AMERICAN ACADEMY OF OPHTHALMOLOGY®

Deep Learning Approaches for Predicting Glaucoma Progression Using Electronic Health Records and Natural Language Processing

Sophia Y. Wang, MD, MS,[1] *Benjamin Tseng,*[1] *Tina Hernandez-Boussard, PhD*[2]

Purpose: Advances in artificial intelligence have produced a few predictive models in glaucoma, including a logistic regression model predicting glaucoma progression to surgery. However, uncertainty exists regarding how to integrate the wealth of information in free-text clinical notes. The purpose of this study was to predict glaucoma progression requiring surgery using deep learning (DL) approaches on data from electronic health records (EHRs), including features from structured clinical data and from natural language processing of clinical free-text notes.

Design: Development of DL predictive model in an observational cohort.

Participants: Adult patients with glaucoma at a single center treated from 2008 through 2020.

Methods: Ophthalmology clinical notes of patients with glaucoma were identified from EHRs. Available structured data included patient demographic information, diagnosis codes, prior surgeries, and clinical information including intraocular pressure, visual acuity, and central corneal thickness. In addition, words from patients' first 120 days of notes were mapped to ophthalmology domain-specific neural word embeddings trained on PubMed ophthalmology abstracts. Word embeddings and structured clinical data were used as inputs to DL models to predict subsequent glaucoma surgery.

Main Outcome Measures: Evaluation metrics included area under the receiver operating characteristic curve (AUC) and F1 score, the harmonic mean of positive predictive value, and sensitivity on a held-out test set.

Results: Seven hundred forty-eight of 4512 patients with glaucoma underwent surgery. The model that incorporated both structured clinical features as well as input features from clinical notes achieved an AUC of 73% and F1 of 40%, compared with only structured clinical features, (AUC, 66%; F1, 34%) and only clinical free-text features (AUC, 70%; F1, 42%). All models outperformed predictions from a glaucoma specialist's review of clinical notes (F1, 29.5%).

Conclusions: We can successfully predict which patients with glaucoma will need surgery using DL models on EHRs unstructured text. Models incorporating free-text data outperformed those using only structured inputs. Future predictive models using EHRs should make use of information from within clinical free-text notes to improve predictive performance. Additional research is needed to investigate optimal methods of incorporating imaging data into future predictive models as well. *Ophthalmology Science 2022;2:100127 © 2022 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org

Glaucoma is the leading cause of irreversible blindness worldwide.[1] The clinical trajectories of patients with glaucoma can be highly variable, with some patients remaining stable with medical treatment, whereas others progress to require invasive surgery.[2] Although intraocular pressure (IOP) is the primary modifiable risk factor for glaucoma progression, many aspects of the wide range of glaucomatous disease phenotypes can contribute to variation in glaucoma disease course, including other examination findings, treatment patterns and medication adherence, and potential underlying secondary causes,[3–5] which makes it difficult to precisely predict whether a particular patient will have a stable or progressive clinical course.

Advances in artificial intelligence have enabled the development of predictive models in many medical fields[6,7] and have begun to enable predictive models in glaucoma, with many studies focused on predicting disease trajectory based on imaging and testing findings.[8–11] A wealth of data now resides within electronic health records (EHRs), and a previously published logistic regression model predicted glaucoma progression to surgery with an area under the receiver operating characteristic curve (AUC) of 0.67.[12] However, uncertainty remains regarding how to integrate the wealth of

https://doi.org/10.1016/j.xops.2022.100127
ISSN 2666-9145/22

clinical information residing in free-text clinical notes, which represent an entirely different method of data that has not yet been integrated into ophthalmology predictive models.

Free-text clinical progress notes in EHRs are difficult to access and compute over, requiring specialized natural language processing techniques, such as the use of neural word embeddings.[13–15] With word embeddings, individual words are mapped onto numeric vectors, such that word "meaning" is encoded within the geometry of vector space and words with similar meanings are clustered in vector space. Mapping words to neural word embeddings can provide an approach to integrating text into predictive models, because this effectively transforms the meaning of words into numeric values that can be computed over. In the field of medicine, word embedding approaches have been used over EHR text to predict unplanned readmission after hospital discharges,[16,17] to automate medical coding,[18] to identify intracranial hemorrhages from radiology reports,[19] and to identify patients with particular clinical phenotypes such as metastatic cancer, substance abuse, or obesity,[20] among other applications. We also previously developed specialized, ophthalmology domain-specific word embeddings that we have used to develop predictive models for low-vision prognosis.[21]

The purpose of this study was to predict glaucoma progression requiring surgery using ophthalmology domain-specific neural word embeddings to represent clinical notes. As a secondary outcome, we compared the performance of models that integrate free-text notes with those that used only structured input data from the EHRs. The results from this study can inform the development of future clinical decision support tools.

