Methods and Reproducibility of Grading Optimized Digital Color Fundus Photographs in the Age-Related Eye Disease Study 2 (AREDS2 Report Number 2)

Ronald P. Danis,1 Amitha Domalpally,1 Emily Y. Chew,2 Traci E. Clemons,3 Jane Armstrong,1 John Paul SanGiovanni,2 and Frederick L. Ferris III2 for the AREDS2 Study Group4

1Department of Ophthalmology and Visual Sciences, University of Wisconsin–Madison, Madison, Wisconsin
2National Institutes of Health, National Eye Institute, Bethesda, Maryland
3EMMES Corporation, Rockville, Maryland
4See the Appendix in the Supplementary Material for the members of the AREDS2 Study Group.

Correspondence: Ronald P. Danis, Department of Ophthalmology and Visual Sciences, University of Wisconsin–Madison, 8010 Excelsior Drive, Madison, WI 53717; rpdanis@wisc.edu.


PURPOSE. To establish continuity with the grading procedures and outcomes from the historical data of the Age-Related Eye Disease Study (AREDS), color photographic imaging and evaluation procedures for the assessment of age-related macular degeneration (AMD) were modified for digital imaging in the AREDS2. The reproducibility of the grading of index AMD lesion components and for the AREDS severity scale was tested at the AREDS2 reading center.

METHODS. Digital color stereoscopic fundus photographs from 4203 AREDS2 subjects collected at baseline and annual follow-up visits were optimized for tonal balance and graded according to a standard protocol slightly modified from AREDS. The reproducibility of digital grading of AREDS2 images was assessed by reproducibility exercises, temporal drift (regrading a subset of baseline annually, n = 88), and contemporaneous masked regrading (ongoing, monthly regrade on 5% of submissions, n = 1335 eyes).

RESULTS. In AREDS2, 91% and 96% of images received replicate grades within two steps of the baseline value on the AREDS severity scale for temporal drift and contemporaneous assessment, respectively (weighted Kappa of 0.73 and 0.76). Historical data for temporal drift in replicate gradings on the AREDS film-based images were 88% within two steps (weighted Kappa = 0.88). There was no difference in AREDS2–AREDS concordance for temporal drift (exact P = 0.57).

CONCLUSIONS. Digital color grading has nearly the same reproducibility as historical film grading. There is substantial agreement for testing the predictive utility of the AREDS severity scale in AREDS2 as a clinical trial outcome. (ClinicalTrials.gov number, NCT00345176.)

Keywords: color photographic imaging, AMD, reproducibility

The Age-Related Eye Disease Study 2 (AREDS2, www.areds2.org, in the public domain) is a 5-year, multicenter, randomized, controlled clinical trial designed to evaluate the efficacy and safety of daily dietary supplementation with lutein/zeaxanthin (10 mg/2 mg) and/or omega-3 long-chain polyunsaturated fatty acids (LCPUFAs; docosahexaenoic acid [DHA] 350 mg + eicosapentaenoic acid [EPA] 650 mg) on the rate of progression to advanced age-related macular degeneration (AAMD) among people with at least intermediate age-related macular degeneration (AMD) at baseline. AAMD is defined as neovascular AMD determined by a standardized grading procedure applied at the AREDS2 reading center (University of Wisconsin–Madison, Fundus Photograph Reading Center) or treatment for choroidal neovascularization (CNV) or geographic atrophy (GA) involving the center of the macula determined by the AREDS2 reading center.1 Findings from the Age-Related Eye Disease Study (AREDS) indicate that nutritional supplementation with a zinc and antioxidant combination formulation could decrease the rate of progression to AAMD by 25%.2 In AREDS2, a second-tier randomization was applied to investigate the impact of four variations of the original AREDS formulation on the study endpoints.

