Augmenting Kalman Filter Machine Learning Models with Data from OCT to Predict Future Visual Field Loss

An Analysis Using Data from the African Descent and Glaucoma Evaluation Study and the Diagnostic Innovation in Glaucoma Study

Mohammad Zhalechian, MS,[1] Mark P. Van Oyen, PhD,[1] Mariel S. Lavieri, PhD,[1]
Carlos Gustavo De Moraes, MD, PhD,[2] Christopher A. Girkin, MD,[3] Massimo A. Fazio, PhD,[3]
Robert N. Weinreb, MD,[4] Christopher Bowd, PhD,[4] Jeffrey M. Liebmann, MD,[2] Linda M. Zangwill, PhD,[4]
Christopher A. Andrews, PhD,[5,6] Joshua D. Stein, MD, MS[5,6,7]

Purpose: To assess whether the predictive accuracy of machine learning algorithms using Kalman filtering for forecasting future values of global indices on perimetry can be enhanced by adding global retinal nerve fiber layer (RNFL) data and whether model performance is influenced by the racial composition of the training and testing sets.

Design: Retrospective, longitudinal cohort study.

Participants: Patients with open-angle glaucoma (OAG) or glaucoma suspects enrolled in the African Descent and Glaucoma Evaluation Study or Diagnostic Innovation in Glaucoma Study.

Methods: We developed a Kalman filter (KF) with tonometry and perimetry data (KF-TP) and another KF with tonometry, perimetry, and global RNFL data (KF-TPO), comparing these models with one another and with 2 linear regression (LR) models for predicting mean deviation (MD) and pattern standard deviation values 36 months into the future for patients with OAG and glaucoma suspects. We also compared KF model performance when trained on individuals of European and African descent and tested on patients of the same versus the other race.

Main Outcome Measures: Predictive accuracy (percentage of MD values forecasted within the 95% repeatability interval) differences among the models.

Results: Among 362 eligible patients, the mean ± standard deviation age at baseline was 71.3 ± 10.4 years; 196 patients (54.1%) were women; 202 patients (55.8%) were of European descent, and 139 (38.4%) were of African descent. Among patients with OAG (n = 296), the predictive accuracy for 36 months in the future was higher for the KF models (73.5% for KF-TP, 71.2% for KF-TPO) than for the LR models (57.5%, 58.0%). Predictive accuracy did not differ significantly between KF-TP and KF-TPO ($P = 0.20$). If the races of the training and testing set patients were aligned (versus nonaligned), the mean absolute prediction error of future MD improved 0.39 dB for KF-TP and 0.48 dB for KF-TPO.

Conclusions: Adding global RNFL data to existing KFs minimally improved their predictive accuracy. Although KFs attained better predictive accuracy when the races of the training and testing sets were aligned, these improvements were modest. These findings will help to guide implementation of KFs in clinical practice. *Ophthalmology Science* 2022;2:100097 © 2022 Published by Elsevier Inc. on behalf of the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Supplemental material available at https://www.ophthalmologyscience.org.

Kalman filtering is a machine learning approach that has been used for decades in the aerospace and aviation industries.[1] Researchers have extended this methodology to forecast the future trajectory of patients with chronic

diseases.[2–4] The Kalman filter (KF) model considers the dynamics of an underlying population of patients with the disease of interest along with the past disease trajectory from the actual patient of interest to generate personalized

forecasts of how the patient's disease will change over time. The more the KF learns from the dynamics of the actual patient, the less it relies on the dynamics of the underlying population when generating future forecasts for the patient of interest. Prior work from our group demonstrated that KFs can predict the future trajectory of visual field (VF) loss better than more conventional approaches for patients with ocular hypertension,[5] moderate to severe open-angle glaucoma (OAG),[4] and normal-tension glaucoma.[6]

The data used to populate our previously developed KFs came from large multicenter, randomized clinical trials such as the Advanced Glaucoma Intervention Study, the Collaborative Initial Glaucoma Treatment Study, and Ocular Hypertension Treatment Study.[7−9] All of these trials collected measurements of tonometry and perimetry at baseline and every 6 months for many years to permit us to train and test our models. However, a major limitation of using data from these trials was that none of them had collected retinal nerve fiber layer (RNFL) measurements as captured on OCT so we could not incorporate such measurements into our models. Given the now widespread use of OCT to permit clinicians to assess for structural evidence of glaucoma progression,[10] we believed it was important to identify other data sources that capture longitudinal measurements of the RNFL and to determine whether inclusion of such data can enhance the predictive performance of our existing KF models. In this study, we explored whether it is possible to enhance the predictive accuracy of our KF models by incorporating OCT data into our models using data from patients enrolled in the African Descent and Glaucoma Evaluation Study (ADAGES)[11] and Diagnostic Innovation in Glaucoma Study (DIGS).[12] The second objective of this study was to assess whether differences in the race used in the training and the testing sets affect the performance of our KF models. Past research in other medical fields revealed that machine learning models that are parameterized and trained on populations of patients with different characteristics than the target population they are to be used on can lead to biased predictions.[13−17] We sought to assess whether this was the case for our KF models as well.

Methods

Data Sources

We used data from ADAGES[11] and DIGS,[12] which are prospective observational cohort studies of patients with OAG, glaucoma suspects, and healthy individuals. Participants in ADAGES were recruited from the Hamilton Glaucoma Center at the Department of Ophthalmology, University of California, San Diego; the New York Eye and Ear Infirmary; and the Department of Ophthalmology, University of Alabama at Birmingham. Diagnostic Innovation in Glaucoma Study participants were recruited from the University of California, San Diego. The protocols for ADAGES were harmonized to be identical to DIGS so that DIGS participants could be used as a comparison group for ADAGES participants.[11] The main objective of these studies was to investigate structural and functional changes associated with glaucoma. Recruitment for DIGS began in April 1995 and recruitment for ADAGES began in January 2003, and both

studies continue to enroll new patients and to follow up those who have been enrolled. All participants underwent a complete ophthalmologic examination, including tonometry and perimetry at baseline, and most patients continue to undergo follow-up testing twice per year. Retinal nerve fiber layer assessment by OCT began in 2009. The University of California, San Diego, Institutional Review Board approved data collection for ADAGES and DIGS, and both of those studies adhere to the tenets of the Declaration of Helsinki. All participants provided their written informed consent to participate. The University of California, San Diego, approved sharing de-identified data from ADAGES and DIGS with researchers at the University of Michigan to carry out the present study, which was approved by the University of Michigan Institutional Review Board.

Sample Selection

All participants in ADAGES[11] and DIGS[12] at baseline had best-corrected visual acuity of 20/40 or better, spherical equivalent refractive error of < 5.0 diopters (D), cylinder correction < 3.0 D, and open angles by gonioscopy in at least 1 eye. We also required all eligible patients to contribute at least 8 Swedish Interactive Threshold Algorithm Standard 24-2 Humphrey visual fields (VFs) using the Humphrey Field Analyzer (Carl Zeiss Meditec), at least 8 intraocular pressure (IOP) measurements, and at least 8 global RNFL thickness measurements using the Spectralis OCT (Heidelberg Engineering) over a span of 2 years or more. This ensured that we could incorporate at least 3 sets of initial measurements into our KFs and evaluate our predictions at least 24 months into the future. Those with a history of other diseases that could affect the VF such as coexisting ocular trauma, retinal disease, uveitis, or nonglaucomatous optic neuropathy were excluded. Healthy persons who did not have glaucoma or were not glaucoma suspects also were excluded. If both eyes were eligible, the eye with more available measurements was chosen.

