AMERICAN ACADEMY OF OPHTHALMOLOGY®

Predicting Glaucoma Progression to Surgery with Artificial Intelligence Survival Models

Shiqi Tao, MS, Rohith Ravindranath, MS, Sophia Y. Wang, MD, MS

Purpose: Prior artificial intelligence (AI) models for predicting glaucoma progression have used traditional classifiers that do not consider the longitudinal nature of patients' follow-up. In this study, we developed survival-based AI models for predicting glaucoma patients' progression to surgery, comparing performance of regression-, tree-, and deep learning–based approaches.

Design: Retrospective observational study.

Subjects: Patients with glaucoma seen at a single academic center from 2008 to 2020 identified from electronic health records (EHRs).

Methods: From the EHRs, we identified 361 baseline features, including demographics, eye examinations, diagnoses, and medications. We trained AI survival models to predict patients' progression to glaucoma surgery using the following: (1) a penalized Cox proportional hazards (CPH) model with principal component analysis (PCA); (2) random survival forests (RSFs); (3) gradient-boosting survival (GBS); and (4) a deep learning model (DeepSurv). The concordance index (C-index) and mean cumulative/dynamic area under the curve (mean AUC) were used to evaluate model performance on a held-out test set. Explainability was investigated using Shapley values for feature importance and visualization of model-predicted cumulative hazard curves for patients with different treatment trajectories.

Main Outcome Measures: Progression to glaucoma surgery.

Results: Of the 4512 patients with glaucoma, 748 underwent glaucoma surgery, with a median follow-up of 1038 days. The DeepSurv model performed best overall (C-index, 0.775; mean AUC, 0.802) among the models studied in this article (CPH with PCA: C-index, 0.745; mean AUC, 0.780; RSF: C-index, 0.766; mean AUC, 0.804; GBS: C-index, 0.764; mean AUC, 0.791). Predicted cumulative hazard curves demonstrate how models could distinguish between patient who underwent early surgery and patients who underwent surgery after $>$ 3000 days of follow-up or no surgery.

Conclusions: Artificial intelligence survival models can predict progression to glaucoma surgery using structured data from EHRs. Tree-based and deep learning-based models performed better at predicting glaucoma progression to surgery than the CPH regression model, potentially because of their better suitability for high-dimensional data sets. Future work predicting ophthalmic outcomes should consider using tree-based and deep learning-based survival AI models. Additional research is needed to develop and evaluate more sophisticated deep learning survival models that can incorporate clinical notes or imaging.

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

Supplemental material available at www.ophthalmologyscience.org.

Glaucoma is a progressive disease of the optic nerve that causes vision loss and irreversible blindness. However, the clinical trajectory of glaucoma can vary dramatically between patients, with some patients progressing quickly to surgery and others remaining stable for many years.[1] Although elevated intraocular pressure (IOP) is a major risk factor for glaucoma progression, many other ancillary factors crucially influence or are indicators of the clinical trajectories of patients with glaucoma (e.g., medication use, eye examination findings, or ancillary testing results).[2–4] Thus, identifying patients who are at high risk and predicting glaucoma progression is complex and requires multifactorial data inputs, rendering the task ripe for artificial intelligence (AI) prediction algorithms.

Some previous work has used AI models to predict glaucoma progression using electronic health records (EHRs). These include traditional machine learning classification models, such as logistic regression, random forest, and support vector machines, for structured data from EHRs,[5,6] as well as deep learning models using natural language processing for free-text notes from EHRs.[7] However, most AI predictive models are classifiers that provide a binary outcome prediction and do not explicitly consider the longitudinal nature of follow-up with patients.

https://doi.org/10.1016/j.xops.2023.100336
ISSN 2666-9145/23

Survival analyses are longitudinal analyses commonly used in traditional inferential studies but are not as common for developing AI prediction models, especially for ophthalmology. Cox regression is the most widely used model for longitudinal analysis. Still, it operates under many restrictive assumptions,[8] such as the assumption of proportional hazards and of uncorrelated features. These restrictions may be difficult to satisfy, especially for large data sets with many features, such as those typically used for AI predictive models. Alternative tree-based survival model approaches, such as random survival forest (RSF) and gradient-boosting survival (GBS), have shown superior performance in diagnosing several diseases, including breast cancer, lung cancer, and brain tumors, such as glioma.[8–11] Deep learning approaches to survival analyses, such as DeepSurv and DeepHit, have also achieved outstanding results in multiple studies.[11,12]

The purpose of the present study was to predict glaucoma progression to surgery using survival-based AI models and comparing the performance of different approaches. In this article, we applied regression-based (Cox regression), tree-based (RSF and GBS), and deep learning–based (DeepSurv) survival AI models to our glaucoma data set to evaluate the performance of these 4 models and their associated analytic approaches.

