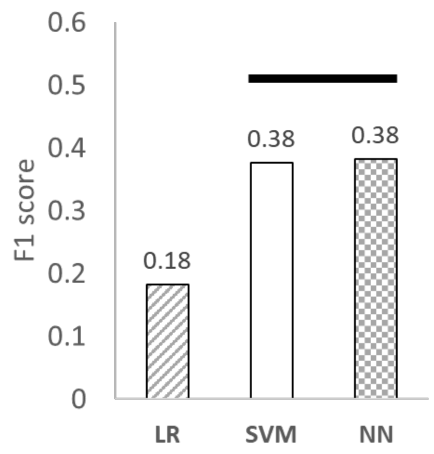


Figure 16. Post hoc Tukey-Kramer multiple comparisons of least squares means for effect of modeling method

For statistical hypothesis 2, three-way ANOVA indicated significant interactions at all levels (**Table 16**). In order to evaluate the effect of A1C variability on the prediction for each microvascular complication, two-way ANOVA was further performed within each cohort, respectively.

Table 16. Three-way ANOVA testing the effect of modeling method, microvascular complication and A1C variability on F1 scores (n=495)

Source	DF	Sum of Squares	Mean Squares	F Value	P
Modeling method (X1)	2	4.762	2.381	902.37	<.0001
Microvascular complication (X2)	2	0.174	0.087	32.88	<.0001
A1C variability (X3)	4	0.010	0.003	0.97	0.423
X1*X2	4	0.785	0.196	74.37	<.0001
X1*X3	8	0.057	0.007	2.71	0.006
X2*X3	8	0.049	0.006	2.33	0.018
X1*X2*X3	16	0.103	0.006	2.44	0.001
Model	44	5.940	0.135	51.16	<.0001
Error	450	1.187	0.003		
Corrected Total	494	7.128			

In the nephropathy cohort, two-way ANOVA indicated significant interaction between modeling method and A1C variability (**Table 17**). There was statistically significant difference in F1 scores between ML and LR models (F=119.03, p<0.0001).

Table 17. Two-way ANOVA testing the effect of modeling method and A1C variability on F1 scores of the nephropathy cohort (n=165)

Source	DF	Sum of Squares	Mean Squares	F Value	P
Modeling method	2	0.632	0.316	63.27	<.0001
A1C variability	4	0.032	0.008	1.61	0.173
Interaction term	8	0.119	0.015	2.98	0.004
Model	14	0.783	0.056	11.20	<.0001
Error	150	0.749	0.005		

Corrected Total	164	1.533			
Contrast: ML vs LR	1	0.595	0.595	119.03	<.0001

An examination of the interaction indicates that the interaction was mainly between the modeling methods of SVM and NN (**Figure 17**).

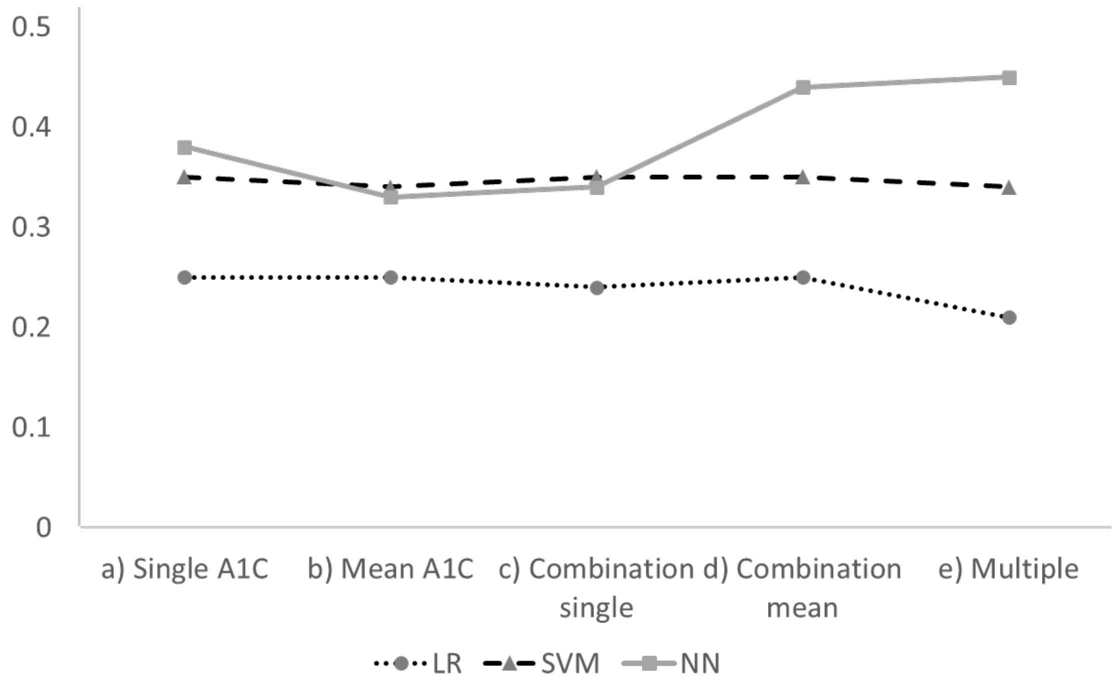


Figure 17. Interaction plot for F1 scores of the nephropathy cohort by microvascular complication and A1C variability

F test indicates that A1C variability had significant effect on F1 score of the nephropathy cohort when the modeling method was NN: F=6.78, p<.0001. Post hoc Tukey-Kramer test indicates that mean F1 scores of the nephropathy cohort from NN models using d) combination mean or e) multiple were significantly higher than using b) mean A1C or c) combination single.

In the retinopathy cohort, two-way ANOVA did not indicate significant interaction between modeling method and A1C variability (**Table 18**). There was statistically significant difference in F1 scores between ML and LR models (F=119.03, p<0.0001). There is no effect of A1C variability on F1 scores of the retinopathy cohort.

Table 18. Two-way ANOVA testing the effect of modeling method and A1C variability on F1 scores of the retinopathy cohort (n=165)

Source	DF	Sum of Squares	Mean Squares	F Value	P
Modeling method	2	3.796	1.898	1652.32	<.0001
A1C variability	4	0.004	0.001	0.98	0.418
Interaction term	8	0.004	0.0006	0.49	0.861
Model	14	3.805	0.272	236.61	<.0001
Error	150	0.172	0.001		
Corrected Total	164	3.977			
Contrast: ML vs LR	1	0.595	0.595	119.03	<.0001

In the neuropathy cohort, two-way ANOVA indicated significant interaction between modeling method and A1C variability (**Table 19**).

Table 19. Two-way ANOVA testing the effect of modeling method and A1C variability on F1 scores of the neuropathy cohort (n=165)

Source	DF	Sum of Squares	Mean Squares	F Value	P
Modeling method	2	1.119	0.559	315.79	<.0001
A1C variability	4	1.119	0.559	315.79	<.0001
Interaction term	8	0.036	0.004	2.58	0.011
Model	14	1.178	0.084	47.50	<.0001
Error	150	0.266	0.002		
Corrected Total	164	1.444			
Contrast: ML vs LR	1	1.118	1.118	631.20	<.0001

An examination of the interaction indicates that the interaction was between the modeling methods of SVM and NN (**Figure 18**).

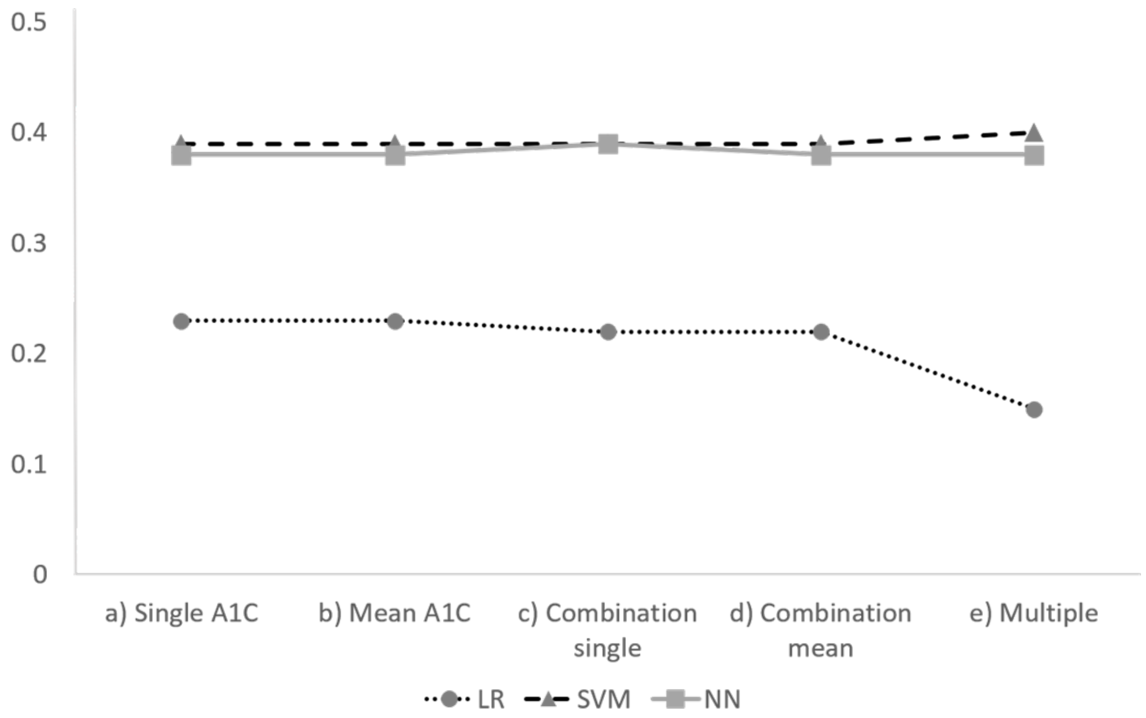


Figure 18. Interaction plot for F1 scores of the neuropathy cohort by microvascular complication and A1C variability

F test indicates that A1C variability had significant effect on F1 score of the neuropathy cohort when the modeling method was LR: $F=8.19$, $p<.0001$. Post hoc Tukey-Kramer test indicates that mean F1 score of the neuropathy cohort from LR models using e) multiple was significantly lower than using other A1C variability measures.

In CHAPTER 6, the results of the study, its implications, strengths, and limitations will be discussed and future research will be recommended.

CHAPTER 6

Discussion, Recommendation, and Conclusions

Discussion

The objectives of this research were to 1) develop predictive models for diabetic nephropathy, retinopathy and neuropathy using conventional LR and two ML methods (SVM and NN) and compare their performance based on F1 score; and 2) evaluate whether ML methods differ from LR in utilizing A1C variability for the prediction of diabetic nephropathy, retinopathy, and neuropathy. This chapter begins with a discussion of the key findings and their implications. Following that, the strengths and limitations of the study as well as recommendation for future research will be discussed. The chapter will end with conclusions.

The study found that mean F1 scores (0.38) of ML models were significantly higher than that of conventional LR models (0.19) across predicted outcomes. Specifically, two-way ANOVA test indicated that SVM and NN models produced higher F1 scores than LR models in predicting diabetic nephropathy, retinopathy, and neuropathy. F1 scores of SVM models and NN models did not differ significantly. The difference in F1 score between ML models and LRs was larger in the cohort of retinopathy and this may be due to worse performance of LR in this cohort.

When different predictor sets were used for prediction, more specifically, when different levels of A1C variability was employed while keeping other covariates the same, no significant difference in F1 score was found between ML and LR methods in utilizing A1C variability for prediction. In the cohort of diabetic nephropathy, F test indicates that A1C variability had significant effect on F1 score when the modeling method was NN. Post hoc Tukey-Kramer test indicates that mean F1 scores of the nephropathy cohort from NN models using d) combination mean or e) multiple were significantly higher than using b) mean A1C or c) combination single. There is no effect of A1C variability on F1 scores of the retinopathy cohort. In the cohort of neuropathy, F test indicates that A1C variability had significant effect on F1 score when the modeling method was LR. Post hoc Tukey-Kramer test indicates that mean F1 score of the neuropathy cohort from LR models using e) multiple was significantly lower than using other A1C variability measures.

This was among the first studies that developed predictive models for microvascular complications in T1D patients with a specific focus on improving the F1 score. This has important clinical implications because F1 scores better represented the model's capability in identifying less represented level of a class – patients who were at risk for a disease. If applied, these models can help clinicians, hospitals and managed care organizations capture high-risk patients and interventions can be followed based on predicted risk scores. Previous predictive models for microvascular complications among T1D patients were based on accuracy or AUC, which were not good indicators for the model's ability in identifying high-risk patients.

This study explicitly compared the performance between ML models and LR models and indicated ML models performed significantly better even when using the same predictor set. Between the two ML methods, SVM and NN, no significant difference was found in their performance across cohorts. The F1 scores of LR models were rather low: between 0.07 and 0.24. ML models improved the average F1 scores to an average of 035-0.39, although still not satisfactory. However, the sensitivity of predictive models increased from the average of 0.04-0.14 of LR models to 0.63-0.72 of SVM models and 0.66-0.75 of NN models. This suggested that conventional LR models could only identify a maximum of 14% of patients who were at risk, whereas ML models improved this number to above 70%. There is much room to improve in terms of precision, with the average precision scores of ML models below 0.3. However, laboratory tests are readily available for T1D patients and those who were suspected can be easily confirmed by a follow-up lab test.

When the last 3 A1C values were considered in the models, the prediction of diabetic nephropathy was significantly improved using the modeling method of NN. Hence, the optimal predictive model for a certain disease would depend on both the predictors used and the modeling method. LR models indicated that A1C variability, more specifically, SD-A1C was significantly associated with diabetic nephropathy and neuropathy among T1D patients. LR models in this study did not indicate a significant association between A1C variability and diabetic retinopathy. There was inconsistent evidence of the association between A1C variability and microvascular complications among T1D patients. Some suggested positive associations (Gorst et al., 2015) whereas other indicate non-association (Lachin et al., 2017). This research would add to current evidence and future research is needed to confirm the relationship. In future, HCPs can record patients' last 3 A1C values and calculate their standard deviations for risk evaluation. Algorithms can be developed to better understand both the magnitude and direction of A1C variability in order to assess how A1C variability affect each microvascular complication in T1D patients.

This study only included patients who had at least 3 A1C values in order to evaluate A1C variability. This may introduce sampling bias as patients who get tested more frequently may take better care of their health and be healthier than those who do not. This may limit the applicability of the developed predictive models to patients who are tested more frequently. It was reported that patients who did not take frequent retest for A1C achieved worse A1C control (Driskell et al., 2014). A study assessed daily blood glucose monitor frequency and glucose control and indicated significant racial health disparity in adolescent patients with T1D (Chalew et al., 2018). Specifically, black T1D patients with less social advantage were less likely to take blood test regularly and manage their blood glucose well. Since our study sample is mainly composed of white patients who had relatively higher social economic status and participated in the T1D exchange clinic registry, the predictive models may not produce accurate prediction in less advantaged patient population and other races such as black.

Predictors selected for each microvascular complication was similar to those used in previous studies (Aspelund et al., 2011; Kazemi et al., 2016; Lagani et al., 2015; Ravizza et al., 2019; Skevofilakas et al., 2010; Vergouwe et al., 2010). This research took into account the different insulin regimens used by patients and did not find significant association between insulin types and microvascular complications. The overall astounding costs for insulin has been heavily debated and there were research suggesting the use of the less expensive intermediate acting Neutral Protamine Hagedorn (NPH) insulins instead of insulin analogs such as detemir and glargine (Cefalu et al., 2018; Lipska, Hirsch, & Riddle, 2017; Luo, Avorn, & Kesselheim, 2015). However, as most T1D patients in this study were using insulin analogs, the effect of insulin NPH insulin on microvascular complications among T1D patients cannot be determined. This research did not find use of other medications for blood glucose control had a significant effect on any microvascular complication among T1D patients, either.

Managing T1D is expensive and two fifths of the costs were reported to be related to managing T1D complications (Joish et al., 2020). Predictive models can serve as a useful tool for healthcare providers and clinicians. For diabetic retinopathy, it usually takes a long time and multiple ophthalmic photography images to confirm its diagnosis. Early prediction can help optometrists consider closer monitoring and preventive interventions for at-risk patients. Many patients with diabetic neuropathy have no symptoms in its early stages and are left undiagnosed until it’s too late. Predictive models with a high capability in identifying patients at risk can enable general doctors to refer patients at risk to neurologists earlier and take proactive interventions. Future research is needed to facilitate the clinical application of predictive models. Moreover, prescriptive analytics should accompany the research in predictive modelling. Prescriptive analytics can tell us what to do once a prediction is made and a diagnosis is confirmed.(Abid, Keshavjee, Karim, & Guergachi, 2017; Islam, Hasan, Wang, Germack, & Noor-E-Alam, 2018). Prescriptive analytics can supplement predictive modelling by guiding clinicians to choose between difference courses of actions given a patient's risk.

