Title

Developing and Validating Models to Predict Progression to Proliferative Diabetic Retinopathy

Journal

Ophthalmology Science, 3(2)

ISSN

2666-9145

Authors

Guo, Yian

Yonamine, Sean

Jian, Chu

et al.

Publication Date

2023-06-01

DOI

10.1016/j.xops.2023.100276

Peer reviewed

AMERICAN ACADEMY OF OPHTHALMOLOGY®

Developing and Validating Models to Predict Progression to Proliferative Diabetic Retinopathy

Yian Guo, MS,[1,2] Sean Yonamine, BA,[1,3] Chu Jian Ma, MD, PhD,[1] Jay M. Stewart, MD,[1] Nisha Acharya, MD, MS,[1,2] Benjamin F. Arnold, PhD,[1,2] Charles McCulloch, PhD,[4] Catherine Q. Sun, MD[1,2]

Purpose: To develop models for progression of nonproliferative diabetic retinopathy (NPDR) to proliferative diabetic retinopathy (PDR) and determine if incorporating updated information improves model performance.

Design: Retrospective cohort study.

Participants: Electronic health record (EHR) data from a tertiary academic center, University of California San Francisco (UCSF), and a safety-net hospital, Zuckerberg San Francisco General (ZSFG) Hospital were used to identify patients with a diagnosis of NPDR, age $\geq$ 18 years, a diagnosis of type 1 or 2 diabetes mellitus, $\geq$ 6 months of ophthalmology follow-up, and no prior diagnosis of PDR before the index date (date of first NPDR diagnosis in the EHR).

Methods: Four survival models were developed: Cox proportional hazards, Cox with backward selection, Cox with LASSO regression and Random Survival Forest. For each model, three variable sets were compared to determine the impact of including updated clinical information: Static_0 (data up to the index date), Static_{6m} (data updated 6 months after the index date), and Dynamic (data in Static_0 plus data change during the 6-month period). The UCSF data were split into 80% training and 20% testing (internal validation). The ZSFG data were used for external validation. Model performance was evaluated by the Harrell's concordance index (C-Index).

Main Outcome Measures: Time to PDR.

Results: The UCSF cohort included 1130 patients and 92 (8.1%) patients progressed to PDR. The ZSFG cohort included 687 patients and 30 (4.4%) patients progressed to PDR. All models performed similarly (C-indices $\sim$ 0.70) in internal validation. The random survival forest with Static_{6m} set performed best in external validation (C-index 0.76). Insurance and age were selected or ranked as highly important by all models. Other key predictors were NPDR severity, diabetic neuropathy, number of strokes, mean Hemoglobin A1c, and number of hospital admissions.

Conclusions: Our models for progression of NPDR to PDR achieved acceptable predictive performance and validated well in an external setting. Updating the baseline variables with new clinical information did not consistently improve the predictive performance.

Financial Disclosure(s): Proprietary or commercial disclosure may be found after the references.

Supplemental material available at www.ophthalmologyscience.org.

Vision loss from diabetic retinopathy remains the leading cause of preventable blindness in working-aged adults in the United States.[1] In many patients, blindness associated with diabetic retinopathy can be prevented with appropriate and timely diagnosis and treatment.[2] There have been many studies aimed at identifying risk factors associated with progression of diabetic retinopathy. Large population-based studies and randomized controlled trials, such as the landmark ETDRS in the 1980s, informed us of the risk factors for diabetic retinopathy progression and helped to establish guidelines for screening and management.[3-10] However, there have been improvements in screening, diagnosis, and treatment for diabetes and diabetic retinopathy since many of those studies were first conducted, and the risk factors for diabetic retinopathy progression should be revisited.

In 2013, a study using health care claims data found that 6.7% of newly diagnosed nonproliferative diabetic retinopathy (NPDR) patients progressed to proliferative diabetic retinopathy (PDR) with a median follow-up time of 1.7 years.[11] Using a similar method to the ETDRS paper, they determined the risk factors for progression to PDR at 5 years were hemoglobin A1c (HbA1c), diabetic nephropathy, and non-healing ulcers.[11] The investigators acknowledged intrinsic limitations to claims data including no clinical examination information (i.e., vitals, laboratory data, eye examination), no information about disease severity and duration using International Classification of Diseases (ICD)

https://doi.org/10.1016/j.xops.2023.100276
ISSN 2666-9145/23

9 codes, and the inclusion of only insured patients.[11] Electronic health record (EHR) data can help overcome many of these limitations. There has been significant adoption of EHRs in the United States in the last decade; they are used in nearly 90% of outpatient physician offices.[12] Concurrently, newer statistical and machine-learning techniques have been developed, allowing researchers to better utilize and analyze the massive amount of EHR data available.[13,14]

The goal of this study was to use time-to-event models with systemic and ocular data from EHRs to predict progression from NPDR to PDR. In contrast to treating the primary outcome as a dichotomous outcome, as in the majority of predictive models for diabetic retinopathy, we used time to PDR as the primary outcome so that time information in the outcome was preserved and utilized, since time to progression is indicative of levels of risk. Furthermore, we wanted to determine if incorporating updated clinical information into the model could improve model performance. Incorporation of updated clinical information helps to simulate disease control and fluctuation to better represent the natural course of disease. The goal is that these models will help identify patients at high risk of progression and allow for earlier intervention.[15] We hypothesized that incorporating updated clinical information could improve model performance for prediction of progression to PDR.