Methods

Study Population and Cohort Construction

We identified from EHRs 4512 patients with glaucoma seen by the Stanford Department of Ophthalmology from 2008 to 2020. These patients included patients who had either undergone incisional glaucoma surgery (Current Procedural Terminology codes 66150, 66155, 66160, 66165, 66170, 66172, 66174, 66175, 66179, 66180, 66183, 66184, 66185, 67250, 67255, 0191T, 0376T, 0474T, 0253T, 0449T, 0450T, 0192T, 65820, 65850, 66700, 66710, 66711, 66720, 66740, 66625, and 66540) and who had ≥ 2 instances of a glaucoma diagnosis but did not undergo glaucoma surgery (International Classification of Diseases [ICD] 9 codes H40- (excepting H40.0-), H42-, Q150-, and their ICD9 equivalents). At least 120 days of baseline follow-up after the first visit (and before surgery for the surgical patients) was required to allow for adequate baseline testing to be gathered on new patients, a process which could take several visits at our center. The cohort was split into training, validation, and test sets in a 6:3:1 ratio. All models were trained on the training set, with hyperparameters tuned on the validation set or by crossvalidation on the training set, and final results were reported on the test set. This study adheres to the tenets of the Declaration of Helsinki and was approved by the Stanford Institutional Review Board with a waiver of informed consent.

Feature Engineering

The structured features considered in the modeling included demographics, eye examination findings, diagnoses, and medication information from the baseline period, defined as the first 120 days after the initial ophthalmology visit. All baseline features were converted into either categorical variables or continuous numeric variables. Categorical variables included all diagnoses, medications, gender, race, and ethnicity. Race and ethnicity were included as defined in the EHR of the patient. Numeric variables included age at baseline, best visual acuity for both eyes during the baseline period, and maximum IOP for both eyes during the baseline period. For categorical variables, features with $< 1\%$ variance were removed; for numeric variables, missing values were filled in using column-mean imputation. A total of 361 features were included in the input data set. The follow-up time was defined as the number of days from the baseline date to either surgery or the last visit.

Modeling

We developed AI survival models using regression-, tree-, and deep learning–based approaches to predict the time of patients with glaucoma progression to surgery. Regression-based models predict outcomes by constructing linear combinations of multiple predictive factors, in contrast with tree-based and deep learning–based models that capture highly nonlinear relationships between predictive factors and predicted outcomes. We also sought to characterize the most important features contributing to the prediction. Two Cox regression models were constructed with principal component analysis (PCA). We built 2 tree-based survival models using RSF and GBS models. A deep learning survival model was also developed and evaluated.

Cox Proportional Hazards Model

The Cox proportional hazards (CPH) model is a regression model that uses hazard rate as the measure of risk or probability of occurrence of a certain event. The CPH model has several important assumptions, including independence of survival times, absence of correlation between features, a multiplicative relationship between the predictors and the hazard, and a constant hazard ratio. The following formula illustrates the associations between risk factors and the outcome:

$$ln\left\{\frac{h(t)}{h_0(t)}\right\} = b_1X_1 + b_2X_2 + ... + b_pX_p$$

where $h(t)$ is the expected hazard at time t; $h_0(t)$ is the baseline hazard; $X_1, X_2, ...X_p$ are the predictors or risk factors; $b_1, b_2 ... b_p$ are regression coefficients to qualify the associations between predictors.

1. CPH: the baseline model was a CPH with regularization, commonly known as penalized Cox regression. Hyperparameters, including the number of iterations and penalty term weight (alpha), were optimized using threefold crossvalidated grid search.
2. Cox proportional hazards model with principal component analysis (PCA_CPH): because there were numerous input features, to reduce the dimensionality of the input feature space, we built a machine learning pipeline with PCA added as the first step. The PCA-derived components were then input into the CPH model. Hyperparameters, including number of principal components, number of iterations, and alpha, were fine-tuned using threefold crossvalidated grid search.

RSF Model

The RSF model, an extension of the random forest model, ensembles a number of survival trees and uses averaging to reduce predictive variance and control overfitting for time-to-event data.[13] We used the RandomSurvivalForest method from the sksurv package (version 0.17.1) to build the RSF model.[14] Using a threefold crossvalidated grid search, we fine-tuned the number of survival trees, the maximum depth of each tree, the minimum number of samples required to split an internal node, and the maximum number of features to consider when looking for the best split.