Strengths & Limitations

Strengths

This study was among the first to target F1 score for risk prediction of diabetic nephropathy, retinopathy, and neuropathy. We aimed to focus on F1 score because F1 score reflected the model's ability to identify patients who were at risk. The study was also among the first that explicitly compared the performance of different modeling methods and different predictor sets, and tested the difference using statistical tests. Whereas previous studies only provided results of one performance metric and did not make statistical inference of the performance between different modeling methods. The prediction models were based on data from the largest registry of T1D patients in the United States. Patients were residing across vast areas of the U.S. They have been receiving standard care from participating hospitals and clinics and followed by the registry once a year. The registry data were updated by information from the participant questionnaire as well as from their electronic medical records and are well-documented in the registry. Thus, the database provides the foundation of valid prediction of long-term microvascular complications. In addition to cross-sectional measures, the dataset also captures longitudinal measures of A1C levels for each individual patient. This permits us to include the variability of A1C.

Limitations

Methodological limitations: Due to lack of time to event measures in our data, we cannot predict time to progression to microvascular complications in this study. As we tried to compare performance of different modelling methods, ensemble of multiple modelling methods were not attempted (Geron, 2017). 'Ensemble learning' refers to the method of aggregating two or more ML algorithms to build even more complex models (Geron, 2017). Ensemble or stacking of ML algorithms will assign a weight to multiple algorithms and yield a weighted average of their outputs. The predictive models of SVM and NN were difficult to interpret. And we were unable to test the predictive models on an external dataset.

Database limitations: As majority of our patients were White, our study cannot be used to identify ethnicity risk groups for any of the three microvascular complications. Nor can we identify risk factors for the three types of microvascular complications among other ethnicity groups due to lack of data. The diagnosis of peripheral neuropathy may be significantly higher than recorded in the database because on one hand, individuals with symptoms of neuropathy may not all have been tested for the disease and on the other hand, current tests for neuropathy may not cover all forms of the disease and the complexity as well as variability of neuropathy symptoms may lead to under- or mis- diagnosis (*Peripheral neuropathy fact sheet*, 2018). Due to limitation of the registry database, we could not obtain information of time in range (TIR) among patients who have been using CGM. Other lifestyle factors of patients such as alcohol use and safe drinking (Viswanathan, 2015), which can contribute to diabetic neuropathy was not captured. Future research is needed to incorporate TIR into the models and see how it impacts the prediction performance of the models. However, we need to bear in mind the possible false positives reported by CGM for prediction of hypoglycemia because the data reported by CGM is also based on mathematical algorithms that are not 'true' patient data (Cichosz et al., 2015).

In CHAPTER 7, a summary of this research will be provided.

CHAPTER 7

Summary

This study used a factorial experimental design that employed real-world registry data to develop predictive models for three types of microvascular complications in T1D patients. Three factors, i.e., modelling method, microvascular complication, and A1C variability were manipulated and their effect on performance measure was evaluated. Specifically, modelling method was operationalized as two levels, conventional statistical method (LR) and ML methods, which are further manipulated into two levels, SVM and NN. Microvascular complication was manipulated as three levels, i.e., diabetic nephropathy, retinopathy, and neuropathy. A1C variability was manipulated as five levels, i.e., a) single A1C, b) mean A1C, c) combination single, d) combination mean, and e) multiple. Performance measure was operationalized as F1 score. A total of 495 models were developed and their performance in terms of F1 score compared.

Factorial ANOVA indicates that ML methods (SVM and NN) performed significantly better than conventional statistical method (LR) irrespective of microvascular complication or A1C variability. There is minor interaction between the two ML methods, i.e., SVM and NN. In other words, SVM and NN had different effect within different levels of microvascular complication and A1C variability. However, the interaction was deemed not important and their performance in terms of mean F1 score was not significantly different.

There is significant difference in model performance for predicting diabetic nephropathy in T1D patients when using different A1C variability measures under the modeling method of NN. However, A1C variability does not have a significant effect for the prediction of diabetic retinopathy or neuropathy, no matter what modeling method was used.

This study provides much needed empirical data on the comparison between ML and conventional statistical methods and implies that ML methods are superior to conventional statistical method in this study and should be used for prediction. NN models were found to utilize A1C variability better for predicting diabetic nephropathy. The last 3 A1C measures of a patient may be considered by clinicians for managing their T1D patients, especially for preventing diabetic nephropathy. Future research is needed to develop algorithms to better calculate A1C variability to monitor T1D progression.

This study focused on predicting microvascular complications in adult T1D patients. Future research may apply predictive models to pediatric population, type 2 diabetes, and other disease areas. Future research is also needed to develop decision support systems that can advise clinicians based on the results from predictive models.

APPENDICES

Appendix 1. Summary of commonly used insulin and its analogues in the United States

Type of Insulin	Generic Name (Brand Name, Company & Year of Initial FDA Approval)	Onset	Duration of Action
Rapid acting	• Lispro (Humalog®, Eli Lilly, 1996; Admelog®, Sanofi-Aventis, 2017) • Aspart (Novolog®, Novo Nordisk, 2000; Fiasp®, Novo Nordisk, 2017) • Glulisine (Apidra®, Aventis, 2004)	10-30 minutes	1-5 hours
Short acting	• Insulin human or regular (R) (Humulin® R, Eli Lilly, 1982; Novolin® R, Novo Nordisk, 1991)	30-60 minutes	5-8 hours
Intermediate acting	• NPH (N) or isophane insulin (Humulin® N, Eli Lilly, 1982; Novolin® N, Novo Nordisk, 1991)	1-3 hours	18-24 hours
Long acting	• Glargine (Lantus® and Lantus® SoloStar®, Sanofi-aventis, 2000; Toujeo® SoloStar®, Sanofi-Aventis, 2015; and Basaglar® KwikPen®, Eli Lilly, 2000) • Detemir (Levemir®, Novo Nordisk, 2005) • Degludec (Tresiba®, Novo Nordisk, 2015)	1-2 hours	20-24 hours (glargine, detemir)
Pre-mixed* or combinations	• NPH and human insulin (Humulin® 70/30 & Humulin® 50/50, Eli Lilly, 1989; Novolin® 70/30, Novo Nordisk, 1991) • Lispro protamine and Lispro: intermediate-rapid insulin mixture: similar to mixing NPH & lispro (Humalog® Mix 75/25 and Humalog® Mix 50/50, Eli Lilly, 1999) • Aspart protamine suspension and insulin aspart (Novolog® Mix 70/30 and Novolog® Mix 50/50, Novo Nordisk, 2001) • Degludec and aspart: long-rapid insulin mixture (Ryzodeg® 70/30, Novo Nordisk, 2015)	10-30 minutes	14-24 hours
*Pre-mixed insulins combine specific amounts of intermediate-acting and short-acting insulin in one bottle or insulin pen. The numbers following the drug brand name indicate the percentage of each type of insulin.			

Appendix 2. Definition of "definite T1D"

"Definite T1D" was assessed by the registry and all the participants in the registry are already confirmed with definite T1D. Participants need to meet one of the following criteria to be classified as having definite Type 1 diabetes:

1. Age less than 10 years at diagnosis;
2. Positive pancreatic autoantibodies at any time (GAD-65, IA-2, ICA, or ZnT8) or positive anti-insulin autoantibody at diagnosis only (within 10 days of starting insulin); or
3. The presence of two or more of the following clinical indicators suggestive of T1D:
   - Age at diagnosis less than 40 years;
   - Non-obese at diagnosis according to body mass index (<95th percentile pediatric and <30 kg/m2 adult);
   - Diabetic ketoacidosis (DKA) at any time;
   - Plasma C-peptide level below 0.8 ng/ml (with blood glucose   80 mg/dl if available) at any time; and
   - Family history of T1D in a first-degree relative (parent, sibling, or child).

Appendix 3. Operational definition of study measures

<table>
<tr><th>STUDY MEASURE</th><th>OPERATIONAL DEFINITION</th></tr>
<tr><td colspan="2" align="center">Outcomes</td></tr>
<tr><td>Nephropathy (yes/no)</td><td>Defined as “yes”
if in the subject file (of either follow-up dataset), the patient indicated having any of the following conditions:
• Albuminuria or macroalbuminuria (Albuminuria/macroalbuminuria is defined as 2 consecutive ACRs >300 mcg/mg or 2 out of the past 3.)
• A glomerular filtration rate <60 ml/min
AND the participants did not indicate if the renal disease was believed to be solely due to a cause other than diabetes.
Or if the medical condition file (of either follow-up dataset) indicated any of the following MedDRA condition for the patient:
• Chronic kidney disease
• Diabetic nephropathy
• Protein urine present
• Proteinuria
Note: Patients with a MedDRA condition of acute renal failure were excluded from the analysis because it can be caused by an injury of kidney other than the progression of diabetes; MedDRA conditions of chronic renal failure, end stage renal disease (ESRD), kidney failure, kidney transplant, renal failure, or renal insufficiency were excluded as well, because these conditions usually take a long time (over 20 years) to develop. Hence, patient who had a diagnosis of any of these conditions should usually already have had diabetic nephropathy at baseline and were excluded. Definitions of kidney failure or ESRD can be found at the K/DOQI Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, classification, and stratification (available at https://www.kidney.org/sites/default/files/docs/ckd_evaluation_classification_stratification.pdf, last accessed 07/27/2019). Patients with a history of kidney function abnormal, renal disease or renal impairment were not considered as indication for diabetic nephropathy, as they may be caused by a reason other than diabetes.</td></tr>
<tr><td>Retinopathy (yes/no)</td><td>Defined as “yes”
if in the subject file (of either follow-up dataset), the patient indicated having any of the following conditions:
• Diabetic macular edema
• Vitreous hemorrhage
• Non-proliferative retinopathy
• Proliferative retinopathy
or receiving any of the following eye treatment:
• Intravitreal injection (such as Lucentis, Avastin, Macugen, or Triamcinolone)
• Vitrectomy
• Other treatment (laser treatment to correct nearsightedness or farsightedness is not included)
Or if the medical condition file (of either follow-up dataset) indicated any of the following MedDRA condition for the patient:
• Diabetic macular edema</td></tr>
</table>

	• Diabetic retinopathy • Macular edema • Non-proliferative diabetic retinopathy • Preproliferative diabetic retinopathy • Proliferative diabetic retinopathy Note: Patients with blindness anytime during the study were excluded from the analysis. Operational definition of diabetic retinopathy was discussed with and confirmed by three clinicians in optometry (Drs. Carolyn R. Carman, Jennifer Tasca, and Joe L. Wheat from University of Houston College of Optometry).
Neuropathy (yes/no)	Defined as "yes" *if in the subject file (of either follow-up dataset), the patient indicated having any of the following conditions:* • Diabetic peripheral neuropathy • Autonomic neuropathy • Gastroparesis *Or if the medical condition file (of either follow-up dataset) indicated any of the following MedDRA condition for the patient:* • Peripheral neuropathy NOS • Neuropathy • Neurogenic bladder • Gastroparesis • Diabetic peripheral neuropathy • Diabetic neuropathy • Diabetic mononeuropathy • Diabetic polyneuropathy (data does not have this indication) • Diabetic gastroparesis • Charcot's joint • Charcot arthropathy Note: Patients who reported to have gastroparesis or who had a history of gastroparesis, peripheral neuropathy NOS, neuropathy, neurogenic bladder, gastroparesis, diabetic gastroparesis, Charcot's joint, or Charcot arthropathy are considered to have diabetic neuropathy during follow-up because these conditions are usually caused by diabetic neuropathy in diabetic patients. This is based on the assumption that these patients did not have any type of these conditions at baseline. But we cannot exclude that patients may already have had these conditions but was not diagnosed at baseline. Patients with a history of numbness in hand, numbness generalized, leg amputation, foot ulcer, foot amputation, diabetic ulcer NOS, diabetic foot ulcer, erectile dysfunction or arm amputation were not considered as indications of diabetic neuropathy because these conditions are more likely to be caused by vascular conditions other than diabetic neuropathy.
Predictors	
Individual Characteristics	
A1C (%)	• Single A1C value: most recent A1C value recorded in clinic chart • A1C variability: For each patient, a total of 3 A1C values that were closest to the consent date were included into the analysis. Each of these A1C values had to be at least 3 months apart. If a patient had multiple A1C values that were measured within the "3-month gap",

	then the mean of these multiple A1C values were calculated to impute the point value and the first date that went beyond the 3-month gap was used to impute the point "HbA1cMonthsFromConsent" value. o Mean-A1C: The average of the most recent (including some "imputed") 3 A1C values was calculated as the mean A1C value for a patient. o SD-A1C: For LR, as correlated A1C values cannot be incorporated into the model directly, A1C variability was defined as standard deviation (SD) of the most recent (including some "imputed") three A1C values that were measured at least three months apart. o CV-A1C: coefficient of variation of A1C, calculated as SD-A1C divided by mean-A1C
Age at baseline	• In years; Defined as age at consent indicated in subject file • Age categories: 18-27 years; 28-37 years; 38-47 years; 48-64 years; ⩾ 65 years based on age distribution
Duration of T1D	In years; Calculated as the difference between age at consent and age of T1D diagnosis indicated in subject file
Gender	Male or female indicated in subject file
Race	Categorized as 1) White non-Hispanic, 2) Black/African American, 3) Hispanic or Latino, or 4) others (including native Hawaiian/other Pacific Islanders, Asian, American Indian/Alaskan native, or more than one race/ethnicity) indicated in subject file
Education level	Categorized as 1) less than bachelor's degree; 2) bachelor's degree; and 3) master's, professional, or doctorate
Insurance coverage	Categorized as 1) Commercial health insurance: private or single service insurance plan; 2) Government-sponsored insurance (Medicare, Medicaid, SCHIP, state, military, Indian, or other government insurance); and 3) others: not indicated as any insurance type above; When incorporated into predictive models, it was divided into two categories: commercial vs others.
Marital status	Categorized as 1) married or living together; or 2) divorced, separated, single (never married), or widowed indicated in subject file.
Annual household income	• Categorized as 1) <$50K, 2) $50K to $100K, or 3) >=$100K as indicated in subject file (self-reported). • When incorporated into predictive models, it was divided into two categories: >=100K vs <100K.
Employment status	• Categorized as 1) Working full time or part-time at baseline; 2) Student or homemaker; or 3) Unemployed, retired, disabled or other • When incorporated into predictive models, it was dummy coded into two variables using 'Working full time or part time' as the reference group.
Body mass index (BMI)	• In kg/m^2; Calculated by clinic chart indication of weight in kilograms divided by height in meters squared. (Height unavailable at baseline was imputed from follow-up datasets when available) • BMI category: 1) under or normal weight, 2) overweight, 3) obese
Blood pressure (mmHg)	Blood pressure indicated in clinic chart: • Systolic blood pressure (SBP) • Diastolic blood pressure (DBP)

