Artificial Intelligence

Evaluation of Multiple Machine Learning Models for Predicting Number of Anti-VEGF Injections in the Comparison of AMD Treatment Trials (CATT)

Rajat S. Chandra[1] and Gui-shuang Ying[2]

[1] Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[2] Center for Preventive Ophthalmology and Biostatistics, Department of Ophthalmology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence: Gui-shuang Ying, 3711 Market Street, Suite 801, Philadelphia, PA 19104, USA. email: gsying@pennmedicine.upenn.edu

Purpose: To apply machine learning models for predicting the number of pro re nata (PRN) injections of antivascular endothelial growth factor (anti-VEGF) for neovascular age-related macular degeneration (nAMD) in two years in the Comparison of AMD (age-related macular degeneration) Treatments Trials.

Methods: The data from 493 eligible participants randomized to PRN treatment of ranibizumab or bevacizumab were used for training (n = 393) machine learning models including support-vector machine (SVM), random forest, and extreme gradient boosting (XGBoost) models. Model performances of prediction using clinical and image data from baseline, weeks 4, 8, and 12 were evaluated by the area under the receiver operating characteristic curve (AUC) for predicting few ($\leq$8) or many ($\geq$19) injections, by R^2 and mean absolute error (MAE) for predicting the total number of injections in two years. The best model was selected for final validation on a test dataset (n = 100).

Results: Using training data up to week 12, the models achieved AUCs of 0.79–0.82 and 0.79–0.81 for predicting few and many injections, respectively, with R^2 of 0.34–0.36 (MAE = 4.45–4.58 injections) for predicting total injections in two years from cross-validation. In final validation on the test dataset, the SVM model had AUCs of 0.77 and 0.82 for predicting few and many injections, respectively, with R^2 of 0.44 (MAE = 3.92 injections). Important features included fluid in optical coherence tomography, lesion characteristics, and treatment trajectory in the first three months.

Conclusions: Machine learning models using loading dose phase data have the potential to predict two-year anti-VEGF demand for nAMD and quantify feature importance for these predictions.

Translational Relevance: Prediction of anti-VEGF injections using machine learning models from readily available data, after further validation on independent datasets, has the potential to help optimize treatment protocols and outcomes for nAMD patients in an individualized manner.

Introduction

Age-related macular degeneration (AMD) is a leading cause of visual loss and blindness worldwide.[1,2] Antivascular endothelial growth factor (anti-VEGF) therapy is a mainstay of treatment for neovascular AMD (nAMD) based on existing clinical practice guidelines, because various studies from different populations and of different design all support its efficacy and safety for nAMD.[3] The Comparison of AMD Treatments Trials (CATT) evaluated the relative efficacy and safety of two anti-VEGF treatments (ranibizumab, bevacizumab) treated monthly or pro re nata (PRN) for macular neovascularization (MNVs).[4] The CATT established that anti-VEGF therapy administered PRN achieved similar gain as monthly treatment in visual acuity (VA) at 1 year and about 2 letters

less gain than monthly treatment at 2 years.[4,5] Patients with nAMD require frequent long-term follow-up and anti-VEGF treatment, and monthly injections for all nAMD patients would be impractical because of the large socioeconomic burden on healthcare systems, as well as patients. For this reason, PRN or treat and extend (T&E) treatment strategies based on the nAMD disease activities evaluated by ocular coherence tomography (OCT) and VA are commonly used to manage patients with nAMD, allowing for patients to maintain or improve their VA without as many injections as monthly treatment.[6,7]

Previous studies found that in patients receiving PRN treatment, the number of anti-VEGF injections varied substantially, indicating differential responses to anti-VEGF treatment in eliminating fluid.[5,8] This variation is likely to be influenced by many factors specific to each patient, including severity of disease, lesion composition, and genetic factors.[8–10] Because PRN and T&E protocols are the most popular approaches for treating nAMD in clinical practice, identifying prognostic factors and being able to predict early-on the number of injections of anti-VEGF therapy a patient would need have great value.[6,8,11] Knowing the number of anti-VEGF injections a patient may require in the long-term is useful in consultations between physicians and patients by allowing for setting patient expectations and planning for the most appropriate regimen, potentially enabling them to make more informed decisions about the best course of treatment and improving their anti-VEGF treatment adherence. Because clinicians may have different experiences and varying judgements, machine learning models may provide useful tool for standardizing predictions. Thus such prediction has the potential to lead to the improved anti-VEGF treatment protocols that can help reduce the burden of anti-VEGF treatments and may result in individualized, flexible treatment regimens.

Machine learning models, including random forest models, natural gradient boosting (NGBoost) models, and neural networks, have demonstrated great promise for predicting the anti-VEGF treatment demand for nAMD and other retinal diseases.[12–15] These previous machine learning models primarily used morphological features extracted from OCT scans taken at baseline and up to the first few months of treatment. These models generally demonstrated poor prediction at baseline, but achieved better predictions with the addition of OCT images in the first two to three months after the initial treatment, reaching their best values for the area under the receiver operating characteristic (ROC) curve (AUC) between 0.68 and 0.85 for predicting low and high numbers of injections.[12–15] However, these machine learning models were based on limited sample sizes and did not include other readily available predictors such as MNV lesion features, VA, anti-VEGF injection trajectory during the loading dose phase, and other clinical variables. Additionally, most of these studies did not validate a final machine learning model on a dataset independently reserved for evaluating the models' performances.

To overcome the limitations of the previous studies, we used the rich CATT data to develop multiple machine learning models and to validate the best machine learning model for predicting both the high and low treatment demand, as well as the total number of injections over the course of 2 years for nAMD patients randomized to PRN treatment with ranibizumab or bevacizumab for 2 years.