Methods

Data Sources

We obtained data from the EHRs of two hospitals: University of California San Francisco (UCSF), a tertiary academic center, and Zuckerberg San Francisco General (ZSFG) Hospital, a safety-net hospital. The UCSF EHR transitioned to Epic in June 2012. Data were accessed from UCSF's de-identified clinical data warehouse, which is based on the Epic Caboodle Data Warehouse and is updated monthly. Dates are shifted by up to 365 days in the de-identified clinical data warehouse and protected health information is removed according to the Safe Harbor Method. The ZSFG EHR system for ambulatory care transitioned to Epic in August 2019. Zuckerberg San Francisco General data were queried from Epic Clarity by the UCSF Clinical and Translational Science Institute. The UCSF data were last accessed on February 8, 2022 using SQL (SQLPro for MSSQL), and the ZSFG data were extracted on December 15, 2021. The Institutional Review Board at UCSF approved this study and issued a waiver of informed consent for all subjects. This study followed the tenets of the Declaration of Helsinki.

Cohorts

Patients with a diagnosis of diabetic retinopathy based on ICD-9 and 10 codes (Table S1, available at www.ophthalmologyscience.org/) who had $\geq$ 1 completed, in-person visit with an eye provider (optometrist or ophthalmologist) at UCSF after June 1, 2012 or at ZSFG after August 1, 2019 were selected (Fig 1A). We included patients who were age $\geq$ 18 years, had $\geq$ 1 coded diagnosis of type 1 or 2 diabetes mellitus (DM), a diagnosis of NPDR coded by an eye provider, and $\geq$ 6 months of eye follow-up data before progression or censoring. Patients with a prior diagnosis of PDR before the index date (the date of first NPDR diagnosis in the EHR) were excluded.

Disease Outcome

The primary event was the progression from NPDR to PDR. Diagnosis of PDR was defined by ICD-9 or 10 code after the index date and had to be coded by an eye provider. The primary outcome was the time from the index date to the first PDR diagnosis for patients who progressed, and time from the index date to the last ophthalmology follow-up visit for patients who were censored.

Variables

We extracted variables from the following categories that were available in the EHR: demographics, eye-related diagnoses and procedures, systemic comorbidities, laboratory data, vital signs, medications, and health care utilization (Table 2). The variables were classified as time-constant or time-varying (Table 2 and Table S3, available at www.ophthalmologyscience.org/). For medications, procedures, and systemic diagnoses, we created indicator variables based on the first occurrence, except for myocardial infarction and stroke for which both indicators and counts were created. Due to the high percentage of missing data in laboratory values and vital signs (e.g., HbA1c, blood pressure), these variables were categorized with missingness as one of the subgroups.

To determine the impact of including updated information, we used three variable sets that reflected different time-related data (Fig 1B). The Static_0 set contained variables that used data up to the index date. The Static_{6m} set contained the same variables that were updated with data up to 6 months after the index date (referred to as 6-month timepoint). The Dynamic set contained the same variables as in Static_0 (up to the index date) plus the time-varying variables that represented the change from the index date to 6 months (Table S3, available at www.ophthalmologyscience.org/).

Model Training and Validation

For each of the 3 variable sets, 4 survival models were constructed and trained: Cox proportional hazards regression (Cox), Cox with backward selection (Cox-BW), Cox with LASSO regression (Cox-LS), and random survival forest (RSF). For the Cox model, variables were selected if they were significant ($P < 0.05$) in univariable Cox or if they had known clinical significance in previous studies (i.e., sex, tobacco use, DM type, presence of nephropathy, presence of neuropathy).[3,11] For the other 3 models, we allowed the models to perform feature selection on all variables. For Cox-BW, we used Akaike information criteria for backward selection (equivalent to dropping a variable when $P \geq 0.157$). For Cox-LS, we utilized a group lasso method to ensure either all or no sublevels of a multicategorical variable were being selected.[16]

The UCSF cohort was randomly split into a training set (80%) and a test set (20%) with the same event and censoring proportion. To train parameters in Cox-LS (regularizing parameter lambda) and RSF (number of trees, number of random splits, number of variables randomly selected for splitting a node and minimum size of terminal nodes), a 10-by-10-fold cross-validation was applied. The training set were divided into 10 folds and 10-fold cross-validation was conducted. This process was repeated 10 times, each time partitioning the training data into different 10 folds, and the evaluation performance were averaged. Tuning of the parameters was performed using grid search.

Harrell's concordance index (C-index) was used to assess the predictive performance of the models.[17] The C-index computes the proportion of all usable patient pairs in which the predictions and outcomes are concordant and ranges from 0 to 1, with 1

