

---

## Research and Applications

# Bias or biology? Importance of model interpretation in machine learning studies from electronic health records

Amanda Momenzadeh <sup>1,2</sup> Ali Shamsa<sup>1</sup>, and Jesse G. Meyer <sup>1,2</sup>

<sup>1</sup>Department of Biochemistry, Medical College of Wisconsin, Milwaukee, Wisconsin, USA, and <sup>2</sup>Department of Computational Biomedicine, Cedars-Sinai, Beverly Hills, California, USA

Corresponding Author: Amanda Momenzadeh, PharmD, Cedars-Sinai, Department of Computational Biomedicine, 121 N San Vicente Blvd, Rm 304, Beverly Hills, CA 90211, USA; [amanda.momenzadeh@gmail.com](mailto:amanda.momenzadeh@gmail.com)

Received 26 March 2022; Revised 27 June 2022; Editorial Decision 30 June 2022; Accepted 4 July 2022

### ABSTRACT

**Objective:** The rate of diabetic complication progression varies across individuals and understanding factors that alter the rate of complication progression may uncover new clinical interventions for personalized diabetes management.

**Materials and Methods:** We explore how various machine learning (ML) models and types of electronic health records (EHRs) can predict fast versus slow onset of neuropathy, nephropathy, ocular disease, or cardiovascular disease using only patient data collected prior to diabetes diagnosis.

**Results:** We find that optimized random forest models performed best to accurately predict the diagnosis of a diabetic complication, with the most effective model distinguishing between fast versus slow nephropathy (AUROC = 0.75). Using all data sets combined allowed for the highest model predictive performance, and social history or laboratory alone were most predictive. SHapley Additive exPlanations (SHAP) model interpretation allowed for exploration of predictors of fast and slow complication diagnosis, including underlying biases present in the EHR. Patients in the fast group had more medical visits, incurring a potential informed decision bias.

**Discussion:** Our study is unique in the realm of ML studies as it leverages SHAP as a starting point to explore patient markers not routinely used in diabetes monitoring. A mix of both bias and biological processes is likely present in influencing a model's ability to distinguish between groups.

**Conclusion:** Overall, model interpretation is a critical step in evaluating validity of a user-intended endpoint for a model when using EHR data, and predictors affected by bias and those driven by biologic processes should be equally recognized.

**Key words:** machine learning, clinical model interpretation, SHAP, diabetes mellitus complication prediction, bias in EHR

---

### LAY SUMMARY

Type 2 diabetes is a major health problem that affects 415 million people worldwide, roughly 1 in 20 people. Diabetes leads to debilitating complications in multiple organ systems, including the heart, eyes, kidneys, and nerves. There is an urgent need to discover factors that delay or promote progression to these complications. One rich source of information that is largely untapped is the electronic health record (EHR), which contains multiple types of information about patients. Demographics, social history, laboratory, prior diagnoses, and vital signs all capture different aspects of a person's health. Machine learning (ML) promises to discover relationships in data and relate them to outcomes without requiring expert knowledge. However, EHR data contain numerous biases, and ML algorithms are known to use biases in data if they are present. It can be hard to know if a ML model has learned something about the underlying disease or if that model is simply memorizing bias. In this work, we present a thorough exploration of what EHR data are most predictive of diabetic complication onset, which type of ML model is best to use, and how to use model interpretation to discover if the model has learned bias or underlying disease biology.

### INTRODUCTION

According to the 2020 National Diabetes Statistics Report, an estimated 34 million (or 13%) of the United States (US) adult population has diabetes,<sup>1</sup> and the prevalence of diagnosed diabetes among US adults is projected to rise to 61 million (or 18%) by the year 2060.<sup>2</sup> Diabetes is the most expensive chronic condition in the United States; one of every 4 US health care dollars is spent on care for people with diabetes.<sup>3</sup> Globally, the direct health expenditure on diabetes in 2019 was \$760 billion, which is projected to rise to \$845 billion in 2045, with the largest expenditure in individuals 60–69 years old.<sup>4</sup> The prevalence of diabetes is highest among adults over 65 years, and the expected rise in diabetes is partially due to a decline in mortality in the diabetes population.<sup>2</sup>

Long-term complications of diabetes are categorized as either microvascular, including nephropathy, neuropathy, and retinopathy, or macrovascular, including cardiovascular and peripheral vascular disease. Diabetes is the leading cause of new cases of blindness and kidney failure in the United States, and was the seventh leading cause of death in 2017.<sup>5</sup> Targeted therapies that delay or inhibit progression of diabetic complications are lacking and there remains a need for a better understanding of the pathophysiology underlying diabetic complications.<sup>6</sup>

Maintaining blood glucose, blood pressure, and cholesterol levels within therapeutic goals is critical to reducing the risk of diabetic-related complications.<sup>7–14</sup> For example, every percentage point reduction in glycosylated hemoglobin (HgbA1c) can reduce the risk for microvascular complications by 40%.<sup>14</sup> However, 21% of US adults who met laboratory criteria for diabetes were unaware of or did not report having diabetes,<sup>1</sup> thus, type 2 diabetes mellitus (T2DM) is often undiagnosed until irreversible complications have developed.<sup>7,14</sup> If detected and treated early, as much as 90% of blindness due to diabetic retinopathy may be preventable.<sup>14</sup> More accurate identification of individuals with T2DM at risk for complications would allow clinicians the chance for early intervention.

Electronic health records (EHR) are a powerful tool in understanding trends in disease development and creating prediction models that allow early interventions or modification of treatment options to improve patient outcomes. With the increasing use of EHR, large-scale patient data have become more accessible. Machine learning (ML) has been a powerful tool aiding in clinical decision-making, identification of patients at risk for diseases (eg, septic shock<sup>15</sup>) as well as repurposing of drugs for new indications.<sup>16</sup> ML algorithms can be trained using a set of patient attrib-

utes (or features) and health outcomes given a clinical scenario, and then used to predict outcomes when provided previously unseen patient profiles. EHR data are highly complex and heterogeneous, and their use in designing a model intended for the real-world clinical setting warrants evaluation of whether the model in fact learned what the user had intended. For example, how do the important features learned by the model align with established risk factors<sup>17</sup>? Most ML studies, however, focus on predictive performance and rarely provide meaningful explanation of their models,<sup>18</sup> that is, patient characteristics that led to the prediction.<sup>19</sup> Due to overwhelming evidence indicting poor reproducibility and reporting of clinical ML models, a 2020 paper made several recommendations for transparent and comprehensible reporting of results from ML studies, including presenting high impact predictors of the model in a summary/tabular format and a narrative focusing on these variables.<sup>20</sup> Additionally, authors should discuss clinical interpretation of these variables with respect to the model outputs, including translation to health care.<sup>20</sup>

SHapley Additive exPlanations (SHAP) is a popular and effective approach published in 2017 for understanding each features' contribution to a model's predictions.<sup>17</sup> SHAP is unique in that it provides insights into the magnitude of importance for a feature as well as the direction a feature shifts a predicted outcome. Of six studies published after 2017 using ML to predict one or more diabetic complications, only one study displayed feature importance results using SHAP.<sup>21</sup> Additionally, this study only summarized the mean absolute SHAP value per feature, forgoing the opportunity to understand whether each feature generally increased or decreased a prediction. The five other studies did not present any analysis of feature importance.<sup>22–26</sup> Studies that do not present model interpretation may be misrepresenting the true output of their model.

In this paper, we describe a study using patient EHR data prior to diagnosis of T2DM to predict a binary outcome of fast versus slow diagnosis of T2DM complications using ML. In other words, at the time a patient is diagnosed with T2DM, can we predict whether that individual will be diagnosed with a diabetic complication faster or slower than 50% of the study population? Our main objectives were to (1) compare the utility of different EHR data types, (2) compare different model architectures, and (3) focus on interpretation of the models. Through SHAP, we identified a models' top predictors, which led us to investigate whether differences in patients' levels of interaction with their healthcare system could be biasing model output. Exploration of the level of medical care received between groups led to the uncovering of an informed pres-

ence bias, namely that those with more medical encounters may have more opportunities to be diagnosed with a complication. Model interpretation is a critical tool that may reveal the effect of inherent biases in models built using EHR data.

