Deep Learning-Based Modeling of the Dark Adaptation Curve for Robust Parameter Estimation

Tharindu De Silva¹, Kristina Hess¹, Peyton Grisso¹, Alisa T. Thavikulwat², Henry Wiley², Tiarnan D. L. Keenan², Emily Y. Chew², Brett G. Jeffrey³, and Catherine A. Cukras¹

¹ Unit on Clinical Investigation of Retinal Disease, National Eye Institute, National Institutes of Health, Bethesda, MD, USA
² Division of Epidemiology & Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, MD, USA
³ Ophthalmic Genetics and Visual Function Branch, National Eye Institute, National Institutes of Health, Bethesda, MD, USA

Correspondence: Catherine A. Cukras, Unit on Clinical Investigation of Retinal Disease, National Eye Institute, National Institutes of Health, Bethesda, MD 20892, USA. e-mail: cukrasc@nei.nih.gov

Purpose: This study investigates deep-learning (DL) sequence modeling techniques to reliably fit dark adaptation (DA) curves and estimate their key parameters in patients with age-related macular degeneration (AMD) to improve robustness and curve predictions.

Methods: A long-short-term memory autoencoder was used as the DL method to model the DA curve. The performance was compared against the classical nonlinear regression method using goodness-of-fit and repeatability metrics. Experiments were performed to predict the latter portion of the curve using data from early measurements. The prediction accuracy was quantified as the rod intercept time (RIT) prediction error between predicted and actual curves.

Results: The two models had comparable goodness-of-fit measures, with root mean squared error (RMSE; SD) = 0.11 (0.04) log-units (LU) for the classical model and RMSE = 0.13 (0.06) LU for the DL model. Repeatability of the curve fits evaluated after introduction of random perturbations, and after performing repeated testing, demonstrated superiority of the DL method, especially among parameters related to cone decay. The DL method exhibited superior ability to predict the curve and RIT using points prior to −2 LU, with 3.1 ± 3.1 minutes RIT prediction error, compared to 19.1 ± 18.6 minutes RIT error for the classical method.

Conclusions: The parameters obtained from the DL method demonstrated superior robustness as well as predictability of the curve. These could provide important advances in using multiple DA curve parameters to characterize AMD severity.

Translational Relevance: Dark adaptation is an important functional measure in studies of AMD and curve modeling using DL methods can lead to improved clinical trial end points.

Introduction

Dark adaptation (DA) tests can reveal early functional changes in patients with age-related macular degeneration (AMD). The biphasic DA curve captures the recovery of retinal sensitivity of the cones and rods over time following exposure to a background light that bleaches rhodopsin.¹ Parameters derived from these DA curves can serve as useful indicators in assessing retinal function, and provide information relevant to disease severity with associations to severe stages including atrophy.² The rod intercept time (RIT), defined by the time taken to reach a criterion threshold, is a commonly used clinical parameter derived from the DA curve.³⁻⁵ RIT has shown particular relevance to AMD,³⁻⁶⁻⁹ with reported associations to AMD disease severity and the presence of certain phenotypes, such as reticular pseudodrusen (RPD).¹⁰⁻¹² Other parameters obtained from the DA test include final thresholds and rates of recovery for the rods and cones, and the time to the rod-cone break.⁸⁻¹³⁻¹⁵ These parameters...
Dark adaptation measures the rate of recovery of retinal sensitivity following exposure to a light that bleaches some fraction of rhodopsin in the photoreceptors. Recent developments have been made to both the instruments and the testing paradigms to facilitate the feasibility of DA testing in the clinical setting. The changes include utilization of a partial bleach, which leads to shorter recovery times, and the application of the RIT parameter as an outcome, which obviates the need to reach the final threshold sensitivity. RIT parameter measures the time to reach a threshold sensitivity within the rod-driven portion of the recovery (portion of the recovery up to approximately 1 log unit [LU] below the cone rod break), thus requiring only limited curve modeling. To estimate parameters in addition to RIT, the raw data acquired in practical clinical test protocols requires curve fitting. The DA curve may be modeled as exponent + linear decay (or exponent + two linear decays), with the cone-mediated portion of the curve represented by an exponential decay and the rod-mediated portion of the curve (S2) represented by a linear decay (and the third component remaining unmeasured in this clinical testing paradigm). Although rod-related variables derived from the dark adaptation curve have shown stronger relationship with AMD disease severity, both cone- and rod-related variables in isolation or in combination could enhance the understanding of disease-related changes further. This requires robust approaches for curve modeling to derive reliable measurements for all parameters computed from the curve.

