AMERICAN ACADEMY OF OPHTHALMOLOGY®

Prediction Models for Glaucoma in a Multicenter Electronic Health Records Consortium: The Sight Outcomes Research Collaborative

Sophia Y. Wang, MD, MS,[1] *Rohith Ravindranath, MS,*[1] *Joshua D. Stein, MD, MS,*[2] *on behalf of the SOURCE Consortium**

Purpose: Advances in artificial intelligence have enabled the development of predictive models for glaucoma. However, most work is single-center and uncertainty exists regarding the generalizability of such models. The purpose of this study was to build and evaluate machine learning (ML) approaches to predict glaucoma progression requiring surgery using data from a large multicenter consortium of electronic health records (EHR).

Design: Cohort study.

Participants: Thirty-six thousand five hundred forty-eight patients with glaucoma, as identified by International Classification of Diseases (ICD) codes from 6 academic eye centers participating in the Sight OUtcomes Research Collaborative (SOURCE).

Methods: We developed ML models to predict whether patients with glaucoma would progress to glaucoma surgery in the coming year (identified by Current Procedural Terminology codes) using the following modeling approaches: (1) penalized logistic regression (lasso, ridge, and elastic net); (2) tree-based models (random forest, gradient boosted machines, and XGBoost), and (3) deep learning models. Model input features included demographics, diagnosis codes, medications, and clinical information (intraocular pressure, visual acuity, refractive status, and central corneal thickness) available from structured EHR data. One site was reserved as an "external site" test set (N = 1550); of the patients from the remaining sites, 10% each were randomly selected to be in development and test sets, with the remaining 27 999 reserved for model training.

Main Outcome Measures: Evaluation metrics included area under the receiver operating characteristic curve (AUROC) on the test set and the external site.

Results: Six thousand nineteen (16.5%) of 36 548 patients underwent glaucoma surgery. Overall, the AUROC ranged from 0.735 to 0.771 on the random test set and from 0.706 to 0.754 on the external test site, with the XGBoost and random forest model performing best, respectively. There was greatest performance decrease from the random test set to the external test site for the penalized regression models.

Conclusions: Machine learning models developed using structured EHR data can reasonably predict whether glaucoma patients will need surgery, with reasonable generalizability to an external site. Additional research is needed to investigate the impact of protected class characteristics such as race or gender on model performance and fairness.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100445*

Supplemental material available at www.ophthalmologyscience.org.

Glaucoma is one of the leading causes of blindness worldwide.[1] Although many patients remain stable on medical therapy for long periods of time without progression, some will progress to vision loss and require invasive surgery.[2] Although there are risk factors known to affect likelihood of glaucoma progression, such as elevated intraocular pressure (IOP), reduced central corneal thickness, and others,[3–6] nevertheless, it is often difficult for eye care professionals to predict whose glaucoma will remain stable or not. If we have a tool that can accurately predict disease stability, it may be possible to focus efforts on the highest risk patients for more frequent follow-up, testing, recruitment to clinical trials, and interventions, while safely relieving some of the burden of follow-up, testing, and patient anxiety for those who are less likely to progress.

Advances in machine learning (ML) and artificial intelligence (AI) have enabled several prediction models for

https://doi.org/10.1016/j.xops.2023.100445
ISSN 2666-9145/23

glaucoma to be developed leveraging electronic health record (EHR) data. Our work and others have developed models predicting which patients with glaucoma would progress to requiring glaucoma surgery using structured EHR inputs.[7–10] Our work has further advanced to make predictions on free-text clinical progress notes and their combination.[7,8,11] However, most previous work has been limited to single-center data and uncertainty exists regarding the generalizability of such models to patients receiving care in other settings.

Challenges to evaluating performance on external datasets exist due to the difficulties of data-sharing for protected health information between health centers and harmonizing data to a common standard. The Sight Outcomes Research Collaborative (SOURCE) was recently established to combine rich granular EHR data from academic ophthalmology departments across the United States who use a common underlying EHR system (Epic Systems), permitting researchers to develop and test AI algorithms on a large, diverse population of patients across multiple health systems. The purpose of this study was to develop and evaluate ML and deep learning algorithms to predict which patients with glaucoma will progress to require glaucoma surgery using data from 6 SOURCE consortium sites.

Methods

Data Source and Study Population

Data Source. Data were derived from the SOURCE Ophthalmology Data Repository (https://www.sourcecollaborative.org/). The SOURCE captures EHR data of all patients receiving any eye care at academic health systems participating in this consortium from the time each site went live on the EHR until the present. This study used data from 6 active SOURCE sites, each of whom contributed 7 to 14 years of data. The SOURCE captures information on patient demographics, diagnoses identified based on International Classification of Diseases (ICD) billing codes, eye examination findings from every clinic visit, along with data on medications, laser and surgical interventions. The data in SOURCE are completely deidentified. However, privacy-preserving software (Datavant Inc) permits researchers to follow patients longitudinally over time and across institutions, while still protecting patients' identities. This study was approved by the University of Michigan and Stanford institutional review boards and adhered to the tenets of the Declaration of Helsinki.