Baseline Characterization of Patients as Glaucoma Suspects or with Open-Angle Glaucoma

All eligible participants had received a diagnosis of glaucoma suspect or OAG. The glaucoma suspect classification comprised participants who at baseline showed normal white-on-white Swedish Interactive Threshold Algorithm Standard 24-2 VFs, but had a history of either elevated IOP (IOP > 21 mmHg or a history of requiring ocular hypotensive treatment) or an optic disc appearance suspicious of glaucoma. The OAG classification comprised participants who at baseline showed glaucomatous VF loss on standard automated perimetry along with optic nerve changes consistent with glaucoma.

Using Kalman Filtering to Forecast Future Values of Mean Deviation, Pattern Standard Deviation, and Global Retinal Nerve Fiber Layer Thickness

Kalman Filter. Kalman filtering is a machine learning approach that forecasts future values of a given variable by integrating population-level data on disease progression dynamics with past readings from the individual patient in question.[18] The more the KF learns about the unique trajectory of a particular patient over time, the less it relies on the underlying population and the more the past readings from that patient influence future predictions for the variable of interest. As such, KFs generate forecasts that are personalized for each particular patient. Unlike many traditional forecasting techniques, KF uses matrix computations

to update its forecasts dynamically based on past prediction errors without rebuilding the progression model with each new set of measurements.

Data Elements. For each eligible patient, we considered demographic information (e.g., age at baseline, sex, and race), IOP measurements, mean deviations (MDs), and pattern standard deviations (PSDs) from VF testing and, for selected models, global RNFL thickness measurements from OCT. Our KF models require measurements to be evenly spaced. Because the intervals between tests in ADAGES and DIGS were not always exactly 6 months, we performed a linear interpolation of IOP, MD, PSD, and global RNFL values to obtain equally spaced readings at 6-month intervals.[6,19] Participants self-identified their race and ethnicity as described in past studies involving ADAGES and DIGS data.[11]

For this project, we developed 2 KF models: a KF model with tonometry and perimetry (KF-TP) and a KF model with tonometry, perimetry, and global RNFL from OCT (KF-TPO). The KF-TP model was tuned and parametrized using demographic information on the patient along with past IOP, MD, and PSD measurements, along with their velocities and accelerations. The KF-TPO model was tuned and parametrized using the same data elements as the first model along with global RNFL thickness measurements and their velocities and accelerations. The KF models were fit using the expectation-maximization algorithm, which is an iterative algorithm to find the maximum likelihood estimates of the parameters.[20]

Training

Our models sought to predict future values of MD, PSD, and global RNFL up to 36 months in the future. The first 6 measurements of IOP, MD, and PSD were used to train the KF-TP model. These same parameters along with the addition of global RNFL were used to train the KF-TPO model. We then used these KFs to predict future values of these parameters and compared them with the observed values of MD, PSD, IOP, and global RNFL obtained in ADAGES and DIGS after 24 and 36 months of follow-up. We used leave-one-out cross-validation to estimate the performance of the KF models. That is, for every patient, a separate KF model was trained using the data from all remaining patients besides the patient of interest for whom we were generating a forecast.

Model Comparisons

Separately for glaucoma suspects and those with OAG, we compared the performance of the KF-TP and KF-TPO models with one another and with 2 LR models. Linear regression models fit a linear relationship between the input and the output. We used linear regression-based models as our reference models because they have been proven empirically to predict future values on perimetry more accurately than some complex nonlinear models.[21] The models are of the form $y = ax + b$, where y is the forecasted value (e.g., MD, PSD, or global RNFL) and the input x is the time in months since the last measurement. The parameters a (slope) and b (intercept) are estimated from the 6 initial measurements. Linear regression model 1 (LR1) is the least squares regression line. Linear regression model 2 (LR2) is an econometric forecasting model that has the same slope as LR1, but passes through the most recent observation.[22]

Performance Measures

We analyzed the prediction errors of KF-TP, KF-TPO, LR1, and LR2 when forecasting our primary outcome (MD) at 24 and 36 months in the future. First, we computed the number and

proportion of patients with prediction errors for MD within clinically relevant boundaries (i.e., 0.5 dB, 1.0 dB, and 2.5 dB from the observed value obtained in ADAGES and DIGS) for all 4 models and compared their performance across the defined thresholds. We studied this separately for the OAG and glaucoma suspect groups. Comparison of the proportion of eyes forecasted at 24 and 36 months in the future between the 2 KF models and among all 4 models was made using Cochran's Q test and Friedman's test, respectively. Next, for each cohort, we computed the number and proportion of patients with prediction errors within the 95% repeatability interval (i.e., computed using the MD values obtained by 5 repeated tests[23]) for all 4 models. To further investigate the distribution of prediction errors for all 4 models at 36 months into the future, we created (1) scatterplots illustrating the forecasted MD values within the 95% repeatability interval and (2) violin plots illustrating the distribution of MD prediction errors. The repeatability interval is used to capture the magnitude of the last observed MD value in identifying a forecasted value as a successful one.[23] Violin plots show the probability density of the prediction error, and thus show the presence of different peaks in distribution errors among the models.

Next, we assessed the magnitude of the prediction errors for MD, PSD, and global RNFL by computing 2 commonly used error measures: root mean square error (RMSE) and mean absolute error (MAE). Root mean square error is the square root of the mean of the squared prediction errors. A small RMSE value indicates a good overall estimation of the observed values. Mean absolute error is the mean of the absolute values of the prediction errors, and again, lower values indicate better performance. After predicting MD at 24 and 36 months in the future, we used 1-sample t tests to investigate the assumption that the mean value of the prediction errors obtained by the KF models is 0. We compared magnitude of the prediction errors of the KF models using paired t tests.

Outlier Analysis

Eyes for which the MD prediction errors at 36 months were outside of the 2 times wider repeatability interval were considered outliers. We calculated the number and proportion of eyes that were outliers in the KF-TP and KF-TPO models when forecasting 36 months into the future. We compared the outlier and nonoutlier groups in terms of demographics at baseline and follow-up readings using 2-sided t tests and Pearson's chi-square tests, as appropriate. For all analyses, P values < 0.05 were considered statistically significant.