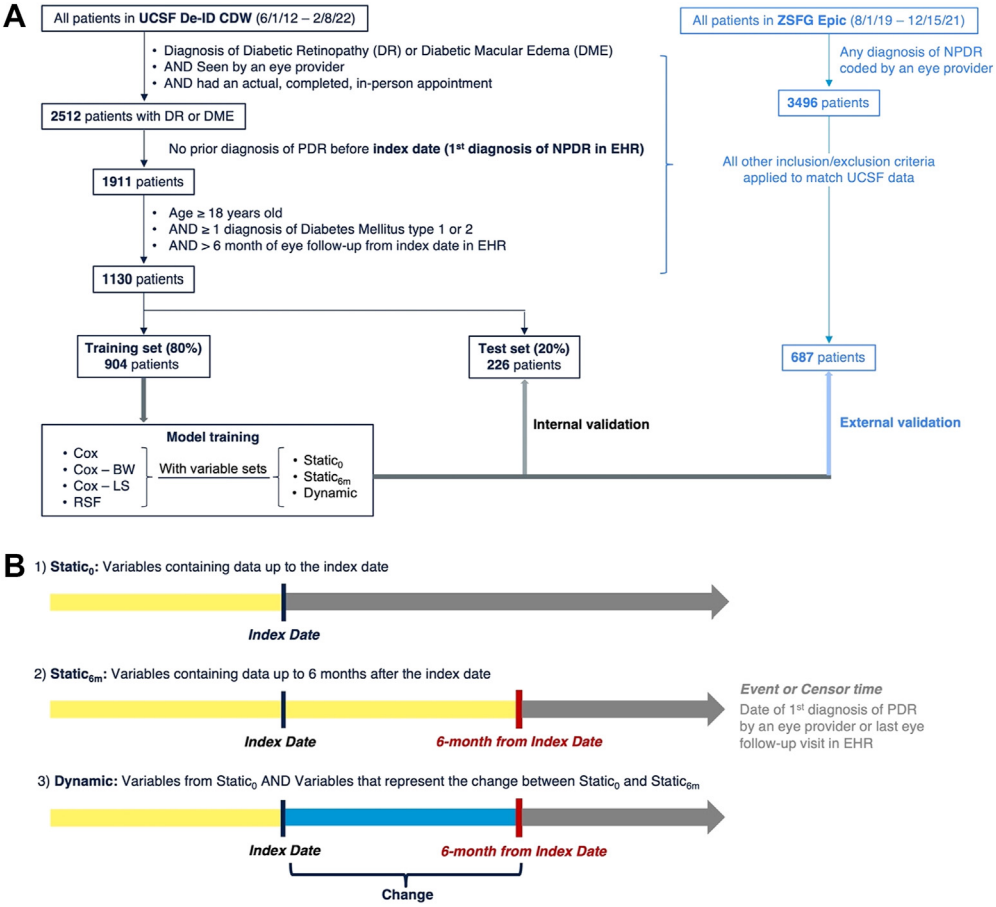

Figure 1. Filtering process for development of University of California San Francisco (UCSF) and Zuckerberg San Francisco General (ZSFG) Proliferative Diabetic Retinopathy (PDR) cohorts with **A,** depiction of inclusion/exclusion criteria, model training, internal/external validation, and **B,** description of variable sets used in prediction models. Cox-BW = Cox with backward selection; Cox-LS = Cox with LASSO regression; De-CWD = de-identified clinical data warehouse; EHR = electronic health record; NPDR = nonproliferative diabetic retinopathy; RSF = random survival forest.

indicating perfect prediction. Another common performance measure for survival prediction is Integrated Brier Score. It gives the integrated squared prediction error over time, and is usually weighted by inverse probability of censoring weights to account for censored observations.[18] Because our event proportion was low, the Integrated Brier Score was nearly indiscernible between model sets during training, and we chose to use C-index as our primary measure of performance.

The selected models from training were internally validated on the UCSF 20% test set. In addition, to assess the robustness and

Table 2. Full List of Variables Used for Model Development

Demographics	**Comorbidity**	**Vitals, Laboratory Values**	**Health care Utilization**
Age at index date	Tobacco use‡	Mean hemoglobin A1c *	No show visit (any, N)*,†
Sex‡	Diabetes mellitus type	Mean body mass index*	Ophthalmology no show visit (any, N)*,†
Race/ethnicity	Hypertension*	Mean systolic blood pressure*	Outpatient encounter (N)*
Insurance*	Hyperlipidemia*		Ophthalmology outpatient encounter (N)*
Eye-Related	Diabetic Nephropathy*,†	**Medications**	Hospital admission (any, N)*,†
Intravitreal injection (any)*	Diabetic Neuropathy*,‡	Insulin use*	Emergency room visit (any, N)*,†
Diabetic macular edema (any)*,‡	Diabetic foot ulcer*	Antihyperglycemic use*	
NPDR severity*	Myocardial infarction (any, N)*,†	ACE inhibitor use*	
	Stroke (any, N) *,†	Statin use*	

ACE = angiotensin-converting enzyme; N = number; NPDR = Non-Proliferative Diabetic Retinopathy.
*Time-varying variables
†Both indicators (any vs. none) and the actual counts (N) were created for these variables.
‡These variables were included due to known clinical significance in previous studies.

generalizability of our prediction models, trained models were evaluated on the entire ZSFG cohort for external validation.

Calibration refers to the extent of bias of a model (i.e., the absolute difference between the predicted and observed probabilities). To assess the calibration of the models, the Kaplan-Meier curve was compared with the predicted survival probability over time.[19] We averaged the predicted survival probability over all patients in the UCSF and ZSFG test sets separately at the observed times and plotted it alongside the actual Kaplan-Meier curve. In addition, we used the median of the linear predictors from the training set as the cut-off point to divide the patients in the test sets into high-risk and low-risk groups. The predicted survival curve and the Kaplan-Meier curve of each risk group were also compared.

Statistical Analysis

The data characteristics between groups were compared using Wilcoxon Rank Sum test for continuous variables or Pearson's chi-squared test for categorical variables. The proportional hazard assumption of the Cox models and Cox-BW models was checked by examining the slope of Schoenfeld residuals against time. The C-index was compared by the method of Kang et al[20] and was adjusted for multiple comparison using the Benjamini-Hochberg method.[21] To evaluate the predictive ability of variables, we looked at variable importance from RSF as well as the number of times they were selected by Cox-BW and Cox-LS. The Breiman-Cutler method for variable importance is associated with the change in prediction error of the original forest grown using a variable compared with a new forest grown without the variable.[22] Large variable importance values indicate variables with strong predictive ability. All analyses were done in R 4.1.0. We used the following packages for model development: 'survival', 'grpreg' and 'randomForestSRC'.