Dark adaptation tests can be long and burdensome to elderly patients and, as with psychophysical measures, could contain noisy measurements, even with the shorter and more targeted testing afforded by focal partial bleach testing. During the DA test, each collected test point aims to estimate sensitivity threshold at that time point in a specific area of the retina using a staircase estimation. Measurement fluctuations could be triggered by premature or laggard responses from the tired patients. Slight differences in fixation could result in responses coming from nonidentical areas of retina. Moreover, some stages of the curve may contain sparse and limited number of measurements causing high uncertainty to parameter estimation. Robustly modeling the curve overcoming these spurious and sparse measurements can thus be a challenge and a limitation to obtaining representative curve-derived parameters reliably. When evaluating how disease (e.g. AMD) affects the DA curve and its ability to capture changes over time, it is important for the curve fitting method to be minimally sensitive to fluctuations and ideally able to distinguish disease-induced deviations from measurement fluctuations.

Different curve-fitting approaches can be used to generate a DA curve from the collected test points. Using parametric representation for the curve, curve-fitting methods, such as nonlinear regression, are traditionally used to estimate the parameters: errors between the curve and the measured data points are minimized using an iterative optimization algorithm. This classical method treats all data points equally, while attempting to minimize the error between the measured data and the estimated curve. Although faster implementations with error bounds have been explored, spurious measurements can still drive to undesired solutions causing high variability in parameters. Alternatively, deep learning (DL) sequence modeling techniques can be used to observe the patterns in the acquired data, potentially identify and suppress spurious data points, and augment sparse regions to robustly estimate the curve. Importantly, this approach does not explicitly specify the curve equation or the number of curve parameters; the model predicts a suitable curve representation after observing the data points for that particular test, as well as learning from trends in the tests of other patients. However, the resulting curve obtained from the DL algorithm can be utilized to fit the classical DA equation, in order to obtain the classical curve parameters.

In this work, we hypothesize that a DL-based approach for modeling the DA curve can have twofold advantages. First, by observing measured sequences collectively across a large number of patients, it can learn to differentiate between spurious and reliable measurements and to mitigate the impact of spurious measurements. Second, within a measured sequence of a DA curve from a single patient, it can learn how the curve progresses over time by appropriately weighing the effect of early versus late measurements. We conjecture that the classical parameters derived from the DL-estimated curve will improve robustness in the presence of noisy/sparse measurements. In addition, we investigate the method’s ability to predict the late phase of the curve, using only partial data from earlier phases. This could be useful in estimating the DA parameters of patients who were not able to complete the test. We additionally perform analysis in associating...
the estimated curve with AMD severity score\textsuperscript{20} and presence of RPD.\textsuperscript{10}

**Methods**

**Materials**

The data were collected from participants of a clinical study of DA with a range of AMD severities, including no AMD (NCT01352975). Dark adaptation was measured using a prototype of the AdaptDx dark adaptometer (MacuLogix, Hummelstown, PA). Details about the testing procedure have been described previously by Jackson et al.\textsuperscript{4} In brief, the patient’s pupil was dilated and the participant was asked to focus on a fixation light. A photoflash producing an equivalent 82% bleach centered at 5 degrees on the inferior visual meridian was performed, and threshold measurements were made at the same location with a 1.7 diameter, 500-nm wavelength circular test spot, using a 3-down/1-up modified staircases threshold estimate procedure. The initial stimulus intensity ($P_0$) was 5 cd/m\textsuperscript{2}. Threshold measurements were continued until the patient’s visual sensitivity recovered to be able to detect a dimmer stimulus intensity of 5 × 10\textsuperscript{−3} cd/m\textsuperscript{2} (a relative decrease of 3 log units [LU], denoted as −3 LU), or until a maximum test duration of 40 minutes was reached, whichever occurred first. The raw test sensitivities ($P$) were extracted from the instrument at each measured time point and the relative decrease is recorded in LU according to log\textsubscript{10} ($P/P_0$). The total number of DA tests performed was 1496. These were from longitudinal follow-up testing of 207 unique patients or healthy volunteers, over a maximum period of 6 years. There were 349 tests repeated within 3 months of the baseline test for the same individual. For additional independent validation, 21 DA tests of 10 patients with AMD were acquired where 11 test pairs were performed within a 60-day interval.