We identified all patients in SOURCE with a glaucoma-related billing code (ICD 365, H40, H42, Q15.0 and their descendants). We excluded patients with only glaucoma suspect codes (H40.0 and ICD 365.0 and their descendants). From among this set of patients, we identified those who underwent glaucoma surgery (including traditional surgery and minimally invasive glaucoma surgery, but excluding selective laser trabeculoplasty or laser peripheral iridotomy) based on Current Procedural Terminology[12] codes or who had $\geq$ 2 separate encounters with a glaucoma diagnosis identified by ICD[13] coding (Table S1).

A final cohort was developed to predict which patients with glaucoma would progress to require glaucoma surgery over the following 12 months, using the previous 4 to 12 months of data, in a similar formulation as our previous work[7] and others.[14] The prediction formulation was designed in this manner so that a future algorithm could be run on any patient with glaucoma at any time in their treatment trajectory, rather than only on new patients using their baseline data. Thus, models trained in this manner would retain maximum flexibility for practical clinical applications. Briefly, a prediction date (or index date) was defined for each patient, which divided the patient's medical timeline into a lookforward period over which the model would predict likelihood of the patient's progression to surgery, and a lookback period of a minimum of 4 months and up to 12 months (if available) from which models' input data were drawn. This prediction formulation allows the resulting model to predict which patients are at highest risk of progression over the next year and allows the flexibility to apply the model to perform a prediction at any point in a patient's follow-up using the most recent clinical data. Patients without $\geq$ 4 months of lookback data were excluded. For surgical patients, the date of first glaucoma surgery in either eye was identified and the prediction date was defined as either 12 months prior to surgery or after the initial 4 months of follow-up (whichever was later). For nonsurgical patients, $\geq$ 12 months of follow-up after the prediction date was required to ascertain that no surgery was performed over the entire lookforward period. Thus, the prediction date was defined as 12 months prior to their last follow-up date. Patients without $\geq$ 4 months of available input data were excluded. A summary of cohort construction timelines with examples is given in Fig S1.

Feature Engineering

An overview of the feature engineering and cohort construction process is illustrated in Fig S2. Input features from the EHRs included demographics, clinical variables, diagnosis codes, and glaucoma and general medication usage. The model is laterality-agnostic, in that it predicts future glaucoma surgery in either eye, using input features from both eyes. This is because many feature categories are inherently at the patient level (such as demographics and systemic medications), and others have missing or ambiguous laterality (ICD codes), and sometimes the decision to proceed to surgery in 1 eye does depend on the status of the contralateral eye. Demographics included age (calculated at the prediction date), gender, race, ethnicity, urbanicity of residence using rural-urban commuting area codes,[15] and distressed communities index score, a measure of the affluence level of the patient's community of residence.[16] Age and distressed communities index score were continuous variables which were scaled by dividing by 100. Missing values for distressed communities index score were filled with column mean imputation. Gender, race, ethnicity, and rural-urban commuting area code were categorical variables which were dummy encoded for model input. Clinical input variables included best recorded visual acuity, IOP, most recent central corneal thickness, and refraction spherical equivalent, all for both eyes from the input time window. Best recorded visual acuity was expressed in logarithm of the minimum angle of resolution units and summarized into best-recorded, worst-recorded, and most recent for both eyes. Intraocular pressures were summarized into max, min, and mean for each eye and standardized to mean 0 and standard deviation of 1. Central corneal thickness was scaled by dividing by 1000, and refraction spherical equivalent was scaled by dividing by 10; both of these had missing values filled by column mean imputation and missing value indicator variables were created. Encounter ICD codes (both ocular and nonocular) were aggregated to the first decimal level, then converted to Boolean vectors such that patients with an encounter in the input period with that ICD code have a "1," 0 otherwise. ICD codes with near-zero variance ($<$ 0.5%) were removed, yielding a total of 92 ICD-based features. Similarly, medications (both ocular and nonocular) from the input period were aggregated by their generic name and turned into Boolean vector inputs as with diagnosis codes. Medication features with near-zero variance ($<$ 2%) were removed, yielding a total of 52 medication features remaining. The total number of structured input features was 179.

The data were split for model training, model validation, and evaluation. Data from 1 site was reserved as an "external site" (N = 1550). The remaining 5 sites of SOURCE data were split by patient in an 80:10:10 ratio for training (N = 27 999), validation (N = 3500), and test sets (N = 3499).

Modeling

Machine Learning. Several classical ML models were fit on the training data using the Python sklearn version 1.2.1 package (Python Software Foundation, open-source software library). These models included penalized logistic regression models (L1, L2, and elastic net penalization), random forest, gradient boosted trees, and XGBoost. Hyperparameters were tuned using threefold cross-validation on the training set to optimize the area under the receiver operating curve (AUROC). Grid search was used to tune penalization for regression models, and random search was used for tree-based models. A summary of hyperparameters is in Table S2.

Deep Learning. Two deep learning models were built and evaluated using the Python tensorflow version 2.11.0 package. The first deep learning model was a fully connected model using the same structured features as the ML models, constructed with an input size of 179 and passed through 2 dense layers, a dropout layer between those 2, and an output layer with a sigmoid activation function (Input (179)- > Dense (512, batch normalization)- > Dropout (0.8)- > Dense (64, batch normalization)- > Output (1, sigmoid). The second deep learning model was an embedding-based model, structured by grouping ICD-related features, medication-related features, and the remaining features (i.e., demographics and eye exam information) as separate inputs into the model. International Classification of Diseases-related features and medication-related features were each individually processed by an embedding layer followed by 2 dense layers. The 2 output vectors are then concatenated with the third input group and goes through a dense layer and a final output layer with a sigmoid activation function. Batch normalization and dropout layers were included throughout the architecture to add regularization. The full architecture of the multi-input model is shown in Figure 3.