Sensitivity Analyses

We performed several sensitivity analyses. First, we assessed whether more complex models of KF-TPO including additional RNFL measurements led to more accurate predictions. We tested a model containing not only the global RNFL, but also mean RNFL in the superior and inferior quadrants. As a second sensitivity analysis, we explored whether shifting the global RNFL measurements for each patient back 6 or 12 months from the date they had been acquired would enhance the predictive accuracy of KF-TPO. Third, we assessed the performance of a null model that forecasts the future MD values by assuming a fixed rate of worsening of the MD (0.05 dB every 12 months[24]) for all eligible patients in the sample. Fourth, we assessed whether a KF model using only IOP and global RNFL data (but no perimetry data) performed better or worse compared with KF-TP and KF-TPO at predicting future MD values.

Assessing Whether Disparities in Race between Members of the Training and Testing Sets Affect the Magnitude of Prediction Errors for Kalman Filter Models

Finally, we investigated whether the race of the patients with OAG in the training set and in the testing set affected the performance of the KF-TP and KF-TPO models. We identified all 93 patients of African descent who had at least 36 months of follow-up (set AD) and also randomly selected 93 patients of European descent (set ED) who had at least 36 months of follow-up. We measured KF performance of predicting MD at 36 months using MAE for each of 4 possible scenarios: train AD, test AD; train ED, test AD; train AD, test ED; and train ED, test ED. Leave-one-out cross-validation was used when the same set was used for training and testing.

Results

Study Sample

Open-Angle Glaucoma. Two hundred ninety-six patients with OAG were eligible. The mean ± standard deviation (SD) age was 71.7 ± 10.2 years, and 141 patients (47.6%) were men and 155 patients (52.4%) were women. The racial composition included 160 patients (54.1%) of European descent, 119 patients (40.2%) of African descent, and 17 patients (5.7%) of other races. The mean ± SD values of MD, PSD, IOP, and global RNFL at baseline were −4.7 ± 5.9 dB, 4.9 ± 4.0 dB, 19.2 ± 6.4 mmHg, and 77.7 ± 17.4 μm, respectively. The mean ± SD number of IOP, VF, and OCT measurements per patient were 20.6 ± 9.4, 19.9 ± 6.6, and 11.7 ± 4.6, respectively (Table 1).

Glaucoma Suspects. Sixty-six glaucoma suspects were eligible. The mean ± SD age was 69.7 ± 11.0 years, and 25 patients (37.9%) were men and 41 patients (62.1%) were women. The racial composition of these patients included 42 patients (63.6%) of European descent, 20 patients (30.3%) of African descent, and 4 patients (6.1%) of other races. The mean ± SD values of MD, PSD, IOP, and global RNFL at baseline were −0.5 ± 1.5 dB, 1.6 ± 0.6 dB, 20.2 ± 5.8 mmHg, and 86.5 ± 12.6 μm, respectively. The mean ± SD number of IOP, VF, and OCT measurements per patient were 20.1 ± 9.8, 19.5 ± 5.6, and 10.8 ± 3.7, respectively (Table 1).

Prediction Errors When Forecasting Mean Deviation

Open-Angle Glaucoma. When MD was forecasted 36 months into the future, the numbers of eyes successfully forecasted within 1.0 dB of the observed value by the KF-TP, KF-TPO, LR1, and LR2 models were 94 (42.9%), 101 (46.1%), 67 (30.6%), and 66 (30.1%), respectively. Although a statistically significant difference was found in the proportions of eyes forecasted within 1.0 dB when comparing all 4 models with one another ($P < 0.001$), with the 2 KF models outperforming the 2 LR models, we did not observe a statistically significant difference between the KF-TP and KF-TPO models ($P = 0.88$). The numbers and percentages of forecasted MD values within clinically relevant boundaries of 0.5 dB, 1.0 dB, and 2.5 dB of the observed value at 24 and 36 months into the future for the KF-TP, KF-TPO, LR1, and LR2 models are shown in Table S1. The violin plots of the distribution of prediction errors when forecasting MD for patients with OAG at 36 months in the future are shown in Figure S1. The somewhat shorter and fatter violin plots of the KF-TP and KF-TPO models indicate that the distribution is closer to the median, whereas longer and thinner plots of the LR1 and LR2 models indicate greater variability.

When MD was forecasted 36 months into the future, the numbers of eyes successfully forecasted within the repeatability interval by the KF-TP, KF-TPO, LR1, and LR2 models were 161 (73.5%), 156 (71.2%), 126 (57.5%), and 127 (58.0%), respectively (Table 2). Although a statistically significant difference was found in the proportions of eyes forecasted within the repeatability interval among the 4 models ($P < 0.001$), with the 2 KF models outperforming the 2 LR models, we did not observe a statistically significant difference between the KF-TP and KF-TPO models ($P = 0.20$). Figure 1 illustrates the forecasted MD values and the 95% repeatability interval when forecasting MD for patients with OAG at 36 months into the future. The repeatability interval becomes larger with increasing glaucomatous damage. Having more points within the repeatability interval in the KF-TP and KF-TPO models compared with the LR1 and LR2 models indicates the better prediction accuracies of the KF models.

Glaucoma Suspects. For glaucoma suspects, when MD was forecasted 36 months into the future, the numbers of eyes successfully forecasted within 1.0 dB of the observed value by the KF-TP, KF-TPO, LR1, and LR2 models were 23 (69.7%), 25 (75.8%), 17 (51.5%), and 17 (51.5%), respectively. Although a statistically significant difference was found in the proportions of eyes forecasted within 1.0 dB when comparing all 4 models with one another ($P = 0.04$), with the 2 KF models outperforming the 2 LR models, we did not observe a statistically significant difference between the KF-TP and KF-TPO models ($P = 0.56$).

When MD was forecasted 36 months into the future, the numbers of eyes successfully forecasted within the repeatability interval by the KF-TP, KF-TPO, LR1, and LR2 models were 27 (81.8%), 28 (84.8%), 20 (60.6%), and 23 (69.7%), respectively (Table 2). A statistically significant difference was found in the proportions of eyes forecasted within the repeatability interval among the 4 models ($P = 0.05$), with the 2 KF models outperforming the 2 LR models, but again, no statistically significant difference was found between the KF-TP and KF-TPO models ($P = 0.70$). Figure 2 shows the forecasted MD values and the 95% repeatability intervals for glaucoma suspects when forecasting MD at 36 months into the future, with the 2 KF models showing better prediction accuracies than the LR1 and LR2 models.