During study visits, all participants underwent a comprehensive ocular examination, including multimodal image acquisitions. Color fundus photography was acquired as part of the protocol\textsuperscript{9,21} and each eye was graded according to age-related eye disease study (AREDS) criteria\textsuperscript{22} for AMD severity (AREDS 9-step scale) and the presence/absence of RPD at annual increments.

**Curve Estimation**

DA curve interpolation methods fit a continuous curve in the time domain from a set of sparse measurements. We explored a DL-based curve interpolation method and compared its performance with the classical approach of nonlinear regression. We have given details of each curve fitting method in the sections below.

**Nonlinear Regression Method**

The biphasic DA curve was described by Equation 1 below, with an exponential decay representing the cone phase and a linear decay for the rod phase:

$$y = ae^{−bx} + c \max(x − e, 0) + d$$

where relative retinal sensitivity $y$ (LU) at time $x$ (minutes [min]) is measured following cessation of the bleach. The five derived parameters were (a, [LU]) the time cone decay intercept, (b, [min\textsuperscript{−1}]) cone decay time constant, (c [LU/min]) rod slope, (d, [LU]) the cone plateau and (e, [min]) the time to the rod-cone break.

Given a set of measurements in a DA test, nonlinear regression is used to estimate the parameters via optimization, such that, at convergence, the mean squared error between the actual measurements and the curve predictions are minimized. Parameters were initialized by computing a curve overlapping with measured data points as follows. Rod-cone break ($e_0$) was set to the halfway time point of the total test duration. Cone plateau ($d_0$) was set to the average of the maximum and minimum intensity $[(LU_{\text{max}} + LU_{\text{min}})/2]$ measurements in the test. After empirical testing parameters $a_0$, $b_0$, and $c_0$ were set according to $LU_{\text{min}} − d_0$, $15/e_0$, and $(LU_{\text{max}} − d_0)/e_0$, respectively. Optimization was performed using a trust-region algorithm with convergence criteria: parameter tolerance $= 10^{-5}$, function tolerance $= 10^{-5}$, maximum function evaluations 600, and maximum iterations $= 400$. These hyperparameters were set empirically during initial development in a subset of the data.

**Deep Learning-Based Method**

The same data were used to fit the curve using the long short-term memory autoencoder (LSTM AE). As the first step, the sparse data points were interpolated within a −1.0 LU and −3.1 LU range, using isotonic regression, to obtain a monotonically-decreasing fit to the data. An LSTM AE model was then devised to estimate the DA curve while minimizing the curve fluctuations due to noise. The input to the network was 161 interpolated points from isotonic regression. Both encoder and decoder consisted of single layer LSTM with embedding dimension $= 64$. The model was trained for 150 epochs using Adam optimizer and L1-norm loss function with a learning rate $= 10^{-2}$. \textsuperscript{3}
Deep Learning-Based DA Curve Modelling

Figure 1. Curve fits obtained for DA test data points (yellow). (A) Classical method (red curve). (B) LSTM output (thin blue curve) and NLR method applied to the LSTM output (thick blue curve).

Figure 2. Comparison of RMSE distributions for NLR and LSTM methods.

These hyperparameters were set empirically during initial development in a subset of the data.

The classical method intrinsically solves for five parameters, whereas the DL method fits a nonparametric curve to the data. For direct comparison, the classical equation (Equation 1) was subsequently applied to the DL curve points and the five parameters were estimated. Figure 1 shows example curve fits using the classical method (nonlinear regression [NLR]), LSTM fit, and the NLR applied to the output of the LSTM fit (LSTM\text{NLR}).

As an additional baseline comparison, we also implemented locally weighted scatterplot smoothing (LOWESS) regression to the data acquired during the DA test with 33% of the data used in estimating each y (LU) value. This locally weighted regression approach can also serve as a noise reduction and smoothing method that can mitigate the effect of spurious measurements during the test. The classical equation was applied subsequently to the LOWESS regression to obtain the curve parameters for comparison (LOWESS\text{NLR}).

The LSTM model was then trained to predict the latter part of the curve using the sequence of measurements up to different early portions of the curve, ranging from $-2.0$ LU to $-2.9$ LU.