The color photograph grading procedures developed for AREDS were based upon the Wisconsin Age-Related Maculopathy Grading System,3 which assessed presence, location, and severity of the abnormalities characteristic of AMD. AREDS color fundus image gradings were used to develop severity scales for AMD.4,5 These include a detailed nine-step scale for individual eyes and a simplified five-step person-based scale. In the development of the eye-based scale, AMD lesion components such as drusen and pigment abnormalities were examined as risk factors for 5-year progression to AAMD in the right eyes of 3212 participants without AAMD in either eye at baseline. Three closely associated drusen characteristics (area, size, and type) were evaluated. Area of drusen within the AREDS macular grid was the strongest and most consistent predictor. Areas of increased and decreased pigmentaion and presence of noncentral GA were assessed and combined in a pigment abnormality assessment. These drusen and pigment subscales were then combined and tested on AREDS AMD endpoints. The 5-year risk of neovascular AMD increased from 0.3% for participants classified as step 1 to 21% in participants classified as step 8. Central GA risk increased from 0.5% in step 4 to 45.5% in step 9.4 Reproducibility of the scale was...
Methods and Grading Procedures for AREDS2 Images

Methods

Subjects

The AREDS2 study enrolled 4203 participants aged 50 to 85 years after evaluating for eligibility criteria: (1) bilateral large drusen or (2) large drusen in one eye and advanced AMD (neovascular AMD or central GA) in the fellow eye. Further details on subject characteristics are available in AREDS2 Report 1.1

Photography

All AREDS2 photographers and clinical site digital camera systems are certified by the reading center. Color stereoscopic fundus photographs are obtained using three photographic fields of the macula and optic nerve with 50° or 35° fundus cameras, as in AREDS. The imaging protocol specifies field position and stereoscopic technique. Red reflex photographs with a fundus camera are also obtained, which are helpful in determining the causes for suboptimal-quality image sets. Seven models of digital fundus cameras were permitted for use in AREDS2. All had a minimum resolution specification of 3 megapixels. For baseline image collection, 20 of 82 clinical sites did not have approved digital fundus cameras and were allowed to use Ektachrome color slide film (Eastman Kodak Co., Rochester, NY) for photography. Subsequently, all clinical sites transitioned to digital color photography. Film photograph submissions are digitized at the reading center using a Nikon Super CoolScan 5000 (Nikon, Inc., Melville, NY) to produce 3400 × 2300 pixel images at 8 bits/channel (which produces a 22.4 megabyte image) so that both film-based and natively digital images may be viewed and managed in the same display environment. Digital images are exported and saved in formats specified in the study procedures and are uploaded to the reading center by secure Web site or mailed to the reading center on computer discs. At the reading center, all camera exports are converted to lossless jpeg images (compression to reduce file size without a visible effect on image detail) and are imported into a database for viewing with Topcon IMAGEnet software (Topcon Medical Systems, Paramus, NJ), using image scale calibration factors determined for each digital camera from the certification process. Image quality is monitored throughout the course of the study; and graders evaluate the image quality of each submission according to the degree to which quality affects the ability to discern important lesion features. All ungradable image sets are then evaluated with a detailed technical assessment. Image quality reports are compiled by the center and by site semiannually, and photographers with more than 5% ungradable images that appear to be due to operator-dependent issues (such as focus, illumination, color balance, field definition, and not related to cataract, small pupil or other media opacity) have a sample of their work reviewed in detail by a reading center photographer to provide assistance in improving the quality of images.

Grading

Grading Tools. Prior to disease evaluation, the tonal properties of the digital image are assessed by optimizing red/green/blue luminance curves through a method developed by Hubbard et al. Briefly, luminance histograms of image submissions are examined with tools in IMAGEnet software (version 2.56; Topcon Medical Systems). If the tonal characteristics are determined to be not acceptable according to standardized criteria, individual color channel curves are adjusted to match an optimal (if possible). In this manner, the variability in illumination and color contrast between digital images is minimized and more closely resemble film quality. Evaluation is performed using both the original and optimized images. Images are accessed from the database using IMAGEnet software (Topcon Medical Systems) for display and measurement on calibrated HP400 monitors (Hewlett Packard, Palo Alto, CA) in calibrated ambient lighting and viewed with a handheld stereoscope (Screen Vu; PS Manufacturing, Portland, OR). Graders are allowed to use limited zoom features in the display software. An electronic Early Treatment Diabetic Retinopathy Study (ETDRS) macular grid, appropriately sized for the magnification of the digital fundus image, is overlaid to specify the location of some macular lesions by grid subfield, similar to the methodology used in AREDS with acetate overlays on color slides. Drusen area circles as employed in AREDS are also scaled to the magnification of the photograph (determined at the time of camera system certification) and overlaid on the digital image as needed (Fig. 1). The display software includes mouse-driven planimetry tools that are used for image measurements with outputs calibrated to the magnification scale.

