*Journal of the American Medical Informatics Association*, 2024, **31(2)**, 456–464
https://doi.org/10.1093/jamia/ocad213
Advance access publication 14 November 2023
**Research and Applications**

Research and Applications

# Prediction of multiclass surgical outcomes in glaucoma using multimodal deep learning based on free-text operative notes and structured EHR data

**Wei-Chun Lin, MD, PhD[1,2], Aiyin Chen, MD[2], Xubo Song, PhD[1], Nicole G. Weiskopf, PhD[1], Michael F. Chiang, MD[3,4], Michelle R. Hribar , PhD[1,2,3,*]**

[1]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, 3181 S.W. Sam Jackson Park Rd, Portland, OR, 97239, United States, [2]Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, 545 SW Campus Dr, Portland, OR, 97239, United States, [3]National Eye Institute, National Institutes of Health, 31 Center Dr MSC 2510, Bethesda, MD, 20892, United States, [4]National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD, 20894, United States

*Corresponding author: Michelle R. Hribar, PhD, Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, 3181 S.W. Sam Jackson Park Road, Portland, OR 97239 (hribarm@ohsu.edu)

## Abstract

**Objective:** Surgical outcome prediction is challenging but necessary for postoperative management. Current machine learning models utilize pre- and post-op data, excluding intraoperative information in surgical notes. Current models also usually predict binary outcomes even when surgeries have multiple outcomes that require different postoperative management. This study addresses these gaps by incorporating intraoperative information into multimodal models for multiclass glaucoma surgery outcome prediction.

**Materials and methods:** We developed and evaluated multimodal deep learning models for multiclass glaucoma trabeculectomy surgery outcomes using both structured EHR data and free-text operative notes. We compare those to baseline models that use structured EHR data exclusively, or neural network models that leverage only operative notes.

**Results:** The multimodal neural network had the highest performance with a macro AUROC of 0.750 and F1 score of 0.583. It outperformed the baseline machine learning model with structured EHR data alone (macro AUROC of 0.712 and F1 score of 0.486). Additionally, the multimodal model achieved the highest recall (0.692) for hypotony surgical failure, while the surgical success group had the highest precision (0.884) and F1 score (0.775).

**Discussion:** This study shows that operative notes are an important source of predictive information. The multimodal predictive model combining perioperative notes and structured pre- and post-op EHR data outperformed other models. Multiclass surgical outcome prediction can provide valuable insights for clinical decision-making.

**Conclusions:** Our results show the potential of deep learning models to enhance clinical decision-making for postoperative management. They can be applied to other specialties to improve surgical outcome predictions.

**Key words:** multimodal model; deep learning; multiclass surgical outcomes; glaucoma; operative notes.

## Introduction

The adoption of electronic health records (EHRs) has generated a large volume of clinical data that has shown significant potential for clinical research.[1–3] Applications for this data include retrospective analysis, comparative effectiveness research, and artificial intelligence (AI) applications. AI techniques have been successfully applied to EHR data in various domains, such as disease screening,[4] diagnosis improvement,[5] decision-making,[6] and treatment outcome predictions.[7]

One area of medicine that has potential to benefit from AI predictions is the anticipation of potential surgical outcomes. Accurately predicting surgical outcomes is difficult, however, due to the complex nature of postoperative recovery. AI techniques have been shown to outperform conventional risk stratification scores such as American Society of Anesthesiologists score and Surgical Apgar Score.[8–11] Further, machine learning algorithms have been used in various specialties to predict surgical outcomes of postoperative complications and mortality.[12] Some examples include predictions for postoperative complications for end-stage renal disease patients, quality of life improvement for degenerative cervical myelopathy, and myopic regression after corneal refractive surgery.[13–15] However, most of these surgical outcome prediction models focused on binary outcomes—success or failure—and included only structured data about patients' pre- and post-op conditions. Surgeries can have multiple outcomes that need different postoperative management.[16–19] In particular, glaucoma surgery postoperative care differs depending on the types of surgical failure and success.[16] However, previous models have only predicted binary outcomes of success or failure and used only pre- or post-op structured EHR data.[20–22]

**Received:** June 13, 2023. **Revised:** October 16, 2023. **Editorial Decision:** October 19, 2023. **Accepted:** October 25, 2023

Glaucoma is a disease of the optic nerve and is a leading cause of blindness worldwide.[23] Glaucoma surgeries attempt to reduce intraocular eye pressure (IOP) that can lead to optic nerve damage and vision loss. The surgeries have 3 long-term outcomes: failure due to elevated IOP, failure due to low IOP, and surgical success.[24] Glaucoma surgery outcomes are highly dependent on postoperative management within the first 3 months following surgery and this care differs according to anticipated long-term outcomes.[16,17] For example, the dosage of steroids, usage of antifibrotic agents, and timing of suture lysis may differ based on the anticipated long term IOP.[16,17] Thus, it is clinically important to predict multiple surgical outcomes immediately following glaucoma surgery.

While most previous surgery prediction models use pre- and post-op structured patient data, there is a wealth of intraoperative information reflecting patients' conditions and care *during the surgery* that are useful for predicting outcomes. The operative note is a clinical document that records intraoperative information such as surgical findings, procedures performed, and the patient's condition during the surgery. Incorporating this information into the prediction model may improve the model's performance but requires extra processing since operative notes are free-text documents. Natural language processing (NLP) can be used to process free-text EHR data with techniques such as deep contextualized word representations,[25–27] information extraction,[28,29] and text classification.[30,31]

Recently, deep learning multimodal models have combined structured EHR data and free-text data to develop prediction models, leveraging flexible deep neural networks that integrate diverse functional blocks in a single model.[32–34] In this study, we hypothesize that a multimodal deep neural network incorporating both structured EHR data and free-text operative notes can enhance the prediction of multiclass glaucoma surgical outcomes. We aim to evaluate the impact of including operative notes in these multimodal models and to identify the most effective methods for extracting information from these notes, comparing transformer encoder blocks, long short-term memory (LSTM) units, and a pretrained large language model (LLM) called Bio-Clinical Bidirectional Encoder Representations from Transformers (BERT).[31,35,36]

## Methods

### Overview

We evaluate the prediction power of operative notes in multimodal models for multiclass glaucoma surgery outcome prediction. We implemented and evaluated 3 groups of models: (1) structured EHR data alone, (2) operative notes alone, and (3) a multimodal combination of both data. We optimized the model architecture and hyperparameter settings of all models.