Methods

Data Source and Study Cohort

From the Stanford Clinical Data Warehouse,[22] we identified 4512 unique adult patients treated from 2009 through 2018 who underwent 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) or who had $\geq$ 2 instances of a glaucoma diagnosis, but did not undergo glaucoma surgery (International Classification of Disease, Ninth Revision, codes H40- [excepting H40.0-], H42-, Q150- and their International Classification of Disease, Ninth Revision, equivalents). Surgical patients must have had at least 120 days of baseline follow-up before surgery. Nonsurgical patients must have had at least 120 days of follow-up. In all, 748 surgical patients and 3764 nonsurgical patients were identified.

This study adhered to the tenets of the Declaration of Helsinki and was approved by the Stanford University Institutional Review Board. A waiver of informed consent was granted by the institutional review board because of the minimal risk posed to participants by review of observational health records and the large number of participants, which would have rendered the study infeasible if individual informed consent were required.

Data Preprocessing and Feature Engineering

Free-Text Clinical Progress Note Inputs. We identified and included up to the first 3 clinical progress notes from within the first 120 days of follow-up. All notes were lower-cased and tokenized (split into separate words), and punctuation and stop words were removed ("a," "all," "also," "an," "and," "are," "as," "at," "be," "been," "by," "for," "from," "had," "has," "have," "in," "is," "it," "may," "of," "on," "or," "our," "than," "that," "the," "there," "these," "this," "to," "was," "we," "were," "which," "who," "with"). Words were mapped to 300-dimensional neural word embeddings customized for ophthalmology, pretrained on PubMed ophthalmology abstracts.[21]

Structured Inputs. Structured features available from the research warehouse were processed either as Boolean (0 or 1) variables or as continuous numeric variables (standardized to mean of 0 and variance of 1). Features input as Booleans included gender, race, ethnicity, billing codes indicating prior diagnoses (n = 101 features), and current medication use (n = 241 features). Medications included any medications (both systemic and ophthalmic) recorded in the EHR medication list during the baseline period, standardized as they are coded in the EHR database, which is based on RxNorm. Features with $<$ 1% variance were removed. Numeric variables included age, best documented visual acuity for both eyes (measured in logarithm of the minimum angle of resolution units), and maximum documented IOP for both eyes. Missing clinical measurements were imputed using column mean imputation, and indicator variables were created to indicate whether an individual clinical measurement was missing. In total, 361 structured input features were included.

Modeling Approach

Overview. Three models were constructed for comparison on predicting whether a patient with glaucoma would progress to require glaucoma surgery (once or more), including a structured model that relied on only structured input features, a text model relying on only free-text clinical notes as input features, and a combination model that used both sets of features. All models were trained with hyperparameters and classification probability threshold tuned through grid search on a validation set (n = 400) to achieve optimal AUC and F1 score, respectively. Threshold tuning prevents all predictions from defaulting to "no surgery" because of the imbalanced nature of the dataset. Final performance was evaluated on a held-out independent test set of 500 patients. Deep learning (DL) models were trained in Python using the tensorflow[23] framework.

Structured Model. The structured model consisted of a densely connected neural network with a final sigmoid output to predict the probability that a patient would progress to require surgery. The model had the following architecture: 361 input features $\rightarrow$ 1024 dimension (rectified linear unit activation function, type L2 regularization) $\rightarrow$ dropout (0.5) $\rightarrow$ 64-dimension layer (rectified linear unit activation function, type L2 regularization) $\rightarrow$ sigmoid activation. Supplementary classical machine learning models were also trained in scikit-learn,[24] including L1 penalized regression, L2 penalized regression, elastic net penalized regression, and gradient boosted trees, with hyperparameters tuned over the validation set using grid search and final parameters shown in Supplemental Table 1.

Text Model. The text model was based on a convolutional neural network architecture[25] that had been previously shown to perform well in other ophthalmology-related[21] natural language tasks. The model architecture is depicted in Figure 1. The most recent 1000 words of clinical documentation were mapped first to previously developed ophthalmology domain-specific word embeddings that were pretrained on abstracts from PubMed relating to ophthalmology.[21] These were passed through a fully connected layer and then through multiple 1-dimensional convolutions before concatenation, max pooling, flattening, and passing