Evaluation

We used standard classification evaluation metrics including sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, F1-score (the harmonic mean of recall and precision), AUROC, and the area under the precision-recall curve, all evaluated on both the test set and the independent external site data. The classification threshold for each model was tuned for the optimum F1-score on the test set. Confidence intervals (CIs) were generated using boostrapping with 1000 replicates. We also performed explainability studies for the structured models using Shapley values, also known as SHapley Additive exPlanations values.[17,18] This technique calculates the importance of the features based on the magnitude of feature attributions, using a game theory approach to explain the results of any ML model and make them interpretable. Shapley values represent the marginal contribution to the model predictions for each feature, calculated over all combinations of subsets of features. We estimated the SHapley Additive exPlanations values on the XGBoost model on the test set.

Results

Study Population

Population characteristics for the entire study cohort of 36 548 patients with glaucoma are summarized in Table 3. Patients who progressed to surgery represented 16.9% (N = 6019) of the population. The rate of surgery ranged from 14.1% to 22.5% across all institutions, with the median rate of 16.1%. The rate of surgery for the institution held out as the external site was 16.9%. The overall mean age was 70.1 years (standard deviation 14.6). The majority of the population was White (63.5%, N = 23 202) and Black (23.0%, N = 8417). There were 1526 Hispanic patients (4.2%). Additional population demographic information by individual site is available in Table S4.

Model Performance

Receiver operator characteristic curves and precision-recall curves for the ML models are depicted in Figure 4A, evaluated both on the test set and the external site data. XGBoost demonstrated the highest AUROC which was 0.771 (95% CI, 0.770−0.772) on the test set and 0.750 (95% CI, 0.749−0.751) on the external site set. Tree-based models (random forest, gradient boosted trees, XGBoost) demonstrated much superior performance compared with penalized regression models. For all models, performance was slightly degraded on the external test site data compared with the test set, although drops in performance were typically < 3% on AUROC. The AUROCs for all models on the test and external site set are shown with 95% CIs in Table S5.

Receiver operator characteristic curves and precision-recall curves for the deep learning models are depicted in Figure 4B, evaluated both on the test set and the external site data. The embedding model outperformed the fully connected model, with AUROC of 0.755 (95% CI, 0.755−0.756) on the test set and 0.741 (95% CI, 0.740−0.742) on the external site. Classification metrics are summarized in Table 6, with individual classification thresholds tuned to maximize F1 score on the validation set. The F1 score is the harmonic mean of the precision and recall, so that higher values (closer to 1) indicate better performance. Under these circumstances, specificity and negative predictive value was generally higher than sensitivity and positive predictive value, respectively. Overall accuracy ranged up to 0.792.

Explainability

We performed explainability analyses to determine which input features contributed most to the model predictions, calculating Shapley values for the XGBoost model (Fig 5), the gradient boosted model (Fig S6A), and the elastic net model (Fig S6B). Shapley values are plotted for the most important features for patients in the test set; for each patient, negative Shapley values indicate that the feature influenced the model toward a prediction of no surgery, and positive Shapley values indicate that the feature influenced the model toward a prediction of surgery. Of note, the purpose of explainability studies is for reassurance that the model is relying on reasonable features rather than spurious associations. The purpose therefore is not to discover new associations and risk factors for glaucoma progression, for which traditional