## MATERIALS AND METHODS

### Study population

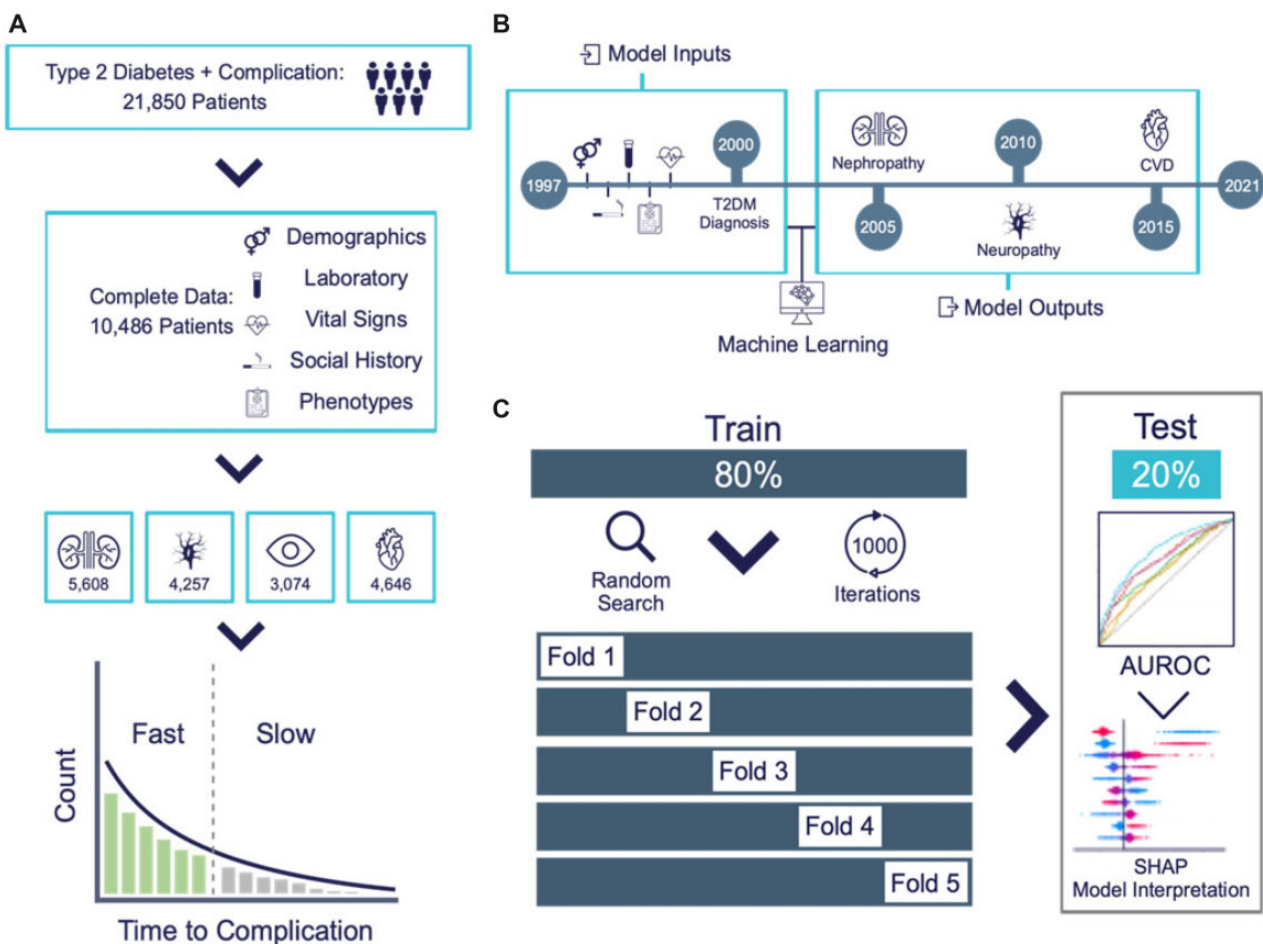
This was a retrospective study across an academic hospital network in the Milwaukee metropolitan area to predict rapid versus delayed diagnosis of diabetic complications in individuals with T2DM. Retrospective, deidentified patient data were queried using the Medical College of Wisconsin (MCW) Clinical Research Data Warehouse using the Froedtert Health System’s Informatics for Integrating Biology and the Bedside (i2b2) Cohort Discovery tool and extracted using the Froedtert Health System Honest Broker. MCW and Froedtert Institutional Review Board (IRB) approval was waived due to the use of deidentified data through the i2b2 Cohort Discovery tool. Data extracted from i2b2 generated 30 854 unique patients with a diabetic complication occurring after the initial T2DM diagnosis.

Extracted data spanned over 24 years from May 1997 to August 2021.

### Data collection

T2DM diagnosis was defined as the date of the first ICD-9 code 250.00 (T2DM without complications) or ICD-10 code E11.9 (T2DM without complications). Due to the transition from ICD-9 to ICD-10 codes in October of 2015,<sup>27</sup> an individual diagnosed prior to 2015 would be coded with 250.00; if the diagnosis occurred after 2015, an E11.9 would have been coded. In order to exclude individuals who had an occurrence of a diabetic complication prior to their first T2DM diagnosis, a temporal query in i2b2 was used to only extract individuals who had a T2DM without complications (250.00 or E11.9) diagnosis that occurred prior to a diabetic complication diagnosis. ICD-9/10 codes and their descriptors used for identifying the occurrence of a diabetic complication are listed in [Supplementary Table S1](#).<sup>22,28</sup>

Patients who had a time-to-complication less than one month were excluded from the study to avoid inclusion of patients who had diabetic complications diagnosed at the same time as their T2DM diagnosis. Per the American Diabetes Association (ADA) guidelines,



**Figure 1.** Flowchart depicting study development and analysis. (A) In total, 21 850 patients with T2DM and at least one diabetic complication occurring at least one month after T2DM diagnosis were selected from the Froedtert & MCW health network i2b2 database and multiple types of EHR were collected for each patient. Patients were divided into groups based on their diabetic complication, then further divided into two groups based on whether they had a time to complication below or above the median. (B) Scheme showing the machine learning task concept with training inputs of EHR data and model outputs for an example patient. (C) Scheme showing the machine learning model training strategy and the model evaluation with the test data.

a one month follow-up visit is advised for diabetes care for all patients with hyperglycemia in the inpatient setting; thus, these individuals would have had a complication code recorded, if present, within their one month appointment.<sup>29</sup> Exclusion of these individuals reduced the cohort size to 21 850 patients (Figure 1A).

The study included data from multiple sources, including demographic information, laboratory results, social-lifestyle history, vital signs, and ICD-9/10 diagnosis codes. The data were linked using deidentified unique encoded patient numbers. Only ICD codes before or on the date of T2DM diagnosis were used as model inputs. ICD-9 and ICD-10 codes were truncated to include codes with only one digit after the decimal due to improved model performance with truncated codes. Total number of codes were reduced from 30 408 to 12 229 after truncation. To unify across all patients, codes were replaced with the corresponding phenotype within the phecode system.<sup>30,31</sup> This also avoided learning unintended associations linked to the longer existence of an ICD-9 versus an ICD-10 code rather than the code itself. Phecodes are distinct diseases or traits that map to ICD-9 or ICD-10 codes as a means to provide consistency across these codes over time as well as overlapping disease states.<sup>32</sup> For example, 401.1 (ICD-9) and I10 (ICD-10) would both map to the phenotype “Essential hypertension.” After mapping, 12 229 unique diagnosis codes were converted to 1721 unique phenotypes.

Demographics information comprised five input features: sex, marital status, employment status, race, and age at diabetes diagnosis (Supplementary Table S2). Demographics information did not change over time.

Because patients had many entries for vitals, social-lifestyle history, and laboratory values, and to simulate how our models might be used in the real world, we used the data collected on the day of the T2DM diagnosis. If the patient did not have data collected on this date, we used the last collected data measured prior to the date of T2DM diagnosis. Vital signs included body mass index (BMI), diastolic blood pressure, systolic blood pressure, pulse, temperature, and respiration rate. Social-lifestyle history consisted of alcohol use, illicit drug use, tobacco use (cigarettes, pipes, and cigars), and smokeless tobacco use (snuff and chew). Laboratory values consisted of aspartate aminotransferase, alanine transaminase, bilirubin, alkaline phosphatase, calcium, glucose, bicarbonate, chloride, sodium, potassium, creatinine, estimated glomerular filtration rate (eGFR), eGFR for African Americans, blood urea nitrogen, anion gap, platelet count, hematocrit, hemoglobin, red blood cell count, white blood cell count, mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume, mean platelet volume (MPV), red cell distribution width (RDW), monocyte percentage, neutrophil percentage, eosinophil percentage, lymphocyte percentage, absolute neutrophil count, absolute lymphocyte count, absolute monocyte count, absolute eosinophil count, total protein, and albumin.

