Drusen Analysis in a Human-Machine Synergistic Framework

R. Theodore Smith, MD, PhD; Mahsa A. Sohrab, MD; Nicole M. Pumariega, MS; Kanika Mathur, BA; Raymond Haans, MD; Anna Blonska, MD; Karl Uy, MD; Dominiek Despriet, MD, PhD; Caroline Klaver, MD, PhD

Objectives: To demonstrate how human-machine intelligence can be integrated for efficient image analysis of drusen in age-related macular degeneration and to validate the method in 2 large, independently graded, population-based data sets.

Methods: We studied 358 manually graded color slides from the Netherlands Genetic Isolate Study. All slides were digitized and analyzed with a user-interactive drusen detection algorithm for the presence and quantity of small, intermediate, and large drusen. A graphic user interface was used to preprocess the images, choose a region of interest, select appropriate corrective filters for images with photographic artifacts or prominent choroidal pattern, and perform drusen segmentation. Weighted statistics were used to analyze the initial concordance between human graders and the drusen detection algorithm; discordant grades from 177 left-eye slides were subjected to exhaustive analysis of causes of disagreement and adjudication.

To validate our method further, we analyzed a second data set from our Columbia Macular Genetics Study.

Results: The graphical user interface decreased the time required to process images in commercial software by 60.0%. After eliminating borderline size disagreements and applying corrective filters for photographic artifacts and choroidal pattern, the weighted \( \kappa \) values were 0.61, 0.62, and 0.76 for small, intermediate, and large drusen, respectively. Our second data set demonstrated a similarly high concordance.

Conclusions: Drusen identification performed by our user-interactive method presented fair to good agreement with human graders after filters for common sources of error were applied. This approach exploits a synergistic relationship between the intelligent user and machine computational power, enabling fast and accurate quantitative retinal image analysis.


A ge-related macular degeneration (AMD) is the most common cause of legal blindness in developed countries. Although the biological basis of the process is still unknown, development of soft drusen is a hallmark of AMD. The current standards for grading fundus photographs in AMD involve the manual estimation of drusen loads and locations. Although these methods have demonstrated important relationships between drusen area and disease progression, efficient digital tools to detect and quantify drusen have been slow to develop and even slower to be adopted.

The fundamental difficulty in detecting drusen is that they tend to appear as low-contrast lesions against a variable background. Our approach used a detailed mathematical model based on the geometry of fundus reflectance reconstructed individually for each image to correct macular background and illumination variability. This method permitted the capture of even low-contrast lesions by uniform thresholds and yielded a method that has been validated and used to quantify the relationship among drusen, autofluorescence, and disease progression. However, speed was still limited by preprocessing steps such as cropping and color balancing, moving between commercial software at various stages of implementation, and image postprocessing in the case of certain confounding structures.

Improving the efficiency of our system required 2 steps; the first involved implementing all algorithms in a graphical user interface (GUI) written and compiled in MATLAB R2007 (The MathWorks Inc, Natick, Massachusetts) as a freestanding executable. The second step

Author Affiliations:
Department of Ophthalmology, Harkness Eye Institute (Drs Smith, Sohrab, Haans, Blonska, and Uy and Mss Pumariega and Mathur), and Department of Biomedical Engineering, Fu Foundation School of Engineering and Applied Science (Dr Smith and Ms Pumariega), Columbia University, New York, New York; and Departments of Epidemiology (Dr Despriet) and Biostatistics and Ophthalmology (Dr Klaver), Erasmus Medical Center, Rotterdam, the Netherlands.
required the conversion from the automated system to a user-interactive system that synergistically combined the best of human expert knowledge and our established drusen detection algorithm (DDA) to rapidly achieve a superior result to that attainable by either by itself. We found that minimal human input to indicate drusen-containing areas was immediately leveraged by the algorithm into an accurate segmentation for which postprocessing was largely eliminated.