Experiments

Curve Fitting

We devised a 20-fold cross validation approach where for each fold approximately 85% (1271 on average) tests were used for training 10% (150 average tests for validation) and 5% (75 average tests for testing). In each model, RIT was measured as the time required to reach $-3$ LU. All methods were implemented in PyTorch (version 1.10) and trained using dual Intel Xeon CPUs (2.2 GHz), 128 GB RAM, and 2 Nvidia GTX Titan V GPUs. The goodness-of-fit was measured as the root mean square error (RMSE) between measured data points and the corresponding curve estimated points.

The robustness of the curve fitting method was evaluated in two different ways. First, each measured data point was randomly fluctuated, introducing an error distributed within a maximum range of $\pm 2$ minutes. The parameters extracted from the original measurements were compared to those from noise-fluctuated measurements for both the NLR and LSTM methods. Second, the repeatability was measured as the absolute difference between curve parameters estimated from the tests spanning 3 months or less. In both methods, after fitting the classical curve, the
sum of squared error (SSE) was measured between the curve and the data point, as well as a simple linear model. If the data were better represented by a linear model with a lower SSE, such curves were removed from the comparison. Tests that did not reach the rod-cone break within the test duration were also removed from this analysis. In the noise simulation experiment, a total of 954 tests were compared; in the repeated test experiment, a total of 238 tests were compared. For additional validation, 11 repeated test pairs of the additional clinical study were assessed – these additional tests were not part of the cross validation set up used for model training.

Curve Prediction

Both curve-fitting models were used to explore the ability to predict later points of the curve from only a portion of the early data points after observing early measurements within a range of $-2.0$ LU to $-2.9$ LU. The RIT prediction error for each approach was assessed at different observed levels of LU thresholds used for predictions, as well as a function of the time elapsed from the beginning of the test.

AMD Severity Scale and RPD Estimation Using Curve Parameters

A random forest (RF) nonlinear regression/classification model was trained to predict AMD severity scale curve (AMDSC; regression) and RPD presence/absence (classification) from values extracted from the curves. RF prediction models’ (both regression and classification) performance was assessed in a 10-fold cross validation setting. First, predictions were made using single LU data point measures extracted from different parts of the curve. Second, predictions were made using a combination of points, where the number of equidistant points was increased gradually (1, 3, 6, 12, 100, and 200). These measured the association of AMDSC and RPD with different combinations of points sampled from the curve.

Table 1. Outlier Percentage Comparison Between the Two Methods

|          | a     | b     | c     | d     | e     |
|----------|-------|-------|-------|-------|-------|
| NLR      | 14.1  | 9.8   | 0.1   | 5.2   | 0.0   |
| LSTM$_{NLR}$ | 2.0   | 0.5   | 0.0   | 1.2   | 0.0   |

Robustness of the obtained parameters is an important indicator when comparing the utility of curve fitting methods. Parameter distributions obtained from each curve fitting method showed outliers falling outside the standard range. For this analysis, we defined an outlier as a value that falls outside the $\text{mean} \pm 3\times\text{SD}$ of the distribution. We quantified the percentage of outliers for each parameter distribution computed from the two methods. Overall, the LSTM$_{NLR}$ method showed lower outlier percentages across all parameters (Table 1). The NLR method exhibited substantially increased outliers in parameters related to cone decay.

To evaluate robustness further, random fluctuations of up to ±2 minutes were added to the data prior to each curve-fitting method. Figure 3 and Table 2 show the robustness of the parameters measured as the error in estimations after introducing the simulated noise fluctuations. Overall, the parameters estimated from the LSTM method exhibited lower error/variability where improvements were more prominent in parameters related to the cone decay. Although the LOWESS method improved upon the NLR method and demonstrated some ability to mitigate the effect of spurious measurements, the LSTM method yielded the best overall performance. Significant differences were observed according to the Wilcoxon signed rank test in all parameters with a significance level $P < 0.001$.