### Data preprocessing

Features were excluded if 50% or more of the values for that feature were missing. MissForest imputation was used to impute missing values.<sup>33</sup> Laboratory and social-lifestyle history variables that were excluded are listed in Supplementary Table S2; no variables from the other three inputs were excluded. Input data were filtered to only include values collected on a visit occurring the day of or prior to the initial T2DM diagnosis in order to mirror a clinical scenario where a clinician only has access to the patient’s baseline health records at the time of T2DM diagnosis. Categorical variables were one-hot encoded,<sup>34</sup> continuous variables were normalized using Min-Max nor-

malization, and counts of phenotypes for each column were binarized (Supplementary Figure S2). Any values  $\pm 3$  standard deviations (SDs) from the mean for a particular feature were set to N/A and then imputed because these values are likely reporting errors. We chose  $\pm 3$  SDs from the mean to define the range for each feature as 99.7% of data occurs within 3 SDs of the mean within a normal distribution. We manually reviewed the range for each variable to ensure values were not too restrictive. For example, the range for BMI that encompassed  $\pm 3$  SDs from the mean was 10.73–61.0 kg/m<sup>2</sup>.

Continuous patient baseline variables were reported as the median (interquartile range) and cohort differences were tested using a 2-sided Mann-Whitney *U* test. Categorical variables were reported as counts (percentages) and compared using chi-square test. Statistical significance was based on a 2-tailed *P* value of  $\leq .05$ .

### Study outcomes

The primary endpoint of the study was classification of a diabetic complication (neuropathy, nephropathy, cardiovascular disease [CVD], or ocular disease) prior to or after the median time to complication (years). Individuals who developed a complication prior to the median time were classified as having fast diagnosis of a complication, those with a time to complication longer than the median were classified as having slow diagnosis. Using the median time as the cut-off between the 2 groups allowed for balanced classification.

### Machine learning

To understand the relative utility of different types of EHR, the six different input data sets collected prior to the date of T2DM (phenotypes, demographics, vital signs, social-lifestyle history, laboratory data, and all inputs combined) were used to train one of six ML classification models (Gradient Boosting Decision Trees [GB], Support Vector Classification [SVC], Random Forest [RF], Extra Trees [ET], Logistic Regression [LR], and Adaptive Boosting [AdaBoost]) to predict one of four diabetic complications. Each model was optimized separately to predict the four complications with one of six input data sets: phenotypes, demographics, vital signs, social-lifestyle history, laboratory data, and all inputs combined. A total of six models were optimized with the six potential inputs for each of the four complications, resulting in a total of 144 model/input/output combinations that were optimized.

Data were split into 20% final test and 80% training data. Using the 80% training split, each model’s hyperparameters were tuned using a random search (RandomizedSearchCV) with 5-fold cross-validation, up to 1000 iterations, and AUROC metric used for scoring. That is, the 80% training data were split into 5-folds and a model was trained with each of those folds as a held-out validation set, meaning up to 5000 models were trained for each ML method. The hyperparameters that maximized the average AUROC values obtained from the random search were used to refit a model on the 80% training data set, and the 20% test set was used to evaluate generalization performance of the best model (Figure 1C). Hyperparameters corresponding to the best model for each input are shown in Supplementary Table S3. AUROC scores reported in this study represent performance of the test set. The input and corresponding model with the best performance for each complication were calibrated via parametric “sigmoid” method and 5-fold cross-validation of the CalibratedClassifierCV class. Model calibration was assessed by plotting calibration curves of the observed versus predicted probabilities for the positive class across 10 evenly partitioned bins. The brier scores for each calibration plot were calcu-

lated using the true class values and the predicted probabilities of the test set.

### Model interpretation

SHAP<sup>17,35</sup> values were used to identify features that contribute most to model prediction. For consistency, the random forest classifier models with all EHR types combined as input were used for SHAP analysis of each complication.

To better understand the relationship between number of medical visits and time of complication diagnosis, we used the patient encounters database to derive individuals' inpatient and outpatient visits between their T2DM and complication diagnoses. The number of each type of medical visit between the T2DM and complication diagnoses divided by patient years was further visualized to assess level of care obtained. Patient years, defined as the sum of the individual patient complication times in each group, was used to account for differences in the total years of follow-up between fast and slow complication diagnosis groups. Between-group differences in diabetic complication illness severity on the day of complication diagnosis were explored by comparing proportion of patients belonging to each stage of chronic kidney disease (CKD) and diabetic retinopathy (DR). CKD included stages 1–5 and DR included mild nonproliferative (NP) DR, moderate NPDR, severe NPDR, and proliferative DR (PDR). Both ICD-9 and 10 codes were used to identify CKD and DR diagnoses.

### Statistical analysis

All data cleaning, analysis, and model training were performed in Python version 3.7.11 (Scikit-Learn,<sup>36</sup> SciPy,<sup>37</sup> SHAP<sup>17</sup>) and R (MisForest<sup>33</sup>).

## RESULTS

### Patient characteristics

Data extracted from i2b2 generated 30 854 unique patients with a diabetic complication occurring after their initial T2DM diagnosis. We then selected 15 987 patients who had complete prediabetes data from all five EHR sources. This data set was further reduced to 10 486 patients who had complete data and at least one diabetic complication. Of these patients, 5608 had nephropathy, 4646 CVD, 4257 neuropathy, and 3074 ocular disease (Figure 1A, Supplementary Figure S1). A patient may have had multiple complications present within the study period.

Key patient characteristics are recorded in Supplementary Tables S4–S7. Known risk factors for diabetic complication progression were distributed unevenly between fast and slow diagnosis groups. Across complications, patients in the slow complication group were diagnosed with T2DM at a significantly younger age and were majority female. The most prevalent race in both groups was Caucasian, followed by African American; the percentage of African Americans was consistently higher in the slow diagnosis group. Across complications, BMI was significantly higher in those with slow diagnosis of nephropathy, neuropathy, and CVD. However, several other patient risk factors for progression of diabetic complications were higher in the fast diagnosis group across all complications, including percentage of patients with essential hypertension, hyperlipidemia, and cigarette use, though not always statistically significant. Random glucose levels were significantly lower in patients with fast diagnosis of nephropathy and ocular disease, and significantly higher in those with fast diagnosis of neuropathy. Ma-

majority of patients did not have smokeless tobacco status recorded in their charts, and the slow group had higher rates of unknown smokeless tobacco status (86.4–87% across complications) compared to the fast group (54.3–64.6% across complications).

Supplementary Figure S3 shows the distribution of time to complication between the two groups. All four complications exhibit skewed distributions with a median of approximately three years. CVD had the shortest time to complication (2.95 years) and neuropathy had the longest time (3.26 years).

### Model performance

Six different input data sets collected prior to the date of T2DM were used to train one of six ML classification models to predict one of four diabetic complications (Figure 1B). Data were split into 80% training and 20% test sets and models were optimized using a random search (Figure 1C). AUROCs across each model and data set combination are shown in Table 1. Using all inputs combine as the model input, RF performed best in predicting nephropathy and neuropathy diagnosis. ET and AdaBoost performed best in predicting CVD and ocular disease diagnosis, respectively. Model calibration was assessed by plotting calibration curves of the observed versus predicted probabilities for the positive class across 10 evenly partitioned bins (Supplementary Figure S4). The brier scores for calibration plots were low, ranging from 0.204 to 0.223, indicating accurate probabilistic predictions.

Figure 2 displays overlaid AUROC plots for each complication with individual lines representing a different data set input. Across all complications, using all data sets combined as an input allowed for the highest model predictive performance compared to using individual data sets alone. Models were most effective in distinguishing between fast versus slow nephropathy diagnosis (AUROC = 0.75) and least effective in distinguishing between fast versus slow CVD diagnosis (AUROC = 0.70). Of the individual data sets, use of social history or laboratory values alone as inputs led to the highest model performance. Using vitals, demographics, or phenotypes alone led to poorer performance. Phenotypes outperformed vitals and demographics in prediction of nephropathy or neuropathy diagnosis, however, the demographics input was the strongest of these three inputs in predicting ocular disease or CVD diagnosis.

### Visualization of feature importance

SHAP was used to investigate how inputs to the model differentially affected the rate of diabetic complication diagnosis (Figure 3). The models predominately leveraged social history and laboratory values in making predictions. The only demographic information that was a top 10 predictor was age at diabetes diagnosis. Phenotype was present only once within the top 10 predictors (ie, hyperlipidemia in predicting ocular disease diagnosis). Vitals, if present, tended to be of lower feature importance.

Known smokeless tobacco use status, higher anion gap, and older age at diabetes diagnosis were associated with a faster diagnosis across all four complications. A lower eGFR and higher MPV were important in predicting fast diagnosis of nephropathy, neuropathy, and CVD, but did not play a role in prediction of ocular disease. Features unique to predicting fast ocular disease diagnosis were a higher monocyte percentage, higher serum calcium, presence of hyperlipidemia, and lower bilirubin. Lower MCHC and higher RDW were associated with faster nephropathy diagnosis. Extended SHAP plots displaying the top 25 predictors are presented in Supplementary Figures S5–S8.