Methods

Study Design

This is a secondary analysis of CATT data publicly available at https://hyperprod.cceb.med.upenn.edu/catt/catt_index.php. Details on the study design and methods of the CATT have been reported previously[4,5] and on ClinicalTrials.gov (NCT00593450). Only the major features related to the evaluation of factors within the first 12 weeks and PRN treatment protocol are described here.

Study Participants

The institutional review board associated with each center approved the study protocol and a written consent form was obtained from each participant. Participants enrolled from 43 clinical centers in the United States were randomized to one of the four treatment groups: (1) ranibizumab monthly; (2) bevacizumab monthly; (3) ranibizumab PRN; and (4) bevacizumab PRN. The study enrollment criteria included age of 50 or older, the study eye (one eye per patient) had untreated active MNV caused by AMD, and VA between 20/25 and 20/320 on electronic VA testing. The presence of active MNV, as seen on fluorescein angiography, and fluid, as seen on time-domain OCT, located either within or below the retina or below the retinal pigment epithelium (RPE) were required to establish the presence of active MNV. Either neovascularization or its sequela (i.e., pigment epithelium detachment, subretinal or sub-RPE hemorrhage, blocked fluorescence, macular

edema, or intraretinal, subretinal, or sub-RPE fluid) needed to be under the fovea.

Study Procedures

During the initial visit, participants provided information on demographic characteristics and medical history. Certified photographers followed a standard protocol to obtain stereoscopic, color fundus photographs, and fluorescein angiograms at baseline, year 1, and year 2. OCT images were acquired for all participants at baseline and weeks 4, 8, 12, 24, 52, 76, and 104. The OCT imaging during year 1 was performed with time-domain OCT using Stratus OCT (Stratus software version 6.0 or higher; Carl Zeiss Meditec, Dublin, CA, USA).[16] Spectral-domain OCT was used for 22.6% of scans during year 2, performed on either Cirrus HDOCT (Cirrus software version 5.2 or higher; Carl Zeiss Meditec) or the Spectralis OCT (Spectralis software version 5.3 or higher; Heidelberg Engineering, Carlsbad, CA, USA).[16]

As described in a previous CATT publication,[17] two masked trained readers in the CATT Fundus Photographic Reading Center independently evaluated lesion characteristics based on baseline fundus photographs and fluorescein angiograms. Qualitative evaluations of lesion characteristics included identification of the lesion location, lesion type, lesion composition, retinal angiomatous proliferation features, hemorrhage contiguous with the lesion, serous retinal pigment epithelial detachment, atrophic or fibrotic scars, any hemorrhage associated with lesion (not necessarily contiguous) and geographic atrophy anywhere in the macula. Quantitative measurements of the MNV area and of the total area of MNV lesion were made using Image J (http://rsbweb.nih.gov/ij/, accessed on August 27, 2022). Discrepancies between two trained readers were adjudicated between the readers and Director of the Photograph Reading Center.[17]

An OCT images evaluating team, composed of two independent trained readers and a senior reader, at the OCT Reading Center, masked to the treatment assignment, evaluated each OCT scan.[18] Any discrepancies between the two readers were adjudicated by an independent senior reader. OCT images were assessed with respect to the presence of fluid, location (intraretinal, subretinal, sub-RPE) of fluid, whether fluid was foveal, RPE elevation, subretinal hyper-reflective material, vitreous attached within central 3 mm (vitreomacular traction), and epiretinal membrane. In addition, trained readers measured the total thickness at the foveal center point, which was subdivided into three measurements: thickness of retina, subretinal fluid, and subretinal tissue complex

(material between Bruch's membrane and outer retina or subretinal fluid, which includes pigment epithelial detachment, MNV, blood and fibrosis).[18]

Certified VA examiners, who were masked to the treatment assignments, used the electronic visual acuity tester to measure VA after refraction in both eyes following the Diabetic Retinopathy Clinical Research Network's protocol.[19] These measurements were taken for participants at baseline and weeks 4, 12, 24, 36, 52, 64, 76, 88, and 104. The VA scores from the electronic visual acuity tester ranged from 0 to 100, corresponding with the Snellen equivalents of worse than 20/800 to 20/10.

PRN Treatment Guidelines

All the CATT participants received baseline injection with ranibizumab or bevacizumab. Every 28 days, participants assigned to PRN treatment groups underwent OCT and were evaluated for retreatment by clinical center study–certified ophthalmologist based on the evidence of active MNV. Signs of active MNV were defined as fluid on OCT, new or persistent hemorrhage, decreased VA as compared with the previous examination, or dye leakage or increased lesion size on fluorescein angiography. Ophthalmologists at each clinical center, who were unaware of drug assignments, made retreatment decisions. Fluorescein angiography was performed at the discretion of the ophthalmologist to aid in retreatment decisions.

The clinical center ophthalmologist may have withheld treatment if a patient experienced a serious adverse event in the study eye after treatment including intraocular inflammation ≥2+, intraocular pressure ≥30 mm Hg, vitreous hemorrhage with a ≥30 letters loss in VA, new sensory rhegmatogenous retinal break or detachment (including macular hole), or local infection. The clinical center ophthalmologist may also have suspended the intravitreal injections of the study drug if, in the best medical judgment of the treating ophthalmologist, it is believed that there is no chance of any benefit to the patient from additional intravitreal injections in terms of preserving vision or retinal anatomy.

Machine Learning Models

Among the eligible patients who were randomized to PRN treatment of ranibizumab or bevacizumab at baseline, we applied machine learning models for predicting the burden of PRN treatment in terms of three outcomes: (1) whether patients would have few (≤8) PRN injections in two years; (2) whether patients would have many (≥19) PRN injections in two years; or (3) the total number of PRN injections in two years.