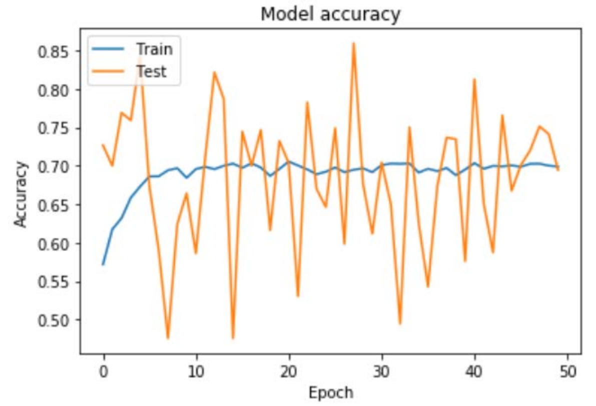
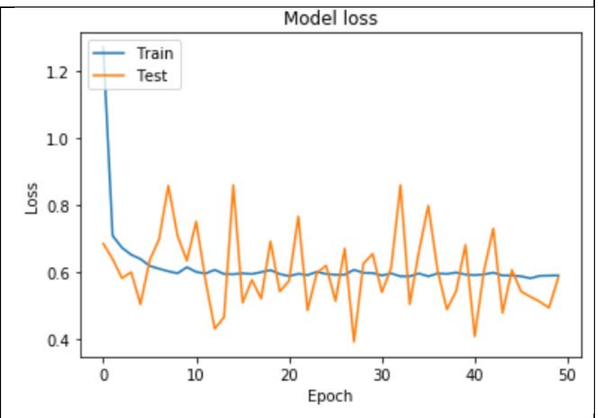
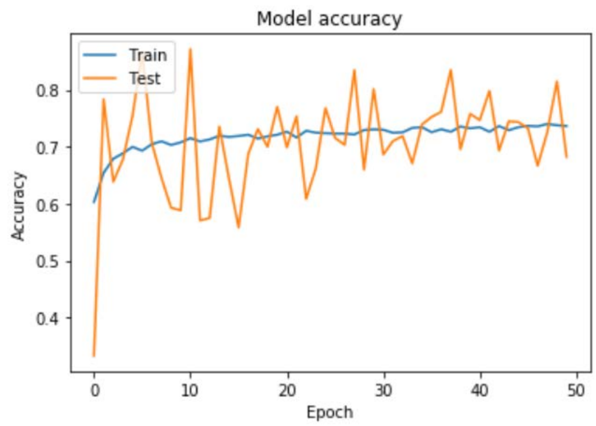
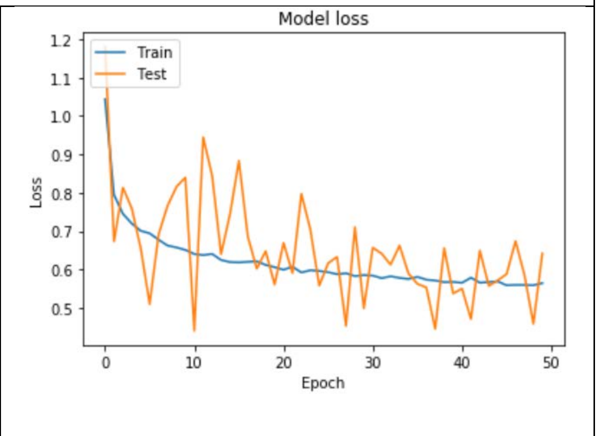
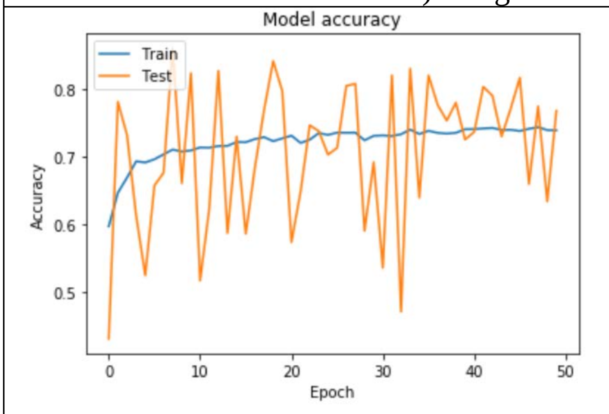
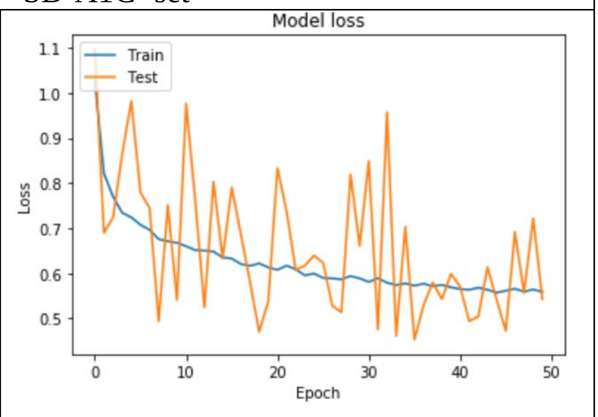
Cholesterol level (mg/dL)	• Low-density lipoprotein (LDL) • High-density lipoprotein (HDL) • Triglyceride • Lipid fasting status (fasting, not fasting, unknown)
Health Behavioural Factors	
Smoking status (yes/no)	• Ever smoked: defined as “yes” if a participant reported to have smoked at least 100 cigarettes (100 cigarettes = 5 packs) in his/her entire life • Smoking at baseline: defined as “yes” if a participant reported to have smoked at least 100 cigarettes (100 cigarettes = 5 packs) in his/her entire life and did not indicate that he/she did not smoke at all anymore
Insulin used at baseline	• Insulin delivery method: 1) Pump only, 2) injections (MDI/basal-bolus or fixed dose) only, or 3) both pump and injections as indicated in subject file. • Name/Type of insulin: insulin lispro (Humalog), insulin aspart (Novolog), insulin detemir (Levemir), insulin glargine (Lantus)
Use of CGM (yes/no)	At the most recent visit or sometime within the 30 days before the visit, was the participant using a continuous glucose monitor (CGM), for real-time, unblinded diabetes management: Yes or no as indicated in subject file.
Use of other medications for blood glucose control (yes/no)	Defined as ‘yes’ if participant reported or indicated in HER the use of other medications for blood glucose control, including dipeptidyl peptidase-4 (DPP4) Inhibitors, glucagon-like peptide-1 (GLP1) agonists, metformin, pramlintide or other medications.
Use of ACE inhibitors or ARBs (yes/no)	• Use of angiotensin-converting enzyme (ACE) inhibitors (yes/no): Use of any of the following medications as indicated in medication file at baseline: o benazepril or benazepril hydrochloride o captopril o enalapril o fosinopril or fosinopril / hydrochlorothiazide o lisinopril o moexipril o perindopril o quinapril o ramipril o trandolapril • Use of angiotensin II receptor blocker (ARBs) (yes/no): Use of any of the following medications as indicated in medication file at baseline: o Candesartan o Eprosartan o Irbesartan o Losartan o Olmesartan o Telmisartan o Valsartan • Use of either ACE inhibitors or ARBs (yes/no): Use of any of the above ACE inhibitors or ARBs as indicated in the medication file or reported by participants questionnaire.

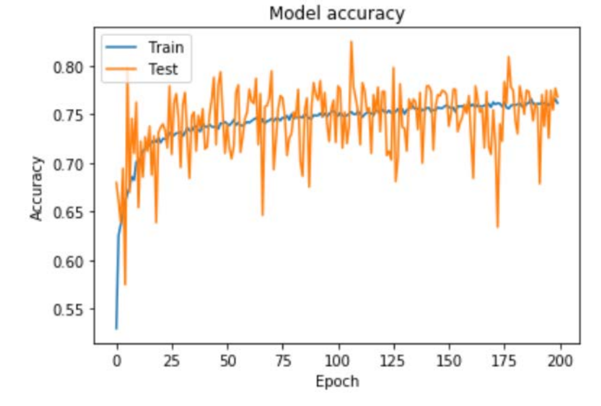
	Note: Use of ACE inhibitors or ARBs may not add up to use of either ACE inhibitors or ARBs because the former two was defined using information contained in the medication file only whereas the latter was defined using information from both the medication file and participant questionnaire.
Comorbidities at Baseline	
Microalbuminuria (yes/no)	Defined as "yes" if a patient indicated an albumin status of microalbuminuria (Microalbuminuria is defined as 2 consecutive albumin/creatinine ratios in the range of 30-300 mcg/mg or 2 out of the past 3 measurements.)
Diabetic microvascular complication	• Diabetic nephropathy: operationalized by measures from the participant questionnaire; Defined as "yes" if a patient had an albumin status of albuminuria/macroalbuminuria, or had a GFR that was below 60, or had a diagnosis of renal failure, or a diagnosis of nephropathy due to other causes, or had received ACE or ARB for diabetic nephropathy. • Diabetic retinopathy: operationalized by measures from the participant questionnaire; Defined as "yes" if a patient was legally blind or had received ACE or ARB for diabetic retinopathy, or had received any treatment for diabetic retinopathy, or had received cataract surgery, or had received surgery for glaucoma. • Diabetic neuropathy: operationalized by measures from both the participant questionnaire and the medical condition file; Defined as "yes" if a patient reported that foot ulcer was present, or had a history of and history of amputation of toe or knee, erectile or sex dysfunction, diabetic neuropathy, Charcot joint, orthostatic hypotension, tachycardia, or gastroparesis from the medical condition file.
Cardiovascular conditions	• Hypertension (yes/no): defined as “yes” if the medical condition file at baseline indicated MedDRA condition of hypertension • Dyslipidemia (yes/no): defined as “yes” if the medical condition file at baseline indicated any of the following MedDRA conditions: o High Triglycerides o Dyslipidemia o Dyslipidemia unspecified o High LDL o Low HDL • CAD (coronary artery diseases, yes/no): defined as “yes” if the medical condition file at baseline indicated any of the following procedures: o Coronary artery bypass graft o Coronary artery angioplasty Or any of the following MedDRA conditions: o Myocardial infarction (MI, heart attack) o Coronary artery disease, without myocardial infarction o Cardiomyopathy o Congestive heart failure • PVD (Peripheral vascular disease): defined as “yes” if the medical condition file at baseline indicated any of the following MedDRA conditions: o Peripheral vascular disease o Peripheral vascular claudication

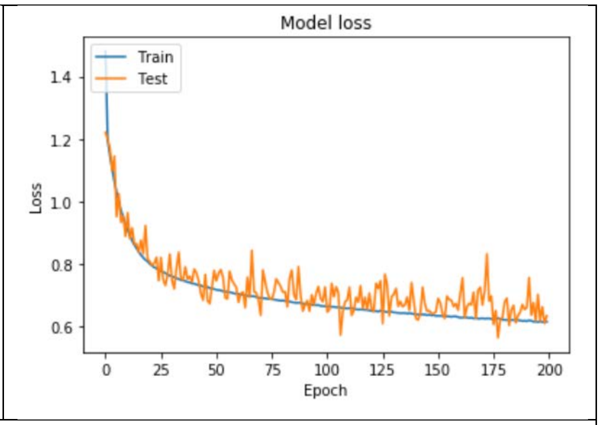
	o Amputation of knee or toe • Cardiac arrhythmia (yes/no): defined as “yes” if the medical condition file at baseline indicated any of the following MedDRA conditions: o Atrial fibrillation o Other cardiac arrhythmia o Cardiac pacemaker • CVA (cerebrovascular accident, yes/no): defined as “yes” if the medical condition file at baseline indicated any of the following MedDRA conditions: o Stroke o Transient ischemic attack (TIA)
Endocrine diseases	• Hypothyroidism (yes/no): defined as “yes” if the medical condition file at baseline indicated any of the following MedDRA conditions: o Hypothyroid o Hashimoto’s disease • Hyperthyroidism (yes/no): defined as “yes” if the medical condition file at baseline indicated any of the following MedDRA conditions: o Hyperthyroid o Grave’s disease • Other endocrine diseases (yes/no): defined as “yes” if the medical condition file at baseline indicated any of the following MedDRA conditions: o Autoimmune adrenal disease (Addison’s disease) o Autoimmune polyendocrine syndrome (type 2) or Schmidt’s syndrome o Polycystic ovarian syndrome (PCOS)
Gastrointestinal diseases (yes/no)	• Defined as “yes” if the medical condition file at baseline indicated any of the following MedDRA conditions: o Celiac disease o Vitamin B12 deficiency/pernicious anemia o Inflammatory bowel disease (IBD, Ulcerative colitis, Crohn’s Disease)
Musculoskeletal/Connective Tissue conditions	• Arthritis (yes/no): defined as “yes” if the medical condition file at baseline indicated any of the following MedDRA conditions: o Rheumatoid arthritis o Osteoporosis/osteopenia • Lupus (yes/no): defined as “yes” if the medical condition file at baseline indicated MedDRA condition of Lupus • Sjogrens (yes/no): defined as “yes” if the medical condition file at baseline indicated MedDRA condition of Sjogrens • Dermatomyositis (yes/no): defined as “yes” if the medical condition file at baseline indicated MedDRA condition of dermatomyositis
Psychiatric conditions	• Depression (yes/no): defined as “yes” if the medical condition file at baseline indicated MedDRA condition of depression • Anxiety (yes/no): defined as “yes” if the medical condition file at baseline indicated MedDRA condition of anxiety • Psychosis (yes/no): defined as “yes” if the medical condition file at baseline indicated MedDRA condition of psychosis • ADHD (yes/no): defined as “yes” if the medical condition file at baseline indicated MedDRA condition of Attention Deficit Hyperactivity Disorder (ADHD)

	• Eating disorder (yes/no): defined as "yes" if the medical condition file at baseline indicated any of the following MedDRA conditions: o Bulimia o Anorexia o Bulimia and Anorexia o Binge eating o Eating disorder not otherwise specified (EDNOS) o Intentional omission/restriction of insulin for weight loss
Skin disorders (yes/no)	Defined as "yes" if the medical condition file at baseline indicated any of the following MedDRA conditions: • Vitiligo • Necrobiosis lipoidica diabeticorum (NLD) • Psoriasis • Alopecia areata

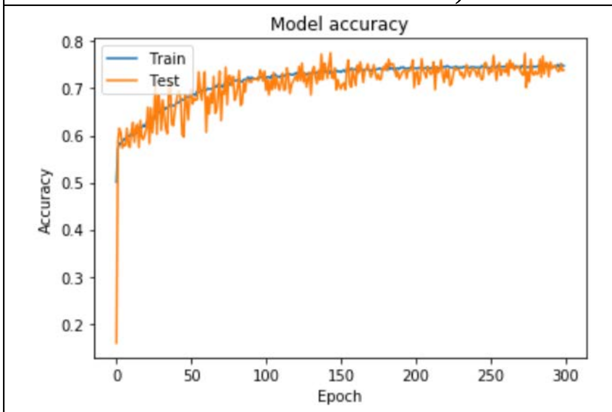
Appendix 4. Examples of accuracy and loss curves of the train and validation set using the 5 predictor sets A through E in cohorts of nephropathy, retinopathy and neuropathy

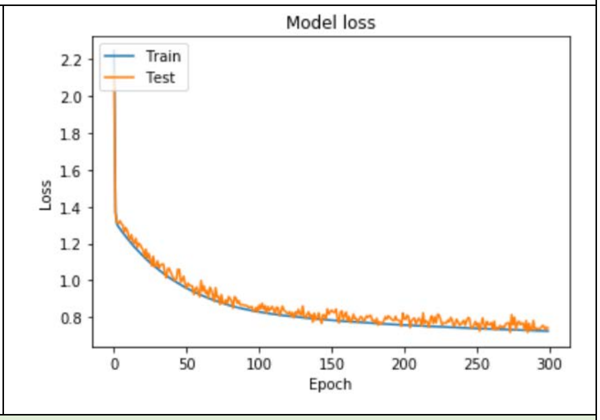
Cohort of Nephropathy	
a) 'Single A1C' set	
	