**Table 1.** Test set AUROCs corresponding to each model input using six different ML models for each complication

Complication	Model	Phenotypes	Demographics	Vitals	Social	Labs	All
Nephropathy	SVC	0.618	0.577	0.579	0.674	0.682	0.730
	GB	0.615	0.581	0.585	0.677	0.661	0.736
	ET	<b>0.633</b>	0.559	0.575	<b>0.684</b>	0.674	0.739
	RF	0.625	0.577	<b>0.593</b>	0.670	<b>0.684</b>	<b>0.747</b>
	AdaBoost	0.589	<b>0.593</b>	0.564	0.673	0.665	0.737
	LR	0.612	0.580	0.567	0.671	0.672	0.728
Neuropathy	SVC	0.633	0.579	0.565	0.632	0.648	0.726
	GB	0.629	0.582	0.559	0.666	0.666	0.713
	ET	0.634	0.583	0.574	<b>0.680</b>	0.661	0.732
	RF	<b>0.638</b>	0.582	0.578	0.674	<b>0.671</b>	<b>0.737</b>
	AdaBoost	0.614	<b>0.590</b>	<b>0.583</b>	0.679	0.664	0.727
	LR	0.624	0.583	0.524	0.677	0.645	0.724
Ocular disease	SVC	<b>0.567</b>	0.608	0.507	<b>0.633</b>	0.662	0.649
	GB	0.539	0.609	0.505	0.610	0.645	0.691
	ET	0.530	0.601	0.520	0.632	0.656	0.682
	RF	0.550	0.598	<b>0.527</b>	0.631	0.624	0.696
	AdaBoost	0.502	0.612	0.508	0.599	0.586	<b>0.707</b>
	LR	0.566	<b>0.619</b>	0.473	0.620	<b>0.669</b>	0.671
CVD	SVC	0.609	0.634	0.538	0.633	<b>0.648</b>	0.672
	GB	0.603	<b>0.637</b>	0.516	0.606	0.634	0.694
	ET	0.609	0.616	<b>0.559</b>	<b>0.642</b>	0.623	<b>0.699</b>
	RF	<b>0.613</b>	0.623	0.542	0.599	0.626	0.693
	AdaBoost	0.597	0.633	0.531	0.626	0.618	0.688
	LR	0.603	0.632	0.539	0.619	0.647	0.679

Note: Best AUROCs for each input (phenotypes, demographics, vitals, social-lifestyle history, laboratory, and all inputs combined) are bolded.

SVC: Support Vector Classification; GB: Gradient Boosting Decision Trees; ET: Extra Trees; RF: Random Forest; AdaBoost: Adaptive Boosting; LR: Logistic Regression; CVD: Cardiovascular Disease.

### Medical care and diabetic complication severity

We further investigated patient engagement in medical care and types of visits sought between the time of T2DM and diabetic complication diagnoses between groups. Across all four complications, the fast diagnosis group had significantly more medical visits per year (Figure 4). Average median visits per year between time of T2DM and complication diagnoses across four complications was 27.3 in the fast diagnosis group and 14.0 in the slow diagnosis group. The most frequent types of visits recorded (Figure 5) were outpatient in nature (eg, telephone, office visit, and therapy) compared to visits necessitating a higher level of care (eg, emergency or inpatient hospital encounter). With respect to illness severity on the day of complication diagnosis, the majority of patients were diagnosed during mild to moderate stages of disease for both CKD and DR, not late in disease. The majority of patients were in CKD stage 3 and the proportion of patients between groups was only different for stage 5/ESRD (slow 0.5%, fast 0.1%,  $P = .013$ ; Figure 6A). However, very few patients were in this stage. Proportions of patients in each stage of DR were only different for mild NPDR (slow 19.3%, fast 11.5%,  $P < .001$ ; Figure 6B).

## DISCUSSION

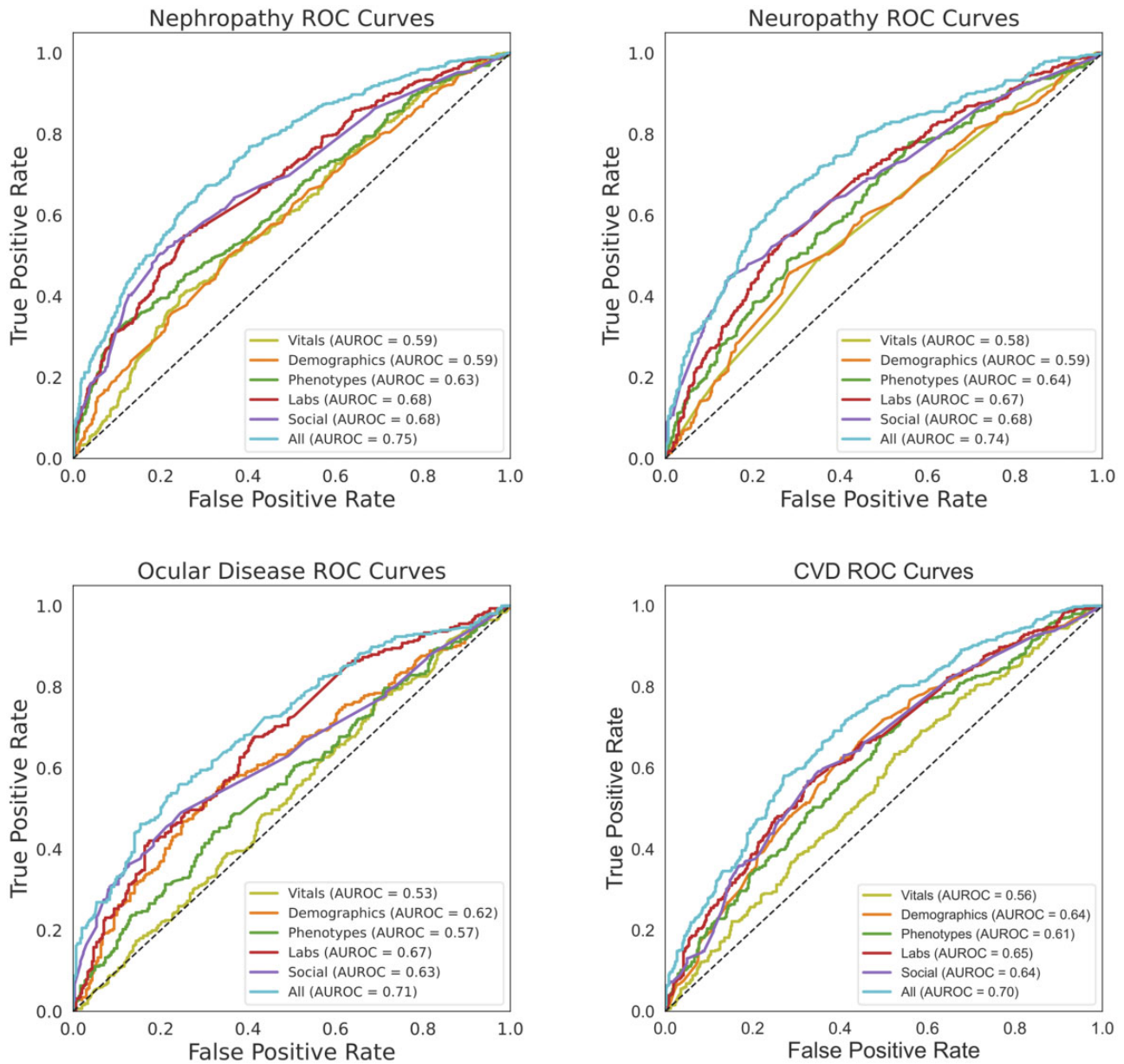
Our study developed well-calibrated models that can predict the fast versus slow diagnosis of a progressive diabetic complication (neuropathy, nephropathy, CVD, or ocular disease). The models performed best in distinguishing between fast and slow diagnosis of nephropathy (AUROC 0.75) and worst in distinguishing fast and

slow diagnosis of CVD (0.70). One strength of the study was the ability of our model to perform with acceptable predictive performance using a smaller cohort relative to similar studies and traditional ML algorithms, which may be more easily implemented in clinical practice requiring less data than deep learning methods.

The combination of all five data sources (vitals, demographics, phenotypes, laboratory, and social history) led to the highest model performance. Of the individual input data sets, social history and laboratory values alone achieved the highest AUROC scores. Laboratory values were most useful in predicting diagnosis of ocular disease and CVD, social history was most useful for predicting diagnosis of neuropathy, and laboratory and social history contributed equally to prediction of diagnosis of nephropathy.

Now we will explore which important variables from SHAP appear to be influenced by bias in the EHR versus disease biology.