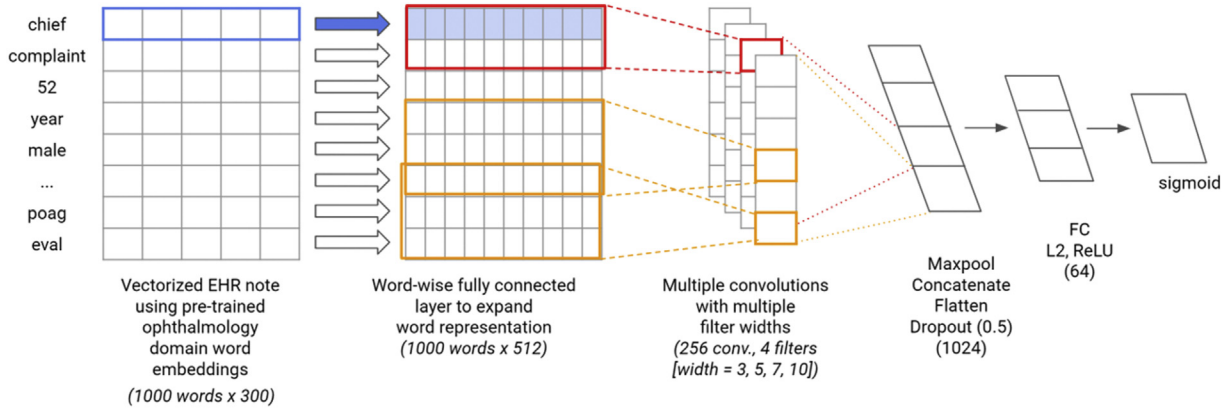

Figure 1. Deep learning model architecture relying on only free-text clinical notes as inputs to predict glaucoma progression to surgery. Architecture for free-text notes based on multiple 1-dimensional convolutions with different filter widths. L2 indicates the type of regularization used. EHR = electronic health record; FC = fully connected layer; ReLU = rectified linear unit activation function.

through another fully connected neural network head with final sigmoid output.

Combination Model. The model architecture that combined both methods of data, the structured input with the clinical progress notes, is shown in Figure 2. The combination of these 2 methods of data followed a late-fusion approach whereby each input was passed through its own model architecture before concatenating the outputs near the end and passing through a sigmoid output layer. The text was mapped to the aforementioned ophthalmology domain-specific word embeddings and then passed through a transformer block with 10-headed self-attention.[26] The structured features were passed through a fully connected layer of size 256. Both model streams were concatenated, with dropout set to 0.5 and passed through an additional fully connected layer of size 64 with type L2 regularization and through the final sigmoid output neuron.

Evaluation

We used as evaluation metrics sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, F1 score (the harmonic mean of recall and precision), and the AUC, all evaluated on the independent held-out test set of patients. In addition, to provide a baseline human-level for prediction performance, a glaucoma specialist (S.Y.W.) reviewed the charts of a sample of 300 patients with glaucoma from the test set to perform

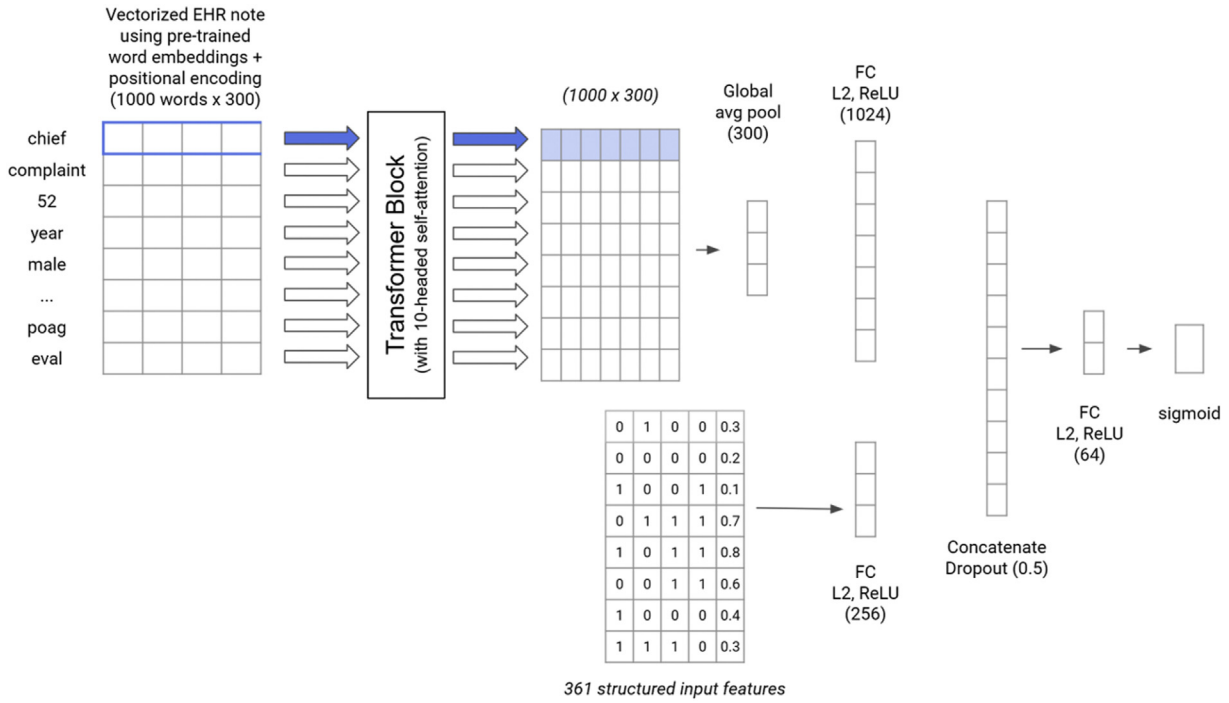

Figure 2. Deep learning model architecture using both free-text clinical notes and structured electronic health records (EHRs) data as inputs to predict glaucoma progression to surgery. Architecture that combines free-text notes and structured data into 1 predictive model. L2 indicates the type of regularization used. FC = fully connected layer; ReLU = rectified linear unit activation function.

clinical predictions on whether they would progress to require surgery using the same data available to the models above. We also performed explainability studies for the structured models using the locally interpretable model-agnostic explanations[27] framework and the Python lime package.[28] Locally interpretable model-agnostic explanation coefficients for structured features were averaged across the entire test set, indicating the magnitude and direction of influence that individual structured features had on the output prediction.