For a rigorous evaluation of the system, we chose a large, fairly difficult, independently graded series of images, some of which contained significant image artifacts and/or were of variable quality, with relatively mild disease. The artifacts and limited disease provided ample opportunity for false-positive errors commonly associated with digital systems based on gray scale. Thus, the challenge in this case was somewhat different than that for images with more extensive disease. In the latter case, the digital drusen grading system provides an accuracy of within 5% with respect to expert drawings from stereo fundus photographs. Many of the images used for this analysis had total drusen areas of less than 5%, making identification of individual lesions and rejection of false-positive results paramount. Initial counts of categorical disagreements between human graders and our user-interactive system were followed by an exhaustive analysis of the causes of these disagreements. After eliminating borderline size disagreements, 2 novel filters were constructed to correct 2 specific sources of error: photographic artifacts and choroidal pattern. Final categorical agreements were assessed by standard weighted \( \kappa \) statistics.

We further validated our drusen segmentation methods using a second data set of images from patients with soft drusen and with more extensive disease. Human graders and the DDA both focused on the central 6000-\( \mu \)m-diameter circle and expressed drusen area measurements as percentages (0%-9.9%, 10.0%-24.9%, 25.0%-49.9%, and 50.0%-100%) of the center, inner, and outer rings of this region. Concordance between human graders and the DDA was assessed by weighted \( \kappa \) statistics. Analysis of disagreements showed most to be false-positive results from more highly reflectant lesions, such as geographic atrophy (GA). The user-interactive method was then used to eliminate such lesions, and weighted \( \kappa \) values were recalculated.

**METHODS**

**STUDY 1**

**Study Participants**

We used data from the Erasmus Rucphen Family, a genetically isolated population in the southwest of the Netherlands. This population was founded by a small number of people in the middle 18th century and has remained in isolation. The subset used for our study, referred to herein as the Netherlands study, consisted of patients with early signs of AMD who were older than 65 years. The research adhered to the tenets of the Declaration of Helsinki and was approved by the Medical Ethics Committee of the Erasmus Medical Center in Rotterdam. A total of 389 fundus photographs were available from a total of 194 patients, most with mild age-related maculopathy.

**Image Analysis**

The 389 fundus photographs of good to fair quality were scanned and digitized, and image analysis was performed by our DDA on a desktop personal computer in 2 implementations. In the first, image preprocessing and user-interactive selection of region of interest were performed in Photoshop CS2, version 9.0.2 (Adobe Systems Inc, San Jose, California) and then exported to MATLAB for drusen segmentation with the DDA (Figure 1). In the second, all steps were performed in a GUI written in MATLAB (Figure 2). A total of 177 left-eye images processed by the GUI were subsequently analyzed exhaustively and adjudicated for causes of the discrepancies between the DDA and previously assigned grades. Drusen sizes were defined as small (<65 \( \mu \)m), intermediate (65-125 \( \mu \)m), and large (>125 \( \mu \)m). The number of drusen of each size was estimated by the human graders and counted by the DDA, and each image was placed in 1 of 4 categories for each drusen size: 0 (no drusen), 1 (1-10 drusen), 2 (11-20 drusen), and 3 (>20 drusen). Thus, there were 3 grades for each of the 177 eyes, yielding a total of 531 grades by the human graders and 531 grades by the DDA. The individual steps are next described as they were implemented in the GUI.

**Graphical User Interface**

Preprocessing by the user is accomplished in 2 clicks: 1 on the foveal center and 1 on a peripapillary point. The image is then automatically cropped to the Wisconsin grading template, a central 6000-\( \mu \)m-diameter circle (the central, middle, and outer...
highly reflectant structures, such as nerve fiber layer bundles at the arcades, peripapillary atrophy, pathologic myopia, retinal pigment epithelium hypopigmentation, exudates, and scars, are more frequently mistaken for drusen by the automated method than by an expert grader, requiring postprocessing steps. Consequently, we developed the more efficient user-interactive method, in which the user initially selects areas of interest from drusen images, excluding unwanted reflectant structures a priori. The algorithm then computes the background model and final drusen segmentation of the macula, recognizing the absence of drusen beyond the region of interest. Unwanted reflectant structures are treated as background and the calculation proceeds as usual, producing a more accurate leveled image for global thresholding (Figures 1 and 2). Thus, a few moments of intelligent human intervention in selecting a region of interest are leveraged into an accurate segmentation of multiple lesions. Dark lesions, such as hyperpigmentation, will automatically be excluded from the drusen segmentation by the global threshold.