Comparison of Magnitude of Prediction Errors for Mean Deviation, Pattern Standard Deviation, Intraocular Pressure, and Global Retinal Nerve Fiber Layer

Patients with Open-Angle Glaucoma. *Mean Deviation.* For patients with OAG, the KF-TP model showed an RMSE improvement of 32.4% (RMSE, 2.16) and the KF-

Table 1. Demographics and Characteristics of Study Sample

Patient Characteristics	Patients with Open-Angle Glaucoma	Glaucoma Suspects	P Value*
Patients, no.	296	66	
Age (yrs)	71.7 ± 10.2	69.7 ± 11.0	0.20
Sex			0.19
Male	141 (47.6)	25 (37.9)	
Female	155 (52.4)	41 (62.1)	
Race			0.32
White	160 (54.1)	42 (63.6)	
Black	119 (40.2)	20 (30.3)	
Other	17 (5.7)	4 (6.1)	
Glaucoma testing			
Initial			
MD (dB)	−4.7 ± 5.9	−0.5 ± 1.5	<0.001
PSD (dB)	4.9 ± 4.0	1.6 ± 0.6	<0.001
IOP (mmHg)	19.2 ± 6.4	20.2 ± 5.8	0.11
Global RNFL (μm)	77.7 ± 17.4	86.5 ± 12.6	<0.001
Measurements per patient			
IOP	20.6 ± 9.4	20.1 ± 9.8	0.58
VF	19.9 ± 6.6	19.5 ± 5.6	0.52
OCT	11.7 ± 4.6	10.8 ± 3.7	0.02
Interval between initial and most recent assessment (yrs)			
IOP	13.6 ± 6.2	13.6 ± 6.0	0.93
VF	12.1 ± 4.1	12.3 ± 3.8	0.70
OCT	5.7 ± 1.9	6.3 ± 1.9	0.29

IOP = intraocular pressure; MD = mean deviation; PSD = pattern standard deviation; RNFL = retinal nerve fiber layer; VF = visual field.
Data are presented as no., no. (%), or mean±standard deviation, unless otherwise indicated.
*P values for sex and race computed using Pearson's chi-square test for independent samples. All other P values computed using a 2-sided *t* test.

TPO model demonstrated an RMSE improvement of 33.1% (RMSE, 2.14) relative to the LR1 model (RMSE, 3.20) at forecasting values of MD 36 months into the future. The KF-TP model showed an MAE improvement of 25.8% (MAE, 1.90) and the KF-TPO model demonstrated an MAE improvement of 25.2% (MAE, 1.91) relative to the LR1 model (MAE, 2.56). Although the KF-TPO model performed slightly better than the KF-TP model, this difference was not statistically significant (*P* = 0.52; confidence interval [CI], −0.06 to 0.03; Table 3; Table S2).

Pattern Standard Deviation. For patients with OAG, when predicting the value of PSD 36 months in the future, the KF-TP, KF-TPO, and LR2 models showed improvements of 11.7% (RMSE, 2.50), 11.4% (RMSE, 2.50), and

4.0% (RMSE, 2.72) relative to the LR1 model (RMSE, 2.83), respectively. The KF-TP model showed an MAE improvement of 25.0% (MAE, 1.36), and the KF-TPO model demonstrated an MAE improvement of 24.1% (MAE, 1.38) relative to the LR1 model (MAE, 1.82). In terms of the absolute prediction errors, the KF-TP model outperformed the KF-TPO model (*P* = 0.009; CI, −0.03 to 0.00; Table 3; Table S2).

Global Retinal Nerve Fiber Layer. The KF-TPO and LR2 models showed improvement of 52.9% (RMSE, 5.07) and 5.7% (RMSE, 10.15), respectively, at predicting global RNFL in patients with OAG at 36 months in the future relative to the LR1 model (RMSE, 10.77). The KF-TPO model demonstrated an MAE improvement of 55.0%

Table 2. Accuracy of 4 Models with Respect to 95% Repeatability Interval for Patients with Open-Angle Glaucoma and Glaucoma Suspects at Forecasting Mean Deviation Values 24 and 36 Months into the Future

Months Forecast Ahead	Patient Cohort	No. of Eyes (%)			
		Kalman Filter with Tonometry and Perimetry Data Model	Kalman Filter with Tonometry, Perimetry, and Global Retinal Nerve Fiber Layer Data Model*	Linear Regression Model 1	Linear Regression Model 2
24	OAG	221 (81.3)	216 (79.4)	184 (67.6)	176 (64.7)
	Glaucoma suspects	38 (73.1)	42 (80.8)	37 (71.2)	42 (80.8)
36	OAG	161 (73.5)	156 (71.2)	126 (57.5)	127 (58.0)
	Glaucoma suspects	27 (81.8)	28 (84.8)	20 (60.6)	23 (69.7)

OAG = open-angle glaucoma.
*No statistical difference was found between Kalman Filter models in any row.
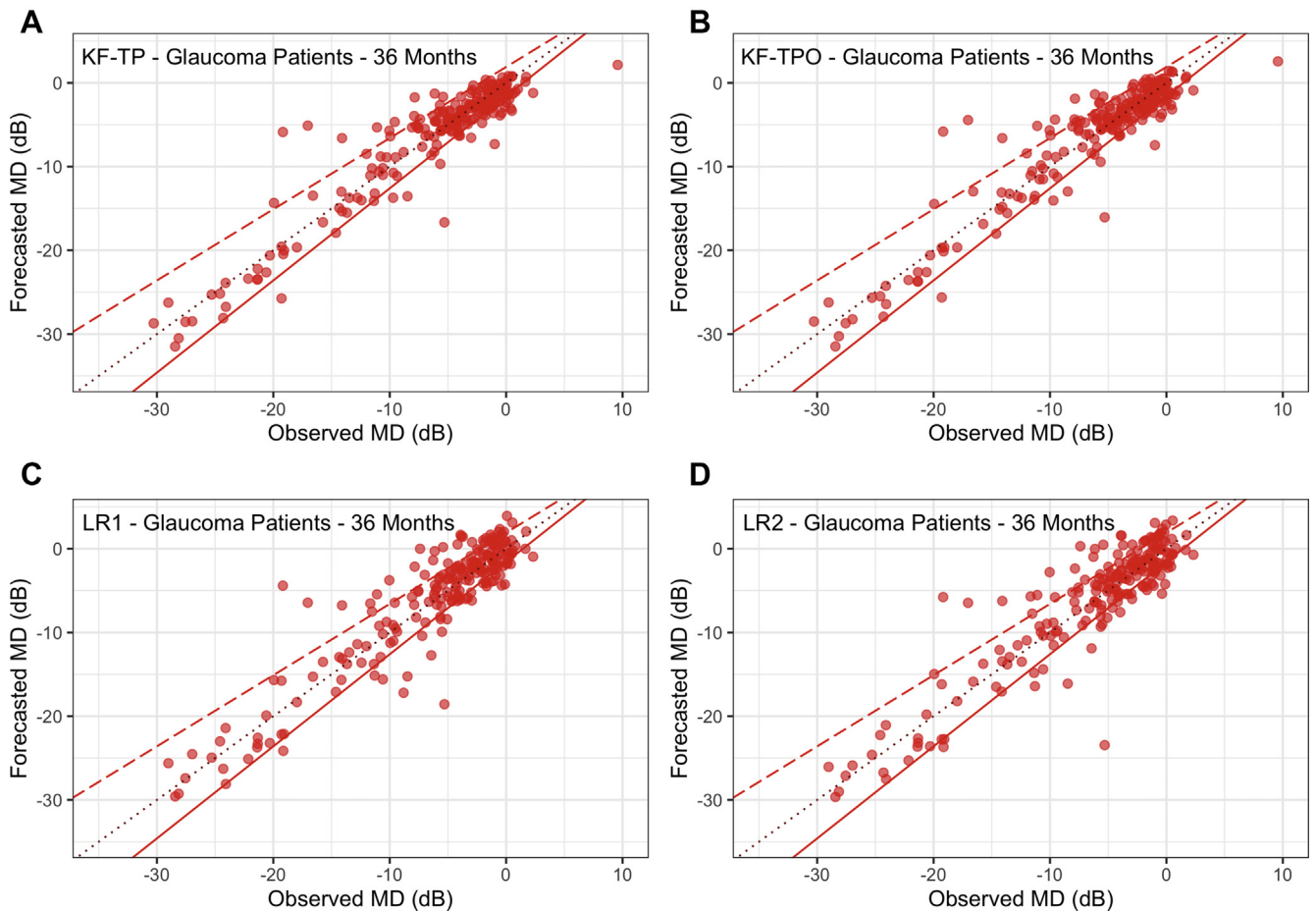
Figure 1. Graphs showing the accuracy of forecasted values of mean deviation (MD) for patients with open-angle glaucoma at 36 months in the future for the 4 models. The dotted line is the alignment line, and the dashed line and the solid line are obtained as the equations $0.94 + 0.86x$ and $-1.23 + 1.10x$, where x is the last observed MD value, respectively. The repeatability interval becomes larger with increasing damage. Having more points within the repeatability interval indicates the better prediction accuracy. Equations used for repeatability interval boundaries obtained from Wall et al.[23] (A) The accuracy of the KF-TP model at 36 months for patients with open-angle glaucoma. (B) The accuracy of the KF-TPO model at 36 months for patients with open-angle glaucoma. (C) The accuracy of the LR1 model at 36 months for patients with open-angle glaucoma. (D) The accuracy of the LR2 model at 36 months for patients with open-angle glaucoma. KF-TP = Kalman filter with tonometry and perimetry data; KF-TPO = Kalman filter with tonometry, perimetry, and global RNFL data; LR1 = linear regression model 1; LR2 = linear regression model 2.