Grading Procedure. The grader first examines the image for characteristics of neovascularization using five photographic features: subretinal fluid, serous pigment epithelial detachment, or retinal edema; intraretinal, subretinal pigment epithelium or subretinal blood characteristic of neovascular AMD; intraretinal lipid exudates (hard exudates); subretinal fibrin or fibrosis; and fibrovascular pigment epithelial detachment. If one out of these five features is definitely present, the eye is graded as “questionable neovascular AMD.” Questionable cases are pooled with the corresponding nonneovascular scale step (the severity for the eye if the single neovascular feature was not present) for data analysis purposes. The presence of two or more such features categorizes an eye as “definite neovascular AMD.” Because in current retinal clinical practice, eyes developing neovascular AMD are usually promptly treated with intravitreal antiangiogenic endothelial growth factor agents such as ranibizumab, bevacizumab, or aflibercept, eyes with neovascular AMD may have no evident signs of neovascularization on color photographs (obtained at an annual AREDS2 study visit. The AREDS2 coordinating center (EMMES Corporation, Rockville, MD) collects information from...
the clinical sites regarding the treatment of CNV, which is used to classify eyes with neovascular AMD.

The nonneovascular or atrophic AMD features documented in the image evaluation process are similar to the grading used in AREDS. Drusen size, type, and area and the presence and severity of hyper- and hypopigmentation are evaluated as in AREDS (Table 1). Maximum drusen size is estimated using standard drusen circles scaled to the magnification of the digital image. Area of drusen within the grid is estimated by mentally moving together all drusen present into the appropriate drusen measurement circle, as described in AREDS. Area of GA is directly measured using computerized planimetry with measurements scaled to the magnification of the digital photograph. GA is considered definitely present if the lesion is at least 433 µm in diameter (greater than or equal to the AREDS circle I-2) and if at least two of the following features are present: absence of RPE pigment, circular shape, or sharp margins. If noncentral, the proximity of the border of the atrophy closest to the center (umbo of the foveola) is documented. In addition, the configuration and severity of the GA are also evaluated. Because drusen and pigmentation changes tend to be small areas and multifocal, planimetry is less useful for quantitative area measurement of these features.

in the digital environment, standard drusen circles and Macular Photocoagulation Study measurement circles are used for area estimations.

Quantitative evaluation is limited to the area within the grid (Fig. 1), which is modified from AREDS. AREDS2 has adopted the modern assumption that the standard disc diameter (DD) equals 1800 µm, rather than retaining the classic assumption that it equals 1500 µm. This change is required to accommodate digital imaging in AREDS2, since most cameras are calibrated with 1800 µm equivalency. Consequently, the grid used in AREDS2 is modified from AREDS. The diameters of the grid circles remain as 1/3 DD, 1 DD, and 2 DD, but are now equivalent to 600 µm, 1800 µm, and 3600 µm, respectively (Table 1, Fig. 1). AREDS2 and AREDS data may be compared by applying a scaling constant to the AREDS data. For linear measurements, this is accomplished by multiplying by 1.20, and for area measurements by multiplying by 1.44.

Baseline AREDS2 images were graded by two independent graders. Grading results were assessed by a software processor, and discrepancies on major questions (component questions for the AREDS2 severity scale) were adjudicated by a third, senior grader (JA). If no grading discrepancies were identified, the first grade was submitted as the grade of record. For annual

**FIGURE 1.** Screen capture with digital AREDS macular grid and drusen circles on color fundus photograph in IMAGEnet. See Table 1 for area definitions of grading circles.
**TABLE 1.** Lesions Graded From Color Photographs in AREDS2