### Setting

Oregon Health & Science University (OHSU) is a large academic medical center in Portland, Oregon. This study was conducted at Casey Eye Institute (CEI), OHSU's ophthalmology department serving all major ophthalmology subspecialties. The department performs over 130 000 outpatient examinations annually and is a major referral center in the Pacific Northwest and nationally. This study adhered to the tenets of the Declaration of Helsinki and was approved by the Institutional Review Board at OHSU (IRB No. STUDY00020203).

### Cohort

The study included patients aged 18 years or older who underwent primary trabeculectomies (Current Procedural Terminology codes 66170 and 66172) from January 1, 2010 to May 31, 2021, at OHSU CEI. We collected structured EHR data and operative notes for the study patients from the enterprise-wide clinical warehouse. Patients were excluded if they: (1) were under 18 years old; (2) did not have a complete operative note; (3) had a trabeculectomy combined with other procedures except phacoemulsification; (4) had less than 1 year of follow-up visit data.

#### Input features: structured EHR data

The structured data features were composed of preoperative EHR features and early postoperative features. The preoperative data included demographic information, glaucoma diagnosis, active medication usage before the surgery, chronic systemic diseases, procedure type, conjunctiva conditions, the best recorded distance visual acuity measures (converted to continuous logMAR values),[37] and the highest IOP recorded 6 months prior to surgery. For demographic information, patient's age, gender, ethnicity, insurance information, and smoking history were included. The postoperative features included IOP measures at multiple time points (at day 1, day 2 to day 14, and day 15 to day 30) and the best visual acuity measured within 30 days (again, converted to logMAR). All categorical features were converted into binary features. Numeric features were normalized, and linear imputation was used to handle the missing data. In addition, all features with less than 2% variance were removed. The final input structured dataset contained 75 features.

#### Input features: free-text operative notes

Operative notes are free-text clinical documents that record detailed information about the surgery, including all steps carried out in the procedure, medications or materials utilized, surgical findings, complications discovered intraoperatively, and the estimated blood loss. In our study, we used the primary trabeculectomy operative notes for each eye. All notes were preprocessed by removing special characters and punctuation, converting to lowercase, removing custom stop words, and tokenizing the text. Tokenization divides text into smaller units called tokens, which can be individual words, punctuation marks, or subword units. Tokenization helps standardize text input for further NLP task. The tokenized operative notes were then mapped to custom 50D word embeddings by training the unsupervised word2vec model[38] with 50D word embeddings. The word2vec model is a neural network-based technique that learns to map words to vectors of real numbers, capturing semantic relationships among them. We optimized the dimensions of word embeddings (50, 100, 300) with the classification models, and the 50D word embeddings showed the best performance. The operative notes were processed to a fixed length of 512 tokens to ensure the consistency of the input text; longer notes ($n = 25$) were truncated, and the shorter notes were padded. The word2vec model was trained on approximately 5000 glaucoma-related operative notes from the CEI data warehouse. We utilized the Gensim toolkit[39] with the Continuous

Bag of Words[40] model to conduct the training in the Python 3.9.0 environment.[41]

### Outcome variable: surgical outcome

Outcomes for trabeculectomy surgeries at 1 year are: (1) surgical success, (2) surgical failure due to elevated IOP, and (3) surgical failure due to low IOP (hypotony). Surgical failure due to elevated IOP was defined as postoperative IOP higher than 21 mmHg or less than 20% IOP reduction from preoperative baseline for 2 consecutive follow-up visits after 3 months, or reoperation for glaucoma due to continuous high IOP. Similarly, surgical failure due to hypotony was defined as postoperative IOP of 5 mmHg or lower on 2 consecutive follow-up visits 3 months after surgery, or reoperation due to hypotony. Eyes not meeting failure criteria were considered surgical successes.

## Models

We developed 3 groups of classification models to identify glaucoma patients with high risk of surgical failures after 30-day postsurgery: (1) models that used structured EHR data alone, (2) models that used unstructured operative notes alone, and (3) multimodal models that combined structured and unstructured data inputs. These models were implemented using Pytorch,[42] Scikit-learn,[43] and Hugging Face's Transformers[44] in the Python environment.[41] We experimented with several architectures for the transformer-based and LSTM-based multimodal neural networks. For the transformer encoder blocks, our primary investigations used architectures akin to BERT. The best configuration comprised 12 transformer layers, 768 hidden units per layer, and 10 attention heads. Additionally, our evaluations of the LSTM models revealed that an architecture with a singular layer, 50 hidden units, and a unidirectional structure yielded the optimal results.

All models were trained with hyperparameters, including optimizers, learning rates, batch size, and dropout rates. During the training process, the models were optimized to minimize the cross-entropy loss value. To circumvent any bias arising from the data imbalance, which could cause the models to consistently predict success, we fine-tuned the models on a validation set by monitoring the macro-averaged area under the receiver operating characteristic curve (AUROC) score and F1 score. We also incorporated class weights, with the weight decay (L2 regularization) set to 1e-5 to handle the overfitting issue. The highest-performing neural network models incorporated the Adam optimizer, utilizing the ReLU activation function. The models were trained with a batch size of 16 and an initial learning rate of 4e-5.

### EHR structured data classification model

We trained an artificial neural network (ANN) and a random forest (RF) with structured input features as the baseline models. The ANN model features 2 dense layers with a dropout (0.5) layer, followed by an output layer with softmax function to predict the probability of surgical outcomes (Dimension: 75D -> 256D -> 64D -> 3D). The RF used a bootstrap aggregating-based ensemble method that is popular in many clinical prediction models.[45] Five-fold cross-validation was used to tune the hyperparameters of the RF model and avoid overfitting.