Table 2. Population Characteristics of 4512 Patients with Glaucoma

	Surgery (n = 748)	No Surgery (n = 3764)	Total (N = 4512)
Median follow-up, days	601	1139	1038
Age, yrs	64.8 ± 17.0	65.0 ± 18.1	65.0 ± 17.9
Gender			
Female	350 (46.8%)	1920 (51.0%)	2270 (50.3%)
Male	398 (53.2%)	1844 (49.0%)	2242 (49.7%)
Race			
White	276 (36.9%)	1616 (42.9%)	1892 (41.9%)
Asian	228 (30.5%)	969 (25.7%)	1197 (26.5%)
Black	48 (6.4%)	168 (4.5%)	216 (4.8%)
Pacific Islander	5 (0.7%)	23 (0.6%)	28 (0.6%)
Native American	2 (0.2%)	7 (0.2%)	9 (0.2%)
Other	177 (23.7%)	805 (21.4%)	982 (21.8%)
Unknown	12 (1.6%)	176 (4.7%)	188 (4.2%)
Ethnicity			
Hispanic/Latino	106 (14.2%)	460 (12.2%)	566 (12.6%)
Non-Hispanic	632 (84.5%)	3159 (83.9%)	3791 (84.0%)
Unknown	10 (1.3%)	145 (3.9%)	155 (3.4%)
Visual acuity (logMAR)			
Right eye	0.43 ± 0.76	0.39 ± 0.74	0.39 ± 0.74
Left eye	0.43 ± 0.76	0.43 ± 0.79	0.43 ± 0.78
IOP (mmHg)			
Right eye	20.1 ± 27.7	18.0 ± 6.2	18.3 ± 12.3
Left eye	21.8 ± 45.8	18.3 ± 6.5	18.8 ± 19.1

IOP = intraocular pressure; logMAR = logarithm of the minimum angle of resolution.
Race and ethnicity were as reported in the electronic health record for each patient.

GBS Model

A GBS model is an extension of traditional gradient-boosting models. Gradient-boosting survival implements gradient boosting with Cox proportional loss with regression trees as the base learner, and the regression tree is fit on the gradient descent of the loss function. We used GradientBoostingSurvivalAnalysis method with partial likelihood loss from the sksurv package,[14] and tuned the learning rate (shrinkage of the contribution of each regression tree) as well as the abovementioned hyperparameters using threefold crossvalidated grid search.

Deep Learning Survival Model

To investigate the performance of deep learning survival models compared with regression-based and tree-based models, we trained DeepSurv,[11] a deep feed-forward neural network that uses multiple fully connected layers, to estimate the cumulative hazard of the outcome. Baseline input features x from the data set are input into multiple hidden layers to get the output layer $\widehat{h}_\theta(x)$ (a single node) with a linear activation equal to the log-risk hazard estimation. In the present study, we trained the DeepSurv model with 2 hidden layers and dropout. Hyperparameters, including the number of nodes, dropout rate, training batch size, and learning rate, were optimized on the validation set.

Evaluation Metrics

Model performance was evaluated using the concordance index (C-index) and mean cumulative/dynamic area under the curve (AUC) score. Concordance index is the standard performance metric for survival models. It measures the rank correlation between predicted risk scores and observed time points[15]; in other words, it gives the probability of concordance between predicted and observed survival. The mean AUC score is the mean value of all time-dependent AUC scores from across the study duration. Because, at any given time point in the study, the number of patients who have experienced the outcome and the number remaining at risk varies, the receiver operating characteristic (ROC) curve is expected to vary among different study time points. Thus, the ROC curve is time dependent. The time-dependent AUC score is the area under the time-dependent ROC curve, which is calculated using cumulative cases and dynamic controls at a given time point t, where cumulative cases are all individuals who underwent glaucoma surgery before or at time t ($ti \leq t$), whereas dynamic controls are those with $ti > t$. By computing the area under the cumulative/dynamic ROC at time t, we can determine how well a model can distinguish patients who require surgery by a given time point ($ti \leq t$) from patients who do not require surgery at or before this time point ($ti > t$).[16]

All models' hyperparameters were tuned for optimal C-index, with the final evaluation performed on the test set. A summary of tuned hyperparameters for each model is shown in Table S1 (available at www.ophthalmologyscience.org).

Explainability and Interpretability

To better explain the models' predictions, we investigated which were the most important features of regression-, tree-, and deep learning–based models using SHapley Additive exPlanations, a model-agnostic method derived from coalitional game theory. Because the calculation of Shapley importance values does not depend on the underlying model architecture, this method enables a fair comparison of important features across different types of models.[17,18]

We also plotted cumulative hazard curves of different models for the same group of patients with glaucoma to investigate how models predict risks for surgical and nonsurgical patients. We selected 3 patients from the test set to highlight the models'

Table 3. Performance of Survival-based AI Models for Predicting Progression to Glaucoma Surgery

Approach	Model	C-index	Mean AUC
Regression-based	CPH	0.720	0.756
Regression-based	PCA_CPH	0.745	0.780
Tree-based	RSF	0.766	0.804
Tree-based	GBS	0.764	0.791
Deep learning	DeepSurv	0.775	0.802

AUC = area under the curve; CPH = Cox proportional hazards; GBS = gradient-boosted survival model; PCA_CPH = Cox proportional hazards with principal component analysis; RSF = random survival forest.

interpretability by plotting cumulative hazard curves: 1 patient who underwent early surgery at day 3 from the baseline time; 1 patient who had late surgery at day 3472; and 1 patient who did not undergo surgery during his follow-up period as of day 3330.