Code Availability

Code for the above-described analyses has been released into a publicly available repository.[29]

Results

Population characteristics are summarized in Table 1. The patients' mean age was 65 years, and the mean IOP for both eyes was near 18 mmHg. Mean logarithm of the minimum angle of resolution visual acuity for both eyes was near 0.4 (Snellen equivalent, approximately 20/50). The population was predominantly White and Asian. Seventeen percent of patients (n = 748) went on to require glaucoma surgery in this cohort.

Receiver operating characteristic and precision recall curves on the held-out test set for the structured, text, and combination models are depicted in Figure 3. The combination model showed the best AUC (0.731) followed by the text model (0.697) and the structured model (0.658). For the area under the precision recall curve, the text model was the best (0.431) followed by the combination model (0.392) and the structured model (0.284).

Performance metrics of F1, sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, and overall accuracy for each of the models and the ophthalmologist's clinical predictions are shown in Table 2. The probability threshold given for optimal F1 score on the validation set was selected. Additionally, the overall proportion of positive predictions (i.e., proportion of patients predicted to progress to surgery) made by each model and the clinical prediction is shown. Of note, the clinical prediction was the most conservative of all the models, predicting only 13% of patients to progress to surgery, compared with the DL models, which ranged from 28% to 51% predictions of progression. A trivial prediction model that predicts that no patient would progress to surgery would have an 83% accuracy, 100% specificity, but 0% sensitivity, because 17% of patients in our cohort progressed to surgery. Within this context, the overall accuracy was best for the clinical prediction (F1 = 0.79) as well as the specificity (F1 = 0.90) and precision (F1 = 0.34), but the F1 was the worst for the clinical prediction (F1 = 0.29). The text model had the highest F1 at 0.42, whereas the combined model had the best sensitivity (F1 = 0.77) and negative predictive value

Table 1. Population Characteristics

Characteristic	Total (n = 4512)	No Surgery (n = 3764)	Progressed to Surgery (n = 748)
Age (yrs)*	65.0 ± 17.9	65.0 ± 18.1	64.8 ± 17.0
IOP (mmHg)			
Right eye	18.3 ± 12.3	18.0 ± 6.2	20.1 ± 27.7
Left eye	18.8 ± 19.1	18.3 ± 6.5	21.8 ± 45.8
VA (logMAR)			
Right eye	0.39 ± 0.74	0.39 ± 0.74	0.43 ± 0.76
Left eye	0.43 ± 0.78	0.43 ± 0.79	0.43 ± 0.76
Female sex†	2270 (50.3)	1920 (51.0)	350 (46.8)
Race			
White	1892 (41.9)	1616 (42.9)	276 (36.9)
Asian/Pacific Islander	1225 (27.1)	992 (26.4)	233 (31.1)
Other/Native American	991 (22.0)	812 (21.6)	179 (23.9)
Black	216 (4.8)	168 (4.5)	48 (6.4)
Unknown	188 (4.2)	176 (4.7)	12 (1.6)
Ethnicity			
Non-Hispanic	3791 (84.0)	3159 (83.9)	632 (84.5)
Hispanic/Latino	566 (12.5)	460 (12.2)	106 (14.2)
Unknown	155 (3.4)	145 (3.9)	10 (1.3)
Common glaucoma medication use			
Latanoprost	1570 (34.8)	1279 (34.0)	291 (38.9)
Bimatoprost	413 (9.2)	311 (8.3)	102 (13.6)
Timolol	1709 (37.9)	1318 (35.0)	391 (52.3)
Dorzolamide	928 (20.6)	678 (18.0)	250 (33.4)
Brimonidine	1212 (26.9)	891 (23.7)	321 (42.9)
Acetazolamide	253 (5.6)	178 (4.7)	75 (10.0)

IOP = intraocular pressure; logMAR = logarithm of the minimum angle of resolution; VA = visual acuity.
Data are presented as mean±standard deviation or no. (%).
*Numeric variables were standardized to a mean of 0 and variance of 1 before input into model, but reported here conventionally for ease of interpretation.
†Categorical variables were separated into a series of Boolean dummy variables before input into model.