A social history variable, smokeless tobacco use, was unexpectedly the most important feature across complications. The proportion of patients with unknown smokeless tobacco use recorded in their charts differed between groups, with the slow group having significantly higher rates of unknown status compared to the fast group (Supplementary Tables S4–S7). Feature sparsity is a known problem that affects the performance of random forest models, and we believe the difference in sparsity of this feature between groups caused the models to preferentially label those with unknown smokeless tobacco as having a slow diagnosis. Many factors determine whether a patient has a measurement recorded in their EHR, and a variable being marked as unknown by a healthcare worker may be due to non-random reasons compared to a variable that is simply missing by chance. For example, is smokeless tobacco use a required question



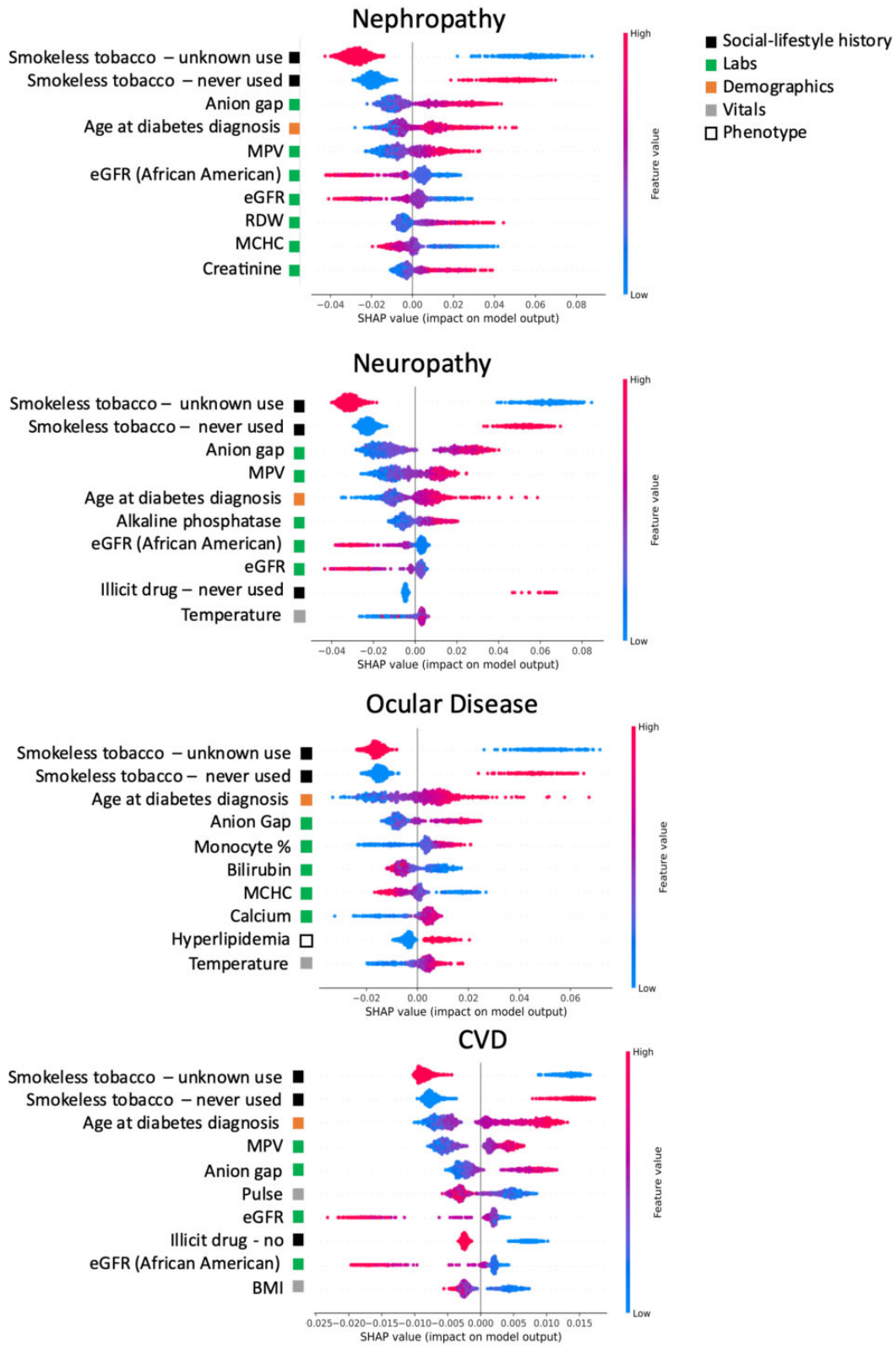
**Figure 2.** Overlaid area under receiver operating characteristic (AUROC) curves representing performance of each data source input for prediction of slow versus fast complication diagnosis. AUROC's corresponding to the best model are plotted for each input. AUROC of 0.5 (diagonal line) corresponds to a model that predicts the output with random chance and 1.0 corresponds to perfect classification.

during all primary care visits? Alternately, are people who interact more with the healthcare system more likely to have a known smokeless tobacco use status?<sup>38</sup> Never users of smokeless tobacco, which was the most dense feature of the three known categories (current, former, and never user), were more likely to have fast diagnosis across complications. Potentially, this variable being marked as a known status could indicate interaction with the healthcare system that generated additional data for that patient. Clinically, lower rates of unknown smokeless tobacco status and higher rates of never users recorded in the charts of individuals with fast complication diagnosis may be an indicator of higher levels of interaction of the fast group with the healthcare system.

Older age at the time of T2DM diagnosis, which was associated with faster diagnosis of all complications, may also be an indicator

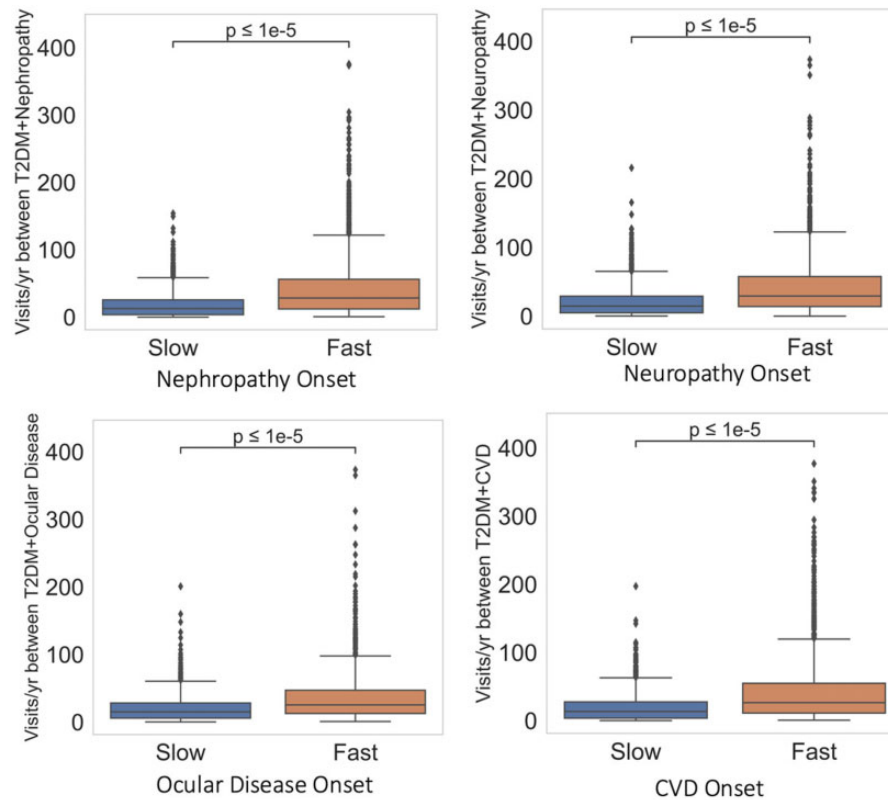
of higher levels of healthcare utilization. On average, older adults have twice as many physicians' office visits compared to those under 65, averaging seven office visits each year.<sup>39</sup> Furthermore, previous studies have established early-onset T2DM to lead to more rapid beta-cell failure and insulin resistance compared to late-onset,<sup>40,41</sup> and large meta-analysis have indicated an inverse relationship between age at diabetes diagnosis and risk of diabetic complications.<sup>42</sup> Since it is established that younger age at diabetes diagnosis is associated with faster onset of complications, the influence of older age at diabetes onset on the models' predictions may be more representative of older patients' higher interaction with the medical care system rather than an underlying biologic process.

Upon an investigation of the level of engagement in medical care, we found the fast diagnosis group had more medical visits (ap-



**Figure 3.** Top 10 features visualized using SHAP. Corresponding data source from which feature is derived is indicated by colored square. Individual patient contributions to the outcome are signified with red dots (high feature values), purple (intermediate), and blue (low). Y axis represents importance of each feature. Dots with x values greater than and less than zero represent patients with a fast and slow complication diagnoses, respectively. MPV: mean platelet volume, eGFR: estimated glomerular filtration rate, RDW: red cell distribution width, MCHC: mean corpuscular hemoglobin concentration, BMI: body mass index.