### Text classification model

We developed a classification model using the operative notes, comparing 2 popular text classification models: transformer encoder block and LSTM neural networks, since both had been previously shown to perform well in text classification tasks.[46] The pretrained word embeddings from our notes were input to the transformer encoder blocks and LSTM layer (50 hidden units), then connected to the 2 dense layers and the softmax output layer (Dimension: 50D -> 256D -> 64D -> 3D). Batch normalization and dropout (0.5) layers were used to prevent gradient vanishing and overfitting. The transformer model consisted of 10 attention heads, 12 layers of transformer blocks, and 768 hidden units.

### Multimodal model

We developed multimodal deep learning models to verify our hypothesis that operative notes can improve the predictive model performance by combing structured input features with unstructured notes. Figure 1 shows the multimodal neural network architecture which combines both structured input features and operative notes using 3 different text processing models: transformer encoder blocks, LSTM models, and a LLM: Bio-Clinical BERT (note: we used this model only in the multimodal model and not in the text classification models). The intermediate-fusion strategy was used to combine the 2 types of models. The operative notes were mapped to the aforementioned custom word embeddings and then passed through transformer encoder blocks (12 layers) with 10 attention heads and 768 hidden units (Figure 1A) and the LSTM layer with 50 hidden units (Figure 1B). A global average pooling layer was added to output the final text vector, which was then concatenated with the structured features. For the Bio-Clinical BERT model (Figure 1C), we utilized the BERT tokenizer to process the operative notes. The structured features were input to the model and concatenated with the text vector and passed through 2 dense layers before reaching the final output layer with a softmax function (Dimension: 125D -> 256D -> 48D -> 3D). Batch normalization and the dropout (0.5) layers were used to prevent the gradient vanishing and overfitting. The code for the transformer-based and Bio-Clinical BERT multimodal models is publicly available on GitHub.

## Model evaluation

The dataset was randomly split on the patient level: 70% of the data was used for training, 10% for validation, and 20% for testing. We used the AUROC, area under precision–recall curve (AUPRC), balanced accuracy, precision, recall, specificity, negative predictive value (NPV) and F1 score as the main evaluation metrics on the test dataset. For the multiclass classification task, we calculated the macro average and One-vs-Rest (ovr) of AUROC and the macro average and per class of precision, recall, specificity, NPV, F1 score, and AUPRC. Additionally, we calculated a precision–recall curve for different thresholds.

# Results

Table 1 shows the descriptive characteristics of the patients in the study cohort. A total of 1540 eyes from 1326 patients who underwent trabeculectomy between January 2010 and May 2021 met the inclusion criteria. At 1 year, 193 (13%) eyes were defined as surgical failure due to hypotony, 183

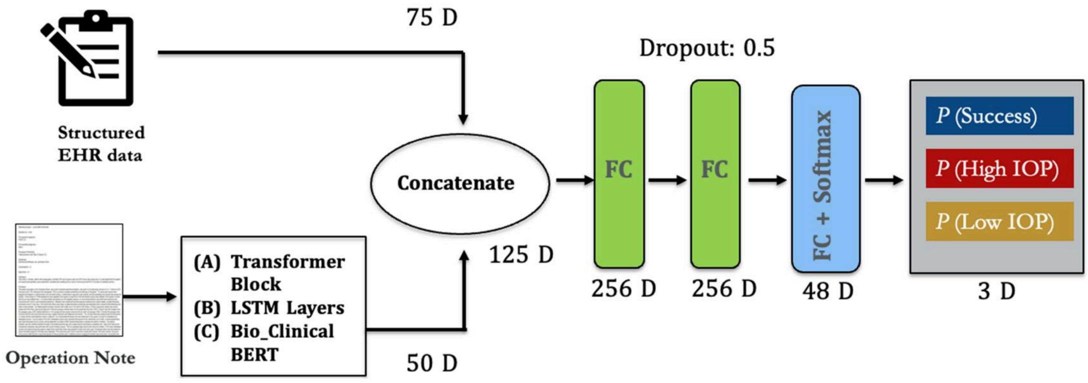

**Figure 1.** Overview of multimodal neural network architectures used: (A) transformer-based model, (B) LSTM-based model, and (C) Bio-Clinical BERT-based model. Abbreviations: LSTM, long short-term memory; BERT, Bidirectional Encoder Representations from Transformers.

**Table 1.** Demographic and clinical characteristics.

| | Total (1540 eyes) |
|---|---|
| Age, years | |
| Mean (SD) | 63.55 (15.69) |
| Sex | |
| Male | 661 (43%) |
| Female | 879 (57%) |
| Race | |
| White | 1329 (86%) |
| Non-White Hispanics | 56 (4%) |
| Black | 50 (3%) |
| Asians | 60 (4%) |
| Others | 45 (3%) |
| Clinical characteristics | |
| Mean intraocular pressure (mmHg) | 21.07 (8.6) |
| Mean visual acuity (logMAR) | 0.25 (0.41) |
| Mean number of glaucoma medications | 2.59 (1.41) |
| Surgical outcomes | |
| Success | 1164 (75%) |
| Low IOP surgical failure | 193 (13%) |
| Elevated IOP surgical failure | 183 (12%) |
| Healthcare insurance | |
| Medicaid | 62 (4%) |
| Medicare | 756 (49%) |
| Commercial insurance | 641 (42%) |
| Unknown | 81 (5%) |

(12%) eyes were defined as surgical failure due to elevated IOP, and 1164 (75%) eyes were defined as surgical success. The patient demographic showed that the majority of patients were Caucasians (86%) and females (57%), with a diagnosis of primary open-angle glaucoma (72%). The patient's average age was 64 years and most patients had Medicaid (49%), followed by commercial insurance (42%). The mean IOP before the surgery was near 21 mmHg and the mean logMAR visual acuity was 0.25, which is equivalent to Snellen of 20/36.