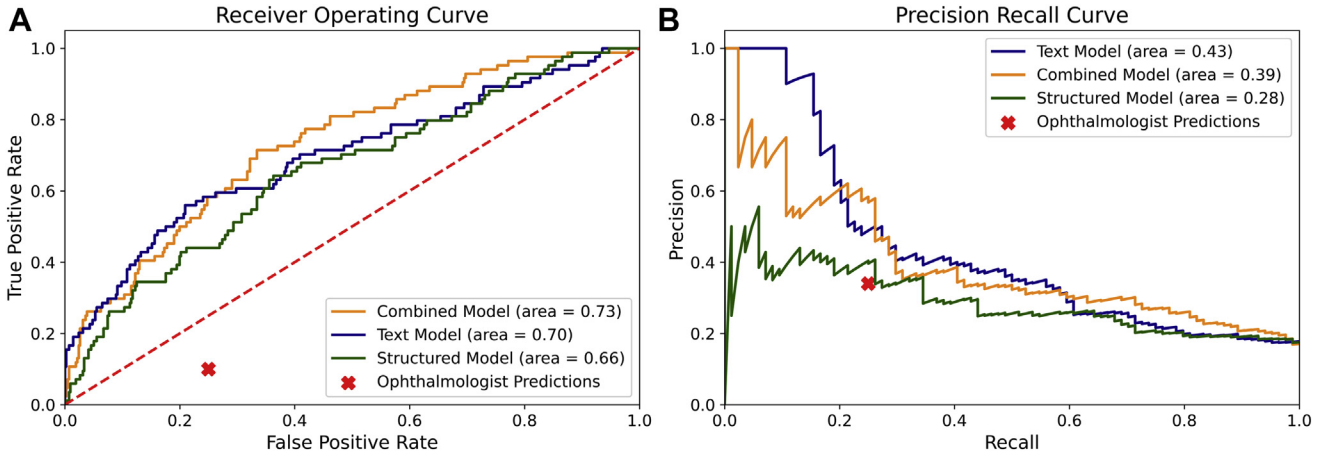

Figure 3. Graphs showing (**A**) receiver operating characteristic curves and (**B**) precision recall curves for the 3 different types of models developed to predict glaucoma progression to surgery based on electronic health records data: models that used structured data only, text data only, or a combination of both. The performance of an ophthalmologist reviewing the health records of these patients and making a clinical prediction is also shown.

(F1 = 0.93). Supplementary classical machine learning models were also trained on the structured inputs to provide general benchmarks for prediction performance on structured input data, with performance metrics presented in the Supplemental Table 1 and receiver operating characteristic and precision recall curves presented in the Supplemental Figure 1.

To gain an understanding of which features were being used by the structured models to make predictions, we used the locally interpretable model-agnostic explanations framework, which calculates the importance of individual features to make predictions on individual example inputs using local linear regressions to approximate the model decision boundary. Figure 4 shows the top 25 most important structured features for predicting surgery or no surgery across the test set. Important structured features include the use or nonuse of various glaucoma medications according to the medication list, presence or absence of important diagnosis codes, and IOP, which are similar to factors that a clinician may take into account when predicting the prognosis of a patient with glaucoma.

Discussion

In this large cohort study, we were able to predict whether patients with glaucoma would need glaucoma surgery in the future based on multiple methods of data from the EHRs at presentation using DL models. Models incorporating text performed better than models using only structured (non–free-text) data. The most important features included the use of various glaucoma medications, cataract or pseudophakia, and IOP. Development of predictive models for patients with glaucoma can be helpful for future clinical decision support tools or for automatically identifying low- or high-risk patients to stratify treatment strategies. In our study, the clinical prediction made by an ophthalmologist showed the best specificity, positive predictive value, and accuracy, but was also the most conservative and had the worst sensitivity, predicting only 13% of patients progressing to surgery. Thus, on the more balanced measure of performance (F1), all models outperformed the clinical prediction.

Previous models to predict glaucoma progression to surgery have focused on using structured data from electronic health records or on using imaging and testing data. Baxter et al investigated several different types of models, including DL and tree-based models, but ultimately found that a logistic regression model achieved the best performance at an AUC of 0.67,[8,23,24] similar to our performance using structured data only. We showed that incorporation of text from EHRs likely improves performance. Text data contains richer and more complete documentation of presenting symptoms, examination findings, and medical history compared with use of billing codes, which is likely to be incomplete, especially for new patients

Table 2. Performance Metrics for Models and Clinical Predictions

Variable	Proportion Predicted to Progress to Surgery	F1	Sensitivity (Recall)	Specificity	Positive Predictive Value (Precision)	Negative Predictive Value	Accuracy	Probability Threshold
Clinical predictions	0.13	0.29	0.25	**0.90**	**0.34**	0.85	**0.79**	—
Structured model	0.51	0.34	0.69	0.53	0.23	0.89	0.56	0.15
Text-only model	0.28	**0.42**	0.56	0.77	0.33	0.90	0.74	0.20
Combined model	0.49	0.40	**0.77**	0.57	0.27	**0.93**	0.60	0.15

— = not applicable.
Boldface indicates best value across all models.