Our prediction is for PRN injections during a two-year period since the CATT participants randomized to PRN treatment regimen were treated and followed-up for two years following the standardized clinical trial protocol.[4,5] Other comparable studies also used a follow-up period of one to two years.[12–15] We prespecified eight PRN injections in two years as the upper bound for having few injections, as that would be equivalent to a rate of at most one injection per quarter. We also prespecified 19 PRN injections in two years as the cutoff for having many injections because the range of 19 injections to a maximum of 26 injections is equivalent to the range for few injections of a minimum of one injection to eight injections.

We applied three machine learning models: (1) the support-vector machine (SVM),[20] (2) random forest,[21] or (3) extreme gradient boosting (XGBoost).[22] We selected these models because they are among the most widely used and are all capable of making predictions for both classification and regression tasks. The SVM model is effective in transforming data to high-dimensional spaces to find a separation between different classes and to predict a continuous response variable with high generalization ability.[20] Random forest models use an ensemble of tree predictors that improves overall results and can prevent overfitting. Random forests have also already been demonstrated to be able to predict levels of treatment demand and the number of injections nAMD patients would receive.[12–15] The XGBoost model makes use of gradient tree boosting with an ensemble of tree predictors like the random forest model but is trained in an additive manner. It has achieved considerable success in machine learning competitions, including Kaggle, and can be applied to a broad range of problems.[22]

For applying machine learning models to the CATT data, we split the data into a training dataset (80% of all samples) for training machine learning models and a test dataset (20% of all samples) for final validation of the best machine learning model identified from the training dataset. We trained machine learning models using participants' features available up to four different time points: baseline, week 4, week 8, and week 12. The data available at baseline included demographics, clinical characteristics, randomized drug group (ranibizumab or bevacizumab), VA at study eye and fellow eye, qualitative and quantitative assessment of lesion characteristics in fundus photos and fluorescein angiograms, and OCT features including presence and location of fluid and thickness. The machine learning models for prediction at weeks 4, 8, and 12 used all the baseline data, and additional data available up to that week (OCT data, VA, and the number of PRN injections up to that week). The data used for machine learn-

ing modelling at each of the four time points is included in Supplementary Table S1.

For each time point, three of each type of machine learning model (SVM, random forest, XGBoost) were trained for predicting each of three outcomes (one for predicting few injections, one for predicting many injections, and one for predicting the total number of injections), for a total of nine unique models at each time point. We performed 10-fold cross-validation using the training dataset for tuning the hyperparameters of our machine learning models. In 10-fold cross-validation at each time point, the training dataset is first divided into 10 nonoverlapping subsets of approximately equal size. Each subset is selected as a validation dataset, and the remaining nine subsets are used to train a model. This model is used to predict on the validation dataset, and this process occurs 10 times, one for each possible validation dataset across the 10 folds, with the mean performance from 10 folds used for model evaluation. This process was repeated for many combinations of hyperparameters to determine the best set of hyperparameters for each model. We tuned hyperparameters based on optimization of the F1 score (the harmonic mean of recall and precision) to train the classification models for predicting whether patients had few and many injections. For the regression models for predicting total number of injections, we tuned hyperparameters based on the optimization of R^2 (a measure for quantifying the amount of variation in number of injections explained by the predictors). Once the hyperparameters were selected in this way, one final model can be fit on the entire training dataset available at each time point using these hyperparameters.

Based on the 10-fold cross-validation results of the machine learning models (SVM, random forest, XGBoost), the best model was selected for final validation on the test dataset. The primary measures for assessing machine learning model performance in the training dataset and test dataset were the AUC for predicting low and high numbers of PRN injections, and R^2 and mean absolute error (MAE) for prediction of the total number of PRN injections in two years.

The importance of each feature was quantified by the permutation importance, defined by the decrease in the model's AUC for classification and R^2 for regression, after shuffling the feature.[21] Feature importance was evaluated using both the training dataset and test dataset for the best machine learning model identified from cross-validation in the training dataset.

All of the machine learning models were implemented using Python 3.9 and its open-source package scikit-learn version 1.1.1. The code for this machine learning analysis can be provided on request to the authors.

Results

PRN Cohort

Among 598 CATT participants randomized to PRN treatment with ranibizumab or bevacizumab at baseline, 497 (83.1%) participants were eligible for this analysis. Participants were excluded from analysis because of death (n = 40), not in the second year of the study (n = 21), treatment futility (n = 6), and missed visits or not treated because of contraindications in more than six out of 26 study visits in two years (n = 34).

Among the 497 participants eligible for analysis, the total number injections over two years (out of 26 maximum injections) ranged from one to 26 injections (median = 13) with a mean (standard deviation) of 13.4 (6.8). In two years, 143 patients (28.8%) had eight or fewer injections, 224 (45.1%) had nine to 18 injections, and 130 (26.2%) had 19 or more injections. Among the 497 eligible patients, four patients did not have baseline OCT grading data because of poor image quality and thus were excluded from the machine learning analysis.

Cross-Validation for Predicting PRN Injections in 2 Years in Training Dataset

We evaluated SVM, random forest, and XGBoost machine learning models for predicting few (≤8) injections, many (≥19) injections, and the total number of injections in two years. The mean results from 10-fold cross-validation of each of these models in the training dataset for predicting at baseline, week 4, week 8, and week 12 are presented in Table 1 and Supplementary Figure S1.

For predicting few (≤8) injections, the mean AUC from the SVM model increases from 0.64 for prediction at baseline, 0.72 at week 4, 0.78 at week 8, and 0.82 at week 12. A similar increase in mean AUC was seen for prediction using the random forest (0.65 at baseline to 0.81 at week 12) and the XGBoost (0.65 at baseline to 0.79 at week 12) (Supplementary Fig. S1A).