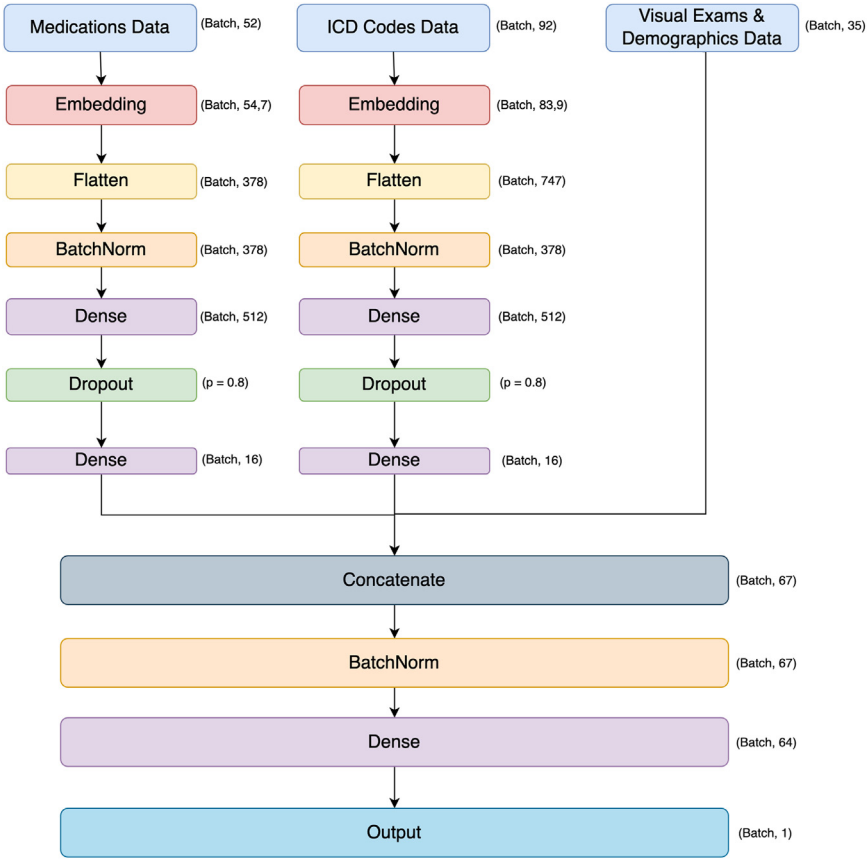

Figure 3. Deep learning model architecture. Depiction of the architecture of deep learning model predicting which patients would progress to surgery within 12 months. This model passes the medication and International Classification of Diseases (ICD) code features through an embedding layer.

hypothesis-driven inferential studies are more appropriate. Model explainability studies showed that the most important features included various demographic and clinical factors including age, maximum IOP, glaucoma diagnosis type, phakic status, and use of certain glaucoma medications. These features are similar to those that would be assessed by a glaucoma specialist when predicting whether a patient may progress to need surgery. Furthermore, as is clinically reasonable, for patients where the IOP values are high, the Shapley values indicate that the IOP feature influenced the model toward a prediction of surgery, and vice versa.

Discussion

This multicenter study developed and evaluated AI algorithms to predict whether patients with glaucoma would require surgery in the coming year, based on data extracted from the EHR. The study compared several ML and deep learning approaches and investigated which features are most important to the predictive models. The clinical features identified by these models align with characteristics that past studies have demonstrated to represent risk factors for glaucoma progression and features that most glaucoma specialists would find clinically relevant. By using data from multiple centers, we were able to explore the performance of our models on a single external test site and found that the generalizability was better preserved with tree-based and deep learning models as compared with regression models.

This study expands on past work developing models predicting which patients with glaucoma would soon require surgery that used EHR data from a single center. Baxter et al[9] explored a variety of ML models and a fully connected deep learning model for structured EHR data, achieving the best performance at predicting need for surgery with a logistic regression model (AUROC 0.67). Our previous work developing models using EHR structured and free-text data achieved AUROC ranging from approximately 0.70 up to 0.90.[7,8,11] Current results using multicenter data demonstrate performance within this range, with an AUROC up to 0.76 on the test data. Although prediction models using EHR data generally do not achieve the 0.99 AUROCs commonly seen in imaging classification models, the baseline human performance on prediction tasks is also lower, as shown in a prior study where a glaucoma specialist predicting likelihood of future surgery achieved only 0.25 sensitivity and 0.34 positive predictive value.[11] In the results, we show performance metrics for a threshold tuned to provide the best F1 score (balancing precision and recall), but with any potential use case, the

Table 3. Population Characteristics

	No Surgery		Surgery		Total	
	N = 30 529		*N* = 6019		*N* = 36 548	
	Mean	*Std Dev*	*Mean*	*Std Dev*	*Mean*	*Std Dev*
Age (years)	70.8	14.4	66.7	15.3	70.1	14.6
Best logMAR VA, OD	0.6	1.1	0.6	1	0.6	1.1
Best logMAR VA, OS	0.7	1.1	0.6	1	0.7	1.1
IOP max, OD (mmHg)	17	6.2	19.7	8	17.5	6.6
IOP max, OS (mmHg)	17.1	6.2	19.9	8.1	17.5	6.7
IOP max of either eye (mmHg)	18.7	7.1	22.3	9.1	19.3	7.6
Spherical equivalent, OD	−1.1	3.5	−1.1	3.6	−1.1	3.5
Spherical equivalent, OS	−1.1	3.4	−1.1	3.5	−1.1	3.5
CCT, OD (um)	550.7	52.3	549.7	53.2	550.5	52.4
CCT, OS (um)	551.3	53.6	549.9	54.9	551.1	53.8
	N	%	N	%	N	%
Female	16 722	54.8%	3231	53.68%	19 953	54.6%
Race						
White	19 579	64.1%	3623	60.2%	23 202	63.5%
Black	6904	22.6%	1513	25.1%	8417	23.0%
Asian	1942	6.4%	386	6.4%	2328	6.4%
American Indian or Hawaiian	91	0.3%	27	0.4%	118	0.3%
Other	1614	5.3%	411	6.8%	2025	5.5%
Unknown	399	1.3%	59	1.0%	458	1.3%
Ethnicity						
Hispanic	1185	3.9%	341	5.7%	1526	4.2%
Non-Hispanic	28 139	92.2%	5547	92.2%	33 686	92.2%
Unknown	1205	3.9%	131	2.2%	1336	3.7%
Rural/urban						
Rural	865	2.8%	231	3.8%	1096	3.0%
Urban	27 158	89.0%	5334	88.6%	32 492	88.9%
Missing	2506	8.2%	454	7.5%	2960	8.1%

CCT = central corneal thickness; IOP = intraocular pressure; logMAR = logarithm of the minimum angle of resolution; OD = right eye; OS = left eye; VA = visual acuity.