We next analyzed the repeatability of the parameters measured from different tests conducted within a short time span (<3 months; Fig. 4 and Table 3). Overall, cone-related parameters estimated from the LSTM$_{NLR}$ method exhibited superior repeatability, compared to those from the LOWESS$_{NLR}$ and NLR methods. All parameters except the rod-cone break (e) and RIT exhibited statistically significant differences ($P < 0.001$, Wilcoxon signed rank test). RIT repeatability was also compared with the values obtained from the AdaptDx machine and found to be comparable across the four different methods compared. However, all methods demonstrated comparable performance when measuring RIT, indicating the relevance of the curve fitting
Figure 3. Parameter estimation robustness from NLR, LOWESSNLIR, and LSTMNLIR methods. DA tests measurements were randomly fluctuated to create up to ±2 minutes error. For each method, parameter estimation error was computed as the absolute difference between original and noise-fluctuated curves (e.g. $\Delta a = |a_{\text{orig}} - a_{\text{noise}}|$).

Table 2. Comparison of Mean ± SD Error for NLR and LSTM$_{NLIR}$ Methods After Randomly Fluctuated to Create up to ±2 Minutes’ Error

|        | $\Delta a$   | $\Delta b$   | $\Delta c$   | $\Delta d$   | $\Delta e$   | $\Delta RIT$ |
|--------|--------------|--------------|--------------|--------------|--------------|--------------|
| NLR    | 0.8 ± 2.4    | 0.8 ± 2.0    | 0.02 ± 0.03  | 0.06 ± 0.29  | 0.8 ± 1.1    | 0.5 ± 0.5    |
| LOWESSNLIR | 0.2 ± 0.4    | 0.1 ± 0.2    | 0.01 ± 0.02  | 0.08 ± 0.30  | 0.6 ± 1.4    | 0.3 ± 0.4    |
| LSTM$_{NLIR}$ | 0.1 ± 0.2    | 0.1 ± 0.2    | 0.01 ± 0.02  | 0.02 ± 0.04  | 0.5 ± 0.8    | 0.5 ± 0.6    |

Figure 4. Parameter estimation repeatability from NLR, LOWESSNLIR, and DL methods (tests performed within 3 months of the same patient). $\Delta RIT$ was also compared to the values obtained from the AdaptDx machine.
Table 3. Comparison of Mean ± SD Error for NLR and LSTM<sub>NLR</sub> Methods After Repeated Tests Within 3 Months

| Method   | \(\Delta a\) | \(\Delta b\) | \(\Delta c\) | \(\Delta d\) | \(\Delta e\) | \(\Delta RIT\) |
|----------|---------------|---------------|---------------|---------------|---------------|---------------|
| NLR      | 0.6 ± 1.7     | 0.7 ± 1.6     | 0.02 ± 0.04   | 0.11 ± 0.20   | 1.9 ± 3.0     | 1.65 ± 3.00   |
| LOWESS<sub>NLR</sub> | 0.4 ± 0.7     | 0.1 ± 0.3     | 0.02 ± 0.03   | 0.36 ± 0.70   | 2.2 ± 3.7     | 1.70 ± 2.90   |
| LSTM<sub>NLR</sub> | 0.2 ± 0.3     | 0.2 ± 0.3     | 0.02 ± 0.03   | 0.07 ± 0.09   | 2.0 ± 3.2     | 1.70 ± 3.00   |
| \(P\) value | 2.5e-18       | 4.1e-24       | 1.2e-15       | 2.5e-10       | 0.26          | 0.39          |

Method when deriving parameters in addition to the RIT form DA tests. Repeated tests capture both noisy random fluctuations in data points as well as any biases resulting in an overall shorter or longer test. Thus, the measured error between parameters, such as RIT, from the tests were larger compared to the random noise experiment described before. Supplementary Figure S1 shows additional validation using 11 independent test set pairs supporting the observations in Figure 4 where the LSTM<sub>NLR</sub> method exhibited superior overall repeatability when all parameters are considered. Supplementary Figure S2 shows 5 representative patients’ cases comparing curve derivations using both methods within 3 months. Supplementary Figure S3 shows convergence plots during training.