Results

Population Characteristics

Out of the 4512 patients with glaucoma included in the study, 748 progressed to require glaucoma surgery. The median follow-up time was 601 days for surgical patients and 1139 days for nonsurgical patients. Population characteristics are summarized in Table 2. White and Asian racial groups constituted the predominant population in this cohort. The patients' mean age was approximately 65 years old, the mean logarithm of the minimum angle of resolution visual acuity for both eyes at baseline was about 0.43, and the mean baseline IOP for both eyes was approximately 18.3 mmHg.

Model Performance

Model performance is summarized in Table 3. In general, tree-based and deep learning–based models performed better than regression-based models, achieving higher C-index and mean AUC.

For the regression-based models, lowering the dimension of the input features via PCA increased the C-index and the time-dependent AUC scores (Table 3). Figure 1A illustrates that PCA_CPH outperformed the original CPH model at almost every time point in terms of AUC score.

Figure 1B shows that the RSF and GBS models had similar time-dependent AUC scores. Although the RSF model had slightly higher C-index and mean AUC scores than the GBS model, the 2 tree-based models both outperformed the regression-based survival models. Figure 1B also more clearly shows that the RSF model had a higher AUC than the GBS model at most time points.

Figure 1C shows that the DeepSurv model had similar performance to the RSF and GBS models. Among these 3 models, RSF had the best time-dependent AUC score, but DeepSurv had the highest C-index, as well as slightly better performance than GBS.

Explainability

Predicted cumulative hazard curves for patients from the test set with different outcomes (no surgery or surgery at different time points) generated by different survival AI models are shown in Figure 2. All models appropriately predicted the steepest rise in cumulative hazard of surgery for the patient who actually underwent surgery early in the follow-up period. Most models were also able to discriminate between a patient who had not yet undergone glaucoma surgery as of their last follow-up time at day 3330 and a patient who underwent surgery at day 3472, with the patient who eventually underwent surgery having a higher predicted cumulative hazard.

We further studied the most important features relied upon by the CPH and GBS models to predict the outcome by calculating the Shapley values of features across the test set. The most important features were similar across models, including features from the broad categories of demographics, medications, diagnoses, and examination components (Table 4). Reassuringly, clinical features, such as IOP and visual acuity, were important predictive factors in these models, as was the usage of many common glaucoma medications.

Discussion

In the present study, we developed and compared the performance of different survival-based AI models to predict glaucoma progression to surgery using structured EHR data from patients with glaucoma. We compared regression- and tree-based survival models, as well as a deep learning AI survival model. According to our evaluation metrics, we found that the deep learning model DeepSurv had the best overall performance, followed by the tree-based RSF and GBS models. DeepSurv, RSF, and GBS models have the advantages of robustness against multicollinearity and the ability to discern highly nonlinear relationships among predictors without prior feature selection. Previous research has shown that deep learning survival models perform better, especially in high-dimensional data sets.[19,20] Although DeepSurv showed the best overall performance, explainability analyses revealed important features that were common for all our model predictions, such as age, visual acuity, and the use of glaucoma medications.

Our work develop and compare survival-based predictive models is relatively novel in the ophthalmology AI literature. Most previous models for predicting glaucoma progression are structured as classification models, whether they operate on EHR data[5–7] or imaging data.[21,22] The outputs of classification models can be interpreted as the probability of experiencing the predicted event, which is relatively simple for users to understand and potentially act upon. However, classification models do not account for the longitudinal nature of patient outcomes. The challenges of modeling for longitudinal data with loss-to-follow-up and censoring are well-known in statistical inference; thus, Kaplan Meier survival analysis and CPHs models are commonly employed for inferential studies that