<table>
<thead>
<tr>
<th>Macular Feature Graded</th>
<th>Grading Data Collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area of hyperpigmentation</td>
<td>4-step scale: Absent, Questionable,* Definite &lt; C-2,† Definite ≥ C-2</td>
</tr>
<tr>
<td>Area of hypopigmentation</td>
<td>4-step scale: Absent, Questionable, Definite &lt; I-2, Definite &lt; 1/2 DA, Definite ≥ 1/2 DA</td>
</tr>
<tr>
<td>Maximum drusen size</td>
<td>6-step scale: Absent, Questionable, Definite &lt; C-0 μm, Definite &lt; C-1 μm, Definite &lt; C-2 μm, Definite ≥ C-2 μm</td>
</tr>
<tr>
<td>Area of drusen within the grid</td>
<td>8-step scale: &lt; C-0, &lt; C-1, &lt; C-2, &lt; I-2, &lt; O-2, &lt; 1/2 DA, &lt; 1 DA, ≥ 1 DA</td>
</tr>
<tr>
<td>Calculated drusen, drusenoid PED, reticular drusen</td>
<td>Presence/absence</td>
</tr>
<tr>
<td>Area of geographic atrophy within the grid</td>
<td>DA (disc areas) by planimetry (computer mouse)</td>
</tr>
<tr>
<td>Involvement of the center of the macula by geographic atrophy</td>
<td>Presence/absence</td>
</tr>
<tr>
<td>Configuration of geographic atrophy</td>
<td>Small single patch &lt; 1 DA, multifocal, horseshoe, ring, solid not center, solid center involved, indeterminate‡</td>
</tr>
<tr>
<td>Neovascular AMD</td>
<td>Presence/absence</td>
</tr>
</tbody>
</table>

* DRUsen pigmenThe AREDS2 grading quality control program consists of several elements. A temporal drift sample of 88 stratified baseline images is regraded annually by the entire grading group; the results are compared to original grades for the same sample. The temporal drift reproducibility exercises allow monitoring the shift due to grader experience, change in grading personnel, and technological advances, particularly in studies with long follow-up such as AREDS2. The contemporaneous quality control includes monthly regrade of a random sample of 5% of submissions. These images are duplicated and passed through the grading process with fictitious identifiers for masked replicate grading. The reproducibility of grading is assessed by calculating percentage agreement and weighted Kappa statistics for ordinal variables and correlation coefficients for continuous area measurements for the entire group. Inspection of scatter plots for outliers and potential systematic bias is performed for temporal drift or problematic questions for targeted retraining if necessary. Regular training exercises are held for the entire grading group with review of difficult cases and reaffirmation of the grading protocol. Reproducibility statistics are also examined for individual graders, and targeted individual retraining is performed if the grader has reproducibility for specific questions below a set threshold. All graders are encouraged to seek out a reading center ophthalmologist for “second opinions” for assistance with unusual presentations or confounding ocular abnormalities. On an ongoing basis, any eyes meeting the study endpoint are reviewed by a reading center ophthalmologist to confirm the endpoint.

**Outcomes**

For categorical variables on an ordinal scale, such as drusen area or AREDS severity scale, agreement is measured in terms of percentage agreement and weighted Kappa. For continuous variables such as area of GA, Bland-Altman plots have been presented.

**RESULTS**

Table 2 shows the results of the temporal drift quality control from the fourth year of the AREDS2 study, contemporaneous quality control data over 4-year period, and published data for AREDS temporal drift. The reproducibility values of grading for AMD lesion components from optimized digital photographs in AREDS2 were essentially equivalent to those from film images in AREDS (Table 2). For AREDS2–AREDS concordance of

**TABLE 2.** AREDS2 Grading Reproducibility for Select Ordinal Variables in Temporal Drift and Contemporaneous Regrading Samples Compared With Published Data From AREDS

<table>
<thead>
<tr>
<th>Feature</th>
<th>AREDS2 Temporal Drift</th>
<th>Contemporaneous Regrades, N = 1335</th>
<th>Historical AREDS Temporal Drift (AREDS Report 6 and 17),‡ N = 119</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regrade Year 4 Compared to BL, N = 88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased pigment within grid</td>
<td>69</td>
<td>0.63 (0.04)</td>
<td>70</td>
</tr>
<tr>
<td>Hypopigmentation within grid</td>
<td>60</td>
<td>0.56 (0.03)</td>
<td>69</td>
</tr>
<tr>
<td>Drusen area within grid</td>
<td>60</td>
<td>0.71 (0.02)</td>
<td>69</td>
</tr>
<tr>
<td>Presence of geographic atrophy</td>
<td>88</td>
<td>0.85 (0.03)</td>
<td>95</td>
</tr>
<tr>
<td>Presence of neovascular AMD*</td>
<td>90</td>
<td>0.79 (0.03)</td>
<td>94</td>
</tr>
<tr>
<td>AREDS severity scale (two-step agreement)</td>
<td>92</td>
<td>0.73 (0.02)</td>
<td>96</td>
</tr>
</tbody>
</table>