(MAE, 3.60) relative to the LR1 model (MAE, 7.99; Table 3; Table S2).

Glaucoma Suspects. *Mean Deviation.* For glaucoma suspects, the KF-TP model showed an improvement of 29.8% (RMSE, 1.05) and the KF-TPO model demonstrated an improvement of 33.2% (RMSE, 1.00) relative to the LR1 model (RMSE, 1.50) at forecasting values of MD 36 months in the future. The KF-TP model showed an MAE improvement of 33.0% (MAE, 0.78), and the KF-TPO model demonstrated an MAE improvement of 36.7% (MAE, 0.74) relative to the LR1 model (MAE, 1.16). Although the KF-TPO model performed slightly better than the KF-TP model, this difference was not statistically significant ($P = 0.62$; CI, -0.13 to 0.22; Table 3; Table S2).

Pattern Standard Deviation. For glaucoma suspects, the KF-TP model showed an improvement of 37.1% (RMSE, 0.54) and the KF-TPO model demonstrated an

improvement of 40.3% (RMSE, 0.51) relative to the LR1 model (RMSE, 0.85) at forecasting values of PSD values 36 months into the future. The KF-TP model showed an MAE improvement of 34.6% (MAE, 0.41), and the KF-TPO model demonstrated an MAE improvement of 44.9% (MAE, 0.35) relative to the LR1 model (MAE, 0.63). Again, although the KF-TPO model performed slightly better than the KF-TP model, this difference was not statistically significant ($P = 0.25$; CI, -0.05 to 0.18; Table 3; Table S2).

Global Retinal Nerve Fiber Layer. For glaucoma suspects, the KF-TPO and LR2 models showed improvement of 41.1% (RMSE, 4.90) and 4.6% (RMSE, 7.93), respectively, at predicting global RNFL at 36 months into the future relative to the LR1 model (RMSE, 8.32). The KF-TPO model demonstrated an MAE improvement of 37.5% (MAE, 3.65) relative to the LR1 model (MAE, 5.84; Table 3; Table S2).
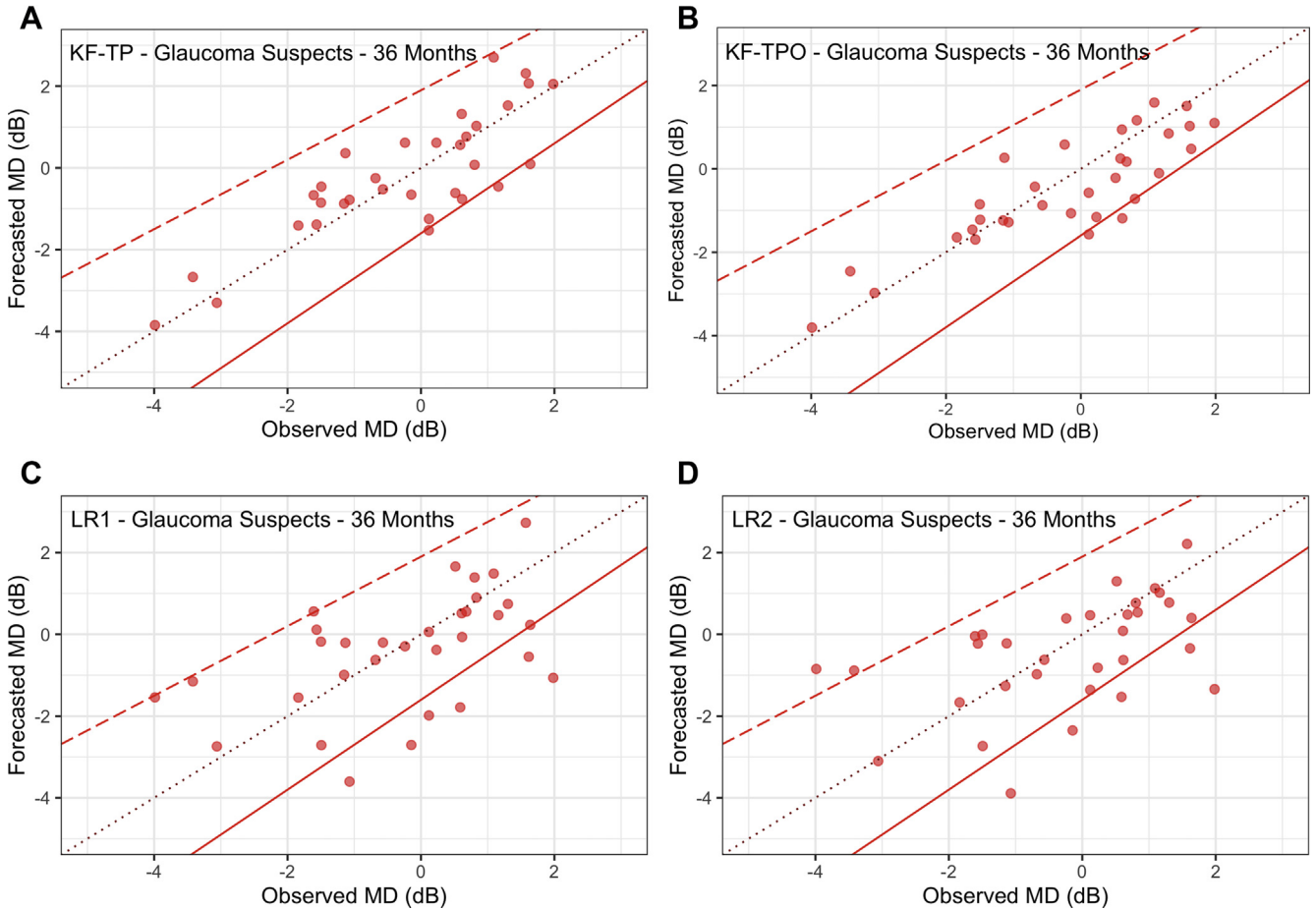
Figure 2. Graphs showing the accuracy of forecasted values of mean deviation (MD) for glaucoma suspects at 36 months into the future for the 4 models. The dotted line is the alignment line, and the dashed line and the solid line are obtained as the equations $0.94 + 0.86x$ and $-1.23 + 1.10x$, where x is the last observed MD value, respectively. The repeatability interval becomes larger with increasing damage. Having more points within the repeatability interval indicates the better prediction accuracy. Equations used for repeatability interval boundaries obtained from Wall et al.[23] (A) The accuracy of the KF-TP model at 36 months for glaucoma suspects. (B) The accuracy of the KF-TPO model at 36 months for glaucoma suspects. (C) The accuracy of the LR1 model at 36 months for glaucoma suspects. (D) The accuracy of the LR2 model at 36 months for glaucoma suspects. KF-TP = Kalman filter with tonometry and perimetry data; KF-TPO = Kalman filter with tonometry, perimetry, and global RNFL data; LR1 = linear regression model 1; LR2 = linear regression model 2.