**Figure 4.** Boxplots comparing number of patient medical visits per year between T2DM and complication diagnoses between fast and slow complication diagnosis group. Horizontal line within each box represents median and the box spans the interquartile range (IQR), extending from the first (Q1) to the third (Q3) quartile for each group's distribution. Box whiskers denote maximum ( $Q3 + 1.5 \times IQR$ ) and minimum ( $Q1 - 1.5 \times IQR$ ); dots outside of whiskers are outliers. Horizontal bar denotes  $P$  value using 2-sided Mann-Whitney  $U$  test.

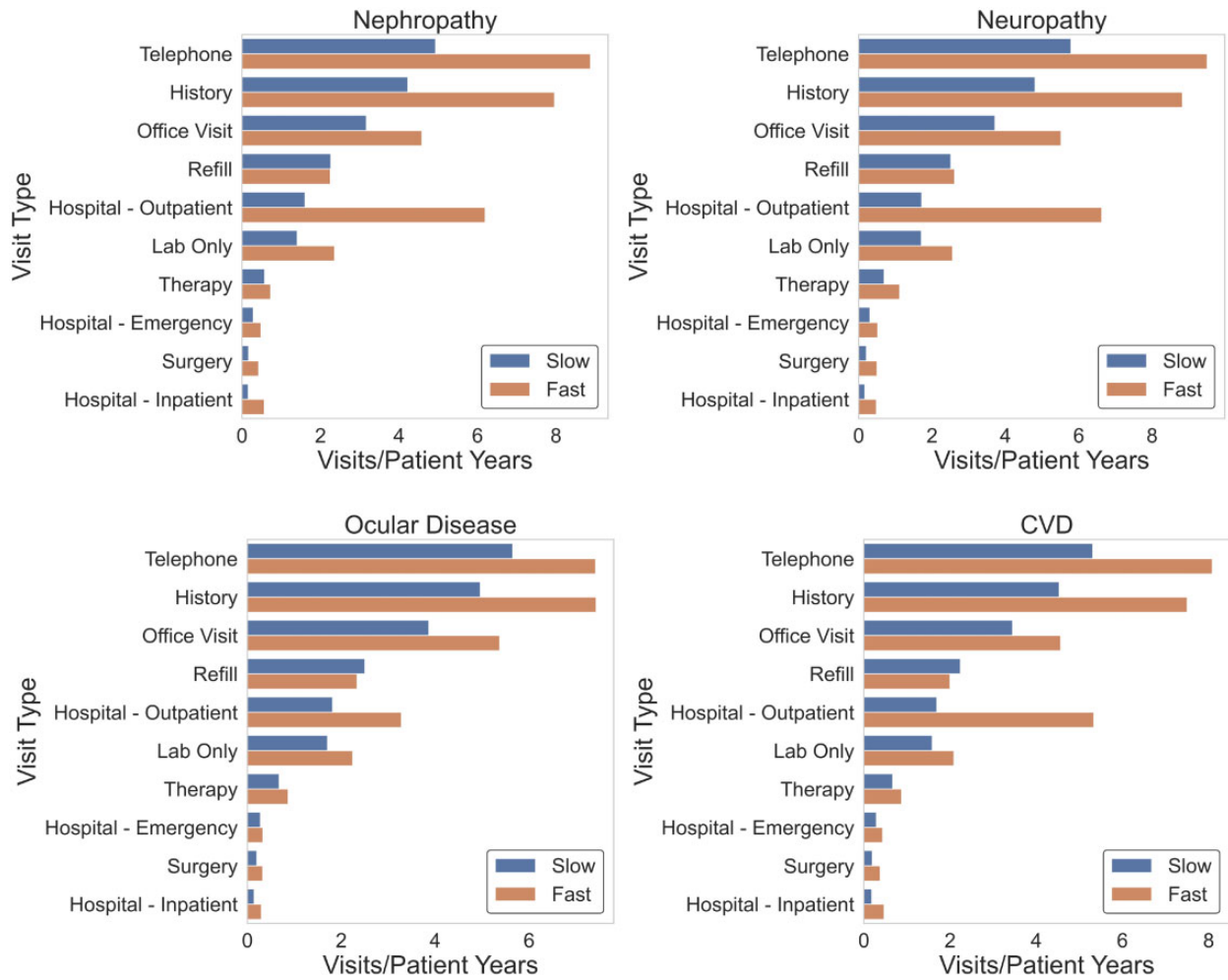
proximately biweekly compared to monthly in the slow diagnosis group) and the majority of visits were within the ambulatory setting. Despite high specificity for accurate diagnosis of a disease, ICD codes are known to have low sensitivity; in other words the presence of a code is a likely indicator of a disease, however, the absence of a code does not reliably indicate absence of that disease.<sup>43</sup> The ICD-9 codes for diabetes with complications (250.1–250.9), for example, have a sensitivity of 63.6% and specificity of 99%.<sup>44</sup> Therefore, since a diagnosis is not captured with high probability, those with more medical encounters are more likely to have the presence of diabetes with complication detected.<sup>45</sup> It is possible the fast group is being diagnosed sooner with a complication due to higher engagement in their healthcare system, not necessarily due to their faster progression of disease. Additionally, a deeper dive into complication severity at the time of complication diagnosis did not show a clear difference between groups, and the majority of patients in both groups were captured within the EHR during mild to moderate stages of their disease. So, the fast group may not be sicker in terms of diabetic complication severity. Due to an ICD code likely representing patient catchment rather than true disease onset, we use the term “diagnosis” rather than “progression” to represent our model's output; previous ML studies have claimed to measure progression without model interpretation may have mischaracterized the patterns learned by their models.

Well-established predictors with biologically validated links to diabetic complication progression included: (1) lower eGFR (or reduced kidney function) was linked to faster diagnosis of nephropathy, (2) higher anion gap (or increase in ketoacids in uncontrolled

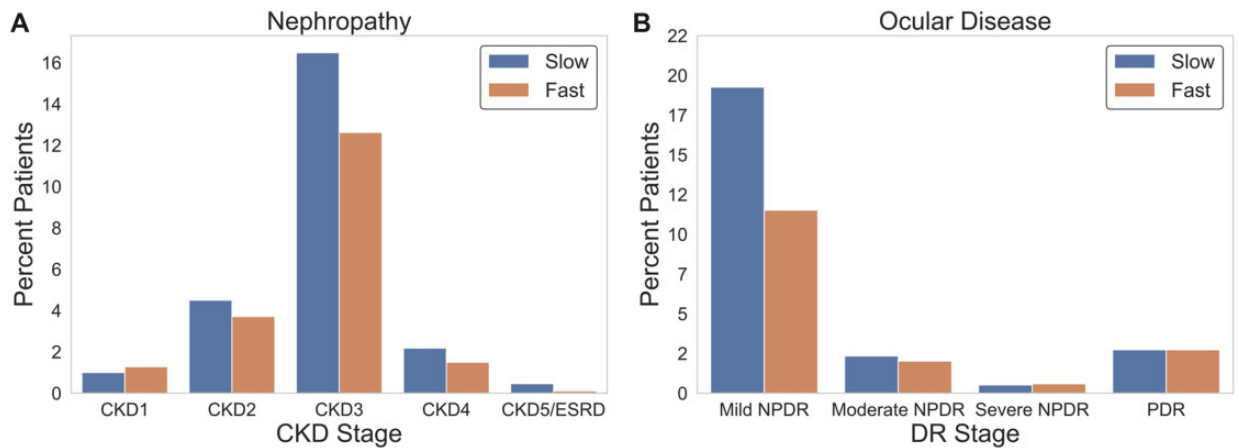
diabetes<sup>46</sup>) was linked to faster diagnosis of all four complications, and (3) hyperlipidemia (an established risk factor for diabetic retinopathy<sup>14</sup>) was linked to faster diagnosis of ocular disease.

Other predictors that did not appear intuitive at first, but have biologic association with diabetic complication progression were also explored. First, higher MPV was associated with faster diagnosis of nephropathy, neuropathy, and CVD in our study. A higher MPV is indicative of larger, more aggregable platelets that produce more procoagulants, leading to thrombogenesis, atherosclerosis, and production of oxidative substances that cause local vascular lesions.<sup>47,48</sup> Small studies have shown that MPV was higher in patients with uncontrolled T2DM ( $HbA1c > 7\%$ ) compared to those with controlled T2DM ( $HbA1c \leq 7\%$ ) and was associated with higher risk of developing diabetic complications.<sup>49,50</sup> Furthermore, improved glycemic control led to recovery in platelet activity.<sup>51</sup> Overall, high MPV is associated with vascular damage in diabetics, and we may be able to prevent this damage through optimizing blood glucose control.