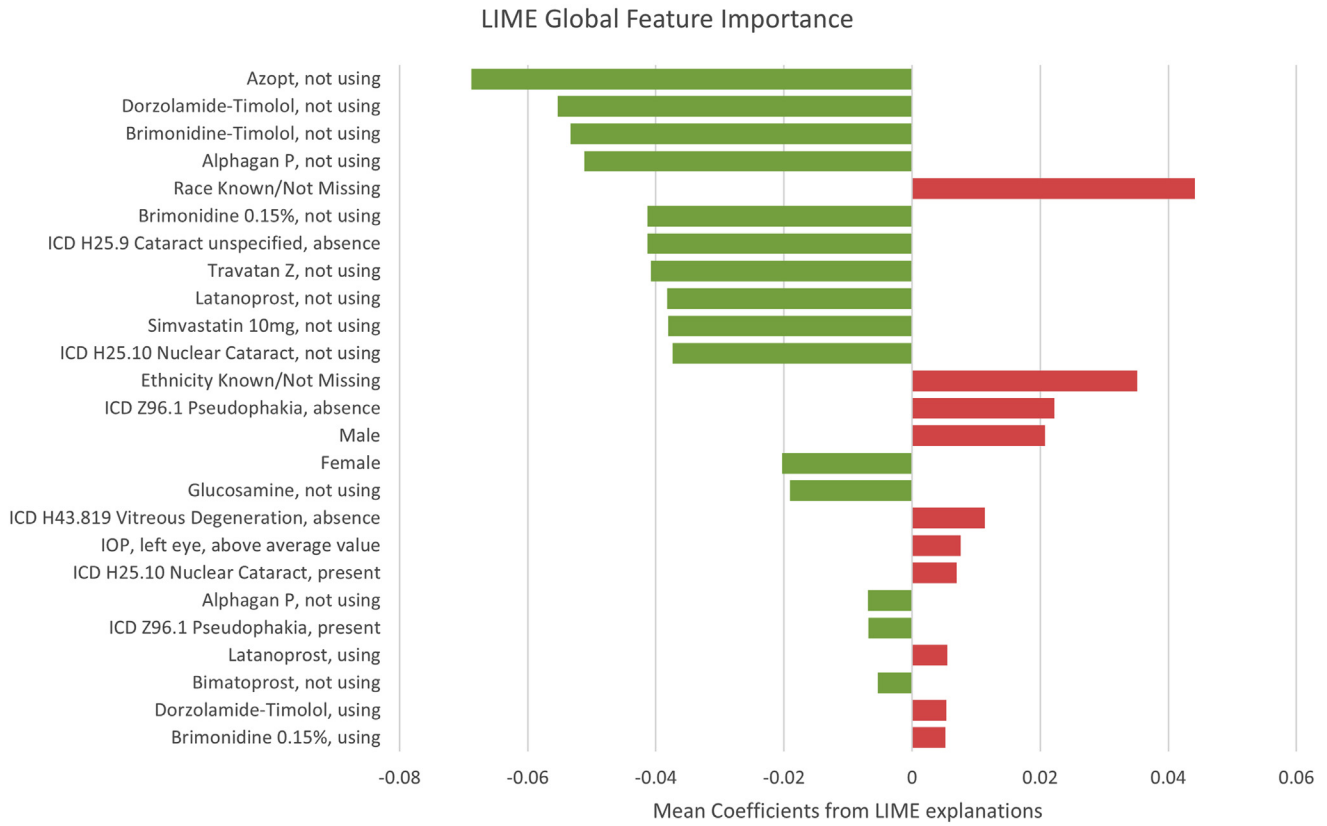

Figure 4. Graph showing the 25 most important features for the structured model predictions. The mean local interpretable model-agnostic explanation (LIME) coefficients of individual structured input features were averaged to produce an overall importance across the entire test set for each feature. Bars in red show features important for predicting that the patient would undergo surgery, whereas bars in green show features important for predicting that the patient would not undergo surgery. ICD = International Classification of Diseases; IOP = intraocular pressure.

without a long prior history in the medical system. Other predictive models for glaucoma progression have focused on using visual field data with or without limited clinical information (e.g., IOP),[8,30,31] retinal nerve fiber layer on OCT,[32] or fundus photographs[18] to predict glaucoma progression. In our models, imaging and testing information was not incorporated directly as input features; rather, their clinical interpretations were incorporated in the form of free text in the notes. Imaging information is yet another completely different method of data that requires future separate study to determine how best to fuse it with structured and textual EHR data into multimodality predictive models.

To provide a baseline performance comparison for our models, a glaucoma specialist (S.Y.W.) provided clinical predictions based on chart review. Only by having a baseline comparison can the models' performances be judged in context. Compared with the clinical predictions, all models showed better F1 score (a balanced measure of positive predictive value and sensitivity), mainly owing to better sensitivity. Thus, a potential role of this type of model in clinical decision support may be in identifying patients at high risk of glaucoma progression with higher sensitivity, although the optimal tradeoff between sensitivity and positive predictive value in the clinical context has yet to be determined. Of note, because the models provide probabilities of surgery as their outputs, the optimal cutoff threshold can be easily tuned in a way that a clinician's judgment cannot. In this case, the model thresholds were chosen to achieve optimal balanced F1, but thresholds could be optimized for other metrics as well, depending on the clinical context of model deployment and the costs of misclassification in each scenario. Furthermore, significant variation exists among glaucoma specialists regarding practice patterns, and variations in clinical judgment are likely to exist; thus, future work may benefit from gathering and comparing multiple clinicians' clinical predictions.