Results

Cohort Characteristics

A total of 1130 patients (median [interquartile range] age 66 [56−75] years, 563 [50%] female) were included in the UCSF cohort after inclusion and exclusion criteria were applied. The ZSFG cohort had 687 patients (median [interquartile range] age 64 [57−71] years, 324 [47%] female; Table 4). The UCSF cohort had a longer median event/censor time of 37 months compared with 17 months for the ZSFG cohort. The rates of PDR progression per person-year were 0.022 for the UCSF cohort and 0.031 for the ZSFG cohort. University of California San Francisco data were available from 2012 whereas ZSFG was only available from 2019. There were differences in sociodemographic factors and disease severity between the UCSF cohort compared with the ZSFG cohort. Specifically, in the UCSF cohort, there were more White patients (33% versus [vs.] 11%), fewer patients of Hispanic descent (14% vs. 38%), fewer patients who were self-pay or had Medicaid insurance (16% vs. 41%), and fewer patients with moderate (9.6% vs. 25%) or severe NPDR (2.9% vs. 7.9%) respectively.

Model Results

Random survival forest with the updated 6-month variable set (RSF_{6m}), and the Cox and Cox-BW with the dynamic variable set ($Cox_{dynamic}$, $Cox\text{-}BW_{dynamic}$) achieved a C-index of 0.70 on the UCSF test set (Table 5). However, they were not significantly different from other models after correcting for multiple comparisons. The best-performing model on the ZSFG cohort was the RSF with updated 6-month variables (RSF_{6m}, C-index 0.76), and it was significantly better than Cox_{6m}, $Cox\text{-}BW_{dynamic}$, and $Cox\text{-}LS_{dynamic}$ (adjusted *P*-values: 0.024, 0.024, 0.024) respectively. Results from the univariable Cox model fit using the UCSF training set are shown in Table S6 (available at www.ophthalmologyscience.org/).

Table 7 lists the top 10 variables ranked by variable importance from the RSF_{6m} model. As shown in Table 7, across models with the updated 6-month variable set, NPDR severity, insurance, and age were consistently predictive of progression to PDR. Among all variables, age and insurance were selected by all Cox models and ranked in the top 10 by variable importance for all RSF models (Table S8, available at www.ophthalmologyscience.org/). Other frequently selected variables were NPDR severity, diabetic neuropathy, number of strokes, mean HbA1c, and number of hospital admissions.

Calibration

The calibration plots of the RSF_{6m} model are shown in Figure 2. To facilitate comparison, we truncated the time span of the UCSF test set to be the same as the ZSFG cohort (28 months). The overall calibration of the ZSFG cohort appeared worse than the UCSF test set. After stratification, the RSF_{6m} model predicted the UCSF low and high-risk groups well, but underpredicted the low-risk group and overpredicted the high-risk group of the ZSFG cohort after 18 months.

Discussion

In this study, we developed survival models for predicting progression from NPDR to PDR using variables at different timepoints, and evaluated the model using EHR data from two different health systems. The C-indices ranged from 0.66 to 0.70 for the UCSF test set (internal validation), and 0.61 to 0.76 for external validation on the ZSFG data, indicating acceptable predictive ability and reasonable generalizability of the models. Updating the variables at 6 months after the index date or including the change in variables during a 6-month period did not consistently improve predictive performance compared with using variables at the index date. Among all variables, age and insurance were selected by all Cox models and ranked in the top 10 by variable importance for all RSF models. Other important predictors included NPDR severity, diabetic neuropathy, number of strokes, mean HbA1c, and number of hospital admissions.

In general, our survival models had good generalizability when evaluated on a second health care system that uses the same EHR system, Epic. The model performance on the ZSFG cohort was comparable or better than the UCSF test set with the $Static_0$ and $Static_{6m}$ variable sets across all 4 types of models. However, the Dynamic models performed worse in external validation than in internal validation, possibly because the Dynamic models

Table 4. Cohort Characteristics

Characteristic	UCSF, N = 1130*	ZSFG, N = 687*	*P*-value†
Time to PDR or last ophthalmology follow-up (months)	37 (19, 65)	17 (12, 23)	< 0.001
Number of patients who progressed to PDR	92 (8.1%)	30 (4.4%)	0.002
Rate of PDR progression per person-year	0.022	0.031	0.133
Age (years)	66 (56, 75)	64 (57, 71)	0.004
Sex			0.271
Female	563 (50%)	324 (47%)	
Male	567 (50%)	363 (53%)	
Race			< 0.001
White	374 (33%)	75 (11%)	
Asian	350 (31%)	229 (33%)	
Black/African American	130 (12%)	84 (12%)	
Hispanic/Latino	154 (14%)	261 (38%)	
Other	122 (11%)	38 (5.5%)	
Insurance			< 0.001
Medicare	536 (47%)	231 (34%)	
PPO/HMO	262 (23%)	71 (10%)	
Self-Pay/Medicaid	178 (16%)	281 (41%)	
Other	154 (14%)	104 (15%)	
Diabetes mellitus Type			< 0.001
Type 1	106 (9.4%)	7 (1.0%)	
Type 2	1024 (91%)	680 (99%)	
NPDR severity			< 0.001
Mild	443 (39%)	201 (29%)	
Moderate	108 (9.6%)	169 (25%)	
Severe	33 (2.9%)	54 (7.9%)	
Unspecified	408 (36%)	53 (7.7%)	
Missing	138 (12%)	210 (31%)	
Diabetic macular edema			< 0.001
No	486 (43%)	274 (40%)	
Yes	178 (16%)	174 (25%)	
Missing	466 (41%)	239 (35%)	
Mean hemoglobin A1c			< 0.001
≤ 6.5	123 (11%)	84 (12%)	
6.5–8	381 (34%)	131 (19%)	
8–10	267 (24%)	114 (17%)	
> 10	89 (7.9%)	68 (9.9%)	
Missing	270 (24%)	290 (42%)	

HMO = Health Maintenance Organization; NPDR = non-proliferative diabetic retinopathy; PDR = proliferative diabetic retinopathy; PPO = Preferred Provider Organization; UCSF = University of California San Francisco; ZSFG = Zuckerberg San Francisco General Hospital.
*Median (interquartile range); n (%).
†Wilcoxon rank sum test; Pearson's chi-squared test.

included more variables than the $Static_0$ and $Static_{6m}$ models (the extra variables representing change during the 6-month period), making it less generalizable to the external dataset.