For predicting many (≥19) injections, the mean AUC from the SVM model increases from 0.70 for prediction at baseline, 0.77 at week 4, 0.78 at week 8, and 0.81 at week 12. The mean AUCs from the random forest (0.63 at baseline to 0.79 at week 12) and XGBoost (0.63 at baseline to 0.80 at week 12) were lower but

Table 1. Tenfold Cross-Validation Results With Mean and SD for Predicting PRN Injections in Two Years Using Demographic and Ocular Characteristics Available at Baseline, Week 4, Week 8, and Week 12 in the Training Dataset

Model	Baseline (n = 393)[*]	Week 4 (n = 387)[*]	Week 8 (n = 374)[*]	Week 12 (n = 352)[*]
AUC for predicting few (≤8) PRN injections in two years, mean (SD)				
SVM	0.64 (0.07)	0.72 (0.08)	0.78 (0.07)	0.82 (0.06)
Random Forest	0.65 (0.06)	0.69 (0.07)	0.75 (0.10)	0.81 (0.07)
XGBoost	0.65 (0.07)	0.70 (0.08)	0.76 (0.08)	0.79 (0.07)
AUC for predicting many (≥19) PRN injections in two years, mean (SD)				
SVM	0.70 (0.06)	0.77 (0.08)	0.78 (0.09)	0.81 (0.09)
Random Forest	0.63 (0.08)	0.68 (0.09)	0.75 (0.07)	0.79 (0.05)
XGBoost	0.63 (0.10)	0.69 (0.06)	0.78 (0.07)	0.80 (0.05)
R^2 for predicting total number of PRN injections in two years, mean (SD)				
SVM	0.06 (0.06)	0.20 (0.07)	0.30 (0.16)	0.35 (0.18)
Random Forest	0.10 (0.07)	0.22 (0.08)	0.31 (0.16)	0.36 (0.17)
XGBoost	0.03 (0.17)	0.20 (0.14)	0.31 (0.15)	0.34 (0.21)
Mean absolute error for predicting total number of PRN injections in two years, mean (SD)				
SVM	5.76 (0.69)	5.27 (0.62)	4.70 (0.58)	4.45 (0.54)
Random Forest	5.59 (0.63)	5.17 (0.60)	4.59 (0.48)	4.58 (0.45)
XGBoost	5.59 (0.70)	5.09 (0.75)	4.61 (0.46)	4.48 (0.52)

SD, standard deviation.

[*]Patients with incomplete data at the specified time point were not included.

Table 2. Final Validation Results With SD From 1000-Fold Bootstrapping for the Selected SVM Model for Predicting Number of PRN Injections in Two Years Using Demographic and Ocular Characteristics Available at Baseline, Week 4, Week 8, and Week 12 in the Test Dataset

	Baseline (n = 100)[*]	Week 4 (n = 98)[*]	Week 8 (n = 95)[*]	Week 12 (n = 89)[*]
AUC (Bootstrapped SD) for predicting few ($\leq$8) PRN injections in two years	0.46 (0.07)	0.63 (0.07)	0.70 (0.07)	0.77 (0.07)
AUC (Bootstrapped SD) for predicting many ($\geq$19) PRN injections in two years	0.57 (0.08)	0.72 (0.06)	0.79 (0.05)	0.82 (0.04)
R^2 (Bootstrapped SD) for predicting total number of PRN injections in two years	0.01 (0.07)	0.16 (0.08)	0.25 (0.10)	0.44 (0.08)
Mean Absolute Error (Bootstrapped SD) for predicting total number of PRN injections in two years	5.23 (0.35)	4.84 (0.33)	4.41 (0.34)	3.92 (0.31)

SD, standard deviation.
[*]Patients with incomplete data at the specified time point were not included.

showed similar trends of increasing over time (Supplementary Fig. S1B).

The SVM model provides the highest mean AUCs in cross-validation for predicting both few injections (AUC = 0.82) and many injections (AUC = 0.81) by using data up to week 12. The ROC curves from each individual fold of the cross-validation for the SVM model in the training data at all four time points are shown in Supplementary Figure S2.

For the prediction of the total number of PRN injections in two years using only baseline data, the R^2 is highest from the random forest model (0.10), and lowest from the XGBoost model (0.03). The R^2 from the three machine learning models increases over time, with R^2 of 0.20-0.22 for the week 4 predic-

tion, 0.30-0.31 for the week 8 prediction, and 0.34-0.36 for the week 12 prediction (Supplementary Fig. S1C). The MAE for the baseline prediction is 5.76 injections for the SVM model and 5.59 injections for both the random forest and XGBoost models. The MAE decreases over time, reaching 4.45 injections in the SVM model, 4.48 injections in the XGBoost model, and 4.58 injections in the random forest model for predictions using data available up to 12 weeks (Supplementary Fig. S1D).