This study is among the first to investigate the usage of free-text clinical notes in predictive models for ophthalmology and to compare performance to models using only the structured, easy-to-access data in EHRs. We have used an innovative natural language processing pipeline tailored specifically for use on ophthalmology-domain language, through the use of specialized word embeddings[21] (which were previously made available to the public). This application of word embeddings to represent and incorporate text into predictive models involves minimal preprocessing of the text of the clinical progress notes and minimal computational power, thus increasing its appeal. This approach has significant advantages over manual

chart review or specific keyword searches of the text for particular terms of interest because manual review is inherently less scalable and keyword searching requires the items of interest to be determined a priori.

Another unique aspect of this study is the investigation of which features the DL models seemed to rely on most for their predictions, using the locally interpretable model-agnostic explanations[28] framework. For the structured model, we were able to generate a list of the most important features contributing to the predictions across the entire test set, which notably included many reasonable candidate features, such as the use of glaucoma medications and IOP. As expected, higher IOP was important for a surgery prediction. Use of many glaucoma medications was also important for surgery predictions, whereas absence of certain glaucoma medications from the medical record contributed to predictions for no surgery, which makes sense because patients who are taking more medications have fewer nonsurgical options for escalating therapy should they require it. Only a few of the top medication-related features were systemic medications; intriguingly, not using simvastatin contributed to a no-surgery prediction. Use of statins has been suggested to have a protective effect against incidence of glaucoma, although their effect on glaucoma progression or IOP is uncertain.[33] Thus, investigation of feature importance in predictive models may be helpful not only to examine the trustworthiness of the model, but also for hypothesis generation to guide further research.

This study has several limitations. Although we used word embeddings as an appealing approach to incorporating free text, other methods in natural language processing may exist that could further improve performance, such as transfer learning using transformer-based DL models pre-trained on massive text corpora.[34,35] However, computation becomes unwieldy as clinical documentation becomes longer, so input note length for free-text models must be limited regardless of how free-text is incorporated. Explainability studies relating individual portions of text to the final model prediction are also difficult to summarize across multiple examples and may represent an area for future research. Another important challenge of this work is the inherent imbalance in the cohort, representing the fact that only a relatively small percentage of patients with glaucoma progressed to require surgery, which poses difficulties for the training of predictive models in that optimizing for overall accuracy often results in a very low percentage of positive predictions. To combat this, the optimal prediction thresholds had to be tuned to achieve an optimal F1 score. Our models also did not use a fixed prediction window for progression to surgery, for example, progression within a 6- or 12-month window, which would further exacerbate the challenges of training on an imbalanced dataset. Future work could better incorporate the temporality of both the input data and the output prediction, taking into account potentially variable amounts of input data over time and outputting predictions over a fixed period. Finally, our study was limited to a single center consisting of a large academic hospital. Patients could pursue some portion of their care, such as surgery, at other institutions, either temporarily or permanently. Developing and validating natural language processing algorithms to identify whether glaucoma surgery was performed based solely on the free-text documentation is a direction for future work. Performing similar studies using multicenter data to capture patients who move between centers also would help to capture more complete clinical data on these mobile patients, as well as improve generalizability. The complex and sensitive nature of data sharing across institutions and to registries for clinical free text containing protected health information remains a challenge to be overcome. However, it must be noted that for truly personalized medical predictions, it actually may be preferable to tune predictive models to the specific setting in which they would be deployed, rather than striving for a one-size-fits all generalizable model for all patients.

In conclusion, we developed and investigated approaches to incorporating clinical free text into multimodality predictive models for glaucoma progression requiring surgery. Compared with models relying only on structured data from EHRs as inputs, the use of free-text data inputs may improve model performance. Future work can continue to explore other methods of incorporating text data into models and incorporating imaging and testing data into models to further improve prediction performance and build a pathway toward clinical deployment.

Footnotes and Disclosures

Originally received: December 7, 2021.
Final revision: January 19, 2022.
Accepted: February 7, 2022.
Available online: February 12, 2022. Manuscript no. D-21-00238.

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

[2] Center for Biomedical Informatics Research, Stanford University, Palo Alto, California.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have no proprietary or commercial interest in any materials discussed in this article.

Supported by the National Eye Institute, National Institutes of Health, Bethesda, Maryland (grant nos.: 1K23EY03263501 [S.Y.W.] and P30-EY026877 [S.Y.W., B.T.]); and Research to Prevent Blindness, Inc., New York, New York (Career Development Award [S.Y.W.] and unrestricted departmental grant [S.Y.W., B.T.]).

Presented in part at: American Glaucoma Society Annual Meeting, March 2021, Virtual.