Sensitivity Analyses

Additional KF-TPO models were developed that included not only the global RNFL, but also the mean RNFL in the superior and inferior quadrants along with its change in velocity and acceleration over time. When tested on patients with OAG and glaucoma suspects, these models containing the additional RNFL measurements did not generate lower RMSEs or MAEs for predicting values of MD or PSD at 36 months into the future relative to the KF-TPO model containing only global RNFL (data not shown). In another sensitivity analysis, shifting back the timing of the global RNFL data by 6 or 12 months relative to the tonometry and perimetry measurements did not achieve any improvement in predictive accuracy either (data not shown). In a third sensitivity analysis, we created a null model in which we assumed that the future MD value all of patients decreased by 0.05 dB every 12 months. We found

that the null model performed worse than all the other models (data not shown). Finally, we considered a KF model that used only IOP and global RNFL (but no perimetry data) to predict future values of MD. When tested on patients with OAG and glaucoma suspects, this model was less accurate compared with the KF-TP and KF-TPO models (data not shown).

Outlier Analysis

When forecasting MD values at 36 months into the future for patients with OAG, 17 outliers (7.8%) and 16 outliers (7.3%) were found (i.e., MD prediction errors were outside of the 2 times wider repeatability interval) in the KF-TP and KF-TPO models, respectively. In the KF-TP and KF-TPO models, the outliers comprised a greater proportion of people of African descent or women than the nonoutliers;

Table 3. Comparison of the Root Mean Square Error of the 4 Models at Forecasting Key Glaucoma Metrics at 24 and 36 Months in the Future for Patients with Open-Angle Glaucoma and Glaucoma Suspects

Months Forecast Ahead	Patient Cohort	Metric	Root Mean Square Error (% Improvement)[*,†,‡]			
			Kalman Filter with Tonometry and Perimetry Data	Kalman Filter with Tonometry, Perimetry, and Global Retinal Nerve Fiber Layer Data Model[§]	Linear Regression Model 1	Linear Regression Model 2
24	OAG	MD	2.22 (23.2)	2.23 (22.8)	2.89	2.73 (5.5)
		IOP	3.56 (28.2)	3.56 (28.2)	4.96	4.87 (1.7)
		PSD	1.59 (22.4)	1.59 (22.4)	2.05	1.92 (6.1)
		RNFL		4.51 (41.9)	7.76	7.14 (8.1)
	Glaucoma suspects	MD	1.26 (1.6)	1.12 (12.9)	1.28	1.24 (3.4)
		IOP	4.10 (31.2)	4.11 (30.9)	5.96	5.65 (5.1)
		PSD	0.47 (8.4)	0.40 (23.0)	0.51	0.54 (-5.6[||])
		RNFL		5.40 (24.1)	7.11	6.94 (2.3)
36	OAG	MD	2.16 (32.4)	2.14 (33.1)	3.20	3.18 (0.7)
		IOP	3.59 (43.9)	3.55 (44.5)	6.40	6.29 (1.8)
		PSD	2.50 (11.7)	2.50 (11.4)	2.83	2.72 (4.0)
		RNFL		5.07 (52.9)	10.77	10.15 (5.7)
	Glaucoma suspects	MD	1.05 (29.8)	1.00 (33.2)	1.50	1.50 (-0.5)
		IOP	4.23 (36.0)	4.20 (36.5)	6.61	6.42 (2.8)
		PSD	0.54 (37.1)	0.51 (40.3)	0.85	0.84 (1.0)
		RNFL		4.90 (41.1)	8.32	7.93 (4.6)

IOP = intraocular pressure; MD = mean deviation; OAG = open-angle glaucoma; PSD = pattern standard deviation; RNFL = retinal nerve fiber layer.

*A low root mean square error value indicates predictions are close to the observed values obtained in the African Descent and Glaucoma Evaluation Study dataset.

†Percentage improvement was measured with respect to the linear regression model 1 and computed as $(RMSE_{LR1} - RMSE_x) / RMSE_{LR1}$, where $RMSE_x$ is the root mean square error corresponding to the Kalman filter with tonometry and perimetry data model, the Kalman filter with tonometry, perimetry, and global retinal nerve fiber layer data model, linear regression model 1, or linear regression model 2. Positive percentage improvement values indicate improved performance compared with linear regression model 1.

‡The root mean square error values for the Kalman filter with tonometry and perimetry data model and Kalman filter with tonometry, perimetry, and global retinal nerve fiber layer data model were estimated using leave-one-out cross-validation.

||Negative improvement indicates that linear regression model 1 performed better in comparison with linear regression model 2.

§The mean difference of squared errors for this model was not statistically different from the Kalman filter with tonometry and perimetry data model in any row.

however, these differences were not statistically significant. (Tables S3 and S4).

Assessing Whether Disparities in Race between Members of the Training and Testing Sets Affected the Magnitude of Prediction Errors for Kalman Filter Models

The KF models trained and tested with data of patients with OAG from the same race (training ED, testing ED or training AD, and testing AD) performed only slightly better than KF models trained and tested using data from different races (training ED, testing AD or training AD, and testing ED). When predicting the MD value 36 months in the future for patients of AD, the model trained on patients of AD performed slightly better than the model trained on patients of ED; the difference of MAEs was only 0.35 dB and 0.64 dB for the KF-TP and KF-TPO models, respectively. Similarly, when predicting the MD value 36 months into the future for patients of ED, the model trained on patients of ED performed slightly better than the model trained on patients of AD; the difference of MAEs was only 0.44 dB and 0.31 dB for the KF-TP and KF-TPO models, respectively. In aggregate, when forecasting MD at 36 months in the future, training with patients of the same racial descent improved the MAE by only 0.39 dB and 0.48 dB for the KF-TP and KF-TPO models, respectively, compared with models in which the race of the training and testing sets differed.