classification threshold can be tuned such that the model performs better on either precision or recall, depending on the desired operational characteristics. Thus, such EHR prediction models may still augment clinical predictions. In addition, as the current model was trained and tested on a much larger dataset across different medical centers, performance is less subject to overfitting on small validation datasets or instability in estimates due to small test datasets. Similar to previous studies,[7] we also show that deep learning and tree-based models generally outperformed simpler penalized regression models. Deep learning still faces many challenges in modeling for EHR data,[19] especially when using structured (tabular) datasets[20] with many features that may be sparse and noisy, such as diagnoses and medications. In such cases, tree-based models remain state-of-the-art,[20] as is consistent with our results, where the XGBoost model slightly outperformed the deep learning embedding model. Given that the differences in performance between the top models are small, the choice of what type of model to actually deploy in the future may ultimately be driven more by other factors such as computational cost.

Explainability studies comprise an important component of model evaluation, to assess whether models rely mostly upon reasonable input features rather than spurious associations. Our explainability studies indicated that the most important features included clinically reasonable features such as age, visual acuity, IOP, glaucoma diagnosis type, and use of certain glaucoma medications such as brimonidine or dorzolamide/timolol. These features are fairly consistent with results of explainability studies on previously single-center work, which also showed that IOP, visual acuity, and medication features were highly important.[7,8] Additionally, for some patients, diagnosis of cataract or phakic status were also important model inputs; this may reflect the fact that the threshold for taking patients to minimally invasive glaucoma surgery might depend more heavily on cataract status. Future iterations of the model could include experiments with predicting different subtypes of surgery as separate outcomes, modeling on subtypes of patients (such as pseudophakic only), or predicting alternative outcomes such as progression on visual fields. Finally, although Shapley values are a convenient mathematical technique for understanding model behavior to some extent, it is crucial to acknowledge that there are limitations to this technique, including prior studies that have suggested that results can sometimes be inaccurate and misleading.[21,22] Therefore, it is still essential to consider explainability results in the context of prior similar studies and overall established medical knowledge.