b) 'Mean-A1C' set	
	
c) 'Single A1C + SD-A1C' set	
	
d) 'Mean-A1C + SD-A1C' set	





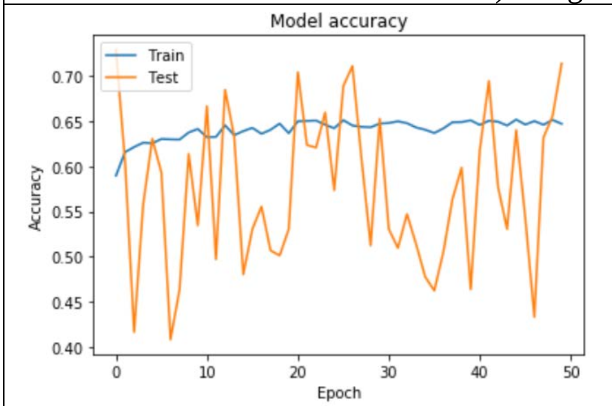
e) ‘3 A1C + SD-A1C’ set

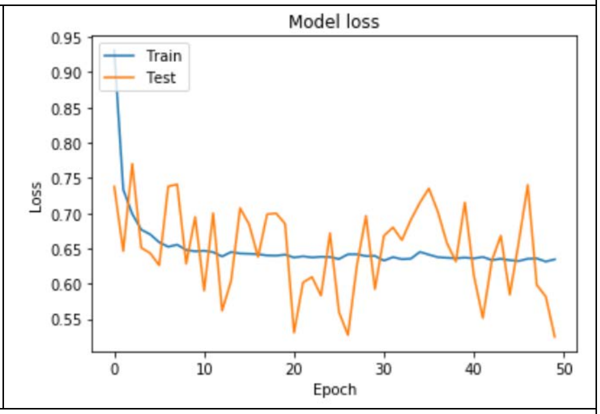




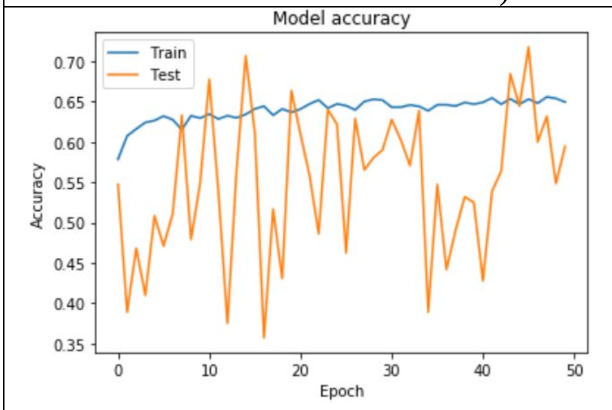
Cohort of Retinopathy

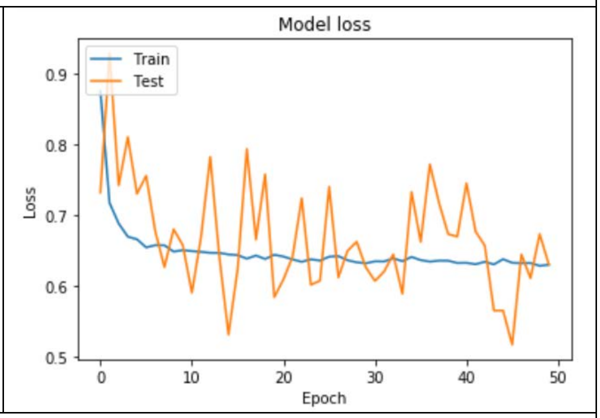
a) ‘Single A1C’ set



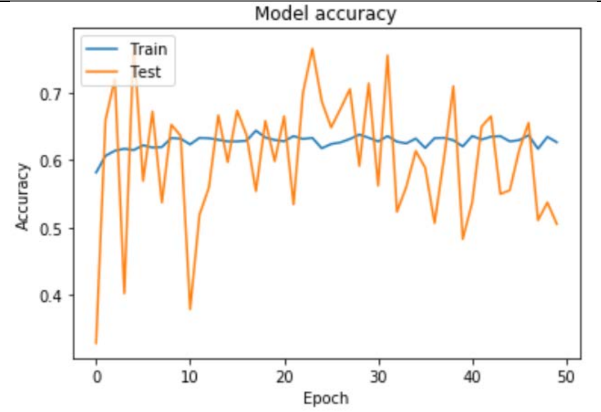


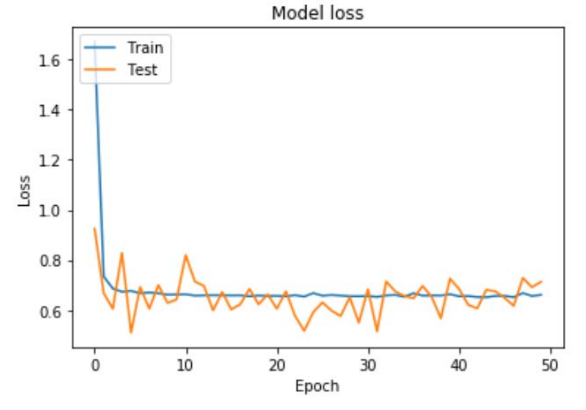
b) ‘Mean-A1C’ set



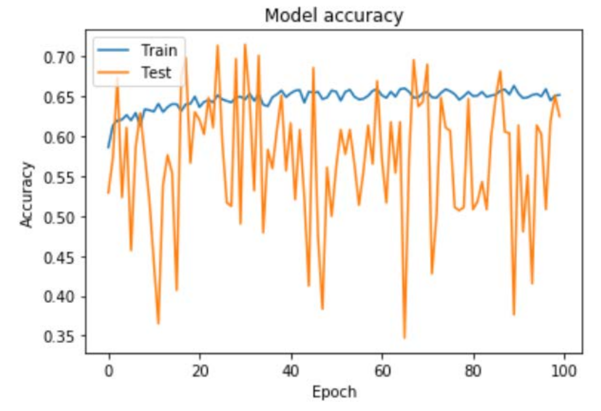


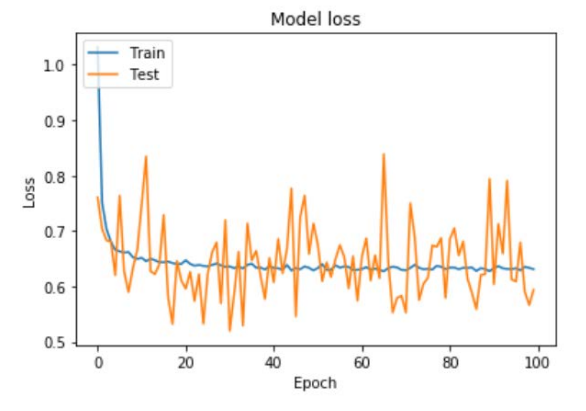
c) ‘Single A1C + SD-A1C’ set





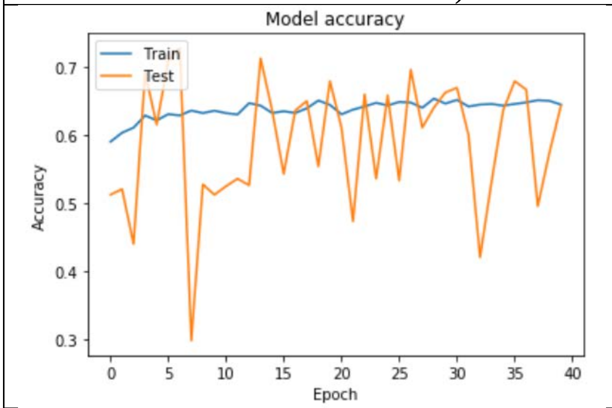
d) 'Mean-A1C + SD-A1C' set

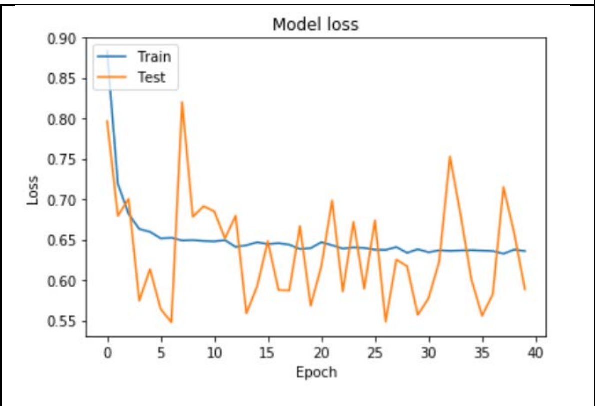




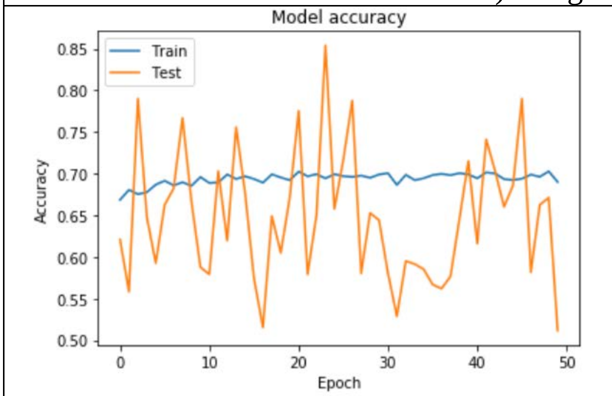
e) '3 A1C + SD-A1C' set

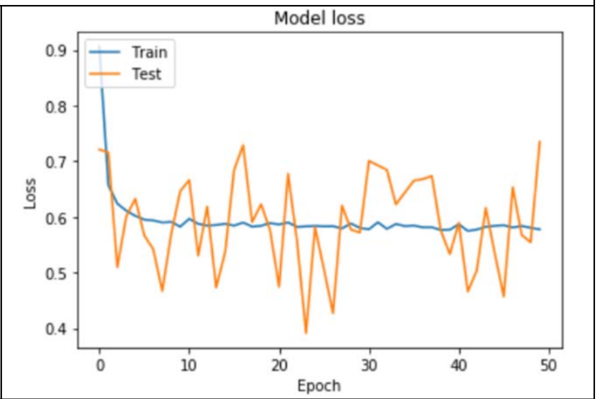
Cohort of Neuropathy



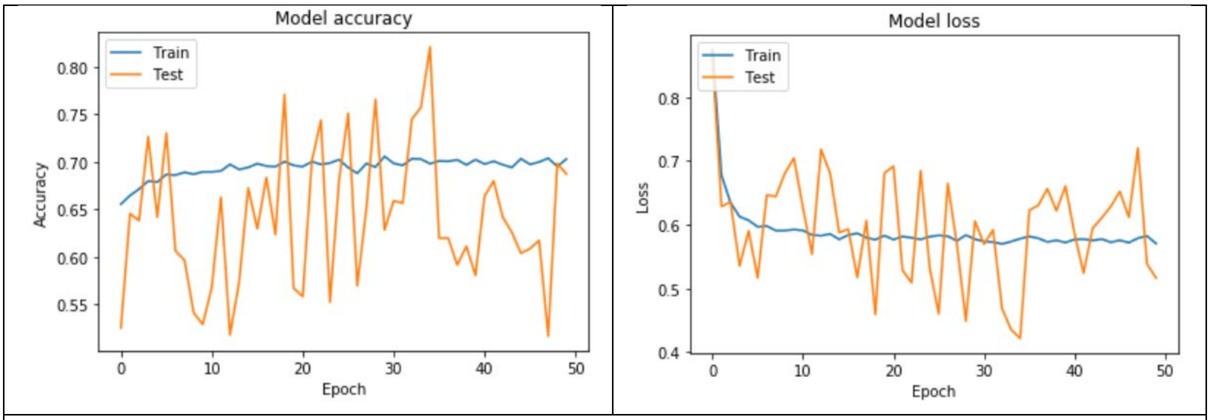


a) 'Single A1C' set

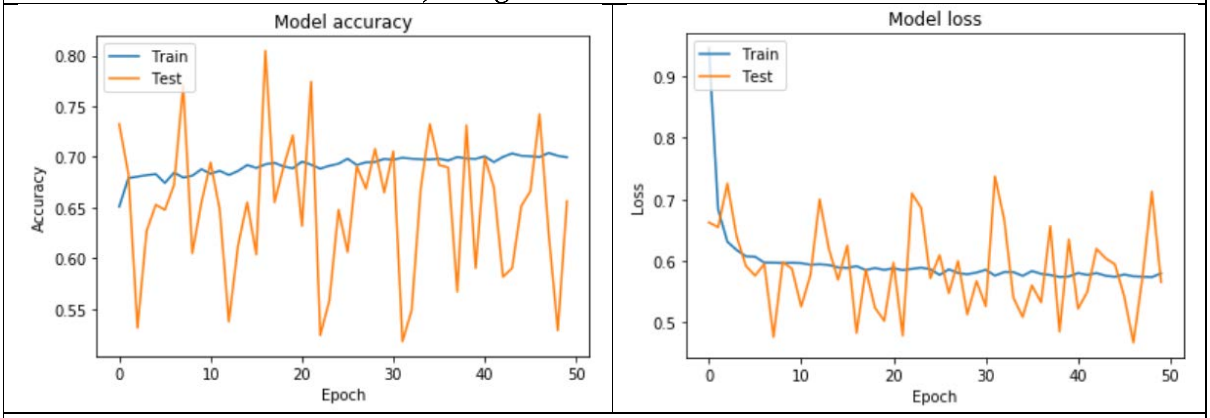




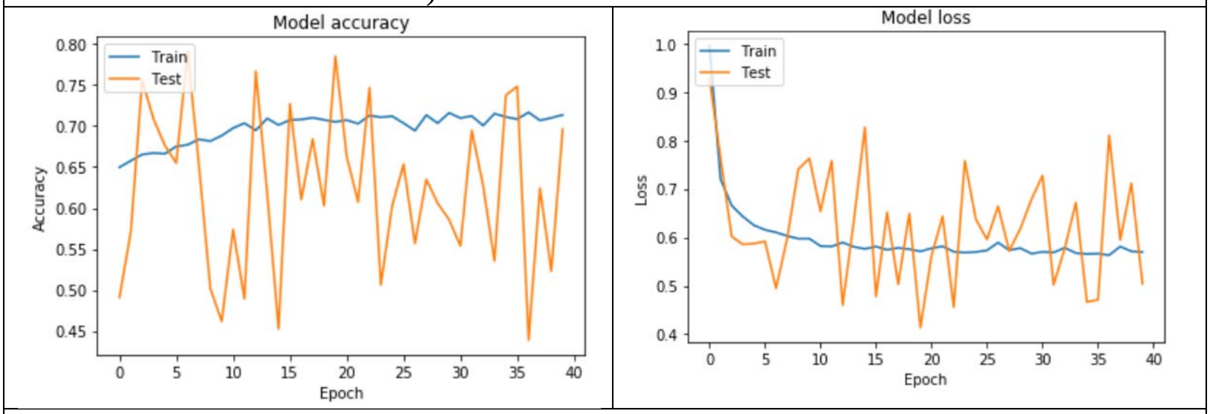
b) 'Mean-A1C' set



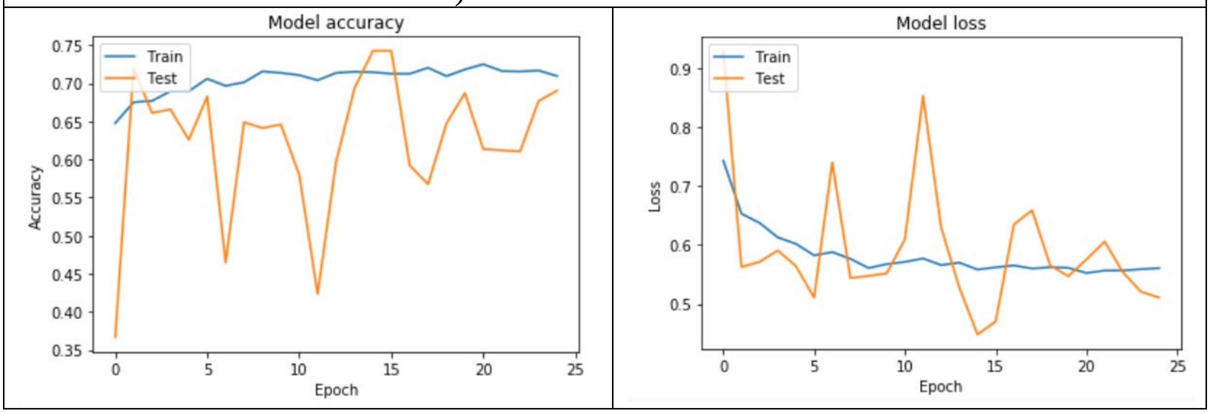
c) ‘Single A1C + SD-A1C’ set



d) ‘Mean-A1C + SD-A1C’ set



e) ‘3 A1C + SD-A1C’ set



Appendix 5. Performance metrics of predictive models of the nephropathy cohort

Obs	Modeling Method	A1C Variability	F1 score	Sensitivity	Precision	Accuracy	Specificity	AUC
1	LR	Single A1C	0.140	0.075	1.000	0.895	1.000	0.771
2	LR	Single A1C	0.178	0.098	1.000	0.896	1.000	0.756
3	LR	Single A1C	0.292	0.171	1.000	0.904	1.000	0.835
4	LR	Single A1C	0.392	0.244	1.000	0.912	1.000	0.819
5	LR	Single A1C	0.217	0.122	1.000	0.898	1.000	0.784
6	LR	Single A1C	0.178	0.098	1.000	0.895	1.000	0.802
7	LR	Single A1C	0.178	0.098	1.000	0.895	1.000	0.740
8	LR	Single A1C	0.392	0.244	1.000	0.913	1.000	0.757
9	LR	Single A1C	0.360	0.220	1.000	0.908	1.000	0.750
10	LR	Single A1C	0.182	0.100	1.000	0.899	1.000	0.731
11	LR	Single A1C	0.226	0.128	1.000	0.900	1.000	0.751
12	LR	Mean A1C	0.140	0.075	1.000	0.895	1.000	0.768
13	LR	Mean A1C	0.217	0.122	1.000	0.899	1.000	0.749
14	LR	Mean A1C	0.292	0.171	1.000	0.904	1.000	0.809
15	LR	Mean A1C	0.392	0.244	1.000	0.912	1.000	0.820
16	LR	Mean A1C	0.217	0.122	1.000	0.898	1.000	0.779
17	LR	Mean A1C	0.178	0.098	1.000	0.895	1.000	0.805
18	LR	Mean A1C	0.136	0.073	1.000	0.893	1.000	0.751
19	LR	Mean A1C	0.392	0.244	1.000	0.913	1.000	0.763
20	LR	Mean A1C	0.360	0.220	1.000	0.908	1.000	0.757
21	LR	Mean A1C	0.182	0.100	1.000	0.899	1.000	0.736
22	LR	Mean A1C	0.211	0.118	1.000	0.899	1.000	0.749
23	LR	Combination single	0.140	0.075	1.000	0.895	1.000	0.771
24	LR	Combination single	0.178	0.098	1.000	0.896	1.000	0.759
25	LR	Combination single	0.292	0.171	1.000	0.904	1.000	0.841
26	LR	Combination single	0.360	0.220	1.000	0.910	1.000	0.815
27	LR	Combination single	0.217	0.122	1.000	0.899	1.000	0.786
28	LR	Combination single	0.178	0.098	1.000	0.895	1.000	0.803
29	LR	Combination single	0.178	0.098	1.000	0.895	1.000	0.742
30	LR	Combination single	0.392	0.244	1.000	0.913	1.000	0.760
31	LR	Combination single	0.360	0.220	1.000	0.908	1.000	0.752
32	LR	Combination single	0.182	0.100	1.000	0.899	1.000	0.735
33	LR	Combination single	0.211	0.118	1.000	0.899	1.000	0.752