An objective of this study was to determine if incorporating updated information into the model could improve model performance. By comparing the results of the $Static_{6m}$ and Dynamic models to the $Static_0$ models, we

Table 5. Model Performances Evaluated by C-indices on the UCSF Test Set and the ZSFG Cohort

	UCSF test set			ZSFG cohort		
	$Static_0$	*$Static_{6m}$*	*Dynamic*	*$Static_0$*	*$Static_{6m}$*	*Dynamic*
Cox proportional hazards regression	0.68	0.69	0.70	0.72	0.67*	0.70
Cox with backward selection	0.67	0.66	0.70	0.68	0.74	0.61*
Cox with LASSO regression	0.67	0.68	0.68	0.72	0.72	0.65*
Random survival forest	0.69	0.70	0.67	0.73	0.76	0.65

UCSF = University of California San Francisco; ZSFG = Zuckerberg San Francisco General Hospital.
*The C-index of random survival forest with the $Static_{6m}$ set was significantly better than these models in external validation (multiple comparison adjusted).

Table 7. The Top 10 Variables for Predicting Progression to PDR Ranked by Variable Importance from the RSF_{6m} Model, and the Associated HRs (95% CIs) in Cox_{6m} and Cox-BW_{6m}, and HR in Cox-LS_{6m}

Variable	RSF_{6m} Rank	Cox_{6m} HR	Cox_{6m} (95% CI)	Cox-BW_{6m} HR	Cox-BW_{6m} (95% CI)	Cox-LS_{6m} HR
NPDR severity (Reference: mild)	1					
Moderate		2.99	(1.37–6.53)	3.60	(1.73–7.49)	2.25
Severe		2.54	(0.84–7.66)	2.84	(1.05–7.63)	2.33
Unspecified		2.07	(0.98–4.39)	1.87	(0.97–3.61)	1.32
Missing		1.50	(0.51–4.43)	1.34	(0.53–3.39)	1.11
Age (years)	2	0.97	(0.95–1.00)	0.96	(0.94–0.99)	0.98
Insurance (Reference: Medicare)	3					
PPO/HMO		2.55	(1.19–5.48)	2.29	(1.08–4.88)	2.00
Self-Pay/Medicaid		2.58	(1.23–5.42)	2.81	(1.31–6.03)	2.28
Other		0.54	(0.15–1.97)	0.57	(0.16–2.08)	0.77
Stroke (N)	4	–	–	1.09	(1.01–1.19)	1.01
Hospital admission (N)	5	1.09	(0.931.28)	1.17	(0.96–1.44)	1.08
Emergency room visit (N)	6	–	–	–	–	–
Ophthalmology outpatient encounter (N)	7	–	–	1.06	(1.01–1.12)	1.04
Mean hemoglobin A1c (Reference: $\leq$ 6.5)	8					
6.5–8		0.47	(0.19–1.21)	0.60	(0.24–1.51)	0.77
8–10		0.52	(0.20–1.36)	0.61	(0.23–1.61)	0.78
> 10		1.43	(0.53–3.83)	1.99	(0.74–5.36)	1.72
Missing		0.67	(0.21–2.09)	1.82	(0.70–4.69)	1.13
Diabetic foot ulcer (Reference: no)	9	1.44	(0.47–4.46)	–	–	1.10
Neuropathy (Reference: no)	10	2.01	(1.02–3.98)	1.79	(0.96–3.34)	1.56

CI = confidence interval; Cox_{6m} = Cox model with the updated 6-month variable set; Cox-BW_{6m} = Cox model with backward selection and the updated 6-month variable set; Cox-LS_{6m} = Cox model with LASSO regression and the updated 6-month variable set; HR = hazard ratio; HMO = Health Maintenance Organization; NPDR = nonproliferative diabetic retinopathy; PPO = Preferred Provider Organization; RSF_{6m} = random survival forest model with the updated 6-month variable set.

found that updating the variables at 6 months from the index date or including the change of variables during the 6-month period did not consistently improve the predictive performance in either the internal or external datasets. One reason might be that 6 months is not long enough for time-varying variables to change. We were not able to extend this time-point to 1 year as it would further reduce our sample size and number of events. In the future, a more extended updated time point of $\geq$ 1 year may demonstrate more distinction between time-varying models.