Discussion

Using data from a racially diverse cohort of patients with glaucoma from the ADAGES and DIGS, we compared the performance of 2 machine learning algorithms involving KF with 2 LR-based algorithms and found that the KF algorithms outperformed LR at predicting MD and PSD values 36 months in the future for patients with OAG and glaucoma suspects. Next, we assessed whether adding structural data on global RNFL to the KF models would enhance their performance and found that the accuracy of the models at predicting MD and PSD improved only negligibly. Finally, we assessed whether alignment of the race of the patients in the training and testing sets affected the predictive accuracy of our models. We learned that the models performed slightly better when the races of the patients in training and testing sets were aligned, but the difference in mean absolute predictive error was small and may not be of much clinical significance.

Although we hypothesized that integrating structural information from OCT into our forecasting algorithms would enhance their ability to forecast future evidence of glaucomatous progression as captured on perimetry, several possible reasons may explain why we did not find much improvement with the KF-TPO model compared with the KF-TP model. First, we used only global RNFL thickness measurements and, as a sensitivity analysis, also tried superior and inferior RNFL measurements. Although it is known that damage to certain regions of the RNFL correspond with certain predictable patterns of visual field loss,[25,26] changes in global values of RNFL may not be as useful at predicting a global index of functional loss such as MD. Kalman filter models incorporating additional quantitative data from OCT may lead to better performance of the KF-TPO model in the future. Second, structural damage to the RNFL can occur ahead of detectable visual field loss,[27] although not all studies have identified such a relationship.[28] To account for the potential lag between changes in RNFL anatomic features and subsequent VF loss, we tried using global RNFL measurements from a prior visit. We tested models using RNFL measurements 6, 12, 18, or 24 months before the current visit (the time at which predictions are made), but this, too, showed a negligible improvement in the predictive accuracy of the KF-TPO model relative to the KF-TP model. Little is known regarding what would be the optimal time lag, if any, between the prior RNFL data and the visit at which the predictions of MD are made. Third, the KF-TPO model seemed to outperform the KF-TP model more when tested on glaucoma suspects relative to patients with OAG. It may be that because glaucoma suspects demonstrated healthier RNFL tissue at baseline, changes in RNFL were more informative to our models relative to those with OAG who already experienced some RNFL thinning, and thus new changes to the thickness of this tissue were less informative to the models.

A second objective of this study was to explore whether our machine learning models performed better when the races of the patients in the training and testing sets were aligned compared with if they were not aligned. These analyses are important because of the growing literature in medicine that highlights how machine learning models can generate biased output if the sample they are trained on differs in important ways from the sample they are asked to predict.[13−17] For a condition like OAG, it is well appreciated that glaucoma is much more likely to develop in patients of AD relative to patients of ED and the former are considerably more likely to progress to blindness from glaucoma.[29−31] Fortunately, we did not observe much of a difference in predictive accuracy of our models when the race of the training and testing sets were similar or dissimilar. We suspect that this may be related to the nature of our KFs, namely, that the models incorporate data from both the underlying population of all patients in the training set along with past readings from the actual patient.[18] As the model acquires more sequential measurements from the actual patient of interest, this carries more weight in future model predictions than data from the population. Furthermore, because the velocity and acceleration of model inputs such

as MD are also integrated into the KF, if a patient is experiencing a rapid decline, or even an increasing rate of decline, he or she will be predicted to be more likely to behave that way in the future, regardless of his or her race.

When comparing our KF models with traditional LR models, we observed that the KF models tended to be more accurate, both for glaucoma suspects as well as patients with OAG. Linear regression models are much more rigid and struggle with predictive accuracy for patients whose glaucoma may be relatively stable for years and then suddenly starts to demonstrate an abrupt worsening. Likewise, LR models assume that all patients behave the same way. By comparison, our KF models can better discern the subset of patients who do not exhibit any further deterioration over time from those who experience a slowly progressive decline in MD, from the small subset of rapid progressors.

Our study has several limitations. First, it may be possible to enhance the performance of our models with larger sample sizes. Because OCT manufacturers are continually upgrading their hardware and software, it can be challenging to identify large cohorts of patients who are followed up longitudinally over many years using the same equipment to train, validate, and test these models. Although ADAGES and DIGS are the largest datasets we had available with longitudinal structural and functional data to perform these analyses, with larger sample sizes and lengthier follow-up for more patients, we expect improvements in the accuracy of both the KF-TP and KF-TPO models. Because KFs learn from each sequential measurement from a given patient, it may be possible that training our models with additional OCT measurements over time could translate into improvements in their predictive accuracy for future results from perimetry. Second, the key structural model input we used in these analyses was global RNFL. We also tried incorporating RNFL data from the superior and inferior quadrants into the KF-TPO model but found that it did not enhance model performance appreciably. We have yet to consider individual clock-hour RNFL data or other quantitative measurements obtained from the OCT, such as ganglion cell layer measurements, as model inputs. The predictive accuracy of the KF-TPO model may be enhanced with these additional inputs. Third, all of these patients underwent testing using the same OCT equipment. It is difficult to know whether our findings apply to other patients tested on other equipment. Finally, in our analyses of racial match of the training and testing sets on model performance, the numbers of eligible patients in the training sets were rather modest, and we had sufficient sample sizes to compare only patients of AD and ED, and not those of other races and ethnicities.

In conclusion, Kalman filtering is a promising machine learning technique that can help to predict the disease trajectory of patients with OAG. The addition of global RNFL measurements to our existing KF models did not yield much improvement in their predictive accuracy. Additional investigation of other OCT parameters may enhance the predictive accuracy of these models. We also learned that although the performance of our KFs were slightly better when congruence existed between the racial composition of

the training and testing sets, this improvement was rather modest. Because the training of KFs incorporate past readings from the actual patient into future predictions, they may be less prone to bias relative to other machine learning algorithms that are trained using data from other patients, rather than the actual patient.

Footnotes and Disclosures

[1] Department of Industrial and Operations Engineering, University of Michigan College of Engineering, Ann Arbor, Michigan.

[2] Bernard and Shirlee Brown Glaucoma Research Laboratory, Edward S. Harkness Eye Institute, Columbia University Irving Medical Center, New York, New York.

[3] Department of Ophthalmology, University of Alabama at Birmingham, Birmingham, Alabama.

[4] Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, University of California, San Diego, La Jolla, California.

[5] Department of Ophthalmology and Visual Sciences, University of Michigan Medical School, Ann Arbor, Michigan.

[6] Center for Eye Policy and Innovation, University of Michigan, Ann Arbor, Michigan.