Curve Prediction

Figure 5 shows the performance of each curve-fitting method in predicting the RIT at different stages of the curve. When predicted at \(-2.5 \text{ LU}\) using the LSTM method, the predicted RIT had error = 1.2 ± 1.3 (mean ± SD) minutes, whereas, at \(-2 \text{ LU}\), the predicted RIT error degraded to 3.2 ± 3.8 minutes. In comparison, the NLR method exhibited statistically significantly larger errors with 6.3 ± 9.9 minutes at \(-2.5 \text{ LU}\) and 19.1 ± 18.6 at \(-2 \text{ LU}\), often showing poor predictability with gross failures. Using the LSTM method, even at \(-2 \text{ LU}\), 78.7% cases reported error <5 minutes, whereas, with the NLR method, only 18.6% of the cases demonstrated error <5 minutes. This shows the LSTM method’s ability to detect trends in the...
curve and make predictions using a limited number of earlier measurements exploiting possible temporal correlations in the data.

Figure 6 shows the RIT prediction error from the LSTM method made at different time points along the curve. Overall, the prediction accuracy increased when making predictions with increased elapsed time from the beginning of the test. The predictions made within 18 to 22 minutes, had good forecast ability with a mean $\pm$ SD = 1.6 ± 1.7 minutes and a range of 0.0 to 12.9 minutes. In comparison, NLR had a poor error $= 12.6 \pm 15.4$ minutes (range = 0.0–59.0 minutes).

Parameter Associations With AMD Severity

The ability to use points on the DA curve to predict AMD severity and the presence/absence of RPD was explored using the curves estimated from the LSTM method. The AMDSC predicted using the latter points in the curve (error $= 2.81$ at $-3.0$ LU) was marginally superior to that predicted using the earlier points (error $= 3.01$ at $-1.0$ LU; Fig. 7A). When using multiple, equally spaced points, the AMDSC prediction error gradually improved from 2.81 (single point at $-3.0$ LU) to 2.44 (using only two points) and 2.38 when using all the points in the curve (Fig. 7B). This suggests that features from the curve other than RIT could supplement the prediction of AMDSC.

The RPD prediction showed excellent performance (accuracy $>0.90$) across different points/point combinations from the curve. RIT alone was a good predictor (accuracy $= 0.90$) for RPD presence and the inclusion of additional curve-derived points did not improve this performance (Fig. 7D). However, it is interesting that even earlier points from the curve demonstrated good predictability (accuracy $= 0.88$ at $-1.0$ LU; Fig. 7C).

Discussion

This study focuses on reliably estimating DA curve parameters from DA test data to improve their utility in associating with early and intermediate AMD. The DA curve parameters, especially RIT, have been shown to provide critical insights into AMD disease severity and progression. However, limitations to maximum utilization of these curves stem both from the patient burden in acquiring the full length of the testing course in the older population and from the variability and sparseness in recorded measurements. Our results indicated that the DL method exhibited superior robustness in the presence of noise fluctuations, especially in the cone decay portion of the curve despite the classical NLR method demonstrating slightly better goodness-of-fit. In addition, the ability to predict later points from the early phase of the DA curve using the LSTM method suggests that early trends in the curve could help estimate these parameters in patients who are unable to complete the test, due to the patient’s limitations or the ceiling time of the test. Although the model does not explicitly use any relationship between the cone- and rod recovery stages, it observes temporal trends in the measured data to predict the latter portion of the DA curve. The prediction accuracy intuitively improves after observing more data points both as a function of elapsed time since the cessation of bleach-
Deep Learning-Based DA Curve Modelling

Figure 7. (A) AMDSC prediction using a single point extracted from the curve. (B) AMDSC predictions from multiple points extracted from the curve. (C) RPD prediction using a single point extracted from the curve. (D) RPD predictions from multiple points extracted from the curve.

ing and LU threshold reached in that time. As the test progresses, information from stages, such as the cone plateau and/or cone rod break, may aid the LSTM method to make a reliable RIT prediction. Our results could also suggest that some trends in the cone decay phase of the curve may aid in the RIT prediction. For example, if cone decay has been impacted due to degeneration, it may be suggestive that rods would also demonstrate delayed adaptation. Such enhancements would yield better overall repeatability of the DA curves and improve reliability in estimations, regardless of variable test ceilings (no matter when the test is ended). These could provide valuable contributions in advancing the utility of the tests, which is especially time-consuming and arduous for elderly patients.

Previous studies have demonstrated that the parameters extracted from dark adaptation testing can provide useful information relevant to AMD disease pathogenesis as they have been shown to correlate with AMD disease severity and progression. RIT has been reported as the most useful and relevant parameter indicative of rod degeneration and the results of this work highlight that this parameter also has the highest reproducibility irrespective of the curve-fitting method used. Thus, the methods presented in this work could benefit in reliably extracting other cone- and rod-related parameters with suboptimal reproducibility. Ideally, the curve fitting methods would derive all dark adaptation curve parameters with excellent reproducibility. Improved reproducibility of curve derived parameters could provide the tools for analyses that could further illuminate the contributions of cone and rod degeneration in patients with AMD.