Our models identified several key predictors of PDR progression that were not shown in previous studies, which we will discuss in further detail here: insurance status, number of previous strokes, and number of previous hospital admissions. Insurance status was selected by all our models, indicating high predictive ability. While insurance status has not been specifically identified as a risk factor of progression from NPDR to PDR, other studies investigating different aspects of diabetic retinopathy care have found insurance-based disparities to be a key risk factor in outcomes. Malhotra et al[23] found that patients on Medicare and private insurance presented with better baseline visual acuity compared with patients on Medicaid when initiating treatment with anti-VEGF therapy for diabetic macular edema. In addition, Cai et al[24] identified that not having a regular primary care provider and having poor housing conditions were associated with poor adherence to diabetic eye examinations. Similarly, Hinkle et al[25] found that patients with Medicaid had reduced odds for following up with an eye provider after an emergent visit for PDR. In our study, both patients with commercial insurance (Preferred Provider Organization/Health Maintenance Organization) and self-pay/Medicaid had higher rates of progression to PDR compared with those with Medicare, after adjusting for variables including age and NPDR severity. We conjecture that there may be other variables that were not captured in our study, such as socioeconomic status, that may be contributing. Given the substantial impact of insurance status on progression to PDR, this is a key area of future investigation to prevent PDR and improve outcomes. Potential future directions include incorporating social determinants of health variables available in the EHR and investigating how different insurance categories impact disease outcome, as well as using smaller subcategories of insurance type, such as separating Preferred Provider Organization from Health Maintenance Organization and self-pay from Medicaid.

The number of previous strokes was another important predictor that was identified in our model. A few recently published papers have found an association between strokes and diabetic retinopathy but there is not conclusive evidence of the relationship.[26] A meta-analysis of 19 cohort studies found that the presence of diabetic retinopathy is associated with an increased risk of stroke in type 2 DM patients, but uncertain for patients with type 1 DM.[27] In a secondary analysis of patients enrolled in the ACCORD Eye study (Action to Control Cardiovascular Risk in Diabetes), diabetic retinopathy was associated with an increased risk of stroke.[26] While these two studies identified diabetic retinopathy as a predictor for the development of strokes,[26,27] our study found that the number of previous

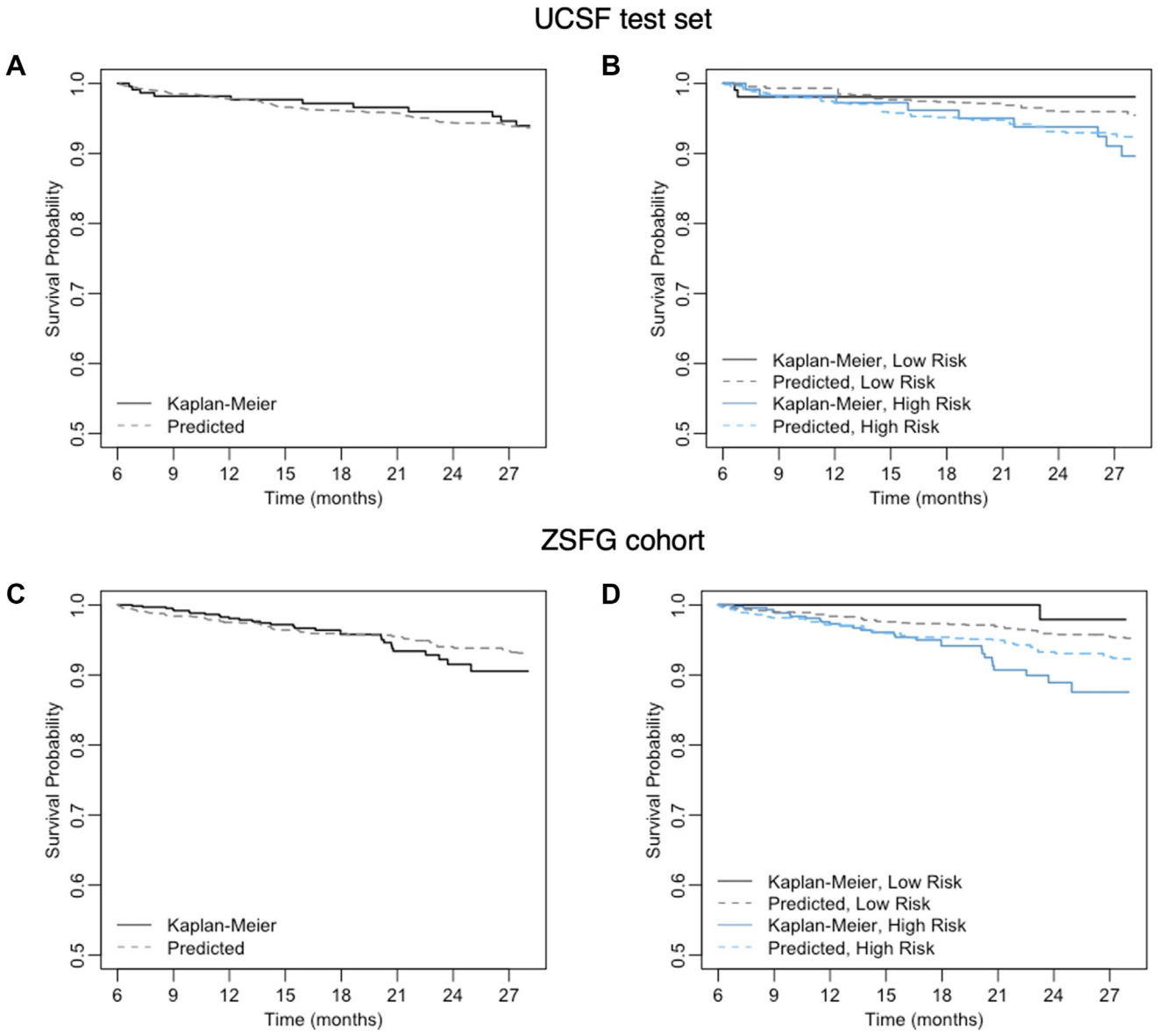

Figure 2. Calibration of survival prediction by the random survival forest model with updated 6-month variable set. Calibration is the absolute difference between the predicted and observed probabilities. **A,** Total University of California San Francisco (UCSF) test set. **B,** UCSF test set stratified into low-risk and high-risk groups. **C,** Total Zuckerberg San Francisco General Hospital (ZSFG) cohort. **D,** ZSFG cohort stratified into low-risk and high-risk groups. Solid lines indicate Kaplan-Meier survival estimates. Dashed lines indicate predicted mean survival probabilities.

strokes was also a predictor for progression of NPDR to PDR. These findings indicate that the microvascular pathology of diabetic retinopathy may have larger macrovascular implications, and needs to be further explored.