Figure 2 presents the macro receiver operating characteristic curves on the test dataset for: (1) the ANN and RF models with structured input features alone, (2) the text classification models (text-transformer and text-LSTM) with operative notes alone, and (3) the transformer, LSTM, and Bio-Clinical BERT multimodal models (MNN-transformer, MNN-LSTM, and MNN-BioClinicalBERT) with both structured EHR data and operative notes. Table 2 shows macro-average evaluation metrics, including the precision, recall, specificity, NPV, balanced accuracy, F1 score, AUPRC, and AUROC for the 6 models. The transformer multimodal neural network had the highest macro-average AUROC (0.750), AUPRC (0.564), and F1 score (0.583), followed by Bio-Clinical BERT multimodal neural network (AUROC = 0.735; AUPRC = 0.538; F1 score = 0.526), and the LSTM multimodal neural network (AUROC = 0.725; AUPRC = 0.537; F1 score = 0.491). The predictive models with structured EHR data or operative notes alone showed lower model performance for AUROC, AUPRC, and F1 score.

ROCs and P–R curves of each class and macro average on the test dataset for the transformer multimodal neural network are shown in Figures 3 and 4. Also in Table 3, the evaluation metrics for each class were depicted for the 3 multimodal neural networks. The model shows the highest AUROC (0.787, ovr) for the elevated IOP surgical failure group, followed by the hypotony surgical failure group (0.756, ovr), and the surgical success group (0.707, ovr). In addition, the model had the highest recall (0.692) for hypotony surgical failure, while the surgical success group had the highest precision (0.884) and F1 score (0.775). Overall, the model showed a better discriminate ability to predict the elevated IOP surgical failure (AUROC = 0.787; AUPRC = 0.463; F1-score = 0.512) than hypotony surgical failure (AUROC = 0.756; AUPRC = 0.373; F1-score = 0.462).

## Discussion

In this study, we used deep learning models to predict multiclass surgical outcomes for glaucoma patients and evaluated the predictive power of operative notes in these models. We also compared methods to extract information from the operative notes in a multimodal deep learning model. We had 3 key findings: (1) operative notes provide useful predictive information, (2) the transformer-based multimodal neural network model outperformed the baseline model, and the other multimodal models using Bio-Clinical BERT and LSTM, and (3) using multiclass surgical outcome prediction for glaucoma patients provides relevant information for clinical decision-making.

Our first key finding was that operative notes can be used to improve surgical outcome prediction by incorporating structured EHR data. To the best of our knowledge, this work is the first study investigating the usage of free-text operative notes in a multimodal predictive model for surgical

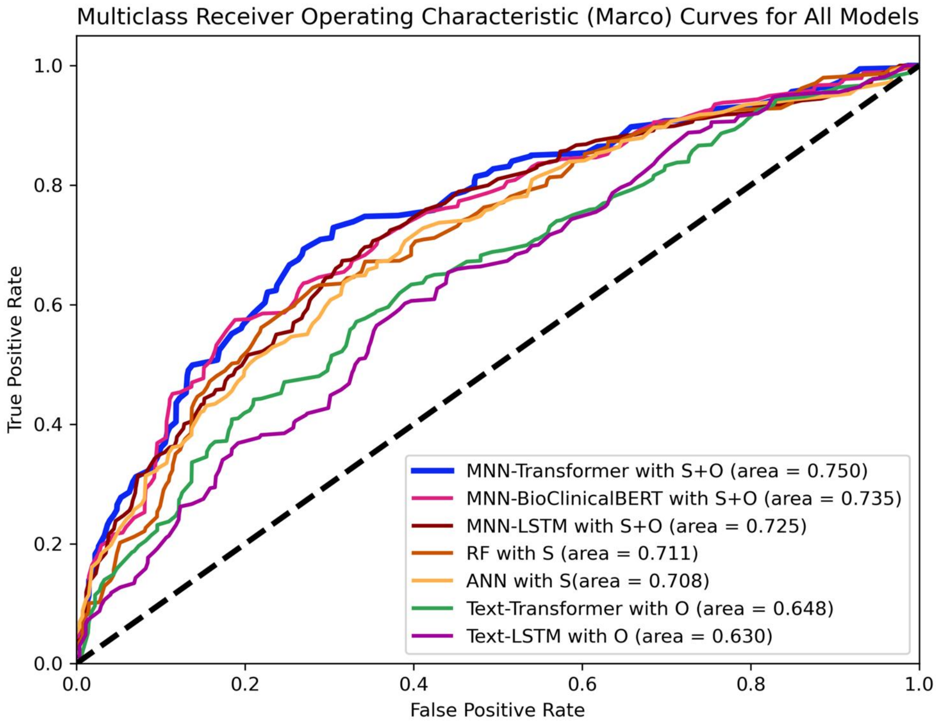

**Figure 2.** Multiclass receiver operating characteristic curves on the test dataset for the ANN, random forest, text-transformer, text-LSTM, transformer-based multimodal model, Bio-Clinical BERT-based model, and LSTM-based multimodal model. Abbreviations: S, structured data; O, operative notes; ANN, artificial neural network; RF, random forest; MNN, Multimodal Neural Network; LSTM, long short-term memory; BERT, Bidirectional Encoder Representations from Transformers.