As shown in Table 1, different numbers of patients were available at each time point in this analysis due to missing data or loss to follow-up, which may bias our evaluation of the performances of machine learning models over time. We performed a
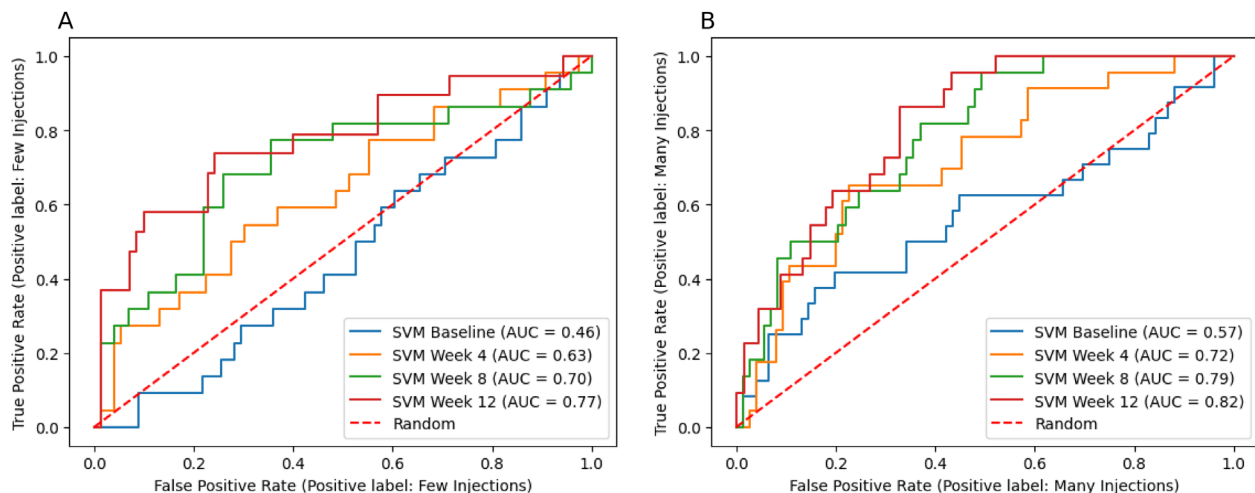

Figure 1. ROC curves for the selected SVM model for predicting (A) few ($\leq$8) and (B) many ($\geq$19) PRN injections in 2 years using demographic and ocular characteristics available at baseline (n = 100), week 4 (n = 98), week 8 (n = 95), and week 12 (n = 89) in the test dataset.
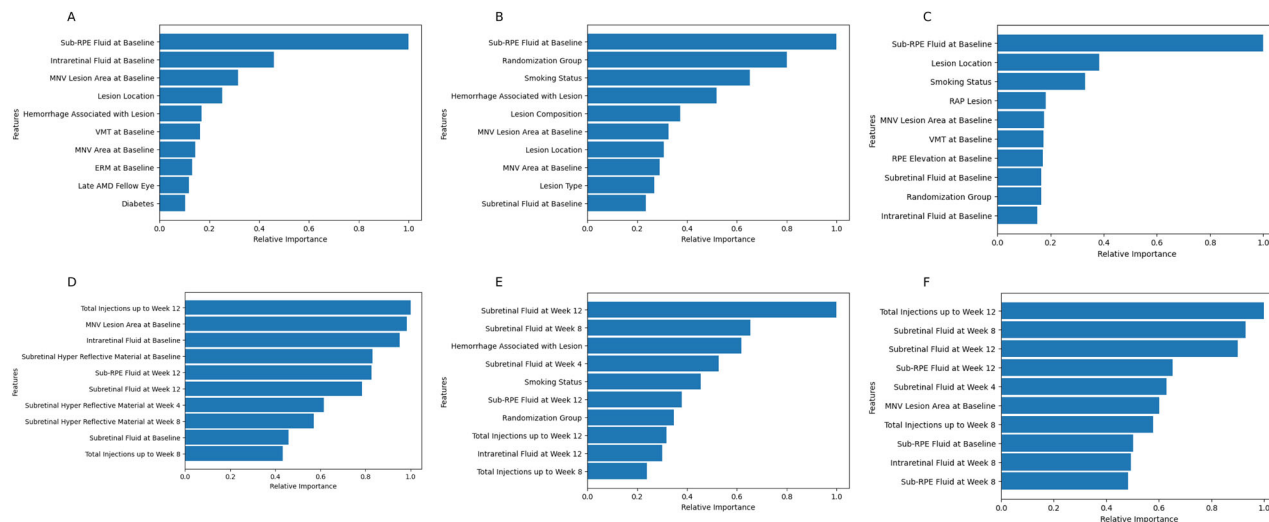
Figure 2. Relative feature importance for the selected SVM model for predicting PRN injections in 2 years using demographic and ocular characteristics available at baseline (n = 393) and week 12 (n = 352) in the training dataset. The top 10 features by relative importance are displayed for the baseline analyses (A, few (≤8) injections; B, many (≥19) injections; C, number of injections) and week 12 analyses (D, few injections; E, many injections; F, number of injections).

sensitivity analysis by restricting to 352 patients who had complete data at each of the time points (baseline and weeks 4, 8, and 12). As shown in Supplementary Table S2, results consistently demonstrated improved model performance at later time points in line with the main cross-validation results in Table 1 that consider all patients with complete data at a specific time point.

Results from Model Validation on Test Dataset

From the three machine learning models, the SVM model was selected for final validation on the test dataset, given its better prediction performance than the random forest and XGBoost models at 12 weeks, with superior performance in the classification tasks and similar prediction in the regression task at other time points compared to the random forest and XGBoost models in 10-fold cross-validation.

The SVM prediction performance in the test dataset was overall consistent with the results in the training dataset (Table 2, Supplementary Fig. S3). The model's best performance occurred using data up to week 12, with AUC values of 0.77 for the prediction of few injections (Fig. 1A, Supplementary Fig. S3A) and 0.82 for predicting many injections (Fig. 1B, Supplementary Fig. S3B). For predicting the total number of injections in two years, the SVM model achieved its greatest R^2 of

0.44 (Supplementary Fig. S3C) and its minimum MAE of 3.92 injections (Supplementary Fig. S3D) when using data up to week 12, consistent with the cross-validation results. Supplementary Figure S4 shows the agreement between the observed and predicted number of injections in the test dataset using data up to week 12.