The number of previous hospital admissions was also an important predictor. Another study identified having more systemic comorbidities to be a risk factor for diabetic retinopathy progression using the Carlson Comorbidity Index score.[28] In our study, more hospital admissions likely indicate more comorbidities. While the finding that worse overall health and comorbidities lead to progression of PDR is not surprising, our use of the number of previous hospital admissions may be a comparable and simpler surrogate marker to capture overall health status for larger data studies. Additional research should be conducted in this area.

There are several limitations to this study. First, there was a high percentage of missing data for vitals and laboratory values, such as HbA1c, in both the UCSF and ZSFG cohorts. Imputation of a large number of missing values across multiple variables would have resulted in bias, especially when variables such as vitals and laboratory values were not missing at random. Therefore, to avoid significantly shrinking our cohort size, we binned the continuous vitals and laboratory values and included missingness as one of the subcategories. This method may have resulted in some loss of power.[29] Given even larger amounts of missing data in less

common laboratory values, we chose to not include other variables that have previously been shown to be related to PDR progression, such as hematocrit and serum albumin.[3]

Second, clinical notes and eye examination data were not available to us for analysis. As such, we did not include variables such as visual acuity or duration of diabetes that have been reported to be significant in previous studies.[3,30] Given the significance of duration of DM in previous studies, we attempted to use the first coded diagnosis of DM in the EHR to determine DM disease duration. However, in the UCSF cohort, only 39% of patients had their first DM diagnosis after their first encounter in the system. Since the EHR system was implemented in 2012 at UCSF and 2019 at ZSFG, it means that we will not know the first date of DM diagnosis for the majority of patients (i.e., 61% in the UCSF cohort) since the EHR has not been around for long enough to capture this information. For this reason, DM duration is a difficult variable to capture reliably from EHR and claims data without using clinical notes. With the release of UCSF's de-identified clinical notes and the availability of new methods and tools for natural language processing of clinical free-text, this is a potential avenue for future work.

With any EHR-based study, there is the inherent limitation of incomplete data if a patient received care outside of the hospital system. Because of the likely incompleteness of EHR data and the fact that we did not include a look-back period before the index date to avoid decreasing sample size, the first NPDR diagnosis in our study does not represent incident NPDR. As such, we expected that our study may have higher rates of progression and shorter time to PDR than in studies that use incident NPDR, such as Nwanyanwu et al.[11] However, we did not find this to be the case with 4.4%−8.1% of patients progressing to PDR in our study compared with 6.7% in their study.[11] While claims data would address the limitation of incompleteness of data that is seen in EHR datasets, the trade-off is the lack of specificity and detail that comes with using claims data alone: only diagnosis codes (ICD-9/10) and billing codes (current procedural terminology, Healthcare Common Procedure Coding System) are available, and there are no uninsured patients, which we found to be a key predictor for progression in our study. A future direction is to combine claims with EHR data for prediction models to better capture a patient's complete medical course.

In conclusion, this was a novel study that built time-to-event models for prediction of NPDR to PDR using data at different timepoints, and evaluated the models on 2 health care systems with different patient sociodemographic factors. Our models demonstrated acceptable predictive ability and generalizability in a different health care system. We identified novel predictors for progression to PDR that warrant further investigation. Specifically, we found that type of insurance was a risk factor for progression. Patients with more previous strokes or more previous hospital admissions were also at a greater risk of progression to PDR. We also confirmed other known predictors that are important for progression, including younger age, greater NPDR severity, presence of diabetic neuropathy, and higher HbA1c. Future directions include additional corroboration at other medical centers, and further investigation into the clinical relevance of these novel predictors.

Footnotes and Disclosures

Originally received: August 6, 2022.
Final revision: December 1, 2022.
Accepted: January 24, 2023.
Available online: February 1, 2023.Manuscript no. XOPS-D-22-00179R1.

[1] Department of Ophthalmology, University of California, San Francisco, California.

[2] F.I. Proctor Foundation, University of California, San Francisco, California.

[3] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland.

[4] Department of Epidemiology & Biostatistics, University of California, San Francisco, California.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s): J.M.S.: Consulting − Genetech

Supported in part by the following grants: National Institutes of Health [NEI K23 EY032637], National Institutes of Health [NIH-NEI P30 EY002162 − UCSF Core Grant for Vision Research], Research to Prevent Blindness unrestricted grant, New York, NY. The sponsor or funding organization had no role in the design or conduct of this research.

HUMAN SUBJECTS: Human subjects were included in this study. The Institutional Review Board at the University of California San Francisco approved this study and issued a waiver of consent for all participants. This study adheres to the tenets of the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Guo, Stewart, Acharya, McCulloch, Sun

Data collection: Guo, Yonamine, Sun

Analysis and interpretation: Guo, Yonamine, Ma, Stewart, Arnold, McCulloch, Sun

Obtained funding: Sun

Overall responsibility: Guo, Yonamine, Ma, Stewart, Acharya, Arnold, McCulloch, Sun

Acronyms and Abbreviations:
C-index = Harrell's Concordance index; **Cox** = Cox proportional hazards regression; **Cox-BW** = Cox with backward selection; **Cox-LS** = Cox with LASSO regression; **DM** = diabetes mellitus; **EHR** = electronic health record; **HbA1c** = hemoglobin A1c; **ICD** = International Classification of Diseases; **NPDR** = nonproliferative diabetic retinopathy; **PDR** = prolifterative diabetic retinopathy; **RSF** = random survival forest; **UCSF** = University of California San Francisco; **vs.** = versus; **ZSFG** = Zuckerberg San Francisco General.