Second, individuals with lower bilirubin, lower MCHC, and higher serum calcium had faster diagnosis of ocular disease in our study. High levels of bilirubin, a breakdown product of hemoglobin,<sup>52</sup> may have the potential to protect against diabetic complication development by suppressing oxidation of lipids and lipoproteins.<sup>53</sup> Several studies, including a meta-analysis of 27 studies, have shown low levels of bilirubin were inversely related to the development of diabetic complications, including retinopathy.<sup>53–55</sup> Next, several observational studies have shown that low hemoglobin levels may accelerate microvascular damage in diabetes. Low he-



**Figure 5.** Barchart exhibiting types of medical visit obtained by fast and slow complication diagnosis groups between T2DM and complication diagnoses. Number of different medical visits per patient year was visualized to assess differences in level of care obtained between groups. Patient year was defined as the sum of the individual times to complication within each group.



**Figure 6.** Barchart exhibiting percent patients belonging to each stage of (A) chronic kidney disease (CKD) and (B) diabetic retinopathy (DR) on the day of nephrology and ocular disease diagnosis, respectively. CKD included stages 1–5 and DR included mild nonproliferative (NP) DR, moderate NPDR, severe NPDR, and proliferative DR (PDR). Both ICD-9 and 10 codes were used to identify CKD and DR diagnoses.

moglobin concentrations are more common in diabetic patients than nondiabetics and hyperglycemia has been shown to decrease red cell survival by 13%.<sup>56</sup> There may be an increased risk of severe diabetic retinopathy in individuals with hemoglobin levels below 12 g/dL,<sup>57,58</sup> although this association diminished after adjusting for diabetes duration in another study.<sup>56</sup> Lastly, a cross-sectional study of over 3000 patients found elevated serum calcium to be a risk factor for vision-threatening diabetic retinopathy,<sup>59</sup> and in vivo histology of the retina revealed elevated serum calcium was associated with retinal photoreceptor apoptosis in diabetic retinopathy.<sup>60</sup> Low bilirubin and MCHC and high serum calcium in T2DM may be indicators of accelerated retinal damage in diabetics, providing clinicians with more personalized information for monitoring and modulating diabetes complication progression.

Third, our findings that higher RDW and lower MCHC are associated with faster diagnosis of nephropathy are also supported by existing studies. RDW, which measures the volume and size of red blood cells, is commonly used to help diagnose different types of anemia.<sup>61</sup> A retrospective study of individuals with biopsy-proven diabetic nephropathy showed that individuals with higher RDW had an increased risk of progression to end-stage renal disease.<sup>61</sup> Diabetic patients with low hemoglobin concentration had more rapid decline in glomerular filtration,<sup>62</sup> and anemia was a risk factor for progression to end stage renal failure.<sup>63</sup> High RDW and low MCHC may be important markers for progression of kidney injury in diabetics.

This study has important limitations. Controlling for the number of inpatient encounters may have been a potential solution to remove the informed decision bias (ie,  $\geq n$  visits to be eligible into the cohort), however, this may incur a selection bias as individuals with fewer visits would be excluded.<sup>43,45</sup> Diagnosis data have acceptable quality due to mandates requiring accurate collection of this data, and certain demographic data (ie, age, gender, and ethnicity/race) are mandated by the Meaningful Use objectives.<sup>64</sup> However, there is no mandated coding system for the remaining nonessential demographic (eg, income, marital status, education), laboratory, vital sign, or social data for EHRs.<sup>64</sup> Social data obtained from EHRs are often of low quality due to incomplete patient responses and the subjective nature of most questions.<sup>64</sup> For example, across complications, the majority of patients had “unknown” smokeless tobacco use. In addition, there may be several laboratory codes in the EHR representing one lab of interest, and challenges arise when different facilities use different lab tests to measure the same analyte.<sup>64</sup> Several labs were excluded due to high missingness in the data set ( $\geq 50\%$ ), including glucose measurements that appeared duplicative (ie, glucose point-of-care and point-of-care whole blood glucose). Random glucose was the only diabetes monitoring lab available with acceptable missingness, however, as this was not specified as collected in a fasted state, this may not have been a reliable measure of true T2DM control. Next, vital signs documented in the EHR may be flawed by human error in recording units of measurement. Another important limitation is the use of diagnosis codes to define the onset of diabetic complications, however, this is our only available proxy for diagnosis information with large scale data and is in line with previous studies.<sup>21,22</sup> Lastly, data were collected from facilities across a single health network, so it is possible the models focused on features that are not as common or important in other institutions. The models should be implemented on a larger scale across different institutions to verify reproducibility.

In this study, ML models were able to accurately predict the diagnosis of 1 of 4 diabetic complications: neuropathy, nephropathy,

ocular disease, and CVD. SHAP provides an interpretation of key features’ contribution to each model, which was affected by a mix of both bias present in the data and biologic pathways that affect true disease progression. Our study is unique in the realm of ML studies as it aimed to explore the predictions learned by a model when given complex EHR data. Predictors affected by bias should not go unrecognized and may be just as important as those driven by biologic processes. In conclusion, SHAP may serve as a critical starting point for evaluating validity of disease prediction models using EHR data.

## FUNDING

This work was supported by startup funds from the Medical College of Wisconsin, the National Center for Research Resources, and the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number UL1TR001436. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. This research was completed in part with computational resources and technical support provided by the Research Computing Center at the Medical College of Wisconsin.

## AUTHOR CONTRIBUTIONS

Conceptualization: AM, AS, JGM. Methodology: AM, AS, JGM. Software: AM, AS, JGM. Formal analysis: AM, AS, JGM. Data curation: AM, AS, JGM. Writing—original draft: AM, AS, JGM. Writing—review and editing: AM, AS, JGM. Visualization: AM, AS, JGM. Supervision: JGM. Project administration: JGM. Funding acquisition: JGM.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

## ACKNOWLEDGMENTS

We thank Eduard Puig for graphic design help. We thank the reviewers for their helpful comments.

## CONFLICT OF INTEREST STATEMENT

None declared.

## MATERIALS AND CORRESPONDENCE

Amanda Momenzadeh

## DATA AVAILABILITY

All code is available from github: <https://github.com/amandamomenzadeh/ML-Biology-or-Bias>.

## REFERENCES

- Centers for Disease Control and Prevention. *National Diabetes Statistics Report, 2020*. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2020.