**Drusen Detection Algorithm**

Our approach uses a mathematical model based on the geometry of fundus reflectance to correct macular background and illumination variability. The technique uses quadratic polynomials in several zones with cubic spline interpolation in blending regions between the zones. As we have previously demonstrated, this model is capable of approximating the global macular background of a normal photograph or autofluorescence image to permit its reconstruction and leveling. After background leveling is performed, the algorithm segments the drusen or autofluorescence abnormalities.

**Digital Filters for Common Sources of Error**

Photographic artifacts, such as camera lens bright spots and dust spots, are abnormally bright in the green channel used for segmentation of drusen and can cause false-positive results. However, these scattering artifacts are even more prominent than drusen in the blue channel, and therefore this channel provides automatic identification of artifacts separate from drusen (Figure 3). In such cases, we filtered the blue channel (gaussian filter, 75 µm half maximum) and applied the 2-threshold, 3-class Otsu method to the filtered image. The Otsu method, in the case of 2 thresholds \( k \) and \( m \), divides the image into 3 classes, \( C_k, C_m, \) and \( C_l \), defined by pixels with levels \([1, \ldots, k]\), \([k+1, \ldots, m]\), and \([m+1, \ldots, L]\), respectively. The criterion for class separability is the total between-class variance. In such cases, the class \( C_l \) selects the areas of excessive blue reflectance (Figure 3). When the green channel has later been leveled for uniform thresholding to locate drusen, the threshold is raised by 4 gray levels for pixels in class \( C_l \) to avoid application of the filter to small bright areas, which potentially represent drusen, the filter is selectively applied to only lesions with an area greater than 0.07 mm\(^2\). The application of this filter requires a single-user choice.

Another source of image variability is retinal pigment epithelium attenuation with increased visibility of the choroidal

**User-Interactive Tool**

Highly reflectant structures, such as nerve fiber layer bundles at the arcades, peripapillary atrophy, pathologic myopia, retinal pigment epithelium hypopigmentation, exudates, and scars, are more frequently mistaken for drusen by the automated method than by an expert grader, requiring postprocessing steps. Consequently, we developed the more efficient user-interactive method, in which the user initially selects areas of interest from drusen images, excluding unwanted reflectant structures a priori. The algorithm then computes the background model and final drusen segmentation of the macula, recognizing the absence of drusen beyond the region of interest. Unwanted reflectant structures are treated as background and the calculation proceeds as usual, producing a more accurate leveled image for global thresholding (Figure 2). The scale is thus predicated on a macula disc distance of 3000 µm, which has been established as a reliable standard of reference in the grading of fundus photographs.

---

**Subfields of the diameters of 1, 3, and 6 mm**, which is followed by automatic initial color balancing (Figure 2). The scale is thus predicated on a macula disc distance of 3000 µm, which has been established as a reliable standard of reference in the grading of fundus photographs.

---

**Figure 2.** Screenshot from the graphical user interface. A, The original image (245805) from the Netherlands study has already been centered and cropped to the 6000-µm region by the 2-click autocrop tool. The image is only of fair quality, and some drusen are poorly defined. B, The user has indicated areas of scattered drusen by broad translucent brushstrokes. C, Outlines of drusen identified by the algorithm in the user-identified regions, within the 6000-µm circle. D, The image is slightly enhanced for visualization of details. Drusen are displayed in green. For complete numerical results, see Table 1.
pattern. This pattern potentially introduces false-positive results in areas of prominent choroidal vasculature, reducing the accuracy of drusen segmentation. However, the vasculature is even more prominent in the red channel, offering a mechanism to reduce choroidal vasculature artifacts (Figure 4). We filtered the red channel in images with a prominent choroidal pattern (gaussian filter, 75 µm half maximum) and applied the 2-threshold Otsu method to the filtered image. In such cases, the class C selects choroidal vessels and C choroidal pigment in the given image (Figure 4). After leveling the green channel for uniform thresholding to locate drusen, the threshold is raised by 4 gray levels for pixels in class C. The application of this filter also requires a single-user choice.