34	LR	Combination mean	0.140	0.075	1.000	0.895	1.000	0.769
35	LR	Combination mean	0.178	0.098	1.000	0.896	1.000	0.754
36	LR	Combination mean	0.292	0.171	1.000	0.904	1.000	0.816
37	LR	Combination mean	0.360	0.220	1.000	0.910	1.000	0.818
38	LR	Combination mean	0.255	0.146	1.000	0.901	1.000	0.781
39	LR	Combination mean	0.178	0.098	1.000	0.895	1.000	0.806
40	LR	Combination mean	0.178	0.098	1.000	0.895	1.000	0.749
41	LR	Combination mean	0.392	0.244	1.000	0.913	1.000	0.763
42	LR	Combination mean	0.360	0.220	1.000	0.908	1.000	0.757
43	LR	Combination mean	0.182	0.100	1.000	0.899	1.000	0.739
44	LR	Combination mean	0.211	0.118	1.000	0.899	1.000	0.750
45	LR	Multiple	0.182	0.100	1.000	0.898	1.000	0.743
46	LR	Multiple	0.093	0.049	1.000	0.890	1.000	0.786
47	LR	Multiple	0.292	0.171	1.000	0.904	1.000	0.799
48	LR	Multiple	0.360	0.220	1.000	0.909	1.000	0.760
49	LR	Multiple	0.136	0.073	1.000	0.893	1.000	0.771
50	LR	Multiple	0.000	0.000	0.000	0.000	0.000	0.000
51	LR	Multiple	0.136	0.073	1.000	0.892	1.000	0.723
52	LR	Multiple	0.360	0.220	1.000	0.910	1.000	0.715
53	LR	Multiple	0.423	0.268	1.000	0.914	1.000	0.741
54	LR	Multiple	0.095	0.050	1.000	0.893	1.000	0.715
55	LR	Multiple	0.179	0.098	1.000	0.897	1.000	0.738
56	SVM	Single A1C	0.331	0.600	0.229	0.729	0.745	0.673
57	SVM	Single A1C	0.340	0.625	0.234	0.729	0.742	0.684
58	SVM	Single A1C	0.350	0.585	0.250	0.751	0.773	0.679
59	SVM	Single A1C	0.362	0.610	0.258	0.754	0.773	0.691
60	SVM	Single A1C	0.377	0.634	0.268	0.760	0.776	0.705
61	SVM	Single A1C	0.336	0.585	0.235	0.735	0.754	0.670
62	SVM	Single A1C	0.290	0.512	0.202	0.712	0.738	0.625
63	SVM	Single A1C	0.405	0.805	0.270	0.729	0.719	0.762
64	SVM	Single A1C	0.348	0.683	0.233	0.707	0.710	0.696
65	SVM	Single A1C	0.329	0.561	0.232	0.737	0.760	0.661
66	SVM	Single A1C	0.373	0.627	0.266	0.760	0.777	0.700
67	SVM	Mean A1C	0.329	0.625	0.223	0.715	0.726	0.676
68	SVM	Mean A1C	0.350	0.700	0.233	0.709	0.711	0.705
69	SVM	Mean A1C	0.372	0.659	0.260	0.746	0.757	0.708
70	SVM	Mean A1C	0.354	0.634	0.245	0.735	0.748	0.691
71	SVM	Mean A1C	0.336	0.610	0.231	0.723	0.738	0.674
72	SVM	Mean A1C	0.354	0.634	0.245	0.735	0.748	0.691

73	SVM	Mean A1C	0.295	0.561	0.200	0.693	0.710	0.635
74	SVM	Mean A1C	0.388	0.805	0.256	0.709	0.697	0.751
75	SVM	Mean A1C	0.320	0.659	0.211	0.679	0.681	0.670
76	SVM	Mean A1C	0.360	0.659	0.248	0.732	0.741	0.700
77	SVM	Mean A1C	0.331	0.698	0.228	0.724	0.741	0.670
78	SVM	Combination single	0.327	0.600	0.224	0.723	0.739	0.669
79	SVM	Combination single	0.351	0.650	0.241	0.732	0.742	0.696
80	SVM	Combination single	0.328	0.537	0.237	0.749	0.776	0.656
81	SVM	Combination single	0.365	0.610	0.260	0.757	0.776	0.693
82	SVM	Combination single	0.364	0.634	0.255	0.746	0.760	0.697
83	SVM	Combination single	0.356	0.634	0.248	0.737	0.751	0.692
84	SVM	Combination single	0.288	0.512	0.200	0.709	0.735	0.624
85	SVM	Combination single	0.425	0.829	0.286	0.743	0.732	0.781
86	SVM	Combination single	0.350	0.683	0.235	0.709	0.713	0.698
87	SVM	Combination single	0.350	0.585	0.250	0.751	0.773	0.679
88	SVM	Combination single	0.360	0.608	0.256	0.754	0.773	0.690
89	SVM	Combination mean	0.329	0.575	0.230	0.737	0.758	0.666
90	SVM	Combination mean	0.342	0.625	0.236	0.732	0.745	0.685
91	SVM	Combination mean	0.326	0.537	0.234	0.746	0.773	0.655
92	SVM	Combination mean	0.385	0.634	0.277	0.768	0.785	0.710
93	SVM	Combination mean	0.373	0.610	0.269	0.765	0.785	0.698
94	SVM	Combination mean	0.372	0.659	0.260	0.746	0.757	0.708
95	SVM	Combination mean	0.308	0.537	0.216	0.723	0.748	0.642
96	SVM	Combination mean	0.397	0.756	0.270	0.737	0.735	0.746
97	SVM	Combination mean	0.321	0.634	0.215	0.693	0.700	0.667
98	SVM	Combination mean	0.329	0.561	0.232	0.737	0.760	0.661
99	SVM	Combination mean	0.339	0.569	0.242	0.748	0.771	0.670
100	SVM	Multiple	0.325	0.650	0.217	0.698	0.704	0.677

101	SVM	Multiple	0.356	0.725	0.236	0.707	0.704	0.715
102	SVM	Multiple	0.378	0.659	0.265	0.751	0.763	0.711
103	SVM	Multiple	0.359	0.634	0.250	0.740	0.754	0.694
104	SVM	Multiple	0.327	0.585	0.226	0.723	0.741	0.663
105	SVM	Multiple	0.340	0.610	0.236	0.729	0.744	0.677
106	SVM	Multiple	0.301	0.561	0.205	0.701	0.719	0.640
107	SVM	Multiple	0.402	0.829	0.266	0.718	0.703	0.766
108	SVM	Multiple	0.345	0.707	0.228	0.693	0.691	0.699
109	SVM	Multiple	0.338	0.610	0.234	0.726	0.741	0.676
110	SVM	Multiple	0.323	0.578	0.224	0.724	0.743	0.660
111	NN	Single A1C	0.361	0.665	0.193	0.611	0.604	0.706
112	NN	Single A1C	0.362	0.810	0.196	0.597	0.571	0.775
113	NN	Single A1C	0.402	0.707	0.211	0.655	0.648	0.757
114	NN	Single A1C	0.421	0.722	0.223	0.650	0.641	0.779
115	NN	Single A1C	0.334	0.790	0.220	0.644	0.625	0.774
116	NN	Single A1C	0.405	0.715	0.196	0.613	0.600	0.744
117	NN	Single A1C	0.354	0.685	0.196	0.599	0.588	0.724
118	NN	Single A1C	0.356	0.968	0.244	0.638	0.595	0.900
119	NN	Single A1C	0.418	0.737	0.185	0.591	0.572	0.737
120	NN	Single A1C	0.397	0.661	0.236	0.705	0.711	0.773
121	NN	Single A1C	0.322	0.675	0.220	0.668	0.667	0.747
122	NN	Mean A1C	0.291	0.703	0.188	0.603	0.590	0.723
123	NN	Mean A1C	0.322	0.770	0.210	0.625	0.607	0.775
124	NN	Mean A1C	0.320	0.732	0.208	0.638	0.626	0.761
125	NN	Mean A1C	0.348	0.724	0.236	0.673	0.667	0.769
126	NN	Mean A1C	0.331	0.754	0.218	0.650	0.636	0.769
127	NN	Mean A1C	0.310	0.693	0.204	0.640	0.633	0.745
128	NN	Mean A1C	0.295	0.754	0.192	0.571	0.548	0.730
129	NN	Mean A1C	0.405	0.934	0.263	0.677	0.643	0.894
130	NN	Mean A1C	0.293	0.712	0.186	0.604	0.590	0.724
131	NN	Mean A1C	0.344	0.734	0.232	0.667	0.658	0.779
132	NN	Mean A1C	0.348	0.600	0.251	0.741	0.759	0.743
133	NN	Combination single	0.291	0.670	0.190	0.622	0.616	0.714
134	NN	Combination single	0.327	0.725	0.219	0.645	0.635	0.773
135	NN	Combination single	0.304	0.710	0.197	0.625	0.614	0.755
136	NN	Combination single	0.341	0.729	0.232	0.659	0.650	0.782
137	NN	Combination single	0.359	0.724	0.241	0.706	0.703	0.775
138	NN	Combination single	0.318	0.676	0.211	0.662	0.660	0.743
139	NN	Combination single	0.318	0.615	0.219	0.695	0.705	0.729
140	NN	Combination single	0.399	0.956	0.253	0.661	0.623	0.897
141	NN	Combination single	0.337	0.649	0.233	0.700	0.706	0.732

142	NN	Combination single	0.342	0.727	0.227	0.672	0.665	0.784
143	NN	Combination single	0.346	0.620	0.245	0.729	0.743	0.748
144	NN	Combination mean	0.329	0.518	0.248	0.766	0.797	0.746
145	NN	Combination mean	0.363	0.525	0.281	0.793	0.827	0.753
146	NN	Combination mean	0.371	0.478	0.314	0.817	0.861	0.787
147	NN	Combination mean	0.409	0.500	0.350	0.836	0.879	0.819
148	NN	Combination mean	0.412	0.373	0.474	0.878	0.944	0.855
149	NN	Combination mean	0.518	0.563	0.506	0.878	0.919	0.885
150	NN	Combination mean	0.439	0.512	0.402	0.853	0.897	0.871
151	NN	Combination mean	0.605	0.690	0.564	0.895	0.921	0.919
152	NN	Combination mean	0.463	0.424	0.546	0.891	0.951	0.877
153	NN	Combination mean	0.564	0.578	0.590	0.899	0.941	0.899
154	NN	Combination mean	0.348	0.551	0.251	0.760	0.787	0.729
155	NN	Multiple	0.340	0.598	0.238	0.740	0.758	0.722
156	NN	Multiple	0.356	0.568	0.260	0.772	0.797	0.765
157	NN	Multiple	0.354	0.500	0.275	0.792	0.830	0.757
158	NN	Multiple	0.440	0.585	0.354	0.830	0.862	0.812
159	NN	Multiple	0.499	0.637	0.414	0.855	0.883	0.857
160	NN	Multiple	0.446	0.549	0.377	0.844	0.883	0.870
161	NN	Multiple	0.466	0.507	0.432	0.866	0.912	0.862
162	NN	Multiple	0.551	0.771	0.429	0.856	0.867	0.904
163	NN	Multiple	0.544	0.649	0.471	0.877	0.906	0.886
164	NN	Multiple	0.545	0.507	0.596	0.905	0.956	0.898
165	NN	Multiple	0.356	0.618	0.250	0.744	0.760	0.765

Appendix 6. Performance metrics of predictive models of the retinopathy cohort

Obs	Modeling Method	A1C Variability	F1 score	Sensitivity	Precision	Accuracy	Specificity	AUC
1	LR	Single A1C	0.000	0.000	1.000	0.814	0.000	0.666
2	LR	Single A1C	0.073	0.038	1.000	0.822	1.000	0.723
3	LR	Single A1C	0.073	0.038	1.000	0.820	1.000	0.689
4	LR	Single A1C	0.037	0.019	1.000	0.817	1.000	0.656
5	LR	Single A1C	0.074	0.038	1.000	0.824	1.000	0.688
6	LR	Single A1C	0.037	0.019	1.000	0.816	1.000	0.724
7	LR	Single A1C	0.107	0.057	1.000	0.826	1.000	0.696
8	LR	Single A1C	0.107	0.057	1.000	0.824	1.000	0.703
9	LR	Single A1C	0.037	0.019	1.000	0.819	1.000	0.741
10	LR	Single A1C	0.074	0.038	1.000	0.822	1.000	0.697
11	LR	Single A1C	0.114	0.061	1.000	0.826	1.000	0.732
12	LR	Mean A1C	0.000	0.000	1.000	0.814	0.000	0.680
13	LR	Mean A1C	0.073	0.038	1.000	0.822	1.000	0.716
14	LR	Mean A1C	0.073	0.038	1.000	0.820	1.000	0.692
15	LR	Mean A1C	0.037	0.019	1.000	0.816	1.000	0.675
16	LR	Mean A1C	0.109	0.058	1.000	0.827	1.000	0.688
17	LR	Mean A1C	0.037	0.019	1.000	0.816	1.000	0.723
18	LR	Mean A1C	0.107	0.057	1.000	0.825	1.000	0.702
19	LR	Mean A1C	0.073	0.038	1.000	0.821	1.000	0.701
20	LR	Mean A1C		0.000	1.000	0.815	0.000	0.738
21	LR	Mean A1C	0.074	0.038	1.000	0.822	1.000	0.696
22	LR	Mean A1C	0.114	0.061	1.000	0.827	1.000	0.734
23	LR	Combination single		0.000	1.000	0.814	0.000	0.672
24	LR	Combination single	0.073	0.038	1.000	0.822	1.000	0.725
25	LR	Combination single	0.073	0.038	1.000	0.820	1.000	0.685
26	LR	Combination single	0.037	0.019	1.000	0.817	1.000	0.658
27	LR	Combination single	0.074	0.038	1.000	0.824	1.000	0.681
28	LR	Combination single	0.037	0.019	1.000	0.816	1.000	0.726
29	LR	Combination single	0.107	0.057	1.000	0.826	1.000	0.698
30	LR	Combination single	0.107	0.057	1.000	0.824	1.000	0.701
31	LR	Combination single	0.037	0.019	1.000	0.819	1.000	0.744
32	LR	Combination single	0.074	0.038	1.000	0.821	1.000	0.701
33	LR	Combination single	0.114	0.061	1.000	0.826	1.000	0.732

34	LR	Combination mean		0.000		0.814	1.000	0.680
35	LR	Combination mean	0.073	0.038	1.000	0.822	1.000	0.717
36	LR	Combination mean	0.073	0.038	1.000	0.820	1.000	0.690
37	LR	Combination mean	0.037	0.019	1.000	0.817	1.000	0.675
38	LR	Combination mean	0.074	0.038	1.000	0.824	1.000	0.684
39	LR	Combination mean	0.037	0.019	1.000	0.816	1.000	0.724
40	LR	Combination mean	0.107	0.057	1.000	0.826	1.000	0.701
41	LR	Combination mean	0.107	0.057	1.000	0.825	1.000	0.700
42	LR	Combination mean		0.000		0.815	1.000	0.738
43	LR	Combination mean	0.074	0.038	1.000	0.822	1.000	0.696
44	LR	Combination mean	0.114	0.061	1.000	0.827	1.000	0.734
45	LR	Multiple	0.050	0.026	1.000	0.761	1.000	0.602
46	LR	Multiple	0.057	0.029	1.000	0.790	1.000	0.600
47	LR	Multiple	0.107	0.056	1.000	0.792	1.000	0.636
48	LR	Multiple	0.049	0.025	1.000	0.758	1.000	0.686
49	LR	Multiple	0.092	0.048	1.000	0.817	1.000	0.683
50	LR	Multiple	0.083	0.043	1.000	0.792	1.000	0.636
51	LR	Multiple	0.094	0.049	1.000	0.822	1.000	0.679
52	LR	Multiple	0.121	0.065	1.000	0.820	1.000	0.670
53	LR	Multiple	0.154	0.083	1.000	0.830	1.000	0.663
54	LR	Multiple	0.094	0.049	1.000	0.819	1.000	0.641
55	LR	Multiple	0.154	0.083	1.000	0.831	1.000	0.718
56	SVM	Single A1C	0.378	0.717	0.257	0.566	0.532	0.624
57	SVM	Single A1C	0.378	0.717	0.257	0.566	0.532	0.624
58	SVM	Single A1C	0.367	0.679	0.252	0.569	0.545	0.612
59	SVM	Single A1C	0.426	0.755	0.296	0.625	0.596	0.675
60	SVM	Single A1C	0.429	0.736	0.302	0.639	0.617	0.676
61	SVM	Single A1C	0.354	0.673	0.240	0.554	0.528	0.600
62	SVM	Single A1C	0.370	0.673	0.255	0.585	0.566	0.620
63	SVM	Single A1C	0.370	0.712	0.250	0.561	0.528	0.620
64	SVM	Single A1C	0.379	0.679	0.263	0.589	0.568	0.624
65	SVM	Single A1C	0.448	0.774	0.315	0.648	0.620	0.697
66	SVM	Single A1C	0.417	0.758	0.287	0.611	0.578	0.670
67	SVM	Mean A1C	0.371	0.679	0.255	0.576	0.553	0.616
68	SVM	Mean A1C	0.380	0.736	0.257	0.559	0.519	0.627
69	SVM	Mean A1C	0.366	0.642	0.256	0.590	0.579	0.610
70	SVM	Mean A1C	0.426	0.736	0.300	0.635	0.613	0.674
71	SVM	Mean A1C	0.420	0.698	0.301	0.646	0.634	0.666
72	SVM	Mean A1C	0.342	0.635	0.234	0.557	0.540	0.588