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. A waiver of informed consent was granted by the Institutional Review Board due to the minimal risk posed to subjects by review of observational health records and due to the large number of subjects, which would have rendered the study infeasible if individual informed consent were required.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Wang, Tseng, Hernandez-Boussard

Analysis and interpretation: Wang, Tseng, Hernandez-Boussard

Data collection: Wang, Tseng, Hernandez-Boussard

Obtained funding: Wang, Tseng

Overall responsibility: Wang, Tseng, Hernandez-Boussard

Abbreviations and Acronyms:

AUC = area under the receiver operating characteristic curve; **DL** = deep learning; **EHR** = electronic health record; **IOP** = intraocular pressure.

Key Words:

Glaucoma, Artificial Intelligence, Deep Learning, Informatics.

Correspondence:

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

References

1. Resnikoff S, Pascolini D, Etya'ale D, et al. Global data on visual impairment in the year 2002. *Bull World Health Organ*. 2004;82:844−851.
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. 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.
4. 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.
5. 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.
6. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18.
7. Gensheimer MF, Henry AS, Wood DJ, et al. Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. *Ann Oncol*. 2018;29:viii548.
8. Garcia G-GP, Nitta K, Lavieri MS, et al. Using Kalman filtering to forecast disease trajectory for patients with normal tension glaucoma. *Am J Ophthalmol*. 2019;199:111−119.
9. Park K, Kim J, Lee J. Visual field prediction using recurrent neural network. *Sci Rep*. 2019;9:8385.
10. Lee T, Jammal AA, Mariottoni EB, Medeiros FA. Predicting glaucoma development with longitudinal deep learning predictions from fundus photographs. *Am J Ophthalmol*. 2021;225:86−94.
11. Medeiros FA, Jammal AA, Mariottoni EB. Detection of progressive glaucomatous optic nerve damage on fundus photographs with deep learning. *Ophthalmology*. 2021;128:383−392.
12. 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.
13. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, et al., eds. *Advances in Neural Information Processing Systems 26*. Red Hook, New York: Curran Associates, Inc.; 2013:3111−3119.
14. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. Available at: https://nlp.stanford.edu/projects/glove/; 2014. Accessed 16.06.20.
15. Khattak FK, Jeblee S, Pou-Prom C, et al. A survey of word embeddings for clinical text. *J Biomed Inform*. 2019;4:100057.
16. Craig E, Arias C, Gillman D. Predicting readmission risk from doctors' notes. *arXiv [statML]*. Available at: http://arxiv.org/abs/1711.10663; 2017. Accessed January 5, 2022.
17. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: a convolutional net for medical records. *IEEE J Biomed Health Inform*. 2017;21:22−30.
18. Patel K, Patel D, Golakiya M, et al. Adapting pre-trained word embeddings for use in medical coding. In: *Biomedical Natural Language Processing Workshop (BioNLP). 2017*. Vancouver: Association for Computational Linguistics; 2017:302−306.
19. Banerjee I, Madhavan S, Goldman RE, Rubin DL. Intelligent word embeddings of free-text radiology reports. *AMIA Annu Symp Proc*. 2017;2017:411−420.
20. Gehrmann S, Dernoncourt F, Li Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One*. 2018;13:e0192360.
21. Wang S, Tseng B, Hernandez-Boussard T. Development and evaluation of novel ophthalmology domain-specific neural word embeddings to predict visual prognosis. *Int J Med Inform*. 2021;150:104464.
22. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE—an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc*. 2009;2009:391−395.
23. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. Available at: https://www.tensorflow.org/; 2015. Accessed January 17, 2022.
24. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825−2830.
25. Kim Y. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics; 2014:1746−1751.
26. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv [csCL]*. Available at: http://arxiv.org/abs/1706.03762; 2017. Accessed March 1, 2021.
27. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. *arXiv [csLG]*. Available at: http://arxiv.org/abs/1602.04938; 2016. Accessed July 15, 2021.
28. Ribeiro MTC. lime. Available at: https://github.com/marcotcr/lime. Accessed 27.05.21.
29. Wang SY, Tseng B. eyelovedata/predictglaucomasurgery-ehr: v1.0.0. Available at: https://zenodo.org/record/5867032; 2022. Accessed January 17, 2022.

30. Wen JC, Lee CS, Keane PA, et al. Forecasting future Humphrey visual fields using deep learning. *PLoS One*. 2019;14:e0214875.
31. Berchuck SI, Mukherjee S, Medeiros FA. Estimating rates of progression and predicting future visual fields in glaucoma using a deep variational autoencoder. *Sci Rep*. 2019;9:18113.
32. Christopher M, Belghith A, Weinreb RN, et al. Retinal nerve fiber layer features identified by unsupervised machine learning on optical coherence tomography scans predict glaucoma progression. *Invest Ophthalmol Vis Sci*. 2018;59:2748−2756.
33. McCann P, Hogg RE, Fallis R, Azuara-Blanco A. The effect of statins on intraocular pressure and on the incidence and progression of glaucoma: a systematic review and meta-analysis. *Invest Ophthalmol Vis Sci*. 2016;57:2729−2748.
34. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234−1240.
35. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. *arXiv [csCL]*. Available at: http://arxiv.org/abs/1904.03323; 2019. Accessed March 16, 2021.