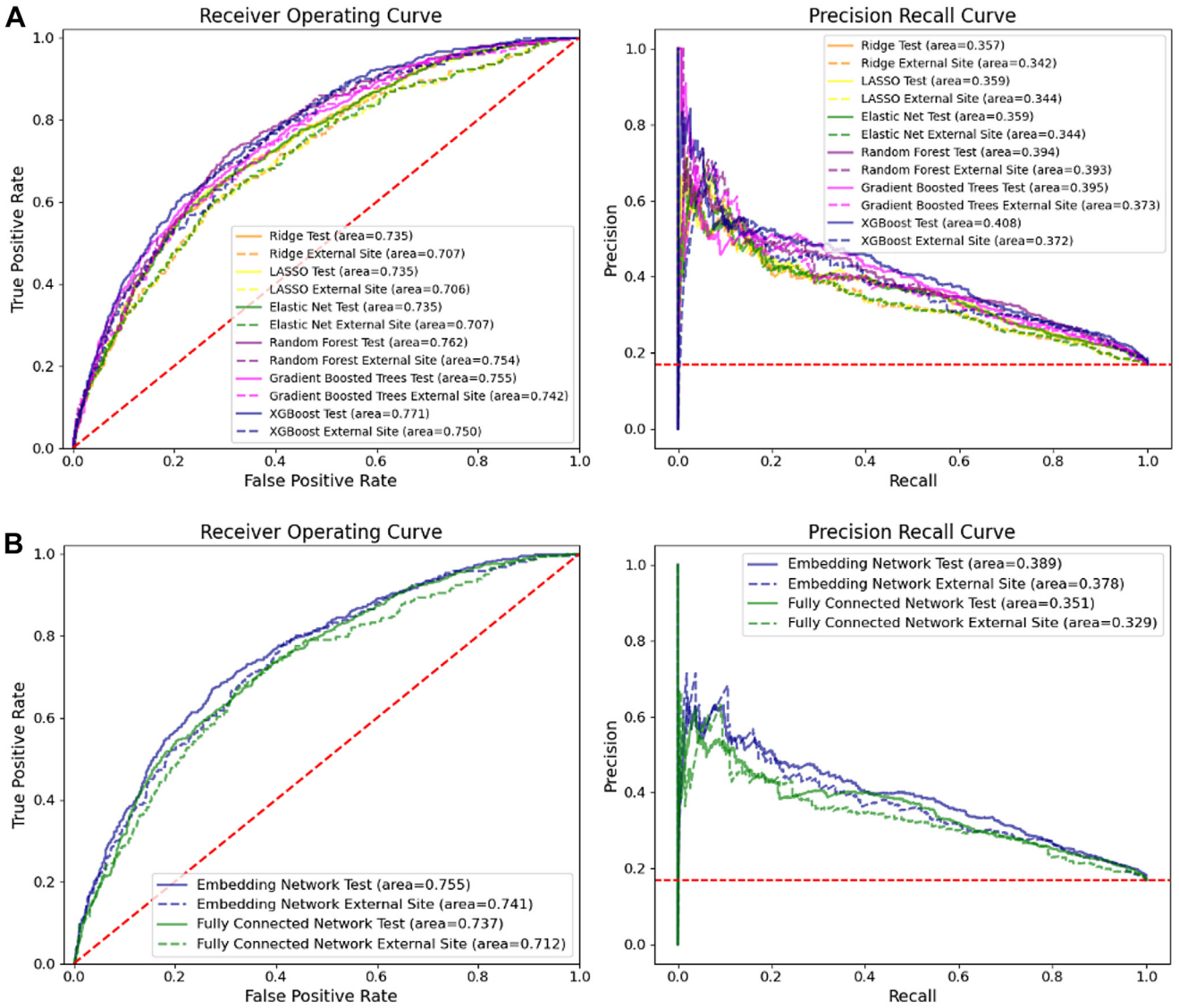

Figure 4. Receiver operating characteristic curves and precision-recall curves for machine learning and deep learning prediction models. This figure depicts receiver operating characteristic curves and precision-recall curves for models predicting glaucoma progression to surgery. **A,** Curves for machine learning models. **B,** Curves for deep learning models. All models were evaluated on the test set (comprising individuals from from the same sites as the training set) and an external test site (comprising individuals from a site that was not included in the training set). LASSO = least absolute shrinkage and selection operator.

Table 6. Model Performance Metrics

Model	Sensitivity (Recall)		Specificity		Positive Predictive Value (Precision)		Negative Predictive Value		F1		Accuracy		Threshold
	Test	*External*	*Test*	*External*	*Test*	*External*	*Test*	*External*	*Test*	*External*	*Test*	*External*	
L2 regression	0.536	0.603	0.808	0.708	0.364	0.296	0.895	0.898	0.434	0.397	0.762	0.690	0.56
L1 regression	0.471	0.538	0.838	0.752	0.373	0.307	0.885	0.889	0.416	0.391	0.775	0.716	0.59
Elastic Net Regression	0.509	0.595	0.818	0.725	0.365	0.306	0.891	0.898	0.425	0.404	0.766	0.703	0.57
Random Forest	0.439	0.611	0.865	0.759	0.399	0.340	0.883	0.905	0.418	0.437	0.792	0.734	0.48
Gradient Boosted Trees	0.494	0.592	0.847	0.759	0.399	0.333	0.891	0.901	0.441	0.426	0.787	0.731	0.21
XGBoost	0.551	0.645	0.822	0.693	0.388	0.300	0.899	0.906	0.456	0.409	0.776	0.685	0.21
Deep Learning Embedding Model	0.501	0.630	0.834	0.727	0.382	0.319	0.891	0.906	0.433	0.423	0.777	0.710	0.61
Deep Learning FCN	0.588	0.706	0.753	0.627	0.328	0.278	0.899	0.913	0.421	0.399	0.725	0.641	0.54

FCN = Fully Connected Model.

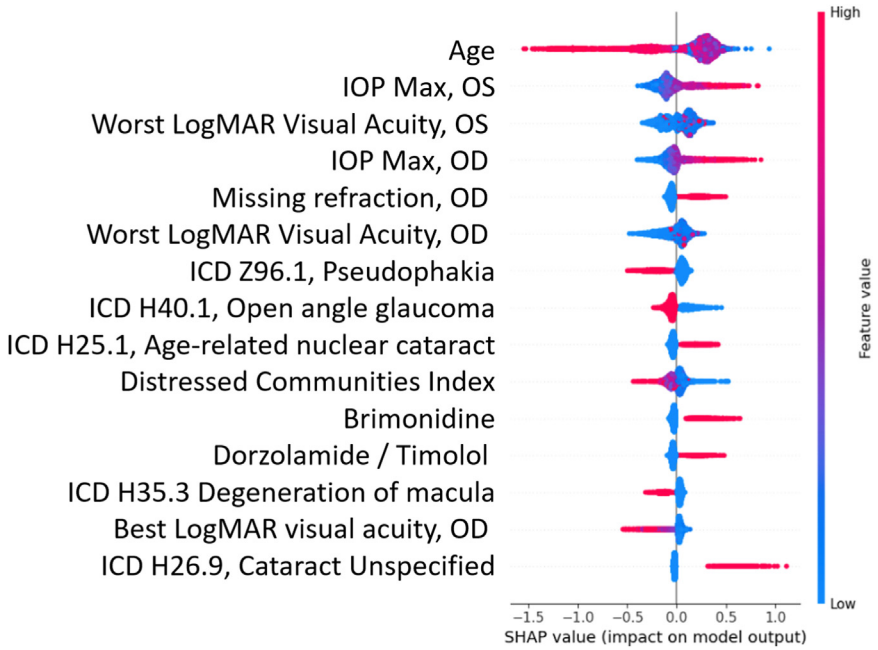

Figure 5. Model explainability with Shapley feature importance. The figure depicts the Shapley value for the top most important features for predicting whether a patient would progress to the point of requiring surgery within the next year, using the XGBoost model and calculated across the test set. Points represent individual observations (patients) in the test. The feature value color of each point indicates whether the value of that feature was high or low for that individual observation. A positive Shapley value for a feature for an individual point indicates influence toward a model prediction of surgery, whereas a negative Shapley value indicates influence toward a model prediction of no surgery. ICD = International Classification of Diseases; IOP = intraocular pressure; logMAR = logarithm of the minimum angle of resolution; OD = right eye; OS = left eye; SHAP = SHapley Additive exPlanations.