2. Lin J, Thompson TJ, Cheng YJ, *et al.* Projection of the future diabetes burden in the United States through 2060. *Popul Health Metr* 2018; 16 (1): 9.
3. The Cost of Diabetes | ADA. <https://www.diabetes.org/resources/statistics/cost-diabetes>. Accessed February 2022.
4. Williams R, Karuranga S, Malanda B, *et al.* Global and regional estimates and projections of diabetes-related health expenditure: results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res Clin Pract* 2020; 162: 108072.
5. US Preventive Services Task Force. Screening for prediabetes and type 2 diabetes: US preventive services task force recommendation statement. *JAMA* 2021; 326: 736–43.
6. Kantharidis P, Wang B, Carew RM, Lan HY. Diabetes complications: the microRNA perspective. *Diabetes* 2011; 60 (7): 1832–7.
7. American Diabetes Association. Standards of medical care in diabetes. *Diabetes Care* 2005; 28: s4–36.
8. Diabetes Control and Complications Trial Research Group; Nathan DM, Genuth S, Lachin J, *et al.* The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 1993; 329: 977–86.
9. UK Prospective Diabetes Study (UKPDS) Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* 1998; 352: 837–53.
10. UK Prospective Diabetes Study (UKPDS) Group. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). *Lancet* 1998; 352: 854–65.
11. Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group; Lachin JM, Genuth S, Cleary P, *et al.* Retinopathy and nephropathy in patients with type 1 diabetes four years after a trial of intensive therapy. *N Engl J Med* 2000; 342: 381–9.
12. Lawson ML, Gerstein HC, Tsui E, Zinman B. *Effect of Intensive Therapy on Early Macrovascular Disease in Young Individuals with Type 1 Diabetes: A Systematic Review and Meta-Analysis. Database of Abstracts of Reviews of Effects (DARE): Quality-Assessed Reviews [Internet]*. York, United Kingdom: Centre for Reviews and Dissemination; 1999.
13. Stratton IM. Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (UKPDS 35): prospective observational study. *BMJ* 2000; 321 (7258): 405–12.
14. Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. *Phys Ther* 2008; 88 (11): 1254–64.
15. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med* 2015; 7 (299): 299ra122.
16. Taubes A, Nova P, Zalocusky KA, *et al.* Experimental and real-world evidence supporting the computational repurposing of bumetanide for APOE4-related Alzheimer's disease. *Nat Aging* 2021; 1 (10): 932–47.
17. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc.; 2017: 4768–4777.
18. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak* 2019; 19 (1): 146.
19. Lundberg SM, Nair B, Vavilala MS, *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; 2 (10): 749–60.
20. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. *Circ Cardiovasc Qual Outcomes* 2020; 13 (10): e006556.
21. Ravaut M, Sadeghi H, Leung KK, *et al.* Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *NPJ Digit Med* 2021; 4 (1): 24.
22. Thomas PB, Robertson DH, Chawla NV. Predicting onset of complications from diabetes: a graph based approach. *Appl Netw Sci* 2018; 3 (1): 1–16.
23. Ljubic B, Hai AA, Stanojevic M, *et al.* Predicting complications of diabetes mellitus using advanced machine learning algorithms. *J Am Med Inform Assoc* 2020; 27 (9): 1343–51.
24. Makino M, Yoshimoto R, Ono M, *et al.* Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Sci Rep* 2019; 9 (1): 11862.
25. Kim E, Caraballo PJ, Castro MR, Pieczkiewicz DS, Simon GJ. Towards more accessible precision medicine: building a more transferable machine learning model to support prognostic decisions for micro- and macrovascular complications of type 2 diabetes mellitus. *J Med Syst* 2019; 43 (7): 185.
26. Song X, Waitman LR, Yu AS, *et al.* Longitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine: retrospective cohort study. *JMIR Med Inform* 2020; 8 (1): e15510.
27. The transition to ICD-10 before October 1 compliance deadline. *The Bulletin*. <https://bulletin.facs.org/2015/06/the-transition-to-icd-10-before-october-1-compliance-deadline/>. 2015.
28. [ftp.cdc.gov/pub/Health\\_Statistics/NCHS/Publications/ICD10CM/2021/](ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/2021/). [https://ftp.cdc.gov/pub/Health\\_Statistics/NCHS/Publications/ICD10CM/2021/](https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/2021/). Accessed February 2022.
29. American Diabetes Association. Diabetes care in the hospital: standards of medical care in diabetes—2019. *Diabetes Care* 2019; 42: S173–81.
30. PheWAS – Phenome Wide Association Studies. [https://phewascatalog.org/phecodes\\_icd10cm](https://phewascatalog.org/phecodes_icd10cm). Accessed February 2022.
31. PheWAS – Phenome Wide Association Studies. <https://phewascatalog.org/phecodes>. Accessed February 2022.
32. Wu P, Gifford A, Meng X, *et al.* Developing and evaluating mappings of ICD-10 and ICD-10-CM codes to phecodes. 462077 <https://www.biorxiv.org/content/10.1101/462077v2>; 2018. doi: 10.1101/462077. Accessed February 2022.
33. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012; 28 (1): 112–8.
34. Okada S, Ohzeki M, Taguchi S. Efficient partition of integer optimization problems with one-hot encoding. *Sci Rep* 2019; 9 (1): 13036.
35. Lundberg SM, Erion G, Chen H, *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020; 2 (1): 56–67.
36. Pedregosa F, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–30.
37. Virtanen P, Gommers R, Oliphant TE, *et al.*; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; 17 (3): 261–72.
38. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. *JAMA Network Open* 2021; 4(2): e210184.
39. Institute of Medicine (US) Committee on the Future Health Care Workforce for Older Americans. *Health Status and Health Care Service Utilization*. Washington, DC: National Academies Press; 2008.
40. Song, SH, Hardisty CA. Early-onset type 2 diabetes mellitus: an increasing phenomenon of elevated cardiovascular risk. Expert review of cardiovascular therapy. <https://pubmed.ncbi.nlm.nih.gov/18327993/>; 2008. doi:10.1586/14779072.6.3.315. Accessed February 2022.
41. Type 2 diabetes mellitus in youth: the complete picture to date – ScienceDirect. <https://www.sciencedirect.com/science/article/abs/pii/S0031395500132X?via%3Dihub>. Accessed February 2022.
42. Nanayakkara N, Curtis AJ, Heritier S, *et al.* Impact of age at type 2 diabetes mellitus diagnosis on mortality and vascular complications: systematic review and meta-analyses. *Diabetologia* 2021; 64 (2): 275–87.
43. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol* 2021; 21 (1): 234.
44. Khokhar B, Jette N, Metcalfe A, *et al.* Systematic review of validated case definitions for diabetes in ICD-9-coded and ICD-10-coded data in adult populations. *BMJ Open* 2016; 6 (8): e009952.
45. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *Am J Epidemiol* 2016; 184 (11): 847–55.

46. Solomon AE-Y, Wankasi MM, Ileimokumo O. Relationship between serum anion gap and diabetes mellitus. *J Diabetes Mellit* 2015; 5: 199–205.
47. Kodiatte TA, Manikyam UK, Rao SB, *et al.* Mean platelet volume in type 2 diabetes mellitus. *J Lab Physicians* 2012; 4 (1): 5–9.
48. Kakourou N, Rade JJ, Kourliouros A, Resar JR. Platelet function in patients with diabetes mellitus: from a theoretical to a practical perspective. *Int J Endocrinol* 2011; 2011: 742719.
49. Radha RKN, Selvam D. MPV in uncontrolled & controlled diabetics – its role as an indicator of vascular complication. *J Clin Diagn Res* 2016; 10 (8): EC22–6.
50. Bali Medical Journal Published by DiscoverSys Inc. 1–10. <https://www.balimedicaljournal.org/index.php/bmj/article/view/806>; 2018. doi:10.15562/bmj.v7i3.806. Accessed February 2022.
51. Demirtunc R, Duman D, Basar M, *et al.* The relationship between glycaemic control and platelet activity in type 2 diabetes mellitus. *J Diabetes Complicat* 2009; 23 (2): 89–94.
52. Link Between Serum Bilirubin and Diabetic Retinopathy in Type 2 Diabetes Patients. Diabetes in control. A free weekly diabetes newsletter for medical professionals. <https://www.diabetesincontrol.com/link-between-serum-bilirubin-and-diabetic-retinopathy-in-type-2-diabetes-patients/>; 2017.
53. Zhu B, Wu X, Ning K, Jiang F, Zhang L. The negative relationship between bilirubin level and diabetic retinopathy: a meta-analysis. *PLoS One* 2016; 11 (8): e0161649.
54. Yasuda M, Kiyohara Y, Wang JJ, *et al.* High serum bilirubin levels and diabetic retinopathy: the Hisayama study. *Ophthalmology* 2011; 118: 1423–8.
55. Karuppanasamy D, Venkatesan R, Thankappan L, Andavar R, Devisundaram S. Inverse association between serum bilirubin levels and retinopathy in patients with type 2 diabetes mellitus. *J Clin Diagn Res* 2017; 11: NC09–12.
56. Chung JO, Cho DH, Chung DJ, Chung MY. Associations between hemoglobin concentrations and the clinical characteristics of patients with type 2 diabetes. *Korean J Intern Med* 2012; 27 (3): 285–92.
57. Qiao Q, Keinänen-Kiukaanniemi S, Läärä E. The relationship between hemoglobin levels and diabetic retinopathy. *J Clin Epidemiol* 1997; 50 (2): 153–8.
58. Traveset A, Rubinat E, Ortega E, *et al.* Lower hemoglobin concentration is associated with retinal ischemia and the severity of diabetic retinopathy in type 2 diabetes. *J Diabetes Res* 2016; 2016: 3674946.
59. Hu Y, Zhou C, Shi Y, *et al.* A higher serum calcium level is an independent risk factor for vision-threatening diabetic retinopathy in patients with type 2 diabetes: cross-sectional and longitudinal analyses. *Endocr Pract* 2021; 27 (8): 826–33.
60. Ankita Saxena S, Nim DK, *et al.* Retinal photoreceptor apoptosis is associated with impaired serum ionized calcium homeostasis in diabetic retinopathy: an in-vivo analysis. *J Diabetes Complicat* 2019; 33: 208–11.
61. Zhang J, Zhang R, Wang Y, *et al.* The association between the red cell distribution width and diabetic nephropathy in patients with type-2 diabetes mellitus. *Renal Failure* 2018; 40 (1): 590–6.
62. Rossing K, Christensen PK, Hovind P, *et al.* Progression of nephropathy in type 2 diabetic patients. *Kidney Int* 2004; 66 (4): 1596–605.
63. Cusick M, Chew EY, Hoogwerf B, *et al.*; Early Treatment Diabetic Retinopathy Study Research Group. Risk factors for renal replacement therapy in the Early Treatment Diabetic Retinopathy Study (ETDRS), Early Treatment Diabetic Retinopathy Study Report No. 26. *Kidney Int* 2004; 66 (3): 1173–9.
64. Ehrenstein V, Kharrazi H, Lehmann H, Taylor CO. *Obtaining Data From Electronic Health Records*. Rockville, MD: Agency for Healthcare Research and Quality; 2019.