The entire image series was first processed without either filter. The optimum gaussian filters, lesion size (in the case of photographic artifacts), and threshold dynamics were then determined empirically by testing on several images of each type in which clear errors were generated. During final analysis of causes of disagreement between results from human graders and those from DDA, the application of these filters was then performed in a user-interactive second pass through the results in which the human observer determined the presence of photographic artifacts and judged the prominence of choroidal circulation. The option to reprocess with either filter selected was left to the user's discretion.

STUDY 2

Study Participants

We used data from the Columbia Macular Genetics Study, a study of the genetic variations in macular degeneration that was approved by the institutional review board of New York Presbyterian Hospital. All patients were white and older than 60 years. To test the DDA on images with more extensive disease, good-quality digital color fundus images of 164 eyes (84 right eyes and 80 left eyes) were selected from 94 patients showing soft drusen with or without GA or drusenoid pigment epithelial detachment but without a clinical history or fundus features, such as hemorrhage, exudates, or subretinal fibrosis, to suggest choroidal neovascularization.

Image Analysis

For the 164 fundus photographs from the Columbia Macular Genetics Study, drusen segmentation was performed by the DDA as previously described for study 1. However, rather than drusen counting, the end points chosen were drusen areas in defined regions, which seemed more appropriate in these eyes with greater disease or confluent drusen. Precisely, the region studied was the central 6000-µm-diameter circle divided into the central, middle, and outer subfields of the diameters of 1, 3, and 6 mm defined by the Wisconsin grading template. All area measurements were stated as percentages of these circles. This macula-disc distance (3000 µm) is established as the constant of reference in clinical macular grading systems. Although this distance varies anatomically, it does not affect area measurements as percentages. Human graders performed the same percentage grading (center, inner, and outer subfields) on the raw images with independent estimation of the macular center. Both sets of results (DDA and human) were categorized (0%-9.9%, 10.0%-24.9%, 25.0%-49.9%, and 50.0%-100%) and analyzed by weighted \( \kappa \) statistics.

RESULTS

STUDY 1

The time required to process all 358 images with our algorithms in commercial software was 2928 minutes (mean [SD] of 491 [88] seconds per image). The total time for
analysis with the GUI was 1165 minutes (mean [SD] of 195 [34] seconds per image). We concluded that the excess time in the first formulation was due to the transfer of image data between Photoshop and MATLAB. In addition, we estimated that most of the time spent on analysis in the GUI related to input and output, cropping, and the user drawing, with only about 12 seconds per image expended on the vectorized version of the main algorithm itself.

Raw agreements between human graders and the DDA in the GUI formulation for 177 left-eye images numbered 301 and disagreements numbered 230. The weighted $\kappa$ values were 0.12, 0.21, and 0.27 for small, intermediate, and large drusen, respectively. Raw disagreements between human graders and the DDA in the GUI formulation for 177 left-eye images numbered 301 and disagreements numbered 230. The weighted $\kappa$ values were 0.12, 0.21, and 0.27 for small, intermediate, and large drusen, respectively.

Table 1. Sample Results for Grading and Adjudication of the Images From Study 1

<table>
<thead>
<tr>
<th>Grading Circle</th>
<th>Drusen Area, mm²</th>
<th>DDA Output Drusen Areas and Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small Intermediate Large</td>
<td></td>
</tr>
<tr>
<td>Inner</td>
<td>0.874</td>
<td>2 2 0</td>
</tr>
<tr>
<td>Middle</td>
<td>0.705</td>
<td>2 1 4</td>
</tr>
<tr>
<td>Outer</td>
<td>0.277</td>
<td>3 8 1</td>
</tr>
<tr>
<td>DDA</td>
<td>1 2</td>
<td>1</td>
</tr>
<tr>
<td>Human</td>
<td>1 1</td>
<td>1</td>
</tr>
</tbody>
</table>

Image 12440S

|                | Small Intermediate Large |
| Inner          | 0.020             | 0 0 2                             |
| Middle         | 0.070             | 0 2 3                             |
| Outer          | 0.425             | 9 14 18                           |
| DDA            | 1 2               | 3                                 |
| Human          | 0 1               | 1                                 |

Image 24580S

Abbreviation: DDA, drusen detection algorithm.