**Table 2.** Comparison of model performance using macro-average metrics.

| | Surgical outcomes predictions (macro) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Model inputs | Precision | Recall | Specificity | NPV | Balanced accuracy | F1 score | AUPRC | AUROC |
| ANN | *S* | 0.492 | 0.470 | 0.723 | 0.730 | 0.597 | 0.476 | 0.521 | 0.708 |
| Random forest | *S* | 0.495 | 0.487 | 0.738 | 0.723 | 0626 | 0.486 | 0.529 | 0.712 |
| Text-Transformer | *O* | 0.461 | 0.410 | 0.707 | 0.752 | 0.411 | 0.414 | 0.416 | 0.648 |
| Text-LSTM | *O* | 0.388 | 0.409 | 0.697 | 0.693 | 0.409 | 0.391 | 0.409 | 0.630 |
| MNN-Transformer | *S* + *O* | **0.559** | **0.659** | **0.811** | **0.774** | **0.735** | **0.583** | **0.564** | **0.750** |
| MNN-BC_BERT | *S* + *O* | 0.509 | 0.581 | 0.760 | 0.738 | 0.670 | 0.526 | 0.538 | 0.735 |
| MNN-LSTM | *S* + *O* | 0.477 | 0.527 | 0.749 | 0.735 | 0.635 | 0.491 | 0.537 | 0.725 |

Abbreviations: *S*, structured data; O, operative notes; ANN, artificial neural network; MNN, Multimodal Neural Network; LSTM, long short-term memory; BC_BERT, Bio-Clinical BERT; NPV, negative predictive value; AUPRC, area under precision–recall curve; AUROC, area under the receiver operating characteristic.
The bold in the table is maximum values of that evaluation metrics.

outcomes. Previous studies mostly used structured data, including patients' pre-existing conditions or early postoperative clinical measures to develop surgical outcome prediction models.[13–15,20,47] The result of this study demonstrated the multimodal neural network that combined structured inputs and unstructured operative notes (transformer multimodal neural network, macro-average AUROC = 0.75; AUPRC = 0.564; F1 score = 0.583) performed better than the model using structured data alone (RF model, macro-average AUROC = 0.712; AUPRC = 0.529; F1 score = 0.486) or operative notes alone (transformer-based text classification model, macro-average AUROC = 0.648; AUPRC = 0.416; F1 score = 0.414). The operative notes included information that was not present in the structured EHR data, such as surgical findings and techniques, medications or materials used, and intraoperative complications, etc. This finding indicates that free-text operative notes and structured inputs can complement each other in predictive modeling, which may lead to performance improvement. This concept can be applied to other surgeries and specialties to improve model performance of surgical outcome prediction.

Our second key finding revealed that the transformer-based multimodal model, supplemented with custom word embeddings, outperformed other methods. In our study, the transformer-based multimodal model outperformed both Bio-Clinical BERT and LSTM-based models in terms of macro AUROC and F1 scores. This superiority is likely attributable to the limitations of LSTMs in handling long text sequences effectively. Additionally, our transformer-based model used custom word embeddings, which seemed to offer

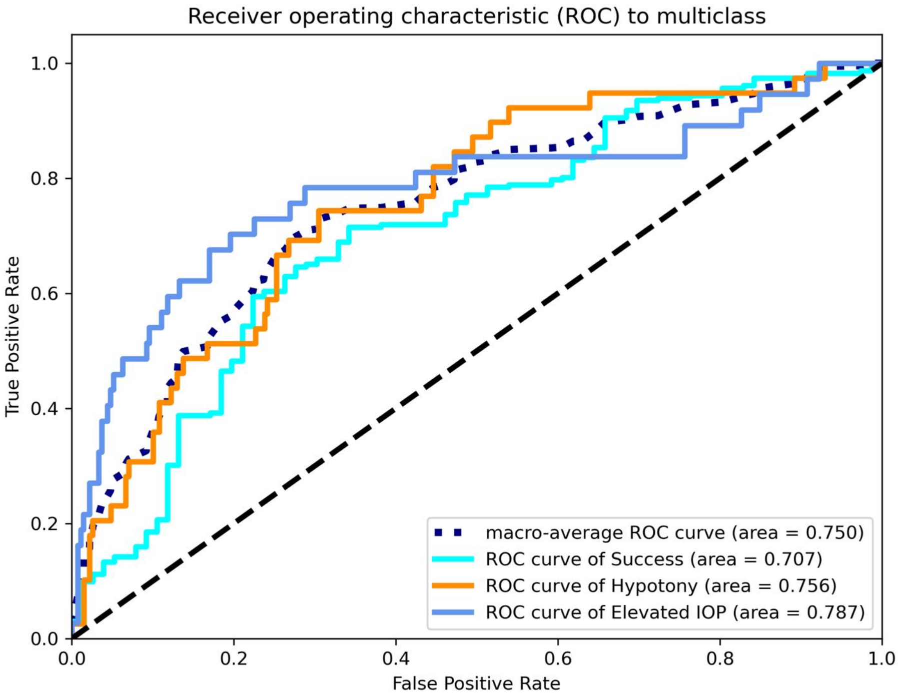

**Figure 3.** Receiver operating characteristic curves of each class and macro average on the test dataset for the transformer multimodal neural network.