Furthermore, the curve fitting methods need to observe measured data points holistically to determine the optimal fit to classical biphasic function, as these classic equations were derived for healthy eyes; disease states may not follow the same assumptions. In doing so, the presented DL approach can use sequential trends in data to mitigate noise and enhance reproducibility. For example, if the cone phase indicates delayed adaptation, it is more likely that the rod phase also exhibits delays.

The dark adaptation testing protocol may not capture ample data points during the cone recovery phase, and cone parameters are thus more susceptible to variability and poor reproducibility. This is evidenced by a larger percentage of outliers from the
NLR method driven by the limited (i.e. sparse) data points acquired during that phase. In this setting, the experiments show that substantial improvements in estimating cone parameters can be achieved with the LSTM method, after exploiting the trends in these sparse data points. Our results indicate that the parameters derived from the LSTM method exhibit better robustness, resulting overall in lower variability and higher repeatability. Such advancements are critical to reliably estimate DA curve parameters in addition to RIT to serve as useful indicators in studying disease correlates/outcomes. Our analysis showed the addition of such parameters could potentially benefit in AMD severity prediction.

The variability observed in the repeated experiments could be due to noisy data points partially triggered by participant burden and tiredness considering the long test duration, patient age, and the impact of disease. Under a short time duration of less than 3 months, we assume there is minimal structural changes and thus the observed variability is an undesirable source, minimizing, which could only benefit the test utility.

The DL method fitted smooth transitions to the curve between the cone and rod phases without identifying an explicit rod-cone break point. This could indicate the challenges in reliably estimating the rod-cone break point using sparse data points. The motivation for the development of RIT as an outcome measure could partly be due to such variability in extracting alternative metrics. Fitting the classical curve to these smooth transitions, however, improved the repeatability as observed in the NLR versus LSTM NLR comparisons performed in this paper. Thus, the LSTM method (see Fig. 3, Fig. 4e) can serve as a noise-reducing, interpolation method to the raw data points recorded during the test.

A particularly useful aspect of the LSTM model is in the learned correlations inherent to the DA curve data. The NLR method does not inherently carry any potential for observing the trends for prediction and only finds the best estimate for the observed set of data points. By contrast, the LSTM method exhibited superior performance in predicting later data points after observing sequential trends in the data. Further analyzing and capitalizing on these sequential and temporal correlations in the data could lead to the development of shorter tests.

Whereas the NLR method was derived by estimating two functional cell kinetic properties, approaching the data agnostic to any cellular function constraints could carry advantages. For example, without assuming an explicit model for the DA curve could be desirable when the disease induces deviations to the classically defined curve. Such deviations can be detected and represented within the curve with the proposed LSTM method. The validation approach and the results presented in this work is limited to the single center setting. External validation using large data sets would further strengthen the generalizability of the method under variable equipment, personnel, and test protocol settings.

Dark adaptation is becoming a more widespread testing modality but, given the psychophysical nature of the test, is limited by its variability. The methods reported in this paper aim to improve the reliable estimation of the DA curve parameters and could lead to improved end points to correlate with disease severity and the development of shorter, effective DA testing protocols.

**Acknowledgments**

Supported by funds from the National Eye Institute, Intramural Research Program, National Institutes of Health (NIH).

Previous Publication: This work was presented at the ARVO 2021 Annual Meeting.