* Component features of neovascular AMD graded in AREDS; cannot be compared with a summary grade in AREDS2.
temporal drift on the two-step agreement in the AREDS severity scale (91.2% vs. 95.6%), an exact test of binomial proportions indicated no difference between AREDS2 and AREDS or in the grading of the component variables of the scale. When masked contemporaneous regrading quality control statistics were compared to the temporal drift samples, the same values for reproducibility were found. Geographic atrophy area measured by planimetry had a mean difference of 0.02 mm$^2$ with Bland–Altman limits of $+1.8$ and $-1.76$ mm$^2$ (Fig. 2). No systematic bias was identified.

**DISCUSSION**

The grading of features of AMD from optimized digital color fundus photographs in AREDS2 has reproducibility equivalent to historical grading from color slides in AREDS. This indicates continuity of AMD lesion classification from AREDS to AREDS2, which is important for comparison between data sets. A secondary outcome of AREDS2 is to evaluate the utility of the AREDS AMD severity scale for disease staging and possibly to predict outcomes. Because the severity scale was developed using color film images, there was question whether the macular lesion characteristics graded with digital photography were sufficient to reasonably justify use of the same grading steps. The agreement within two steps on the AREDS severity scale between the AREDS and AREDS2 contemporaneous sample is 93.6% and 95.6%, respectively.

Figure 2: Bland–Altman plot for reproducibility of area of geographic atrophy ($n = 288$ eyes). Area measured by two graders using planimetry had a mean difference of 0.02 mm$^2$ and Bland–Altman limits of $+1.8$ and $-1.76$ mm$^2$.

Our results indicate that overall there are no gross differences (Table 2). The grading of hypopigmentation from color film images, there was question whether the area of geographic atrophy measured by planimetry had a mean difference of 0.02 mm$^2$ with Bland–Altman limits of $+1.8$ and $-1.76$ mm$^2$ (Fig. 2). No systematic bias was identified.

Advances in technology may make automated grading from color photographs feasible, which is desirable not only to decrease costs but also because machine grading usually has higher reproducibility than human grading. In addition, there has been developing interest in use of automated lesion detection has been employed with some success to measure drusen extent from fundus photographs in eyes with AMD. Advances in technology may make automated grading from color photographs feasible, which is desirable not only to decrease costs but also because machine grading usually has higher reproducibility than human grading. In addition, there has been developing interest in use of
technology alternative to color photographs for the measurement and classification of drusen, such as spectral/Fourier domain optical coherence tomography (OCT). Fundus autofluorescence patterns may be particularly useful for the semi-automated measurement of area of GA, although measurements are, again, sometimes discrepant with color photographs. Autofluorescence imaging is another ancillary study within AREDS2 that should help clarify its predictive value. Advantages to the classification of AMD from color photographs, apart from its validation in epidemiologic studies and clinical trials, include the fact that it does not require technology beyond common clinical practice (such as a scanning laser ophthalmoscope for autofluorescence or a spectral domain OCT) and that it is the image type most similar to the view of the physician examining a patient. This makes extrapolation from clinical trials to clinical practice somewhat easier. As potential therapeutics target earlier stages of AMD, a classification system based upon outcomes validated with a large prospective cohort is increasingly important.

Acknowledgments
Supported by National Eye Institute Grant HHS-N-260-2005-00007-C (TEC).
Disclosure: R.P. Danis; None; A. Domalpally; None; E.Y. Chew; None; T.E. Clemons; None; J. Armstrong; None; J.P. SanGiovanni; None; F.L. Ferris III; None

References
23. Iwama D, Tsujikawa A, Ojima Y, et al. Relationship between retinal sensitivity and morphologic changes in eyes with...