Keywords:
Nonproliferative diabetic retinopathy, Proliferative diabetic retinopathy, Prediction, Time-to-event models.

Correspondence:
Catherine Q. Sun, MD, University of California, San Francisco, Department of Ophthalmology, San Francisco, CA 94131. E-mail: catherine.sun@ucsf.edu.

References

1. Klein R, Klein BE. Vision disorder in diabetes. In: Harris MI, Cowie CC, Stern MP, et al, eds. Diabetes in America. 2nd ed. Bethesda, MD: National Institutes of Diabetes and Digestive and Kidney Diseases, National Institute of Health; 1995: 293−338.
2. Stefánsson E. Prevention of diabetic blindness. *Br J Ophthalmol*. 2006;90(1):2−3.
3. Davis MD, Fisher MR, Gangnon RE, et al. Risk factors for high-risk proliferative diabetic retinopathy and severe visual loss: Early Treatment Diabetic Retinopathy Study Report #18. *Invest Ophthalmol Vis Sci*. 1998;39(2):233−252.
4. Klein R, Klein BE, Moss SE, et al. The Wisconsin epidemiologic study of diabetic retinopathy. III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. *Arch Ophthalmol*. 1984;102(4):527−532.
5. Diabetes Control and Complications Trial Research Group. Progression of retinopathy with intensive versus conventional treatment in the Diabetes Control and Complications Trial. *Ophthalmology*. 1995;102(4):647−661.
6. Diabetes Control and Complications Trial Research Group. The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the diabetes control and complications trial. *Diabetes*. 1995;44(8): 968−983.
7. 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(9131):837−853.
8. Kohner EM, Stratton IM, Aldington SJ, et al. Relationship between the severity of retinopathy and progression to photocoagulation in patients with Type 2 diabetes mellitus in the UKPDS (UKPDS 52). *Diabet Med*. 2001;18(3):178−184.
9. Wong TY, Liew G, Tapp RJ, et al. Relation between fasting glucose and retinopathy for diagnosis of diabetes: three population-based cross-sectional studies. *Lancet*. 2008;371(9614):736−743.
10. Kilpatrick ES, Rigby AS, Atkin SL, Frier BM. Does severe hypoglycaemia influence microvascular complications in Type 1 diabetes? An analysis of the Diabetes Control and Complications Trial database. *Diabet Med*. 2012;29(9):1195−1198.
11. Harris Nwanyanwu K, Talwar N, Gardner TW, et al. Predicting development of proliferative diabetic retinopathy. *Diabetes Care*. 2013;36(6):1562−1568.
12. Office of the National Coordinator for Health Information Technology. 'Office-based Physician Electronic Health Record Adoption,' Health IT Quick-Stat #50. dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php. January 2019.
13. Norgeot B, Glicksberg BS, Trupin L, et al. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA Network Open*. 2019;2(3):e190606−e. https://doi.org/10.1001/jamanetworkopen.2019.0606.
14. Kazemian P, Lavieri MS, Van Oyen MP, et al. Personalized Prediction of Glaucoma Progression Under Different Target Intraocular Pressure Levels Using Filtered Forecasting Methods. *Ophthalmology*. 2018;125(4):569−577.
15. Soyiri IN, Reidpath DD. An overview of health forecasting. *Environ Health Prev Med*. 2013;18(1):1−9.
16. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat Comput*. 2015;25(2):173−187.
17. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361−387.
18. Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometr J*. 2006;48(6):1029−1040.
19. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13(1):33. https://doi.org/10.1186/1471-2288-13-33.
20. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Stat Med*. 2015;34(4):685−703.
21. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B (Methodol)*. 1995;57(1):289−300.
22. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841−860.
23. Malhotra NA, Greenlee TE, Iyer AI, et al. Racial, ethnic, and insurance-based disparities upon initiation of anti−vascular endothelial growth factor therapy for diabetic macular edema in the US. *Ophthalmology*. 2021;128(10):1438−1447.
24. Cai CX, Li Y, Zeger SL, McCarthy ML. Social determinants of health impacting adherence to diabetic retinopathy examinations. *BMJ Open Diabetes Res Care*. 2021;9(1):e002374. https://doi.org/10.1016/j.oret.2020.08.004.
25. Hinkle JW, Flynn HW, Banta JT, Vanner EA. Patients presenting emergently with proliferative diabetic retinopathy: follow-up and factors associated with compliance. *Retina*. 2020;40(5):928−935.
26. Wong K-H, Hu K, Peterson C, et al. Diabetic retinopathy and risk of stroke. *Stroke*. 2020;51(12):3733−3736.
27. Hu K, Jiang M, Zhou Q, et al. Association of diabetic retinopathy with stroke: a systematic review and meta-analysis. *Front Neurol*. 2021;12.
28. Grauslund J, Thykjær AS, Kawasaki R, et al. Identification and characterization of patients with rapid progression of diabetic retinopathy in the Danish National Screening Program. *Diabetes Care*. 2020;44(1):e1−e3. https://doi.org/10.2337/dc20-1534.
29. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med Res Methodol*. 2012;12(1):21. https://doi.org/10.1186/1471-2288-12-21.
30. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye Vision*. 2015;2(1):17. https://doi.org/10.1186/s40662-015-0026-2.