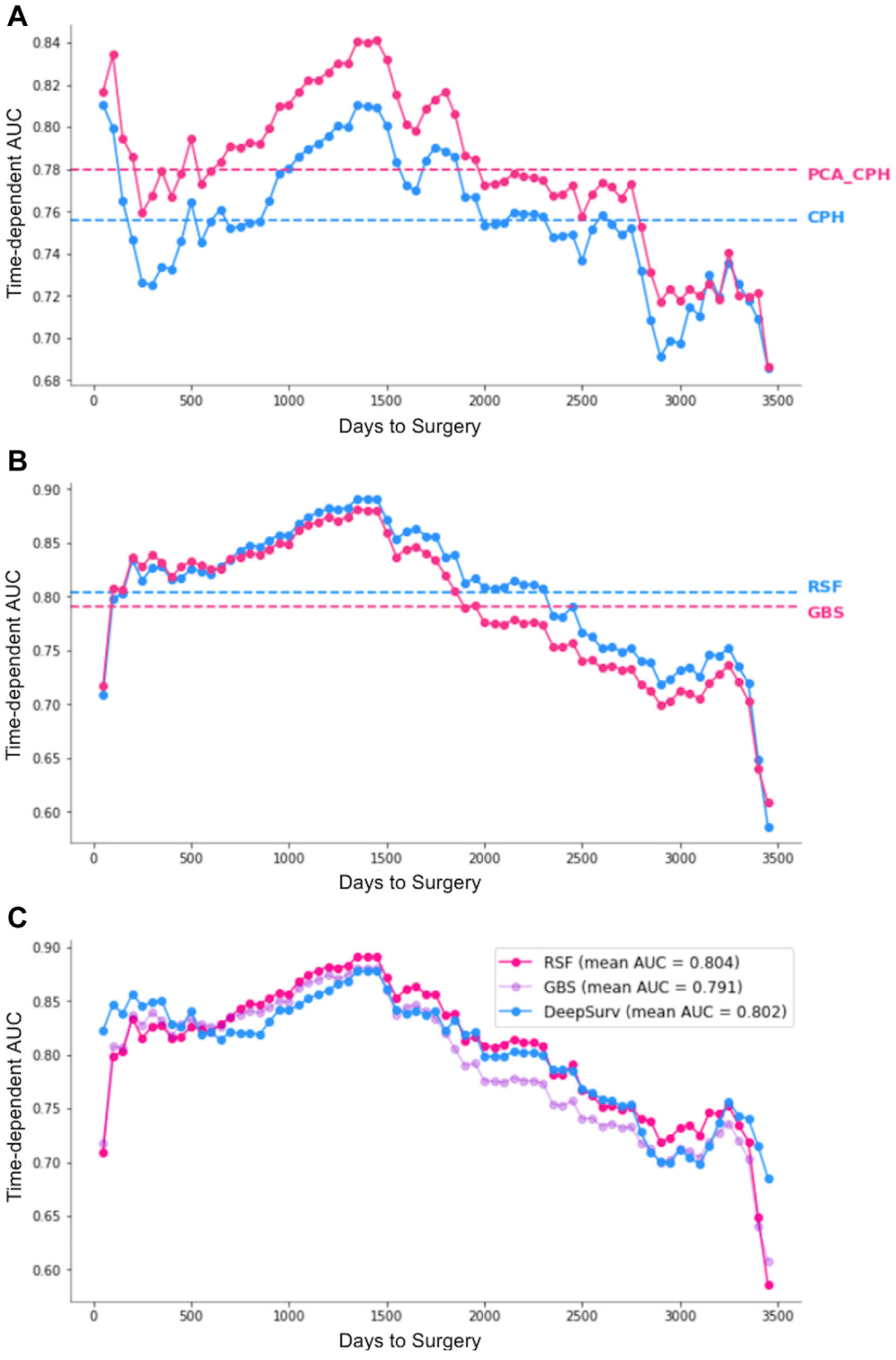

Figure 1. Time-dependent area under the curve (AUC) and mean AUC scores. **A,** Regression models: time-dependent AUC (connected dots) and mean AUC (horizontal dashed line) are shown for the Cox proportional hazards (CPH) model and the Cox proportional hazards model with principal component analysis (PCA_CPH). **B,** Tree-based models: time-dependent AUC (connected dots) and mean AUC (horizontal dashed line) are shown for the random survival forest (RSF) and gradient-boosted survival (GBS) models. **C,** Deep learning models: time-dependent AUC (connected dots) and mean AUC (horizontal dashed line) are shown for the deep learning survival model (DeepSurv) alongside the previous RSF and GBS models for comparison.