Feature Importance

From the SVM model, the top 10 important features based on their relative importance (calculated by dividing each feature importance by the maximum feature importance) in the training dataset are displayed in Figure 2 for the baseline and week 12 analyses. Relative feature importance is shown in Supplementary Figure S5 for week 4 and week 8 analyses using the training dataset, and in Supplementary Figure S6 for baseline and weeks 4, 8, and 12 in the test dataset. Fluid presence (intraretinal, subretinal, or sub-RPE) assessed in OCT was frequently among the features with the greatest importance, with baseline lesion characteristics and the number of injections received up to the specified time point also having relatively high importance. Baseline lesion characteristics include MNV lesion area, lesion location (subfoveal or non-subfoveal), lesion composition (considering lesions such as MNV, hemorrhage, blocked fluorescence, and serous retinal pigment epithelial detachment), and lesion type (occult only, minimally classic, or predominantly classic).

Discussion

In this secondary analysis of CATT data, we assessed the ability of multiple machine learning models for predicting anti-VEGF treatment demand of nAMD patients including few (≤ 8) anti-VEGF injections, many (≥ 19) injections, and the total number of injections in two years. Notably, the rich CATT data enabled us to include additional data beyond OCT features in our analysis that are also generally available in real-world settings, such as demographics, baseline lesion features in fundus photographs, VA, treatment trajectory in the loading phase (first three months), and other clinical variables. Our results showed that machine learning models using baseline data did not provide good prediction for the level of treatment demand or number of injections, but the inclusion of data in the first three months of anti-VEGF treatment lead to substantial improvement in prediction performance.

Understanding the injection demand in terms of whether patients will receive few or many injections, along with measures of confidence for these values, is valuable for broadly gauging required treatment frequency for patients and providers in everyday clinical practice. Although providing a probability for every possible value for potential number of injections over the course of a 2-year period would be overwhelming and less practical in the clinical setting, we have predicted the precise number of injections as a supplementary piece of information for even further detail. Not only could this information be of general interest to patients, but with these predictions obtained in an objective manner, patients and providers can also better plan around the projected, long-term treatment course necessary to achieve the desired therapeutic effect, which may lead to the better treatment adherence. Furthermore, this information enables physicians and patients to consider other treatment options early-on if the expected anti-VEGF injection burden exceeds what the patient would be willing to tolerate.

Of the three types of machine learning models on which we trained (SVM, random forest, XGBoost), we selected the SVM model to evaluate on the test dataset for final validation. Although the SVM model is relatively simple, it allows for high generalization ability by controlling the trade-off between complexity and error rate, making it a useful model for classification and regression tasks.[20] We found that the final validation results from the test dataset are consistent with the cross-validation results in the training dataset. The SVM model ultimately achieved strong cross-validation and final validation results when evaluated

in the context of existing studies that trained machine learning models for similar tasks.[12–15]

Using data from 317 participants of the HARBOR trial (ClinicalTrials.gov number, NCT00891735), Bogunović et al.[12] evaluated random forest models for predicting low treatment demand (≤ 5 injections) and high treatment demand (≥ 16 injections) with ranibizumab PRN for nAMD in two years. When trained primarily on patients' OCT features from the first two months in the clinical trial, these models achieved AUCs of 0.70 and 0.77 from 10-fold cross-validation for predicting low and high treatment demand, respectively. However, their model performance was not validated on a separate test dataset.

Using real-world data, Gallardo et al.[13] similarly trained random forest models for predicting the treatment demand for patients on a T&E regimen of anti-VEGF therapy for retinal diseases including nAMD. These models using demographic information and OCT morphological features from the first three visits for 340 nAMD patients achieved AUCs of 0.79 from 10-fold cross-validation for predicting both low (average treatment interval of ≥ 10 weeks) and high (average treatment interval of ≤ 5 weeks) treatment demand for nAMD patients in one year. Similarly, the performance of these models was not validated on a separate test dataset.

Using features extracted from real-world OCT scans of 96 nAMD patients treated with PRN or T&E protocols, Pfau et al.[14] trained several machine learning models (LASSO, principal component, random forest, NGBoost), to predict the total number of injections, as well as to predict low (≤ 4 injections) and high (≥ 10 injections) treatment demand in one year. The random forest model yielded the greatest R^2 of 0.39 from nested cross-validation. The random forest and NGBoost models had the greatest AUCs of 0.68 for predicting low treatment demand, whereas the principal component and random forest models had the highest AUCs of 0.70 for predicting high treatment demand.

Additionally, Romo-Bucheli et al.[15] have even explored an end-to-end deep learning architecture for predicting anti-VEGF treatment requirements from longitudinal retinal OCT scans for nAMD patients. After being trained using OCT scans from the first two months after initial treatment for 281 patients treated PRN, in the test dataset of 69 patients this approach yielded AUCs of 0.85 and 0.81 in predicting low (≤ 5 injections) and high (≥ 16 injections) treatment demand, as well as an R^2 of 0.22 for total number of injections in two years. Although this architecture performs well for predicting low and high treatment demands and is not limited to using only the

prespecified extracted features from OCT scans, it does require that patients' OCT scans be available to make its predictions after being trained through a more computationally intensive process. Additionally, this approach only considers features from OCT scans and does not consider other easily available data such as demographics, VA, past treatment trajectory, and other clinical characteristics.