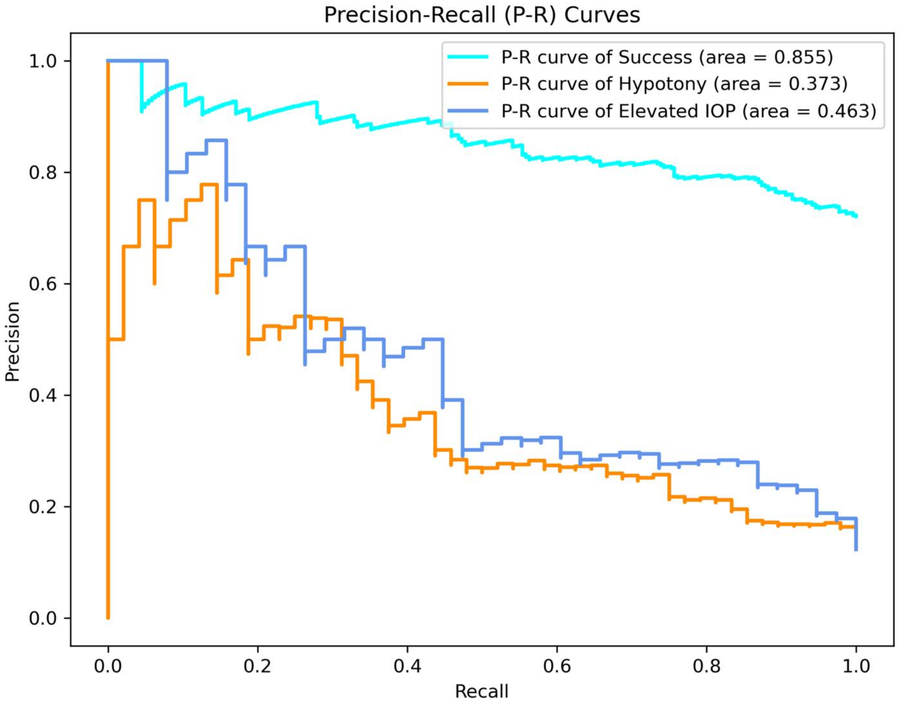

**Figure 4.** Precision–recall curves of each class on the test dataset for the transformer multimodal neural network.

advantages in information extraction for specific tasks, compared to pretrained BERT models trained on broader, larger medical text corpora. Before incorporating Bio-Clinical BERT into the multimodal architecture, we also experimented with text classification using Bio-Clinical BERT on operative notes alone and found comparable results with the transformer-based model (not shown).

In recent years, LLMs have attracted attention due to their emergent properties and multimodal capabilities. These features give them an advantage in processing and understanding complex textual information. Several LLMs, trained on domain-specific clinical data like Bio-Clinical BERT, are publicly available. These models have found multiple applications in clinical and healthcare settings, including clinical

**Table 3.** Performance metrics for each class for the multimodal neural networks.

**Transformer multimodal neural network**

| | Precision | Recall | Specificity | NPV | F1 score | AUPRC | AUROC |
|---|---|---|---|---|---|---|---|
| Success | 0.884 | 0.689 | 0.724 | 0.433 | 0.775 | 0.855 | 0.707 |
| Low IOP | 0.346 | 0.692 | 0.810 | 0.948 | 0.462 | 0.373 | 0.756 |
| High IOP | 0.449 | 0.595 | 0.901 | 0.943 | 0.512 | 0.463 | 0.787 |
| Macro average | 0.559 | 0.659 | 0.811 | 0.774 | 0.583 | 0.564 | 0.750 |

**Bio-clinical BERT multimodal neural network**

| | Precision | Recall | Specificity | NPV | F1 score | AUPRC | AUROC |
|---|---|---|---|---|---|---|---|
| Success | 0.863 | 0.672 | 0.575 | 0.353 | 0.745 | 0.847 | 0.673 |
| Low IOP | 0.264 | 0.514 | 0.804 | 0.924 | 0.349 | 0.322 | 0.742 |
| High IOP | 0.426 | 0.556 | 0.901 | 0.939 | 0.482 | 0.446 | 0.779 |
| Macro average | 0.509 | 0.581 | 0.760 | 0.738 | 0.526 | 0.538 | 0.735 |

**LSTM multimodal neural network**

| | Precision | Recall | Specificity | NPV | F1 score | AUPRC | AUROC |
|---|---|---|---|---|---|---|---|
| Success | 0.783 | 0.649 | 0.524 | 0.384 | 0.709 | 0.819 | 0.654 |
| Low IOP | 0.314 | 0.458 | 0.860 | 0.916 | 0.373 | 0.382 | 0.710 |
| High IOP | 0.333 | 0.474 | 0.862 | 0.905 | 0.391 | 0.409 | 0.797 |
| Macro average | 0.477 | 0.527 | 0.749 | 0.735 | 0.491 | 0.537 | 0.725 |

Abbreviations: LSTM, long short-term memory; NPV, negative predictive value; AUPRC, area under precision–recall curve; AUROC, area under the receiver operating characteristic.

decision support, EHR data processing, radiology reporting, and patient-interaction chatbots.[48–50] However, it is important to note that these models, while promising, have limitations. They are prone to issues such as hallucination, sensitivity to prompt variations, and brittleness.[51]

In this study, we used the word2vec-based custom word embeddings to preprocess the free-text operative notes. Previous studies have shown that transformer blocks with pretrained embedding gave promising results compared to classical deep learning models in text classification tasks.[36,46,52] However, those studies often relied on large datasets comprising over a million sentences. In contrast, our study utilized a relatively smaller dataset—around 150 000 sentences—to develop custom pretrained word embeddings. Given that data scarcity is a common challenge, our findings suggest that transformer encoder, when combined with custom word embeddings, can be effective even with limited sample sizes.

Our final key finding was that our result demonstrates that machine learning algorithms can be utilized to predict multiclass outcomes, which is crucial for postoperative care decision-making for surgeries with multiple outcomes such as glaucoma. In a previous study, machine learning algorithms with structured EHR data were used to predict the binary outcome of success or failure for long-term surgical outcomes for glaucoma patients with AUROC of 0.74.[20] In our study, we focused on the multiclass outcome of success, failure due to high IOP, or failure from hypotony, which provides more practical assistance for clinical decision-making, since postoperative treatment varies drastically according to the anticipated long-term surgical outcome.17,[53] For example, patients at risk for high IOP failure need early laser suture-lysis or suture adjustment, antifibrotic injections, higher steroid dosing, and even mini-procedures such as in-office needling to break down early fibrosis. Needling performed within 4-6 months from the surgery was reported to be more successful.[54] On the other hand, patients at risk for hypotony should not have laser suture-lysis, anti-fibrotic agent, and a steroid taper should be considered. All of these management options are performed during the postoperative period at the doctor's discretion. Therefore, the model's ability to predict surgical failure due to either high or low IOP can help guide the doctor's management to create a personalized postoperative plan. In our study, with the reasonable threshold settings, the optimized multimodal neural network achieved a recall score of 0.691 for the hypotony surgical failure cohort and 0.543 for the elevated IOP surgical failure cohort. This result indicated that the model could correctly identify more than half of the patients with surgical failures within 1 year. With this result, the prediction model might be able to inform physicians of the potential risk for specific surgical failures so that the patients could receive more appropriate therapy.