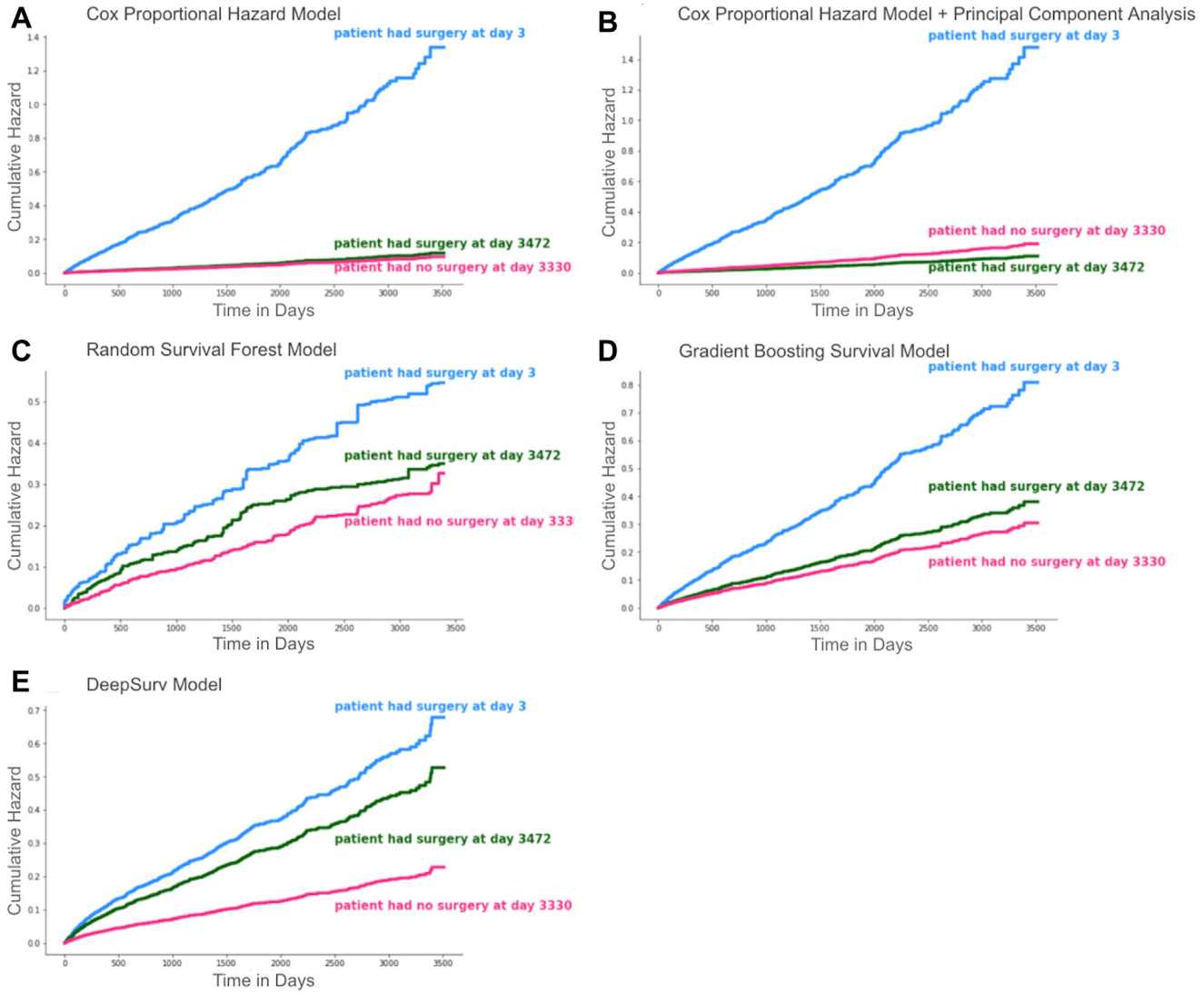

Figure 2. Cumulative hazard functions for sample patients with different outcomes. The cumulative hazard functions as predicted by different models are shown for patients who had early surgery (day 3), patients who had late surgery (day 3472), and patients who did not have surgery even after a long period of follow-up (day 3330). Predicted cumulative hazard curves are shown for the **(A)** Cox proportional hazards (CPH), **(B)**, CPH with principal components analysis, **(C)**, random survival forest, **(D)** gradient-boosted survival, and **(E)** DeepSurv models.

focus on quantifying the relationship between a predictor and an outcome. Survival-based prediction models have begun to be explored for predicting outcomes in other medical domains, including cancer survival prediction[19,23] and dementia prediction.[20] Overall, the performance range of our survival AI models was comparable to similar studies utilizing inputs from EHRs, including Kim et al[19] for oral cancer (C-index, 0.694–0.781) and Spooner et al[20] for dementia prediction (C-index up to 0.828). In addition, our results were similar to those from studies of survival-based AI models in other medical fields. We also found that tree-based methods and deep learning models outperformed regression-based models, potentially because of their suitability for high-dimensional data sets.[19,20,23,24] Similar reasons can explain the improved performance after adding PCA to the CPH model, further illustrating that dimensionality reduction is crucial for prediction models using complex input features from EHRs.

A potential drawback of survival-based AI models is that their prediction outputs seem less interpretable to the user than the simple probabilities of experiencing the outcome that classification models provide. Thus, to provide better insight into model outputs, we showed the cumulative hazard predicted by different models (e.g., patients who underwent early, late surgery, or no surgery). These curves illustrate the stark differences in the predicted cumulative hazard curves between a high-risk patient who underwent surgery early in their clinical trajectory (steeply rising cumulative hazard curve) and patients who did not undergo surgery even after long periods of follow-up (slowly rising cumulative hazard curves). Incredibly, most models also were able to distinguish between a patient who had surgery

Table 4. The Most Important Features Contributing to Model Predictions

	Penalized CPH Model	Gradient-boosted Survival Model	DeepSurv
Demographics	Race (White) Ethnicity (Hispanic)	Age Race (White)	Age Race (Native American) Gender (Female)
Medications	Latanoprost/Xalatan Brimonidine/Alphagan Dorzolamide-Timolol Brinzolamide Dorzolamide Timolol Travoprost	Dorzolamide-Timolol Dorzolamide Brimonidine/Alphagan Timolol Xalatan	Dorzolamide Brimonidine Dorzolamide-Timolol Brinzolamide Xalatan Timolol Hydrocodone
Diagnoses	Z96.1 Presence of intraocular lens H25.10 Age-related nuclear cataract, unspecified eye H25.13 Age-related nuclear cataract, bilateral H26.9 Unspecified cataract H40.009 Preglaucoma, unspecified eye	H25.13 Age-related nuclear cataract, bilateral H25.10 Age-related nuclear cataract, unspecified eye H43.819 Vitreous degeneration, unspecified eye H40.1193 Primary open angle glaucoma, severe stage	H26.9 Unspecified cataract Z96.1 Presence of intraocular lens H47.239 Glaucomatous optic atrophy, unspecified eye H35.30 Unspecified macular degeneration H27.0 Aphakia
Examination components	Best-corrected visual acuity, OS	Best-corrected visual acuity, OD Best-corrected visual acuity, OS IOP max, OS	Best-corrected visual acuity, OD IOP max, OD IOP max, OS