73	SVM	Mean A1C	0.394	0.712	0.272	0.603	0.579	0.645
74	SVM	Mean A1C	0.381	0.692	0.263	0.592	0.570	0.631
75	SVM	Mean A1C	0.387	0.660	0.273	0.613	0.603	0.631
76	SVM	Mean A1C	0.446	0.736	0.320	0.662	0.645	0.691
77	SVM	Mean A1C	0.416	0.727	0.291	0.625	0.602	0.660
78	SVM	Combination single	0.384	0.736	0.260	0.566	0.528	0.632
79	SVM	Combination single	0.374	0.717	0.253	0.559	0.523	0.620
80	SVM	Combination single	0.360	0.679	0.245	0.556	0.528	0.603
81	SVM	Combination single	0.409	0.717	0.286	0.618	0.596	0.656
82	SVM	Combination single	0.435	0.755	0.305	0.639	0.613	0.684
83	SVM	Combination single	0.367	0.692	0.250	0.568	0.540	0.616
84	SVM	Combination single	0.379	0.712	0.259	0.578	0.549	0.630
85	SVM	Combination single	0.368	0.712	0.248	0.557	0.523	0.617
86	SVM	Combination single	0.379	0.698	0.261	0.578	0.551	0.625
87	SVM	Combination single	0.459	0.792	0.323	0.655	0.624	0.708
88	SVM	Combination single	0.417	0.765	0.287	0.608	0.573	0.670
89	SVM	Combination mean	0.378	0.698	0.259	0.576	0.549	0.624
90	SVM	Combination mean	0.371	0.717	0.250	0.552	0.515	0.616
91	SVM	Combination mean	0.370	0.642	0.260	0.597	0.587	0.614
92	SVM	Combination mean	0.426	0.736	0.300	0.635	0.613	0.674
93	SVM	Combination mean	0.413	0.698	0.294	0.635	0.621	0.660
94	SVM	Combination mean	0.351	0.635	0.243	0.575	0.562	0.598
95	SVM	Combination mean	0.398	0.712	0.276	0.610	0.587	0.649
96	SVM	Combination mean	0.377	0.692	0.259	0.585	0.562	0.627
97	SVM	Combination mean	0.385	0.660	0.271	0.610	0.598	0.629
98	SVM	Combination mean	0.453	0.736	0.328	0.672	0.658	0.697
99	SVM	Combination mean	0.415	0.720	0.291	0.628	0.607	0.660
100	SVM	Multiple	0.382	0.717	0.260	0.573	0.540	0.629

101	SVM	Multiple	0.392	0.755	0.265	0.569	0.528	0.641
102	SVM	Multiple	0.358	0.679	0.243	0.552	0.523	0.601
103	SVM	Multiple	0.425	0.774	0.293	0.615	0.579	0.676
104	SVM	Multiple	0.453	0.811	0.314	0.639	0.600	0.706
105	SVM	Multiple	0.363	0.712	0.243	0.547	0.511	0.611
106	SVM	Multiple	0.404	0.750	0.277	0.599	0.566	0.658
107	SVM	Multiple	0.394	0.769	0.265	0.571	0.528	0.648
108	SVM	Multiple	0.367	0.679	0.252	0.568	0.543	0.611
109	SVM	Multiple	0.441	0.811	0.303	0.620	0.577	0.694
110	SVM	Multiple	0.422	0.780	0.289	0.608	0.570	0.680
111	NN	Single A1C	0.396	0.819	0.263	0.539	0.476	0.690
112	NN	Single A1C	0.382	0.751	0.260	0.557	0.513	0.666
113	NN	Single A1C	0.343	0.706	0.228	0.506	0.461	0.618
114	NN	Single A1C	0.395	0.770	0.269	0.562	0.515	0.681
115	NN	Single A1C	0.420	0.834	0.283	0.576	0.518	0.725
116	NN	Single A1C	0.344	0.638	0.238	0.562	0.545	0.633
117	NN	Single A1C	0.384	0.746	0.261	0.568	0.529	0.683
118	NN	Single A1C	0.388	0.738	0.269	0.571	0.534	0.693
119	NN	Single A1C	0.390	0.781	0.262	0.549	0.497	0.682
120	NN	Single A1C	0.424	0.887	0.280	0.552	0.476	0.763
121	NN	Single A1C	0.413	0.775	0.285	0.589	0.547	0.716
122	NN	Mean A1C	0.393	0.770	0.265	0.559	0.512	0.690
123	NN	Mean A1C	0.381	0.794	0.255	0.524	0.463	0.672
124	NN	Mean A1C	0.341	0.632	0.237	0.553	0.535	0.625
125	NN	Mean A1C	0.405	0.792	0.275	0.568	0.517	0.694
126	NN	Mean A1C	0.419	0.851	0.280	0.562	0.497	0.723
127	NN	Mean A1C	0.335	0.610	0.232	0.566	0.557	0.639
128	NN	Mean A1C	0.387	0.733	0.266	0.583	0.550	0.695
129	NN	Mean A1C	0.386	0.783	0.260	0.539	0.485	0.667
130	NN	Mean A1C	0.390	0.768	0.263	0.560	0.512	0.669
131	NN	Mean A1C	0.458	0.815	0.323	0.638	0.598	0.748
132	NN	Mean A1C	0.416	0.795	0.285	0.583	0.535	0.725
133	NN	Combination single	0.399	0.791	0.268	0.560	0.508	0.687
134	NN	Combination single	0.381	0.775	0.256	0.534	0.480	0.668
135	NN	Combination single	0.335	0.613	0.232	0.555	0.542	0.618
136	NN	Combination single	0.389	0.813	0.258	0.525	0.460	0.682
137	NN	Combination single	0.412	0.770	0.290	0.603	0.566	0.730
138	NN	Combination single	0.356	0.685	0.241	0.550	0.520	0.634
139	NN	Combination single	0.380	0.746	0.259	0.559	0.517	0.683
140	NN	Combination single	0.377	0.800	0.250	0.512	0.449	0.693
141	NN	Combination single	0.385	0.726	0.268	0.574	0.539	0.688

142	NN	Combination single	0.459	0.851	0.316	0.627	0.576	0.764
143	NN	Combination single	0.415	0.776	0.287	0.594	0.554	0.690
144	NN	Combination mean	0.388	0.723	0.269	0.583	0.551	0.690
145	NN	Combination mean	0.367	0.774	0.246	0.511	0.452	0.669
146	NN	Combination mean	0.332	0.613	0.234	0.554	0.541	0.621
147	NN	Combination mean	0.403	0.787	0.275	0.561	0.511	0.686
148	NN	Combination mean	0.434	0.792	0.302	0.619	0.580	0.728
149	NN	Combination mean	0.352	0.679	0.247	0.552	0.524	0.636
150	NN	Combination mean	0.387	0.687	0.273	0.610	0.593	0.699
151	NN	Combination mean	0.393	0.762	0.267	0.568	0.526	0.681
152	NN	Combination mean	0.380	0.745	0.257	0.553	0.509	0.671
153	NN	Combination mean	0.453	0.840	0.313	0.617	0.567	0.737
154	NN	Combination mean	0.423	0.783	0.294	0.604	0.564	0.718
155	NN	Multiple	0.397	0.717	0.277	0.600	0.573	0.693
156	NN	Multiple	0.390	0.760	0.263	0.562	0.517	0.670
157	NN	Multiple	0.339	0.681	0.227	0.516	0.478	0.629
158	NN	Multiple	0.405	0.738	0.285	0.601	0.571	0.687
159	NN	Multiple	0.408	0.823	0.275	0.558	0.498	0.723
160	NN	Multiple	0.346	0.638	0.239	0.567	0.551	0.643
161	NN	Multiple	0.393	0.765	0.270	0.577	0.536	0.698
162	NN	Multiple	0.380	0.767	0.255	0.540	0.490	0.696
163	NN	Multiple	0.381	0.711	0.261	0.574	0.542	0.681
164	NN	Multiple	0.447	0.843	0.305	0.611	0.558	0.736
165	NN	Multiple	0.429	0.752	0.303	0.629	0.602	0.723

Appendix 7: Performance metrics of predictive models of the neuropathy cohort

Obs	Modeling Method	A1C Variability	F1 score	Sensitivity	Precision	Accuracy	Specificity	AUC
1	LR	Single A1C	0.296	0.174	1.000	0.882	1.000	0.808
2	LR	Single A1C	0.264	0.152	1.000	0.876	1.000	0.777
3	LR	Single A1C	0.351	0.213	1.000	0.885	1.000	0.806
4	LR	Single A1C	0.264	0.152	1.000	0.877	1.000	0.760
5	LR	Single A1C	0.160	0.087	1.000	0.869	1.000	0.790
6	LR	Single A1C	0.259	0.149	1.000	0.876	1.000	0.810
7	LR	Single A1C	0.231	0.130	1.000	0.874	1.000	0.762
8	LR	Single A1C	0.157	0.085	1.000	0.865	1.000	0.768
9	LR	Single A1C	0.160	0.087	1.000	0.871	1.000	0.755
10	LR	Single A1C	0.196	0.109	1.000	0.870	1.000	0.777
11	LR	Single A1C	0.188	0.103	1.000	0.871	1.000	0.779
12	LR	Mean A1C	0.327	0.196	1.000	0.885	1.000	0.811
13	LR	Mean A1C	0.264	0.152	1.000	0.877	1.000	0.782
14	LR	Mean A1C	0.321	0.191	1.000	0.882	1.000	0.806
15	LR	Mean A1C	0.264	0.152	1.000	0.877	1.000	0.760
16	LR	Mean A1C	0.122	0.065	1.000	0.866	1.000	0.791
17	LR	Mean A1C	0.259	0.149	1.000	0.875	1.000	0.806
18	LR	Mean A1C	0.264	0.152	1.000	0.878	1.000	0.762
19	LR	Mean A1C	0.157	0.085	1.000	0.865	1.000	0.769
20	LR	Mean A1C	0.160	0.087	1.000	0.871	1.000	0.766
21	LR	Mean A1C	0.196	0.109	1.000	0.871	1.000	0.779
22	LR	Mean A1C	0.188	0.103	1.000	0.871	1.000	0.778
23	LR	Combination single	0.264	0.152	1.000	0.879	1.000	0.803
24	LR	Combination single	0.296	0.174	1.000	0.879	1.000	0.780
25	LR	Combination single	0.321	0.191	1.000	0.882	1.000	0.806
26	LR	Combination single	0.264	0.152	1.000	0.877	1.000	0.759
27	LR	Combination single	0.160	0.087	1.000	0.869	1.000	0.791
28	LR	Combination single	0.226	0.128	1.000	0.873	1.000	0.804
29	LR	Combination single	0.264	0.152	1.000	0.877	1.000	0.770
30	LR	Combination single	0.157	0.085	1.000	0.864	1.000	0.769
31	LR	Combination single	0.160	0.087	1.000	0.871	1.000	0.766
32	LR	Combination single	0.160	0.087	1.000	0.868	1.000	0.780
33	LR	Combination single	0.188	0.103	1.000	0.871	1.000	0.782

34	LR	Combination mean	0.327	0.196	1.000	0.885	1.000	0.806
35	LR	Combination mean	0.264	0.152	1.000	0.876	1.000	0.782
36	LR	Combination mean	0.259	0.149	1.000	0.876	1.000	0.803
37	LR	Combination mean	0.264	0.152	1.000	0.877	1.000	0.760
38	LR	Combination mean	0.122	0.065	1.000	0.866	1.000	0.792
39	LR	Combination mean	0.226	0.128	1.000	0.873	1.000	0.802
40	LR	Combination mean	0.264	0.152	1.000	0.878	1.000	0.769
41	LR	Combination mean	0.192	0.106	1.000	0.868	1.000	0.768
42	LR	Combination mean	0.160	0.087	1.000	0.871	1.000	0.772
43	LR	Combination mean	0.196	0.109	1.000	0.871	1.000	0.780
44	LR	Combination mean	0.198	0.103	1.000	0.871	1.000	0.781
45	LR	Multiple	0.231	0.130	1.000	0.876	1.000	0.780
46	LR	Multiple	0.122	0.065	1.000	0.864	1.000	0.715
47	LR	Multiple	0.192	0.106	1.000	0.870	1.000	0.777
48	LR	Multiple	0.196	0.109	1.000	0.872	1.000	0.710
49	LR	Multiple	0.122	0.065	1.000	0.867	1.000	0.755
50	LR	Multiple	0.157	0.085	1.000	0.865	1.000	0.711
51	LR	Multiple	0.083	0.043	1.000	0.864	1.000	0.739
52	LR	Multiple	0.120	0.064	1.000	0.863	1.000	0.777
53	LR	Multiple	0.231	0.130	1.000	0.877	1.000	0.730
54	LR	Multiple	0.083	0.043	1.000	0.863	1.000	0.742
55	LR	Multiple	0.067	0.034	1.000	0.861	1.000	0.737
56	SVM	Single A1C	0.391	0.391	0.263	0.666	0.650	0.705
57	SVM	Single A1C	0.378	0.378	0.263	0.687	0.689	0.682
58	SVM	Single A1C	0.372	0.372	0.248	0.647	0.632	0.686
59	SVM	Single A1C	0.405	0.405	0.276	0.684	0.671	0.716
60	SVM	Single A1C	0.356	0.356	0.244	0.656	0.656	0.658
61	SVM	Single A1C	0.408	0.408	0.266	0.635	0.595	0.734
62	SVM	Single A1C	0.362	0.362	0.241	0.632	0.616	0.670
63	SVM	Single A1C	0.381	0.381	0.262	0.680	0.677	0.687
64	SVM	Single A1C	0.398	0.398	0.267	0.665	0.645	0.714
65	SVM	Single A1C	0.391	0.391	0.268	0.683	0.677	0.697
66	SVM	Single A1C	0.431	0.776	0.298	0.708	0.697	0.740
67	SVM	Mean A1C	0.395	0.739	0.270	0.681	0.671	0.705
68	SVM	Mean A1C	0.390	0.696	0.271	0.693	0.693	0.694
69	SVM	Mean A1C	0.373	0.717	0.252	0.660	0.650	0.684
70	SVM	Mean A1C	0.400	0.761	0.271	0.678	0.664	0.713
71	SVM	Mean A1C	0.349	0.638	0.240	0.656	0.659	0.649
72	SVM	Mean A1C	0.396	0.809	0.262	0.644	0.616	0.712