Although deep learning, and artificial neural networks more generally, are very powerful modeling tools, there are several reasons why we did not evaluate them in our study because they are not as well suited for our analysis. Neural networks are typically used for more complex tasks, such as those with much larger training datasets or for use with image data.[15,23] Additionally, neural networks can be likened to a black-box model that complicates interpretation and ascertaining feature importance.[15] Understanding feature importance was a valuable aspect of our analysis, as we used the rich CATT data with many predictors not previously evaluated for the prediction of anti-VEGF treatment demand, including MNV lesion features, VA, and anti-VEGF injection trajectory during the loading dose phase. Furthermore, our relatively simple SVM model that incorporates these additional features achieved similar AUCs for predicting low and high injections, as well as a larger R^2 for predicting the total number of injections compared to Romo's more computationally intensive deep learning architecture.[15] Nonetheless, further evaluation of the ability of neural networks to predict anti-VEGF treatment demand would be valuable for future research, especially for cases in which more data are available for training these models.

In comparison to previous studies, our SVM model predictions at 12 weeks achieved highest AUCs up to 0.82 and 0.81 for predicting few and many PRN injections, respectively, and an R^2 up to 0.35 for predicting the total number of PRN injections in two years based on the cross-validation results. Similarly, when evaluated on the test dataset for final validation, our SVM model achieved AUCs up to 0.77 and 0.82, respectively, for predicting few and many injections, and an R^2 of 0.44 for predicting the total number of PRN injections in two years. The similar AUCs and improved R^2 of our models for long-term PRN injection prediction when compared to those achieved by Bogunović et al.,[12] Gallardo et al.,[13] Pfau et al.,[14] and Romo-Bucheli et al.[15] support that models like ours can supplement those from the previous studies in clinical application with predictive power gained from considering additional readily available data beyond those just from OCT images. Predictors from the CATT data that we used in training our models, including lesion features,

VA, and treatment trajectory, can be easily obtained and have the potential to enhance prediction accuracy in the clinical setting.

Based on both our cross-validation and final validation results, the SVM model was able to better predict whether a patient would need many injections than whether a patient would need few injections at earlier time points. The SVM model achieved a mean AUC for cross-validation and AUC for final validation of at least 0.70 using data available at baseline and at week 4, respectively, for predicting many injections, but required an additional four weeks of data to achieve similar performance for predicting few injections. However, when using data available at 12 weeks, the model's performances for predicting few and many injections were more similar.

For predicting the total number of injections in the two years, the SVM model, like the other models, performed poorly using data at baseline. Adding features from subsequent weeks allowed the models to substantially improve their predictions, increasing the mean R^2 by almost 0.30 in cross-validation when using the data up to the first 12 weeks. A more dramatic increase was seen in final validation from 0.01 using baseline data to 0.44 using week 12 data, underscoring the value of the patients' features collected over time for predicting the number of injections. A similar improvement was observed in MAE between the SVM model trained only using baseline data and the SVM model trained using data available at week 12.

The ability to adequately predict anti-VEGF treatment demand for nAMD patients can have important implications for clinical practice. In the real-world setting, patients are typically treated using PRN or T&E protocols for nAMD, which have shown promise in improving patients' VA with a reduced number of visits and injections.[24] Given good predictions from using the first three months of data for injection demand patients may need in the long-term, it may be possible to refine a treatment plan that has the potential to improve efficacy with fewer injections in the context of PRN and T&E regimens, as well as to better set expectations for patients.

Our machine learning models provide measures of confidence (i.e., probability) for the classification of few or many injections as well, which can prove valuable in clinical decision-making. As an example, for one patient in the training dataset with 21 injections over two years, the random forest models in cross-validation using baseline data predicted the patient would have 14.6 injections, had a 21% probability of receiving few injections, and had a 33% probability of receiving many injections. The random forest models cross-validation predictions improved substantially when incorporating

data up to week 12, predicting 20.5 injections, with a 3% probability of receiving few injections and a 77% probability of receiving many injections. Notably, this patient had injections at weeks 4, 8, and 12, in addition to their baseline injection. Although this patient had subretinal and intraretinal fluid on OCT at baseline, week 4, week 8, and week 12, this patient had no sub-RPE fluid at baseline and week 4 but did have sub-RPE fluid at weeks 8 and 12.

Determining the most important features used by machine learning models to make predictions also enables providers to consider these specific features when anticipating how individual patients may respond to anti-VEGF therapy. The SVM model indicates OCT features including intraretinal, subretinal, and sub-RPE fluid had high importance, along with MNV area, MNV lesion size, and the number of injections already received, suggesting that these features may help inform providers of how many injections their patients would require.

The strengths of our study include the large sample size, comprehensive high-quality CATT data for prediction, and using a test dataset for model validation that is entirely separate from the training dataset. One limitation of our study is that it relies on OCT grading data by trained readers, which is a manual process that requires expertise. However, other studies have demonstrated methods for automated extraction of information from OCT images to use in predicting the number of anti-VEGF injections nAMD patients require, including the Iowa Reference Algorithms and deep learning.[12–15] These methods can be used to obtain features for use in machine learning models to make these predictions, although the accuracy of the extraction process would need to be ensured. Another limitation of our study is its use of clinical trial data instead of real-world data, which tends to be more heterogenous. However, we have shown that the CATT data lends itself well to training machine learning models to predict the number of injections patients would need, which can be viewed as being more representative of an ideal case given the high-quality data generated from the controlled environment of a clinical trial. Furthermore, real-world data can be augmented with this CATT data to increase the sample size for training, which can theoretically improve the performance of machine learning models.

In conclusion, we have evaluated the ability of machine learning models to predict anti-VEFG treatment demand in two years for nAMD patients and assessed the importance of different features in making these predictions. We have shown the improvement in prediction using data from the first three months of injections (e.g., treatment in the loading dose phase).