A unique strength of this multicenter study was the ability to test whether models trained on data from 1 set of sites could generalize to another site not used at all in training. Every model was thus evaluated on a test set (comprising data from individuals at the same sites as used for training) and an external held-out site. In almost all cases, the performance on the external held-out site was slightly decreased, although the decrease in AUROC was $< 3\%$. Our embedding-based deep learning model actually had improved performance on the external site (AUROC 0.76 external vs. 0.74 test). Overall, generalizability was remarkably preserved on the external test site, with tree-based ML models and deep learning models performing best. These generalizability results are consistent with or improved over prior work in other medical domains. A study predicting heart failure using data from > 400 sites showed approximately 3.6% reduction in AUROC when applied to external hospitals,[23] while a study predicting sepsis across 5 different emergency departments showed variable reductions in AUROC up to approximately 10% when trained and tested on different sites.[24] The authors postulated in their case that models trained on tertiary care emergency departments might not be expected to generalize well to critical-access emergency departments.[24] In our study, generalizability may have been enhanced by the fact that the rate of surgery at the external site was close to the mean overall rate of surgery, despite some variation in surgical patterns as evidenced by the range of surgery rates across sites. Additionally, different centers shared the same underlying EHR system and were all academic ophthalmology centers, despite being in different areas of the country and with patients of varying sociodemographic profiles. Model performance in other settings, such as individual or small group private practices, may be different. Regardless, a perfectly generalizable prediction algorithm is likely not a reasonable goal; models would benefit from transfer learning and fine-tuning[25] or complete retraining on local site data in service of achieving the best predictive performance and precision health for local patient populations.[26]

This study has several limitations. Centers participating in SOURCE were academic and typically serve as referral centers; as such, clinical data from prior to referral could not be incorporated as input features into our models. This lack of distinction between new and previously treated patients with glaucoma may be a limitation, as prior exposure to medications or other treatments may indicate their glaucoma severity upon inclusion into the study and could influence decisions for further surgery. Future research may be performed to evaluate how the model may perform differently for newly diagnosed versus follow-up glaucoma patients. In addition, glaucoma was defined by ICD coding, which can have inaccuracies and be subject to different patterns across different providers and institutions, likely leading to a somewhat heterogeneous population upon which the model was trained. However, this nature of the training data and cohort can translate into a strength for future deployment, where such models could potentially be used on patients who do have imperfectly coded diagnoses in their EHRs. Another limitation is that input features were derived from structured data within the EHRs and did not include imaging or free text. Our previous models combining structured data with free-text clinical notes suggest that incorporation of free-text

may improve performance.[7,8] Possibly, incorporation of free-text could capture some of the data from prior to referral which may be recorded into patients' initial progress notes, or it could capture information on other factors such as medication adherence. Similarly, we hypothesize that incorporation of imaging parameters either from optic nerve photos or retinal nerve fiber layer OCT could be of additional prognostic value, as may data from standard automated perimetry. Collecting, deidentifying, and harmonizing such text and imaging data across multiple sites are part of large-scale ongoing efforts to further develop the SOURCE repository. Once this data is available, we plan to integrate it into our models. Furthermore, there are multiple approaches to developing models that incorporate the time-horizon in prediction; our approach predicts progression to surgery in the next 12 months, but other survival-based AI approaches can also be tried. The temporal nature of EHR inputs (features) also have multiple potential representations[27] which will be an area of future comparison and studies. Finally, future studies will include a more comprehensive investigation of algorithm performance and fairness in subgroups of patients, such as by race, ethnicity, gender, and other characteristics.

In conclusion, we have shown that machine- and deep learning models can predict reasonably well whether patients with glaucoma will soon progress to requiring surgery using structured data from EHRs. We learned that tree-based and deep learning-based models outperformed the regression-based models for this complex high-dimensional data. We demonstrated that our model performance was relatively well preserved when tested on data from a new site, suggesting they can be applied to patients across multiple health systems. Explainability studies showed that important input features were clinically reasonable. Future studies could incorporate imaging or text data to further improve model performance.

Footnotes and Disclosures

Originally received: September 19, 2023.
Final revision: November 22, 2023.
Accepted: December 1, 2023.
Available online: December 6, 2023. Manuscript no. XOPS-D-23-00229.

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

[2] Department of Ophthalmology & Visual Sciences, University of Michigan Kellogg Eye Center, Ann Arbor, Michigan.

*Members of the SOURCE Consortium and their site PIs include: Henry Ford Health System: Sejal Amin, Paul A. Edwards; Johns Hopkins University: Divya Srikumaran, Fasika Woreta; Montefiore Medical Center: Jeffrey S. Schultz, Anurag Shrivastava; Medical College of Wisconsin: Baseer Ahmad, Judy Kim; Northwestern University: Paul Bryar, Dustin French; Scheie Eye Institute: Brian L. Vanderbeek; Stanford University: Suzann Pershing, Sophia Y. Wang; University of Colorado: Anne M. Lynch; Jenna Patnaik; University of Maryland: Saleha Munir, Wuqaas Munir; University of Michigan: Joshua Stein; Lindsey DeLott; University of Utah: Brian C. Stagg, Barbara Wirostko; University of West Virginia: Brian McMillian; Washington University: Arsham Sheybani. The SOURCE Data Center is located at the University of Michigan. The Chief Data Officer of SOURCE is Joshua Stein. The Lead Statistician of SOURCE is Chris Andrews. More information about SOURCE is available at https://www.sourcecollaborative.org/.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

Supported by the National Eye Institute (K23EY03263501; S.Y.W.); unrestricted departmental grant from Research to Prevent Blindness (S.Y.W., R.R.); and departmental grant from the National Eye Institute (P30-EY026877 [S.Y.W., R.R.], R01EY032475 [J.D.S.], and R01EY034444 [J.D.S.]).