Disclosure: T. De Silva, None; K. Hess, None; P. Grisso, None; A.T. Thavikulwat, None; H. Wiley, None; T.D.C. Keenan, None; E.Y. Chew, None; B.G. Jeffrey, None; C.A. Cukras, None

**References**

1. Lamb TD, Pugh EN. Dark adaptation and the retinoid cycle of vision. *Prog Retin Eye Res*. 2004;23(3):307–380.
2. Higgins BE, Taylor DJ, Binns AM, Crabb DP. Are Current Methods of Measuring Dark Adaptation Effective in Detecting the Onset and Progression of Age-Related Macular Degeneration? A Systematic Literature Review. *Ophthalmol Ther*. 2021;10(1):21–38.
3. Owsley C, Clark ME, McGwin G. Natural History of Rod-Mediated Dark Adaptation over 2 Years in Intermediate Age-Related Macular Degeneration. *Transl Vis Sci Technol*. 2017;6(3):15.
4. Jackson GR, Edwards JG. A short-duration dark adaptation protocol for assessment of age-related maculopathy. *J Ocul Biol Dis Infor*. 2008;1(1):7–11.
5. Jacobson SG, Cideciyan A V, Wright E, Wright AF. Phenotypic marker for early disease detection
4. Mcgwin G, Jackson GR, Owsley C. Using non-linear regression to estimate parameters of dark adaptation. Behav Res Methods, Instruments, Comput. 1999;31(4):712–717.

5. Baker HD. Foveal dark adaptation, photopigment regeneration, and aging. Vis Neurosci. 1992;8(1):27–39.

6. Owsley C, Jackson GR, White M, Feist R, Edwards D. Delays in rod-mediated dark adaptation in early age-related maculopathy. Ophthalmology. 2001;108(7):1196–1202.

7. Owsley C, Jackson GR, White M, Feist R, Edwards D. Delays in rod-mediated dark adaptation in early age-related maculopathy. Ophthalmology. 2001;108(7):1196–1202.

8. Dimitrov PN, Guymer RH, Zele AJ, Anderson AJ, Vingrys AJ. Measuring rod and cone dynamics in age-related maculopathy. Invest Ophthalmol Vis Sci. 2008;49(1):55–65.

9. Flynn OJ, Cukras CA, Jeffrey BG. Characterization of rod function phenotypes across a range of age-related macular degeneration severities and subretinal drusenoid deposits. Invest Ophthalmol Vis Sci. 2018;59(6):2411–2421.

10. Flamendorf J, Agrón E, Wong WT, et al. Impairments in Dark Adaptation Are Associated with Age-Related Macular Degeneration Severity and Reticular Pseudodrusen. Ophthalmology. 2015;122(10):2053–2062.

11. Dimitrov PN, Robman LD, Varsamidis M, et al. Relationship between clinical macular changes and retinal function in age-related macular degeneration. Invest Ophthalmol Vis Sci. 2012;83(9):5213–5220.

12. Dimitrov PN, Robman LD, Varsamidis M, et al. Visual function tests as potential biomarkers in age-related macular degeneration. Invest Ophthalmol Vis Sci. 2011;52(13):9457–9469.

13. Jackson GR, Owsley C, McGwin G. Aging and dark adaptation. Vision Res. 1999;39(23):3975–3982.

14. Mcgwin G, Jackson GR, Owsley C. Using non-linear regression to estimate parameters of dark adaptation. Behav Res Methods, Instruments, Comput. 1999;31(4):712–717.

15. Baker HD. Foveal dark adaptation, photopigment regeneration, and aging. Vis Neurosci. 1992;8(1):27–39.

16. Owsley C, Jackson GR, White M, Feist R, Edwards D. Delays in rod-mediated dark adaptation in early age-related maculopathy. Ophthalmology. 2001;108(7):1196–1202.

17. Pugh EN. Rushton’s paradox: rod dark adaptation after flash photolysis. J Physiol. 1975;248(2):413–431.

18. McGwin G, Jackson GR, Owsley C. Using non-linear regression to estimate parameters of dark adaptation. Behav Res Methods, Instruments, Comput. 1999;31(4):712–717.

19. Murray IJ, Rodrigo-Diaz E, Kelly JMF, et al. The role of dark adaptation in understanding early AMD. Prog Retin Eye Res. 2021;88:101015.

20. Ferris FL, Davis MD, Clemons TE, et al. A simplified severity scale for age-related macular degeneration: AREDS report no. 18. Arch Ophthalmol. 2005;123(11):1570–1574.

21. Flamendorf J, Agrón E, Wong WT, et al. Impairments in dark adaptation are associated with age-related macular degeneration severity and reticular pseudodrusen. Ophthalmology. 2015;122(10):2053–2062.

22. Davis MD, Gangnon RE, Lee LY, et al. The age-related eye disease study severity scale for age-related macular degeneration: AREDS report no. 17. Arch Ophthalmol. 2005;123(11):1484–1498.