[7] Department of Health Management and Policy, University of Michigan School of Public Health, Ann Arbor, Michigan.

Abbreviations and Acronyms:

AD = African descent; **ADAGES** = African Descent and Glaucoma Evaluation Study; **CI** = confidence interval; **D** = diopter; **DIGS** = Diagnostic Innovation in Glaucoma Study; **ED** = European descent; **IOP** = intraocular pressure; **KF** = Kalman filter; **KF-TP** = Kalman filter with tonometry and perimetry data; **KF-TPO** = Kalman filter with tonometry, perimetry, and global retinal nerve fiber layer data; **LR1** = linear regression model 1; **LR2** = linear regression model 2; **MAE** = mean absolute error; **MD** = mean deviation; **OAG** = open-angle glaucoma; **PSD** = pattern standard deviation; **RMSE** = root mean square error; **RNFL** = retinal nerve fiber layer; **SD** = standard deviation; **VF** = visual field.

Correspondence:

Joshua D. Stein, MD, MS, W.K. Kellogg Eye Center, Department of Ophthalmology and Visual Sciences, University of Michigan, 1000 Wall Street, Ann Arbor, MI 48105. E-mail: jdstein@med.umich.edu.

References

1. Leffens EJ, Markley FL, Shuster MD. Kalman filtering for spacecraft attitude estimation. *J Guid Control Dyn*. 1982;5(5): 417−429.

2. Wang Q, Molenaar P, Harsh S, et al. Personalized state-space modeling of glucose dynamics for type 1 diabetes using continuously monitored glucose, insulin dose, and meal intake: an extended Kalman filter approach. *J Diabetes Sci Technol*. 2014;8(2):331−345.

3. Eberle C, Ament C. The unscented Kalman filter estimates the plasma insulin from glucose measurement. *Biosystems*. 2011;103(1):67−72.

4. Schell GJ, Lavieri MS, Helm JE, et al. Using filtered forecasting techniques to determine personalized monitoring schedules for patients with open-angle glaucoma. *Ophthalmology*. 2014;121(8):1539−1546.

5. Garcia G-GP, Lavieri MS, Andrews C, et al. Accuracy of Kalman filtering in forecasting visual field and intraocular pressure trajectory in patients with ocular hypertension. *JAMA Ophthalmol*. 2019;137(12):1416−1423.

6. Garcia G-GP, Nitta K, Lavieri MS, et al. Using Kalman filtering to forecast disease trajectory for patients with normal tension glaucoma. *Am J Ophthalmol*. 2019;199:111−119.

7. The Advanced Glaucoma Intervention Study (AGIS): 1. Study design and methods and baseline characteristics of study patients. *Control Clin Trials*. 1994;15(4):299−325.

8. Musch DC, Lichter PR, Guire KE, Standardi CL. The collaborative initial glaucoma treatment study: study design, methods, and baseline characteristics of enrolled patients. *Ophthalmology*. 1999;106(4):653−662.

9. Gordon MO, Kass MA. The Ocular Hypertension Treatment Study: design and baseline description of the participants. *Arch Ophthalmol*. 1999;117(5):573−583.

10. Stein JD, Talwar N, Laverne AM, et al. Trends in use of ancillary glaucoma tests for patients with open-angle glaucoma from 2001 to 2009. *Ophthalmology*. 2012;119(4):748−758.

11. Sample PA, Girkin CA, Zangwill LM, et al. The African Descent and Glaucoma Evaluation Study (ADAGES): design and baseline data. *Arch Ophthalmol*. 2009;127(9):1136−1145.

12. Sample PA, Medeiros FA, Racette L, et al. Identifying glaucomatous vision loss with visual-function−specific perimetry in the Diagnostic Innovations in Glaucoma Study. *Invest Ophthalmol Vis Sci*. 2006;47(8):3381−3389.

13. Rajkomar A, Hardt M, Howell MD, et al. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866−872.

14. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. 2018;378(11):981−983.

15. Gijsberts CM, Groenewegen KA, Hoefer IE, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One*. 2015;10(7):e0132321.

16. Cahan EM, Hernandez-Boussard T, Thadaney-Israni S, Rubin DL. Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digit Med*. 2019;2(1):1−6.

17. Bozkurt S, Cahan EM, Seneviratne MG, et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. *J Am Med Inform Assoc*. 2020;27(12):1878−1884.

18. Catlin DE. *Estimation, Control, and the Discrete Kalman Filter*. Vol 71. New York, NY: Springer Science & Business Media; 2012.

19. DeRoos L, Nitta K, Lavieri MS, et al. Comparing perimetric loss at different target intraocular pressures for patients with high-tension and normal-tension glaucoma. *Ophthalmol Glaucoma*. 2021;4(3):251−259.

20. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol*. 1977;39(1):1−22.

21. McNaught AI, Hitchings RA, Crabb DP, Fitzke FW. Modelling series of visual fields to detect progression in normal-tension glaucoma. *Graefes Arch Clin Exp Ophthalmol*. 1995;233(12):750−755.

22. Krajewski LJ, Ritzman LP, Malhotra MK. *Operations Management: Processes and Supply Chains*. Vol 1. Upper Saddle River, NJ: Pearson; 2013.

23. Wall M, Doyle CK, Zamba KD, et al. The repeatability of mean defect with size III and size V standard automated perimetry. *Invest Ophthalmol Vis Sci*. 2013;54(2):1345−1351.

24. Chauhan BC, Malik R, Shuba LM, et al. Rates of glaucomatous visual field change in a large clinical population. *Invest Ophthalmol Vis Sci*. 2014;55(7):4135−4143.

25. Harwerth RS, Vilupuru AS, Rangaswamy NV, Smith EL. The relationship between nerve fiber layer and perimetry measurements. *Invest Ophthalmol Vis Sci*. 2007;48(2):763−773.

26. Alasil T, Wang K, Yu F, et al. Correlation of retinal nerve fiber layer thickness and visual fields in glaucoma: a broken stick model. *Am J Ophthalmol*. 2014;157(5):953−959.

27. Medeiros FA, Zangwill LM, Bowd C, et al. The structure and function relationship in glaucoma: implications for detection of progression and measurement of rates of change. *Invest Ophthalmol Vis Sci*. 2012;53(11):6939−6946.

28. Gardiner SK, Mansberger SL, Fortune B. Time lag between functional change and loss of retinal nerve fiber layer in glaucoma. *Invest Ophthalmol Vis Sci*. 2020;61(13):5.

29. Tielsch JM, Sommer A, Katz J, et al. Racial variations in the prevalence of primary open-angle glaucoma: the Baltimore Eye Survey. *JAMA*. 1991;266(3):369−374.

30. Leske MC, Connell AMS, Schachat AP, Hyman L. The Barbados Eye Study: prevalence of open angle glaucoma. *Arch Ophthalmol*. 1994;112(6):821−829.

31. Budenz DL, Barton K, Whiteside-de Vos J, et al. Prevalence of glaucoma in an urban West African population: the Tema Eye Survey. *JAMA Ophthalmol*. 2013;131(5):651−658.