Of the 164 images, initial raw agreements between the DDA and the human graders were 343, and disagreements numbered 149. The corresponding weighted $\kappa$ values were 0.43, 0.49, and 0.49 for the center, inner, and outer circles, respectively. Of the 164 images, initial raw agreements between the DDA and the human graders were 343, and disagreements numbered 149. The corresponding weighted $\kappa$ values were 0.43, 0.49, and 0.49 for the center, inner, and outer circles, respectively. Of the 164 images, initial raw agreements between the DDA and the human graders were 343, and disagreements numbered 149. The corresponding weighted $\kappa$ values were 0.43, 0.49, and 0.49 for the center, inner, and outer circles, respectively. Analysis of the raw disagreements revealed that most were false-positive results from the DDA due to GA, hypopigmentation, peripapillary atrophy or hypopigmentation, or pigment epithelial detachment. Other sources were centration of the center circle or small percentage dis-
agreements causing categorical shifts. When the GUI was then used in user-interactive mode as described in study 1 to eliminate lesions obviously causing false-positive results, the agreements between the DDA and the human graders were 415, and disagreements numbered 77. The weighted χ values for the center, inner, and outer circles became 0.76, 0.71, and 0.76, respectively, consistent with differences to be expected between manual human grading. We then eliminated small percentage disagreements that caused categorical disagreements by adjusting the categorical boundaries by up to 3 percentage points, increasing the agreements between the DDA and the human graders to 448 (Figure 3) and decreasing disagreements to 44. The corresponding weighted χ values for the center, inner, and outer circles were 0.86, 0.84, and 0.80, respectively (Table 2).

The purpose of the user-interactive system is to merge synergistically the best of expert human knowledge with machine computational power. Although the result is not superior to what a human can do, the system certainly does what few humans want to do, which is to draw all the drusen in a macular image in accordance with standard grading techniques. The machine, however, performs this task in a few seconds. The result is also better than what the algorithm can produce without user input; the experienced grader can easily eliminate reflective areas of the macula that are clearly not drusen, such as marked retinal pigment epithelium hypopigmentation, GA, pathologic myopia, or peripapillary atrophy, whereas such reflective structures can confound a computer program. Dark lesions such as hyperpigmentation are excluded from the drusen segmentation automatically. The result, therefore, is a pure drusen segmentation, which means that any other important lesions, such as GA or hyperpigmentation, would therefore need to be identified independently by the grader. Whereas completely automated image analyses are a desirable future goal, the method we demonstrate here begins to realize the digital promise of fast and accurate quantitative retinal image analysis with intelligent human input. Similarly, this concept has been used with different mathematical algorithms to produce user-friendly solutions for segmentation of GA in autofluorescence scans of AMD25 and rings of hyperautofluorescence in Stargardt disease,26 as well as multimodal retinal image registration.27

We found the images used in study 1 appropriate for evaluating our method due to variable quality and limited disease, both of which provided ample opportunity for false-positive errors. The challenge for human graders and the computer model was to detect true lesions in this setting and estimate their extent. On analysis, however, we discovered that the largest group of disagreements occurred because of the inherent difficulty in comparing categorical grading of continuous variables, such as drusen size and number. For example, if the human grader found 12 small drusen but the DDA found 9 small drusen and 3 intermediate drusen, then initial χ statistics would assign disagreement in the categories of small and intermediate drusen. In reality, although the grader and the DDA found exactly the same lesions, the DDA measured 3 of them as 10 µm larger than the cutoff for small drusen. We concluded that these differences should not be counted as errors for the human or the DDA. Thus, if we found that differences in drusen size of less than 10 µm resolved a disagreement, we considered it as agreement for the next set of χ statistics. Similarly, disagreements arose when a nodular lesion was interpreted as confluent intermediate drusen by the human grader and a large soft drusen by the DDA, whereas in reality both identified the same lesion. With all these inessential disagreements recategorized as agreements, total disagreements were substantially reduced from 230 to 72.

Essential disagreements due to photographic artifacts and choroidal pattern visibility numbered 25. All these errors were false-positive findings detected by the DDA. Because it would be difficult in some cases for the user-interactive method to eliminate them, we sought to remove them based on automated spectral and morphologic criteria. To our knowledge, these filters are novel in the retinal imaging literature. For photographic artifacts, the blue channel filter exploits their excess scatter of blue light relative to green light (Figure 3). For images with a prominent choroidal pattern, the red channel filter exploits the excess red light reflectance of choroidal vessels (Figure 4). These filters were combined with the basic DDA as a single-user choice and resulted in reduction of false-positive results in all 25 cases. These filters can be applied only to color images because they rely on information in the red and blue channels not available in gray scale or red-free images.