73	SVM	Mean A1C	0.355	0.702	0.237	0.632	0.620	0.661
74	SVM	Mean A1C	0.382	0.717	0.260	0.671	0.663	0.690
75	SVM	Mean A1C	0.398	0.783	0.267	0.665	0.645	0.714
76	SVM	Mean A1C	0.412	0.739	0.286	0.702	0.695	0.717
77	SVM	Mean A1C	0.435	0.750	0.306	0.723	0.718	0.730
78	SVM	Combination single	0.391	0.717	0.268	0.684	0.679	0.698
79	SVM	Combination single	0.388	0.674	0.272	0.699	0.704	0.689
80	SVM	Combination single	0.386	0.717	0.264	0.678	0.671	0.694
81	SVM	Combination single	0.391	0.739	0.266	0.675	0.664	0.702
82	SVM	Combination single	0.357	0.638	0.248	0.669	0.674	0.656
83	SVM	Combination single	0.365	0.702	0.246	0.647	0.638	0.670
84	SVM	Combination single	0.330	0.617	0.225	0.638	0.642	0.629
85	SVM	Combination single	0.360	0.630	0.252	0.683	0.692	0.661
86	SVM	Combination single	0.393	0.739	0.268	0.677	0.667	0.703
87	SVM	Combination single	0.407	0.717	0.284	0.705	0.703	0.710
88	SVM	Combination single	0.441	0.759	0.311	0.726	0.721	0.740
89	SVM	Combination mean	0.395	0.739	0.270	0.681	0.671	0.705
90	SVM	Combination mean	0.390	0.696	0.271	0.693	0.693	0.694
91	SVM	Combination mean	0.375	0.717	0.254	0.663	0.654	0.685
92	SVM	Combination mean	0.393	0.761	0.265	0.669	0.654	0.707
93	SVM	Combination mean	0.363	0.660	0.250	0.666	0.667	0.663
94	SVM	Combination mean	0.398	0.809	0.264	0.647	0.620	0.714
95	SVM	Combination mean	0.348	0.681	0.234	0.632	0.624	0.652
96	SVM	Combination mean	0.367	0.674	0.252	0.671	0.670	0.672
97	SVM	Combination mean	0.407	0.804	0.272	0.668	0.645	0.725
98	SVM	Combination mean	0.412	0.739	0.286	0.702	0.695	0.717
99	SVM	Combination mean	0.441	0.750	0.312	0.729	0.725	0.740
100	SVM	Multiple	0.396	0.783	0.265	0.663	0.643	0.713

101	SVM	Multiple	0.386	0.696	0.267	0.687	0.686	0.691
102	SVM	Multiple	0.372	0.739	0.248	0.647	0.632	0.686
103	SVM	Multiple	0.416	0.804	0.280	0.681	0.661	0.733
104	SVM	Multiple	0.366	0.681	0.250	0.660	0.656	0.668
105	SVM	Multiple	0.371	0.766	0.245	0.626	0.602	0.684
106	SVM	Multiple	0.351	0.702	0.234	0.626	0.613	0.658
107	SVM	Multiple	0.391	0.739	0.266	0.674	0.663	0.701
108	SVM	Multiple	0.391	0.783	0.261	0.655	0.634	0.709
109	SVM	Multiple	0.412	0.761	0.282	0.692	0.681	0.721
110	SVM	Multiple	0.426	0.759	0.296	0.709	0.701	0.730
111	NN	Single A1C	0.360	0.717	0.244	0.635	0.621	0.755
112	NN	Single A1C	0.366	0.570	0.277	0.728	0.754	0.750
113	NN	Single A1C	0.387	0.848	0.268	0.598	0.558	0.834
114	NN	Single A1C	0.418	0.761	0.294	0.699	0.689	0.823
115	NN	Single A1C	0.347	0.628	0.241	0.660	0.665	0.700
116	NN	Single A1C	0.403	0.840	0.267	0.640	0.606	0.784
117	NN	Single A1C	0.346	0.732	0.229	0.597	0.575	0.739
118	NN	Single A1C	0.361	0.698	0.252	0.644	0.635	0.744
119	NN	Single A1C	0.404	0.861	0.265	0.637	0.600	0.812
120	NN	Single A1C	0.426	0.713	0.308	0.722	0.724	0.785
121	NN	Single A1C	0.384	0.799	0.256	0.628	0.600	0.783
122	NN	Mean A1C	0.354	0.743	0.233	0.616	0.595	0.759
123	NN	Mean A1C	0.391	0.648	0.283	0.718	0.730	0.754
124	NN	Mean A1C	0.409	0.857	0.271	0.644	0.609	0.831
125	NN	Mean A1C	0.414	0.728	0.299	0.707	0.703	0.821
126	NN	Mean A1C	0.328	0.689	0.216	0.588	0.571	0.706
127	NN	Mean A1C	0.403	0.851	0.265	0.636	0.599	0.790
128	NN	Mean A1C	0.348	0.674	0.239	0.632	0.625	0.738
129	NN	Mean A1C	0.361	0.709	0.246	0.642	0.632	0.750
130	NN	Mean A1C	0.410	0.824	0.274	0.660	0.633	0.807
131	NN	Mean A1C	0.407	0.722	0.288	0.695	0.691	0.781
132	NN	Mean A1C	0.390	0.797	0.260	0.640	0.614	0.785
133	NN	Combination single	0.354	0.780	0.229	0.596	0.566	0.755
134	NN	Combination single	0.388	0.691	0.273	0.689	0.689	0.755
135	NN	Combination single	0.433	0.822	0.297	0.689	0.667	0.834
136	NN	Combination single	0.405	0.850	0.270	0.644	0.610	0.824
137	NN	Combination single	0.329	0.626	0.229	0.636	0.637	0.700
138	NN	Combination single	0.412	0.823	0.277	0.660	0.632	0.787
139	NN	Combination single	0.357	0.660	0.247	0.653	0.652	0.735
140	NN	Combination single	0.360	0.709	0.244	0.642	0.630	0.748
141	NN	Combination single	0.404	0.852	0.266	0.641	0.606	0.810

142	NN	Combination single	0.421	0.717	0.302	0.716	0.716	0.786
143	NN	Combination single	0.406	0.747	0.284	0.685	0.675	0.787
144	NN	Combination mean	0.358	0.724	0.239	0.633	0.619	0.758
145	NN	Combination mean	0.389	0.693	0.273	0.689	0.689	0.754
146	NN	Combination mean	0.416	0.837	0.285	0.645	0.614	0.832
147	NN	Combination mean	0.410	0.815	0.275	0.668	0.644	0.820
148	NN	Combination mean	0.337	0.689	0.226	0.607	0.593	0.705
149	NN	Combination mean	0.392	0.917	0.250	0.589	0.533	0.791
150	NN	Combination mean	0.355	0.706	0.244	0.625	0.612	0.738
151	NN	Combination mean	0.365	0.665	0.253	0.674	0.675	0.751
152	NN	Combination mean	0.403	0.839	0.267	0.645	0.613	0.807
153	NN	Combination mean	0.402	0.713	0.285	0.694	0.690	0.784
154	NN	Combination mean	0.404	0.771	0.275	0.673	0.657	0.786
155	NN	Multiple	0.356	0.763	0.234	0.608	0.583	0.753
156	NN	Multiple	0.377	0.693	0.261	0.674	0.670	0.744
157	NN	Multiple	0.399	0.865	0.261	0.625	0.585	0.826
158	NN	Multiple	0.409	0.774	0.279	0.684	0.669	0.813
159	NN	Multiple	0.349	0.664	0.239	0.645	0.642	0.708
160	NN	Multiple	0.396	0.813	0.263	0.641	0.613	0.767
161	NN	Multiple	0.344	0.657	0.235	0.637	0.633	0.726
162	NN	Multiple	0.353	0.674	0.240	0.650	0.646	0.733
163	NN	Multiple	0.392	0.802	0.261	0.646	0.620	0.793
164	NN	Multiple	0.365	0.735	0.245	0.634	0.617	0.771
165	NN	Multiple	0.404	0.746	0.280	0.684	0.674	0.780

REFERENCES

Aathira, R., & Jain, V. (2014). Advances in management of type 1 diabetes mellitus. *World J Diabetes, 5*(5), 689-696. doi:10.4239/wjd.v5.i5.689

Abadi, M., AshishBarham, PaulBrevdo, EugeneChen, ZhifengCitro, CraigCorrado, GregDavis, AndyDean, JeffreyDevin, MatthieuGhemawat, SanjayGoodfellow, IanHarp, AndrewIrving, GeoffreyIsard, MichaelJia, YangqingJozefowicz, RafalKaiser, LukaszKudlur, ManjunathLevenberg, JoshMane, DanMonga, RajatMoore, SherryMurray, DerekOlah, ChrisSchuster, MikeShlens, JonathonSteiner, BenoiySutskever, IlyaTalwar, KunalTucker, PaulVanhoucke, VincentVasudevan, VijayViegas, FernandaVinyals, OriolWarden, PeteWattenberg, MartinWiche, MartinYu, YuanZheng, Xiaoqiang. (2015). Large-scale machine learning on heterogeneous distributed systems. In.

Abid, S., Keshavjee, K., Karim, A., & Guergachi, A. (2017). What we can learn from Amazon for clinical decision support systems. *Stud Health Technol Inform, 234*, 1-5.

Akosa, J. S. (2017). *Predictive accuracy: A misleading performance measure for highly imbalanced data*.

Al-Geffari, M. (2012). Comparison of different screening tests for diagnosis of diabetic peripheral neuropathy in Primary Health Care setting. *International journal of health sciences, 6*(2), 127-134.

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLoS One, 12*(7), e0179805. doi:10.1371/journal.pone.0179805

Alleman, C. J., Westerhout, K. Y., Hensen, M., Chambers, C., Stoker, M., Long, S., & van Nooten, F. E. (2015). Humanistic and economic burden of painful diabetic peripheral neuropathy in Europe: A review of the literature. *Diabetes Res Clin Pract, 109*(2), 215-225. doi:10.1016/j.diabres.2015.04.031

Ameh, O. I., Okpechi, I. G., Agyemang, C., & Kengne, A. P. (2019). Global, Regional, and Ethnic Differences in Diabetic Nephropathy. In J. J. Roelofs & L. Vogt (Eds.), *Diabetic Nephropathy: Pathophysiology and Clinical Aspects* (pp. 33-44). Cham: Springer International Publishing.

Andersen, R. M. (2008). National Health Surveys and the Behavioral Model of Health Services Use. *Medical Care, 46*(7), 647-653.

Aspelund, T., Thornórisdóttir, O., Olafsdottir, E., Gudmundsdottir, A., Einarsdóttir, A. B., Mehlsen, J., . . . Stefánsson, E. (2011). Individual risk assessment and information technology to optimise screening frequency for diabetic retinopathy. *Diabetologia, 54*(10), 2525-2532. doi:10.1007/s00125-011-2257-7

Association, A. D. (2018a). 6. Glycemic Targets:. *Diabetes Care, 41*(Suppl 1), S55-S64. doi:10.2337/dc18-S006

Association, A. D. (2018b). 10. Microvascular Complications and Foot Care:. *Diabetes Care, 41*(Suppl 1), S105-S118. doi:10.2337/dc18-S010

Association, A. D. (2019a). 5. Lifestyle Management: Standards of Medical Care in Diabetes—2019. *Diabetes Care, 42*(Supplement 1), S46. doi:10.2337/dc19-S005

Association, A. D. (2019b). 6. Glycemic Targets:. *Diabetes Care, 42*(Suppl 1), S61-S70. doi:10.2337/dc19-S006

Association, A. D. (2019c). 7. Diabetes Technology:. *Diabetes Care, 42*(Suppl 1), S71-S80. doi:10.2337/dc19-S007

Association, A. D. (2019d). 11. Microvascular Complications and Foot Care:. *Diabetes Care, 42*(Suppl 1), S124-S138. doi:10.2337/dc19-S011

Atkinson, M. A., Eisenbarth, G. S., & Michels, A. W. (2014). Type 1 diabetes. *Lancet, 383*(9911), 69-82. doi:10.1016/S0140-6736(13)60591-7

Beck, R. W., Tamborlane, W. V., Bergenstal, R. M., Miller, K. M., DuBose, S. N., Hall, C. A., & Network, T. D. E. C. (2012). The T1D Exchange clinic registry. *J Clin Endocrinol Metab, 97*(12), 4383-4389. doi:10.1210/jc.2012-1561

Bjornstad, P., Cherney, D., & Maahs, D. M. (2014). Early diabetic nephropathy in type 1 diabetes: new insights. *Curr Opin Endocrinol Diabetes Obes, 21*(4), 279-286. doi:10.1097/MED.0000000000000074

Bluestone, J. A., Herold, K., & Eisenbarth, G. (2010). Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature, 464*(7293), 1293-1300.

Boulton, A. J. M., Vinik, A. I., Arezzo, J. C., Bril, V., Feldman, E. L., Freeman, R., . . . Ziegler, D. (2005). Diabetic Neuropathies. *Diabetes Care, 28*(4), 956. doi:10.2337/diacare.28.4.956

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci., 16*(3), 199-231. doi:10.1214/ss/1009213726

Bringsjord, S., & Govindarajulu, N. (2018). *Artificial intelligence* (E. N. Zalta Ed. Fall 2018 ed.): Metaphysics Research Lab, Stanford University

Brown, J. J., Pribesh, S. L., Baskette, K. G., Vinik, A. I., & Colberg, S. R. (2017). A Comparison of Screening Tools for the Early Detection of Peripheral Neuropathy in Adults with and without Type 2 Diabetes. *J Diabetes Res, 2017*, 1467213. doi:10.1155/2017/1467213

Byrne, M. M., Parker, D. F., Tannenbaum, S. L., Ocasio, M. A., Lam, B. L., Zimmer-Galler, I., & Lee, D. J. (2014). Cost of a Community-Based Diabetic Retinopathy Screening Program. *Diabetes Care, 37*(11), e236. doi:10.2337/dc14-0834

Byrne, M. M., Parker, D. F., Tannenbaum, S. L., Ocasio, M. A., Lam, B. L., Zimmer-Galler, I., & Lee, D. J. (2014). Cost of a community-based diabetic retinopathy screening program. *Diabetes Care, 37*(11), e236-237. doi:10.2337/dc14-0834

Callaghan, B., McCammon, R., Kerber, K., Xu, X., Langa, K. M., & Feldman, E. (2012). Tests and expenditures in the initial evaluation of peripheral neuropathy. *Arch Intern Med, 172*(2), 127-132. doi:10.1001/archinternmed.2011.1032

Candrilli, S. D., Davis, K. L., Kan, H. J., Lucero, M. A., & Rousculp, M. D. (2007). Prevalence and the associated burden of illness of symptoms of diabetic peripheral neuropathy and diabetic retinopathy. *J Diabetes Complications, 21*(5), 306-314. doi:10.1016/j.jdiacomp.2006.08.002

Cefalu, W. T., Dawes, D. E., Gavlak, G., Goldman, D., Herman, W. H., Van Nuys, K., . . . Yatvin, A. L. (2018). Insulin access and affordability working group: Conclusions and recommendations. *Diabetes Care, 41*(6), 1299. doi:10.2337/dci18-0019

Chalew, S., Gomez, R., Vargas, A., Kamps, J., Jurgen, B., Scribner, R., & Hempe, J. (2018). Hemoglobin A1c, frequency of glucose testing and social disadvantage: Metrics of racial health disparity in youth with type 1 diabetes. *J Diabetes Complications, 32*(12), 1085-1090. doi:10.1016/j.jdiacomp.2018.02.008

Chawla, N. V., Bowyer, K. W., Hall , L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. doi:10.1613/jair.953

Chen, E., Looman, M., Laouri, M., Gallagher, M., Van Nuys, K., Lakdawalla, D., & Fortuny, J. (2010). Burden of illness of diabetic macular edema: literature review. *Curr Med Res Opin, 26*(7), 1587-1597. doi:10.1185/03007995.2010.482503

Chiang, J. L., Kirkman, M. S., Laffel, L. M. B., & Peters, A. L. (2014). Type 1 Diabetes Through the Life Span: A Position Statement of the American Diabetes Association. *Diabetes Care, 37*(7), 2034.

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract, 138*, 271-281. doi:10.1016/j.diabres.2018.02.023

Cichosz, S. L., Johansen, M. D., & Hejlesen, O. (2015). Toward Big Data Analytics: Review of Predictive Models in Management of Diabetes and Its Complications. *J Diabetes Sci Technol, 10*(1), 27-34. doi:10.1177/1932296815611680

Contreras, I., & Vehi, J. (2018). Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. *J Med Internet Res, 20*(5), e10775. doi:10.2196/10775

Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., . . . Bellazzi, R. (2018). Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol, 12*(2), 295-302. doi:10.1177/1932296817706375

de Jong, P. E., & Curhan, G. C. (2006). Screening, monitoring, and treatment of albuminuria: Public health perspectives. *J Am Soc Nephrol, 17*(8), 2120-2126. doi:10.1681/ASN.2006010097

Donnelly, R., Emslie-Smith, A. M., Gardner, I. D., & Morris, A. D. (2000). Vascular complications of diabetes. *BMJ, 320*(7241), 1062-1066. doi:10.1136/bmj.320.7241.1062

Driskell, O. J., Holland, D., Waldron, J. L., Ford, C., Scargill, J. J., Heald, A., . . . Fryer, A. A. (2014). Reduced testing frequency for glycated hemoglobin, HbA1c, is associated with deteriorating diabetes control. *Diabetes Care, 37*(10), 2731-2737. doi:10.2337/dc14-0297

Fong, D. S., Aiello, L. P., Ferris, F. L., & Klein, R. (2004). Diabetic Retinopathy. *Diabetes Care, 27*(10), 2540. doi:10.2337/diacare.27.10.2540

Foster, N. C., Beck, R. W., Miller, K. M., Clements, M. A., Rickels, M. R., DiMeglio, L. A., . . . Garg, S. K. (2019). State of Type 1 Diabetes Management and Outcomes from the T1D Exchange in 2016-2018. *Diabetes Technol Ther, 21*(2), 66-72. doi:10.1089/dia.2018.0384

Fowler, M. J. (2008). Microvascular and Macrovascular Complications of Diabetes. *Clinical Diabetes, 26*(2), 77. doi:10.2337/diaclin.26.2.77

Geisser , S. (1993). *Predictive inference: An introduction*. New York, NY: Chapman

and Hall.