Although these models have not reached ideal performance (ie, AUROC and F1 scores close to 1), they still represent advances in model development and prediction and are consistent with other similar models. For example, a recently published study using multimodal machine learning approaches to predict myopic regression after corneal refractive surgery showed the performance with AUROC of 0.75.[15] Similarly, prior research has shown that machine learning algorithms, using preoperative surgical data, could identify higher-risk groups for glaucoma surgical failure with AUROC ranging from 0.64 to 0.74 for binary surgical outcomes.[20] With recent LLM advancements, there is potential for future model performance improvements in multiple ways: (1) data augmentation with LLMs like GPT-4 can generate synthetic operative notes to mitigate data imbalances[55]; (2) clinical domain LLMs could derive custom embeddings that would enhance the predictive capabilities of the model[56]; and (3) multimodal capabilities of LLMs like GPT-4 may improve prediction accuracy by capturing richer contextual information.[57]

Despite these innovative findings, there were several limitations in our study. First, an important challenge of this work is the naturally inherent imbalanced dataset in our study cohort. In our study, less than 25% of glaucoma patients were considered surgical failures, making it difficult to train the prediction model. Also, 86% of our patient population was Caucasian, our findings might have limited applicability to more diverse populations. Further, our model did not include ocular imaging data, which was not available for most of the patients in the study cohort. Integrating imaging data into the multimodality predictive model might improve the model performance and is a future direction for this research. Next, the outcome variables in our study related only to postoperative IOP and reoperations, but there are other surgical outcomes such as visual deterioration that occur despite well-maintained IOP. Additionally, our research did not provide an exhaustive explainability analysis for the multimodal models. Moreover, due to the nature of the retrospective study, we were not able to collect expert judgment to ensure a thorough comparison. A prospective clinical study for glaucoma surgeons will be needed in the future. Furthermore, although we collected one of the largest clinical observation datasets with detailed information about trabeculectomy outcomes, the sample size and diversity are still limited due to the study being from a single institution. Future studies will ideally include data from multiple institutions. Our focus was limited trabeculectomy; future studies will include other glaucoma surgeries. Lastly, given our study's emphasis on intraoperative and early postoperative data, our findings are more germane to postoperative decision-making than preoperative patient selection.

## Conclusion

In this study, we developed multimodal prediction models for multiclass surgical outcomes of glaucoma surgery using structured EHR data and free-text operative notes to address the need for effective postoperative management. Our result demonstrated that intraoperative information in operative notes improved the prediction model's performance. Also, we explored a better method to extract information from operative notes in a multimodal prediction model. We believe that our work can be helpful for clinical decision-making for postoperative care in glaucoma surgeries, and the implications of the study can be extended to other surgical specialties to improve outcome predictions. In the future, we plan to incorporate imaging data as well as multisite data to improve the model performance.

## Author contributions

All listed authors (W.-C.L., A.C., X.S., N.G.W., M.F.C., and M.R.H.) meet the 4 criteria for authorship as they have significantly contributed to both the conception and design of this study, interpretation of results, and the drafting and finalization of the manuscript. M.R.H. and W.-C.L. have additionally fulfilled the role of acquiring data for the study. W.-C.L. has further performed in the analysis of the collected data and the development of the underlying model.

## Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## Funding

This work was supported by grant numbers R01LM013426, R21LM013937 (formerly R21EY031443), T15LM007088, and P30EY10572 from the National Institutes of Health (Bethesda, MD). It was also partly funded by the Intramural Research Program of the National Institutes of Health, National Eye Institute (Bethesda, MD) (M.F.C.), and received unrestricted departmental funding from Research to Prevent Blindness (New York, NY).

## Conflicts of interest

None declared.

## Data availability

The code and the models used in this study are available at the links below. The data used in this study cannot be shared publicly due to privacy concerns since it includes free-text intraoperative notes.

1) https://github.com/WeiChunLin/Custom_Transformer_Multimodal_Model
2) https://github.com/WeiChunLin/Bio_Clinical_BERT_Multimodal_Model
3) https://github.com/WeiChunLin/Word2Vec-Ophthalmology-operative-notes

## References

1. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc*. 2007;14(1):1-9.
2. Charles D, King J, Patel V, et al. ONC Data Brief No. 9 March 2013 Adoption of Electronic Health Record Systems among US Non-federal Acute Care Hospitals: 2008-2012. Furukawa, PhD The Health Informat, 2013.
3. Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. *J Intern Med*. 2013;274(6):547-560.
4. Kar A, Wreesmann VB, Shwetha V, et al. Improvement of oral cancer screening quality and reach: the promise of artificial intelligence. *J Oral Pathol Med*. 2020;49(8):727-730.
5. Huang S, Yang J, Fong S, et al. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett*. 2020;471:61-71.
6. Magrabi F, Ammenwerth E, McNair JB, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. *Yearb Med Inform*. 2019;28(1):128-134.
7. Shamout F, Zhu T, Clifton DA. Machine learning for clinical outcome prediction. *IEEE Rev Biomed Eng*. 2021;14:116-126.
8. Kim JS, Merrill RK, Arvind V, et al. Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. *Spine (Phila Pa 1976)*. 2018;43(12):853-860.
9. Kim JS, Arvind V, Oermann EK, et al. Predicting surgical complications in patients undergoing elective adult spinal deformity procedures using machine learning. *Spine Deform*. 2018;6(6):762-770.
10. Hill BL, Brown R, Gabel E, et al. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. *Br J Anaesth*. 2019;123(6):877-886.