CPH = Cox prportional hazards; IOP = intraocular pressure; OD = oculus dexter (right eye); OS = oculus sinister (left eye).

after approximately 10 years versus a patient who did not have surgery after 10 years of follow-up, predicting a slightly higher cumulative hazard in the former. Thus, although it may seem simpler to interpret a predicted probability for glaucoma progression to surgery in traditional classification models, this information does not provide an expected time horizon, and there may not be any inherent relationship between a predicted probability of surgery and its temporal nearness. One potential method of incorporating temporal information into a classification model could be through a multiclassification approach that provides probabilities of glaucoma surgery occurring over discrete future time windows. However, this approach may not naturally account for censoring and may produce probabilities for adjacent time windows that may not be related, plausible, or easily interpretable. Future research could focus on developing classification models that address these limitations or on combining classification with survival models. A cumulative hazard output of a survival AI model may therefore be beneficial for clinical decision support tools that predict future events.

In addition, although traditionally criticized as being opaque and unexplainable, AI tree-based and even deep learning–based models can retain the explainability benefits of the more commonly favored Cox regression models. It is important to note that, although we can shed light onto which features exert a stronger influence on prediction in different models, that does not necessarily suggest a true biologic relationship between the features and the outcome, as would be the goal in a hypothesis-driven inference study. Nevertheless, it is striking that among the top most important predictors of glaucoma surgery are factors that clinicians themselves would consider important, such as visual acuity, age, and the use of various glaucoma medications, including second- and third-line medications, such as dorzolamide and brimonidine. These reassuring explainability studies serve to increase the apparent trustworthiness of these AI prediction models.

Despite the above advantages of this study, there are several limitations. The data we used are from a single clinical center, and models may not generalize well with data from other sites. However, in service of the goal of personalized algorithms to deliver personalized medicine, a fully generalized algorithm that applies universally is not likely to be the goal. Rather, these approaches can and should be fine-tuned to each population they may be deployed upon. Another limitation of single-center data is that patients may seek care at other institutions. In our study, the models were designed to predict the first glaucoma surgery performed at our institution for new patients. To address this limitation, future research could explore the use of natural language processing to extract external surgery information from clinical notes. Single-center longitudinal studies may also be limited by censoring events, such as death or patient departure from our clinical center. Additional challenges in this study included the imbalance in the ratio of surgical to nonsurgical patients with glaucoma, which in our data set was approximately 1:5 and posed challenges to our models. We also did not incorporate time-varying features in this analysis containing hundreds of inputs. In addition to the resultant challenges in cohort construction with multiple time-varying features, this approach would also reduce the ability to perform dimensionality reduction using PCA or elimination of near-zero variance features

and would introduce assumptions during inference that may not be tenable. In addition, our analysis only included structured input data from EHRs. Although this included important measures, such as visual acuity and IOP, measures, such as corneal thickness and refractive error, had a high degree of missingness. Furthermore, unstructured data, such as images and clinical notes, contain a wealth of information about a patients' prognosis. Further studies combining features from these 2 additional modalities of data can be undertaken, using approaches, such as embedding data extracted from images or text into the baseline features.

In conclusion, identifying which patients with glaucoma are at high risk of progressing is an important aspect of clinical care. In our study, observational clinical data were collected from a single academic center, and multiple survival AI models were developed to predict which patients progress to glaucoma surgery. After a comparison of evaluation results across different models, we concluded that the neural network model DeepSurv and tree-based survival AI models outperformed regression-based models. Future research can be conducted to explore larger and more diverse data sets from multiple clinics and integrate multiple modalities of input data, such as text or imaging.

Footnotes and Disclosures

Originally received: January 25, 2023.
Final revision: May 16, 2023.
Accepted: May 17, 2023.
Available online: May 25, 2023. Manuscript no. XOPS-D-23-00016R2.

Byers Eye Institute, Department of Ophthalmology, Stanford University, Palo Alto, California.

Disclosures:

All authors have completed and submitted the ICMJE disclosures form.

The authors have no proprietary or commercial interest in any materials discussed in this article.

Supported by the National Eye Institute (grant no.: K23EY03263501 [S.Y.W.]); Career Development Award from Research to Prevent Blindness (S.Y.W.); unrestricted departmental grant from Research to Prevent Blindness (S.Y.W., R.R., S.T.); National Eye Institute (departmental grant no.: P30-EY026877 [S.Y.W., R.R., S.T.]).