Importantly, our machine learning models incorporated easily available predictors beyond OCT characteristics, including demographics, treatment trajectory, lesion characteristics in fundus images, VA, and other clinical data. Our machine learning models have the potential for clinical use that would be beneficial to both physicians and patients for clinical decision-making. Our machine learning models may provide standardized tools for assessing the expected burden of anti-VEGF injections, equipping physicians and patients to plan the best treatment course that can be tailored at the individual level. Future works are needed to further validate the machine learning model on independent real-word data, as well as identify other useful predictors to enhance the prediction of anti-VEGF treatment demand, before implementation of such machine learning models in clinical settings.

References

1. Congdon N, O'Colmain B, Klaver CC, et al. Causes and prevalence of visual impairment among adults in the United States. *Arch Ophthalmol*. 2004;122:477–485.
2. Pascolini D, Mariotti SP, Pokharel GP, et al. 2002 global update of available data on visual impairment: a compilation of population-based prevalence studies. *Ophthalmic Epidemiol*. 2004;11:67–115.
3. Han X, Chen Y, Gordon I, et al. A systematic review of clinical practice guidelines for age-related macular degeneration [published online ahead of print April 13, 2022]. *Ophthalmic Epidemiol*, https://doi.org/10.1080/09286586.2022.2059812.
4. Martin DF, Maguire MG, Ying GS, Grunwald JE, Fine SL, Jaffe GJ. Ranibizumab and bevacizumab for neovascular age-related macular degeneration. *N Engl J Med*. 2011;364:1897–1908.

translational vision science & technology

5. Martin DF, Maguire MG, Fine SL, et al. Ranibizumab and bevacizumab for treatment of neovascular age-related macular degeneration: two-year results. *Ophthalmology*. 2012;119:1388–1398.

6. Rosenberg D, Deonarain DM, Gould J, et al. Efficacy, safety, and treatment burden of treat-and-extend versus alternative anti-VEGF regimens for nAMD: a systematic review and meta-analysis [published online ahead of print April 8, 2022]. *Eye (Lond)*, https://doi.org/10.1038/s41433-022-02020-7.

7. Li E, Donati S, Lindsley KB, Krzystolik MG, Virgili G. Treatment regimens for administration of anti-vascular endothelial growth factor agents for neovascular age-related macular degeneration. *Cochrane Database Syst Rev*. 2020;5:CD012208.

8. Mehta H, Tufail A, Daien V, et al. Real-world outcomes in patients with neovascular age-related macular degeneration treated with intravitreal vascular endothelial growth factor inhibitors. *Prog Retin Eye Res*. 2018;65:127–146.

9. Holz FG, Korobelnik JF, Lanzetta P, et al. The effects of a flexible visual acuity-driven ranibizumab treatment regimen in age-related macular degeneration: outcomes of a drug and disease model. *Invest Ophthalmol Vis Sci*. 2010;51:405–412.

10. Ashraf M, Souka A, Adelman RA. Age-related macular degeneration: using morphological predictors to modify current treatment protocols. *Acta Ophthalmol*. 2018;96:120–133.

11. Kaiser SM, Arepalli S, Ehlers JP. Current and future anti-VEGF agents for neovascular age-related macular degeneration. *J Exp Pharmacol*. 2021;13:905–912.

12. Bogunović H, Waldstein SM, Schlegl T, et al. Prediction of anti-VEGF treatment requirements in neovascular AMD using a machine learning approach. *Invest Ophthalmol Vis Sci*. 2017;58:3240–3248.

13. Gallardo M, Munk MR, Kurmann T, et al. Machine learning can predict anti-VEGF treatment demand in a treat-and-extend regimen for patients with neovascular AMD, DME, and RVO associated macular edema. *Ophthalmol Retina*. 2021;5:604–624.

14. Pfau M, Sahu S, Rupnow RA, et al. Probabilistic forecasting of anti-VEGF treatment frequency in neovascular age-related macular degeneration. *Transl Vis Sci Technol*. 2021;10(7):30.

15. Romo-Bucheli D, Schmidt-Erfurth U, Bogunović H. End-to-end deep learning model for predicting treatment requirements in neovascular AMD from longitudinal retinal OCT imaging. *IEEE J Biomed Health Inform*. 2020;24:3456–3465.

16. Folgar FA, Jaffe GJ, Ying G-S, Maguire MG, Toth CA. Comparison of optical coherence tomography assessments in the comparison of age-related macular degeneration treatments trials. *Ophthalmology*. 2014;121:1956–1965.e2.

17. Grunwald JE, Daniel E, Ying GS, et al. Photographic assessment of baseline fundus morphologic features in the Comparison of Age-Related Macular Degeneration Treatments Trials. *Ophthalmology*. 2012;119:1634–1641.

18. Decroos FC, Toth CA, Stinnett SS, Heydary CS, Burns R, Jaffe GJ. Optical coherence tomography grading reproducibility during the Comparison of Age-related Macular Degeneration Treatments Trials. *Ophthalmology*. 2012;119:2549–2557.

19. Beck RW, Moke PS, Turpin AH, et al. A computerized method of visual acuity testing: adaptation of the early treatment of diabetic retinopathy study testing protocol. *Am J Ophthalmol*. 2003;135:194–205.

20. Cortes C, Vapnik V. Support-vector networks. *Mach Learning*. 1995;20:273–297.

21. Breiman L. Random forest. *Mach Learning*. 2001;45:5–32.

22. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794.

23. Uhrig RE. Introduction to artificial neural networks. In: *Proceedings of IECON'95-21st Annual Conference on IEEE Industrial Electronics*. 1995;1:33–37.

24. Banaee T, Alwan S, Kellogg C, Kornblau I, El-Annan J. PRN Treatment of neovascular AMD with cycles of three monthly injections. *J Ophthalmic Vis Res*. 2021;16:178–186.