11. Lee CK, Hofer I, Gabel E, et al. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology*. 2018;129(4):649-662.
12. Elfanagely O, Toyoda Y, Othman S, et al. Machine learning and surgical outcomes prediction: a systematic review. *J Surg Res*. 2021;264:346-361.
13. Jeong Y-S, Kim J, Kim D, et al. Prediction of postoperative complications for patients of end stage renal disease. *Sensors*. 2021;21(2):544.
14. Merali ZG, Witiw CD, Badhiwala JH, et al. Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. *PLoS One*. 2019;14(4):e0215133.
15. Kim J, Ryu IH, Kim JK, et al. Machine learning predicting myopic regression after corneal refractive surgery using preoperative data and fundus photography. *Graefes Arch Clin Exp Ophthalmol*. 2022;260(11):3701-3710.
16. Murdoch I. Post-operative management of trabeculectomy in the first three months. *Commun Eye Health*. 2012;25(79-80):73.
17. Vijaya L, Manish P, Ronnie G, et al. Management of complications in glaucoma surgery. *Indian J Ophthalmol*. 2011;59(Suppl 1):S131-S140.
18. Shen Y, Sardar ZM, Chase H, et al. Predicting bone health using machine learning in patients undergoing spinal reconstruction surgery. *Spine (Phila Pa 1976)*. 2023;48(2):120-126.
19. Peretto G, Durante A, Limite LR, et al. Postoperative arrhythmias after cardiac surgery: incidence, risk factors, and therapeutic management. *Cardiol Res Pract*. 2014;2014:615987.
20. Banna HU, Zanabli A, McMillan B, et al. Evaluation of machine learning algorithms for trabeculectomy outcome prediction in patients with glaucoma. *Sci Rep*. 2022;12(1):2473-2411.
21. Esfandiari H, Pakravan M, Loewen NA, et al. Predictive value of early postoperative IOP and bleb morphology in mitomycin-C augmented trabeculectomy. *F1000Res*. 2017;6:1898.
22. Nesaratnam N, Sarkies N, Martin KR, et al. Pre-operative intraocular pressure does not influence outcome of trabeculectomy surgery: a retrospective cohort study. *BMC Ophthalmol*. 2015;15(1):17-18.
23. Allison K, Patel D, Alabi O. Epidemiology of glaucoma: the past, present, and predictions for the future. *Cureus*. 2020;12(11):e11686.
24. Gedde SJ, Schiffman JC, Feuer WJ, et al.; Tube Versus Trabeculectomy Study Group. Three-year follow-up of the tube versus trabeculectomy study. *Am J Ophthalmol*. 2009;148(5):670-684.
25. Sarzynska-Wawer J, Wawer A, Pawlak A, et al. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res*. 2021;304:114135.
26. Yang X, Bian J, Hogan WR, et al. Clinical concept extraction using transformers. *J Am Med Inform Assoc*. 2020;27(12):1935-1942.
27. Devlin J, Chang M-W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*. 2019;1:2.
28. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. 2018;77:34-49.
29. Lin W-C, Chen JS, Kaluzny J, et al. Extraction of active medications and adherence using natural language processing for glaucoma patients. *AMIA Annu Symp Proc*. 2021;2021:773-782.
30. del Carmen Legaz-García M, Martínez-Costa C, Menárguez-Tortosa M, et al. A semantic web based framework for the interoperability and exploitation of clinical models and EHR data. *Knowl Based Syst*. 2016;105:175-189.
31. Luan Y, Lin S. Research on text classification based on CNN and LSTM. *IEEE ICAICA*. 2019;2019:352-355.
32. Kiela D, Firooz H, Mohan A, et al. The hateful memes challenge: detecting hate speech in multimodal memes. *Adv Neural Inf Process Syst*. 2020;33:2611-2624.
33. Zhang D, Yin C, Zeng J, et al. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak*. 2020;20(1):280-211.
34. Mahajan SM, Ghani R. Combining structured and unstructured data for predicting risk of readmission for heart failure patients. *MedInfo*. 2019;264:238-242.
35. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019:72-78.
36. Tezgider M, Yildiz B, Aydin G. Text classification using improved bidirectional transformer. *Concurr Comput Pract Exp*. 2022;34(9):e6486.
37. Ricci F, Cedrone C, Cerulli L. Standardized measurement of visual acuity. *Ophthal Epidemiol*. 1998;5(1):41-53.
38. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*. 2013.
39. Řehůřek R, Sojka P. *Gensim—python framework for vector space modeling*. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic. 2011;3(2):2.
40. Rong X. word2vec parameter learning explained. arXiv preprint, arXiv:1411.2738. 2014.
41. Van Rossum G, Drake FL. Python 3 *Reference Manual*. CreateSpace; 2009.
42. Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst*. 2019;32:8024-8035.
43. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
44. Wolf T, Debut L, Sanh V, et al. Huggingface's transformers: state-of-the-art natural language processing. arXiv preprint, arXiv:1910.03771. 2019.
45. Biau G, Scornet E. A random forest guided tour. *Test*. 2016;25(2):197-227.
46. Soyalp G, Alar A, Ozkanli K, et al. Improving text classification with transformer. *IEEE UBMK*. 2021;2021:707-712.
47. Jalali A, Lonsdale H, Do N, et al. Deep learning for improved risk prediction in surgical outcomes. *Sci Rep*. 2020;10(1):9289-9213.
48. Liu Z, Zhong A, Li Y, et al. Radiology-GPT: a large language model for radiology. arXiv preprint, arXiv:2306.08666. 2023.
49. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940.
50. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180.
51. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J R Soc Med*. 2023;116(5):181-182.
52. Shaheen Z, Wohlgenannt G, Filtz E. Large scale legal text classification using transformer models. arXiv preprint, arXiv:2010.12871. 2020.
53. Rong SS, Meng HL, Fan SJ, et al. Can intraoperative intraocular pressure during primary trabeculectomy predict early postoperative pressure? *J Glaucoma*. 2014;23(9):653-657.
54. Gutiérrez-Ortiz C, Cabarga C, Teus MA. Prospective evaluation of preoperative factors associated with successful mitomycin C needling of failed filtration blebs. *J Glaucoma*. 2006;15(2):98-102.
55. Silva K, Can B, Sarwar R, et al. Text data augmentation using generative adversarial networks – a systematic review. *J Comput Appl Linguist*. 2023;1:6-38.
56. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. *PLoS Digit Health*. 2022;1(12):e0000168.
57. Wu S, Fei H, Qu L, et al. NExT-GPT: any-to-any multimodal LLM. arXiv preprint, arXiv:2309.05519. 2023.