Of the 47 disagreements remaining after choroidal and blue filter corrections, we discovered that many errors by the DDA occurred in detecting the smallest lesions. Hard, small drusen visible to a human observer were sometimes simply overlooked by the DDA. More errors occurred in analyzing images of lesser quality with scratches, poor focus, and media opacity. Remarkably, however, the DDA did well in most instances of fair image quality (Figures 1 and 2). Overall, final adjudication of these disagreements was evenly divided between human grader (25 correct) and DDA (22 correct). We therefore interpreted the χ statistics as evidence that both methods have fair to good agreement, that agreement was best for the largest and arguably most important lesions, and that neither method was “more correct” in the opinion of another human expert (R.T.S.).

Study 2 was performed using data from the Columbia Macular Genetics Study. This population was comparable in age to our first data set but had more advanced disease. The largest group of raw disagreements came from sources of false-positive findings, such as GA. In these cases, the user-interactive mode was then particularly effective in excluding such sources (Figure 6). As in study 1, we also found a large group of disagreements owing to the inherent difficulty in comparing categorical grades of a continuous variable, in this case, the drusen area. When such borderline disagreements (within 3 percentage points) were removed, there was good to very good concordance between the human graders and the user-interactive implementation of the DDA, dem-
Figure 6. The flowcharts of our analyses of disagreements between the drusen detection algorithm (DDA) and human graders for study 1 and study 2 onstrating the robustness of this method in more advanced cases (Figure 5).

In conclusion, the main strength of the user-interaction method for drusen quantification is the synergy between human expert knowledge and the automated algorithm. The DDA’s ease of use, ability to deal quickly and effectively with common sources of error, and automated, quantitative output broaden the potential scope of application to large clinical series. However, we also recognize realistically that deployment of the system at other institutions would require training, software integration, and further validation. An additional strength is that because the output is numerical and a spatially segmented image, the system can perform longitudinal comparisons of serial, accurately registered images on a lesion-by-lesion basis for detailed studies of phenomena, such as drusen remodeling.23,28 Likewise, multimodal image analysis with registered image pairs (eg, color photographs and autofluorescence scans) can be used spatially for lesion colocalization and disease insights.29 With further advances in retinal imaging, including spectral domain optical coherence tomography30-32 and hyperspectral imaging,33-35 effective and efficient techniques for multimodal image analysis are becoming increasingly relevant.

Submitted for Publication: April 1, 2010; final revision received April 8, 2010; accepted April 8, 2010.

Correspondence: R. Theodore Smith, MD, PhD, Columbia University Harkness Eye Institute, 160 Fort Washingt on Ave, Room 509C, New York, NY 10032 (rts1@columbia.edu).

Financial Disclosure: None reported.

Funding/Support: This study was supported by grants from The New York Community Trust (New York, New York), National Eye Institute grant RO1 EY015520 (Bethesda, Maryland), and unrestricted funds from Research to Prevent Blindness (New York, New York).

Role of the Sponsors: The funding organizations had no role in the design or conduct of this research.

Additional Contributions: Editorial assistance was provided by Jennifer L. Dalberth, BA, Department of Ophthalmology, Harkness Eye Institute, Columbia University, New York, New York.

REFERENCES

19. Uy K, Smith RT, Busuioc M, Klaver C, Despriet D. Automating drusen analysis with
user-friendly graphical interfaces with or without artifact correction [e-abstract 5715].
Invest Ophthalmol Vis Sci. 2006;A47.

---

**From the Archives of the Archives**

1. Symptoms, prognosis and treatment of diphtheritic conjunctivitis.
2. Serpent ulcer of the cornea.
3. Diagnosis and treatment of iritis.
4. (a) Mechanism of accommodation. (b) How is it affected by age?
5. A stone-cutter at work feels something strike his eye. A small fresh fracture is found on his chisel. How to determine if the fragment has entered the eye?

—Medical School of Harvard University. February 1899.