HUMAN SUBJECTS: Human subjects were included in this study. This study adheres to the tenets of the Declaration of Helsinki and was approved by the Stanford Institutional Review Board with a waiver of informed consent.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Tao, Wang

Data collection: Tao, Wang

Analysis and interpretation: Tao, Ravindranath, Wang

Obtained funding: Wang

Overall responsibility: Tao, Ravindranath, Wang

Abbreviations and acronyms:
AI = artificial intelligence; **AUC** = area under the curve; **C-index** = concordance index; **CPH** = Cox proportional hazards; **EHR** = electronic health record; **GBS** = gradient-boosting survival; **ICD** = International Classification of Diseases; **IOP** = intraocular pressure; **PCA** = principal components analysis; **PCA_CPH** = Cox proportional hazards model with principal component analysis; **ROC** = receiver operating characteristic; **RSF** = random survival forest.

Keywords:
Artificial intelligence, Deep learning, Electronic health records, Glaucoma, Machine Learning.

Correspondence:
Sophia Wang, MD, MS, 2370 Watson Ct, Palo Alto CA 94303. E-mail: sywang@stanford.edu.

References

1. Chauhan BC, Malik R, Shuba LM, et al. Rates of glaucomatous visual field change in a large clinical population. *Invest Ophthalmol Vis Sci*. 2014;55:4135–4143.
2. Friedman DS, Wilson MR, Liebmann JM, et al. An evidence-based assessment of risk factors for the progression of ocular hypertension and glaucoma. *Am J Ophthalmol*. 2004;138:S19–S31.
3. Newman-Casey PA, Niziol LM, Gillespie BW, et al. The association between medication adherence and visual field progression in the collaborative initial glaucoma treatment study. *Ophthalmology*. 2020;127:477–483.
4. Rivera JL, Bell NP, Feldman RM. Risk factors for primary open angle glaucoma progression: what we know and what we need to know. *Curr Opin Ophthalmol*. 2008;19:102–106.
5. Baxter SL, Marks C, Kuo TT, et al. Machine learning-based predictive modeling of surgical intervention in glaucoma using systemic data from electronic health records. *Am J Ophthalmol*. 2019;208:30–40.
6. Baxter SL, Saseendrakumar BR, Paul P, et al. Predictive analytics for glaucoma using data from the all of us research program. *Am J Ophthalmol*. 2021;227:74–86.
7. Wang SY, Tseng B, Hernandez-Boussard T. Deep learning approaches for predicting glaucoma progression using electronic health records and natural language processing. *Ophthalmol Sci*. 2022;2:100127.
8. Qiu X, Gao J, Yang J, et al. A comparison study of machine learning (random survival forest) and classic statistic (Cox proportional hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. *Front Oncol*. 2020;10:551420.
9. Wieczorkowska A, Jarmulski W. Optimizing C-index via gradient boosting in medical survival analysis. In: Appice A, Ceci M, Loglisci C, et al., eds. *Complex Pattern Mining: New Challenges, Methods and Applications*. Cham: Springer International Publishing; 2020:33–45.

10. Ma B, Yan G, Chai B, Hou X. XGBLC: an improved survival prediction model based on Xgboost. *Bioinformatics*. 2022;38:410−418.
11. Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18:24.
12. Lee C, Zame W, Yoon J, van der Schaar M. DeepHit: a deep learning approach to survival analysis with competing risks. *AAAI*. 2018;32.
13. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2:841−860.
14. Pölsterl S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *J Mach Learn Res*. 2020;21:1−6.
15. Steck H, Krishnapuram B, Dehing-Oberije C, et al. On ranking in survival analysis: bounds on the concordance index. *Adv Neural Inf Process Syst*. 2007;20.
16. Schmid M, Kestler HA, Potapov S. On the validity of time-dependent AUC estimators. *Brief Bioinform*. 2015;16:153−168.
17. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, eds. *Advances in Neural Information Processing Systems*. vol 30. Curran Associates, Inc.; 2017. https://github.com/slundberg/shap. Accessed May 16, 2023.
18. Lundberg S. shap. Github
19. Kim DW, Lee S, Kwon S, et al. Deep learning-based survival prediction of oral cancer patients. *Sci Rep*. 2019;9:6994.
20. Spooner A, Chen E, Sowmya A, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci Rep*. 2020;10:20410.
21. Li F, Su Y, Lin F, et al. A deep-learning system predicts glaucoma incidence and progression using retinal photographs. *J Clin Invest*. 2022:132.
22. Thakur A, Goldbaum M, Yousefi S. Predicting glaucoma before onset using deep learning. *Ophthalmol Glaucoma*. 2020;3:262−268.
23. Mohammed M, Mboya IB, Mwambi H, et al. Predictors of colorectal cancer survival using Cox regression and random survival forests models based on gene expression data. *PLoS One*. 2021;16:e0261625.
24. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. *J Mach Learn Res*. 2019;20:1−30.