HUMAN SUBJECTS: Human subjects data were included in this study. This study was approved by the University of Michigan and Stanford institutional review boards and adhered to the tenets of the Declaration of Helsinki. This is a retrospective study using de-identified subject details. Informed consent was not obtained.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Wang, Ravindranath

Data collection: Wang, Ravindranath, Stein

Analysis and interpretation: Wang, Ravindranath, Stein

Obtained funding: Wang, Stein

Overall responsibility: Wang, Ravindranath, Stein

This work is selected as an AAO meeting paper for the upcoming 2023 meeting.

Abbreviations and Acronyms:
AI = artificial intelligence; **AUROC** = area under the receiver operating curve; **CI** = confidence interval; **EHR** = electronic health records; **ICD** = International Classification of Diseases; **IOP** = intraocular pressure; **ML** = machine learning; **SOURCE** = Sight Outcomes Research Collaborative.

Keywords:
Machine learning, Glaucoma, Multicenter study, Deep learning.

Correspondence:
Sophia Y. Wang, MD, MS, Department of Ophthalmology, Byers Eye Institute, Stanford University, 2370 Watson Ct, Palo Alto, CA 94303. E-mail: sywang@stanford.edu.

References

1. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol*. 2006;90:262–267.
2. 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.
3. Actis AG, Dall'Orto L, Penna R, et al. An internal medicine perspective review of risk factors for assessing and progression of primary open angle glaucoma. *Minerva Med*. 2013;104:471–485.
4. Jain V, Jain M, Abdull MM, Bastawrous A. The association between cigarette smoking and primary open-angle glaucoma: a systematic review. *Int Ophthalmol*. 2017;37:291–301.

5. Grzybowski A, Och M, Kanclerz P, et al. Primary open angle glaucoma and vascular risk factors: a review of population based studies from 1990 to 2019. *J Clin Med Res*. 2020;9:761.
6. Kass MA, Heuer DK, Higginbotham EJ, et al. Assessment of cumulative incidence and severity of primary open-angle glaucoma among participants in the ocular hypertension treatment study after 20 years of follow-up. *JAMA Ophthalmol*. 2021;139:1−9.
7. Jalamangala Shivananjaiah SK, Kumari S, Majid I, Wang SY. Predicting near-term glaucoma progression: an artificial intelligence approach using clinical free-text notes and data from electronic health records. *Front Med (Lausanne)*. 2023;10:1157016.
8. 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.
9. Baxter SL, Marks C, Kuo T-T, 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.
10. 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.
11. Hu W, Wang SY. Predicting glaucoma progression requiring surgery using clinical free-text notes and transfer learning with transformers. *Transl Vis Sci Technol*. 2022;11:37.
12. American Medical Association. *CPT 2019 Professional Edition*. Chicago, IL: American Medical Association; 2018.
13. World Health Organization. *ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision*. Genève, Switzerland: World Health Organization; 2004.
14. Avati A, Jung K, Harman S, et al. Improving palliative care with deep learning. *BMC Med Inform Decis Mak*. 2018;18:122.
15. Cromartie J. Rural-urban commuting area codes. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/. Accessed April 18, 2023.
16. Kesler P. Distressed communities. Economic Innovation Group 2022. https://eig.org/distressed-communities/. Accessed April 18, 2023.
17. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st Interational Conference of Neural Information Processing Systems, Long Beach, California, USA*. 2017:4768−4777.
18. Lundberg S. shap. Github. https://github.com/slundberg/shap. Accessed May 16, 2023.
19. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2018;25:1419−1428.
20. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? arXiv [csLG] 2022. http://arxiv.org/abs/2207.08815. Accessed June 12, 2023.
21. Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S. Problems with Shapley-value-based explanations as feature importance measures. In: Iii HD, Singh A, eds. *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Cambridge, MA: Proceedings of Machine Learning Research. PMLR; 2020:5491−5500.
22. Huang X, Marques-Silva J. The inadequacy of Shapley values for explainability. arXiv [csLG] 2023. http://arxiv.org/abs/2302.08160. Accessed June 15, 2023.
23. Rasmy L, Wu Y, Wang N, et al. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J Biomed Inform*. 2018;84:11−16.
24. Ryu AJ, Romero-Brufau S, Qian R, et al. Assessing the generalizability of a clinical machine learning model across multiple emergency departments. *Mayo Clin Proc Innov Qual Outcomes*. 2022;6:193−199.
25. Wardi G, Carlile M, Holder A, et al. Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Ann Emerg Med*. 2021;77:395−406.
26. Futoma J, Simons M, Panch T, et al. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health*. 2020;2:e489−e492.
27. Xie F, Yuan H, Ning Y, et al. Deep learning for temporal data representation in electronic health records: a systematic review of challenges and methodologies. *J Biomed Inform*. 2022;126:103980.