Geron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (N. Tache Ed. First Edition ed.). Sebastopol, CA: O'Reilly Media, Inc.

Global report on diabetes. (ISBN 978 92 4 156525 7 (NLM classification: WK 810)). (2016). Geneva, Switzerland: WHO Library Cataloguing-in-Publication Data Retrieved from https://www.who.int/diabetes/global-report/en/

Gordois, A., Scuffham, P., Shearer, A., Oglesby, A., & Tobian, J. A. (2003). The Health Care Costs of Diabetic Peripheral Neuropathy in the U.S. *Diabetes Care, 26*(6), 1790. doi:10.2337/diacare.26.6.1790

Gorst, C., Kwok, C. S., Aslam, S., Buchan, I., Kontopantelis, E., Myint, P. K., . . . Mamas, M. A. (2015). Long-term Glycemic Variability and Risk of Adverse Outcomes: A Systematic Review and Meta-analysis. *Diabetes Care, 38*(12), 2354-2369. doi:10.2337/dc15-1188

Group, T. D. C. a. C. T. R. (1995). The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the diabetes control and complications trial. *Diabetes, 44*(8), 968-983. doi:10.2337/diab.44.8.968

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). *The "big data" revolution in healthcare. Accelerating value and innovation*.

Hardin, J. W. (2005). Generalized estimating equations (GEE). *Encyclopedia of Statistics in Behavioral Science*. doi:doi:10.1002/0470013192.bsa250

10.1002/0470013192.bsa250

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med, 15*(4), 361-387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Hastie, T., Robert., T., & Friedman, J. (2009). *The elements of statistical learning*. Heidelberg, Germany: Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. doi:10.1109/TKDE.2008.239

Herman, W. H., Braffett, B. H., Kuo, S., Lee, J. M., Brandle, M., Jacobson, A. M., . . . Lachin, J. M. (2018). What are the clinical, quality-of-life, and cost consequences of 30 years of excellent vs. poor glycemic control in type 1 diabetes? *J Diabetes Complications, 32*(10), 911-915. doi:10.1016/j.jdiacomp.2018.05.007

Hsieh, M. H., Sun, L. M., Lin, C. L., Hsieh, M. J., Hsu, C. Y., & Kao, C. H. (2019). The Performance of Different Artificial Intelligence Models in Predicting Breast Cancer among Individuals Having Type 2 Diabetes Mellitus. *Cancers (Basel), 11*(11). doi:10.3390/cancers11111751

Hébert, H. L., Veluchamy, A., Torrance, N., & Smith, B. H. (2017). Risk factors for neuropathic pain in diabetes mellitus. *Pain, 158*(4), 560-568. doi:10.1097/j.pain.0000000000000785

Islam, M. S., Hasan, M. M., Wang, X., Germack, H. D., & Noor-E-Alam, M. (2018). A systematic review on healthcare analytics: Application and theoretical perspective of data mining. *Healthcare (Basel), 6*(2). doi:10.3390/healthcare6020054

Jiang, R., Law, E., Zhou, Z., Yang, H., Wu, E. Q., & Seifeldin, R. (2018). Clinical Trajectories, Healthcare Resource Use, and Costs of Diabetic Nephropathy Among Patients with Type 2 Diabetes: A Latent Class Analysis. *Diabetes Ther, 9*(3), 1021-1036. doi:10.1007/s13300-018-0410-8

Jiao, Y., & Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology, 4*(4), 320-330. doi:10.1007/s40484-016-0081-2

Joish, V. N., Zhou, F. L., Preblick, R., Lin, D., Deshpande, M., Verma, S., . . . Pettus, J. (2020). Estimation of Annual Health Care Costs for Adults with Type 1 Diabetes in the United States. *J Manag Care Spec Pharm, 26*(3), 311-318. doi:10.18553/jmcp.2020.26.3.311

Juarez, D. T., Ma, C., Kumasaka, A., Shimada, R., & Davis, J. (2014). Failure to reach target glycated a1c levels among patients with diabetes who are adherent to their antidiabetic medication. *Popul Health Manag, 17*(4), 218-223. doi:10.1089/pop.2013.0099

Kahn, H. S., Morgan, T. M., Case, L. D., Dabelea, D., Mayer-Davis, E. J., Lawrence, J. M., . . . Group, S. f. D. i. Y. S. (2009). Association of type 1 diabetes with month of birth among U.S. youth: The SEARCH for Diabetes in Youth Study. *Diabetes Care, 32*(11), 2010-2015. doi:10.2337/dc09-0891

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

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

Kilpatrick, E. S., Maylor, P. W., & Keevil, B. G. (1998). Biological Variation of Glycated Hemoglobin: Implications for diabetes screening and monitoring. *Diabetes Care, 21*(2), 261. doi:10.2337/diacare.21.2.261

Kuhn , M., & Johnson , K. (2013). *Applied predictive modeling*: Springer-Verlag New York.

Kähm, K., Laxy, M., Schneider, U., & Holle, R. (2019). Exploring Different Strategies of Assessing the Economic Impact of Multiple Diabetes-Associated Complications and Their Interactions: A Large Claims-Based Study in Germany. *Pharmacoeconomics, 37*(1), 63-74. doi:10.1007/s40273-018-0699-1

Lachin, J. M., Bebu, I., Bergenstal, R. M., Pop-Busui, R., Service, F. J., Zinman, B., . . . Group, D. E. R. (2017). Association of glycemic variability in type 1 diabetes with progression of microvascular outcomes in the diabetes control and complications trial. *Diabetes Care, 40*(6), 777-783. doi:10.2337/dc16-2426

Lagani, V., Chiarugi, F., Thomson, S., Fursse, J., Lakasing, E., Jones, R. W., & Tsamardinos, I. (2015). Development and validation of risk assessment models for diabetes-related complications based on the DCCT/EDIC data. *J Diabetes Complications, 29*(4), 479-487. doi:10.1016/j.jdiacomp.2015.03.001

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

Li, Q., & Mao, Y. (2014). A review of boosting methods for imbalanced data classification. *Pattern Analysis and Applications, 17*(4), 679-693. doi:10.1007/s10044-014-0392-8

Lipska, K. J., Hirsch, I. B., & Riddle, M. C. (2017). Human insulin for type 2 diabetes: An effective, less-expensive option. *JAMA, 318*(1), 23-24. doi:10.1001/jama.2017.6939

Luo, J., Avorn, J., & Kesselheim, A. S. (2015). Trends in Medicaid reimbursements for insulin from 1991 through 2014. *JAMA Internal Medicine, 175*(10), 1681-1687. doi:10.1001/jamainternmed.2015.4338

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big Data: The Next Frontier for Innovation, Comptetition, and Productivity*.

Marshall, S. M. (2012). Diabetic nephropathy in type 1 diabetes: has the outlook improved since the 1980s? *Diabetologia, 55*(9), 2301-2306. doi:10.1007/s00125-012-2606-1

Maser, R. E., Mitchell, B. D., Vinik, A. I., & Freeman, R. (2003). The Association Between Cardiovascular Autonomic Neuropathy and Mortality in Individuals With Diabetes. *Diabetes Care, 26*(6), 1895. doi:10.2337/diacare.26.6.1895

Mazzanti, M., Shirka, E., Gjergo, H., & Hasimi, E. (2018). Imaging, Health Record, and Artificial Intelligence: Hype or Hope? *Curr Cardiol Rep, 20*(6), 48. doi:10.1007/s11886-018-0990-y

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115-133. doi:10.1007/BF02478259

Molitch, M. E., DeFronzo, R. A., Franz, M. J., Keane, W. F., Mogensen, C. E., Parving, H. H., . . . Association, A. D. (2004). Nephropathy in diabetes. *Diabetes Care, 27 Suppl 1*, S79-83.

Moltchanova, E. V., Schreier, N., Lammi, N., & Karvonen, M. (2009). Seasonal variation of diagnosis of Type 1 diabetes mellitus in children worldwide. *Diabet Med, 26*(7), 673-678. doi:10.1111/j.1464-5491.2009.02743.x

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA, 309*(13), 1351-1352. doi:10.1001/jama.2013.393

Nalysnyk, L., Hernandez-Medina, M., & Krishnarajah, G. (2010). Glycaemic variability and complications in patients with diabetes mellitus: evidence from a systematic review of the literature. *Diabetes Obes Metab, 12*(4), 288-298. doi:10.1111/j.1463-1326.2009.01160.x

Nathan, D. M., Bebu, I., Hainsworth, D., Klein, R., Tamborlane, W., Lorenzi, G., . . . Group, D. E. R. (2017). Frequency of Evidence-Based Screening for Retinopathy in Type 1 Diabetes. *N Engl J Med, 376*(16), 1507-1516. doi:10.1056/NEJMoa1612836

Nathan, D. M., Kuenen, J., Borg, R., Zheng, H., Schoenfeld, D., Heine, R. J., & Group, A. c.-D. A. G. S. (2008). Translating the A1C assay into estimated average glucose values. *Diabetes Care, 31*(8), 1473-1478. doi:10.2337/dc08-0545

Ontario, H. Q. (2018). Continuous Monitoring of Glucose for Type 1 Diabetes: A Health Technology Assessment. *Ont Health Technol Assess Ser, 18*(2), 1-160.

Orsi, E., Solini, A., Bonora, E., Fondelli, C., Trevisan, R., Vedovato, M., . . . Group, R. I. a. C. E. R. S. (2018). Haemoglobin A1c variability is a strong, independent predictor of all-cause mortality in patients with type 2 diabetes. *Diabetes Obes Metab, 20*(8), 1885-1893. doi:10.1111/dom.13306

Ostman, J., Lönnberg, G., Arnqvist, H. J., Blohmé, G., Bolinder, J., Ekbom Schnell, A., . . . Nyström, L. (2008). Gender differences and temporal variation in the incidence of type 1 diabetes: results of 8012 cases in the nationwide Diabetes Incidence Study in Sweden 1983-2002. *J Intern Med, 263*(4), 386-394. doi:10.1111/j.1365-2796.2007.01896.x

Pasquel, F. J., Hendrick, A. M., Ryan, M., Cason, E., Ali, M. K., & Narayan, K. M. (2015). Cost-effectiveness of Different Diabetic Retinopathy Screening Modalities. *J Diabetes Sci Technol, 10*(2), 301-307. doi:10.1177/1932296815624109

Pearce, I., Simó, R., Lövestam-Adrian, M., Wong, D. T., & Evans, M. (2018). Association between diabetic eye disease and other complications of diabetes: Implications for care. A systematic review. *Diabetes Obes Metab*. doi:10.1111/dom.13550

Pedregosa, F., Varoquaux, G., Alexandre, G., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Peripheral neuropathy fact sheet. (NIH Publication No. 18-NS-4853). (2018). Bethesda, MD 20892: Office of Communications and Public Liaison

National Institute of Neurological Disorders and Stroke

National Institutes of Health Retrieved from https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Peripheral-Neuropathy-Fact-Sheet

Pop-Busui, R., Boulton, A. J., Feldman, E. L., Bril, V., Freeman, R., Malik, R. A., . . . Ziegler, D. (2017). Diabetic Neuropathy: A Position Statement by the American Diabetes Association. *Diabetes Care, 40*(1), 136-154. doi:10.2337/dc16-2042

Ravizza, S., Huschto, T., Adamov, A., Böhm, L., Büsser, A., Flöther, F. F., . . . Petrich, W. (2019). Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nature Medicine, 25*(1), 57-59. doi:10.1038/s41591-018-0239-8

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Risk factors for complications. (2018). Retrieved from https://www.cdc.gov/diabetes/data/statistics-report/risks-complications.html

Russel, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd Edition ed.). Saddle River, NJ: Prentice Hall.

Sacks, D. B. (2011). A1C versus glucose testing: a comparison. *Diabetes Care, 34*(2), 518-523. doi:10.2337/dc10-1546

Sadosky, A., Mardekian, J., Parsons, B., Hopps, M., Bienen, E. J., & Markman, J. (2015). Healthcare utilization and costs in diabetes relative to the clinical spectrum of painful diabetic peripheral neuropathy. *J Diabetes Complications, 29*(2), 212-217. doi:10.1016/j.jdiacomp.2014.10.013

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*: Cambridge University Press.

Skevofilakas, M., Zarkogianni, K., Karamanos, B. G., & Nikita, K. S. (2010). A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus. *Conf Proc IEEE Eng Med Biol Soc, 2010*, 6713-6716. doi:10.1109/IEMBS.2010.5626245

Steyerberg, E. W. (2019). *Clinical prediction models*: Springer, Cham.

Steyerberg, E. W., Eijkemans, M. J. C., Harrell, F. E., & Habbema, J. D. F. (2001). Prognostic Modeling with Logistic Regression Analysis: In Search of a Sensible Strategy in Small Data Sets. *Medical Decision Making, 21*(1), 45-56. doi:10.1177/0272989X0102100106

Tao, B., Pietropaolo, M., Atkinson, M., Schatz, D., & Taylor, D. (2010). Estimating the cost of type 1 diabetes in the U.S.: a propensity score matching method. *PLoS One, 5*(7), e11501. doi:10.1371/journal.pone.0011501

Todd, J. A. (2010). Etiology of type 1 diabetes. *Immunity, 32*(4), 457-467. doi:10.1016/j.immuni.2010.04.001

Tripathi, S. K. (2016). How to increase business efficiency with machine learning. Retrieved from https://www.kelltontech.com/kellton-tech-blog/how-increase-business-efficiency-machine-learning

Type 1 Diabetes. (2019). Retrieved from http://www.diabetes.org/diabetes-basics/type-1/?loc=util-header_type1,%20accessed%2001/29/2019.

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. *PLoS One, 9*(1), e84217. doi:10.1371/journal.pone.0084217

Vergouwe, Y., Soedamah-Muthu, S. S., Zgibor, J., Chaturvedi, N., Forsblom, C., Snell-Bergeon, J. K., . . . Moons, K. G. (2010). Progression to microalbuminuria in type 1 diabetes: development and validation of a prediction rule. *Diabetologia, 53*(2), 254-262. doi:10.1007/s00125-009-1585-3

Viswanathan, V. (2015). Preventing microvascular complications in type 1 diabetes mellitus. *Indian J Endocrinol Metab, 19*(Suppl 1), S36-38. doi:10.4103/2230-8210.155382

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2012). *Regression methods in Biostatistics: Linear, logistic, survival, and repeated measures models* San Francisco, CA, USA: Springer US.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol, 63*(8), 826-833. doi:10.1016/j.jclinepi.2009.11.020

Xu, Q., Wang, L., & Sansgiry, S. S. (2019). A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning. *Journal of Medical Artificial Intelligence; Online First*.

Yarnoff, B. O., Hoerger, T. J., Simpson, S. K., Leib, A., Burrows, N. R., Shrestha, S. S., . . . on behalf of the Centers for Disease Control and Prevention, C. K. D. I. (2017). The cost-effectiveness of using chronic kidney disease risk scores to screen for early-stage chronic kidney disease. *BMC Nephrology, 18*(1), 85. doi:10.1186/s12882-017-0497-6

Zhang, P., Brown, M. B., Bilik, D., Ackermann, R. T., Li, R., & Herman, W. H. (2012). Health utility scores for people with type 2 diabetes in U.S. managed care health plans: results from Translating Research Into Action for Diabetes (TRIAD). *Diabetes Care, 35*(11), 2250-2256. doi:10.2337/dc11-2478

Zhang, P., Engelgau, M. M., Valdez, R., Benjamin, S. M., Cadwell, B., & Narayan, K. M. (2003). Costs of screening for pre-diabetes among US adults: a comparison of different screening strategies. *Diabetes Care, 26*(9), 2536-2542.

Zhou, Z., Chaudhari, P., Yang, H., Fang, A. P., Zhao, J., Law, E. H., . . . Seifeldin, R. (2017). Healthcare Resource Use, Costs, and Disease Progression Associated with Diabetic Nephropathy in Adults with Type 2 Diabetes: A Retrospective Observational Study. *Diabetes Ther, 8*(3), 555-571. doi:10.1007/s13300-017-0256-5