Variability and Repeatability Tests of ARMD Assessment Using the AD3RI Tool

André D Mora[1,2], José M Fonseca[1,2], Pedro M Vieira[3],

[1] CTS - Uninova, Campus da FCT-UNL, 2829-516 Caparica, Portugal
[2] Dep. of Electrical Engineering, FCT-UNL, Campus da FCT-UNL, 2829-516 Caparica, Portugal
3 Dep. of Physics, FCT-UNL, Campus da FCT-UNL, 2829-516 Caparica, Portugal
{atm, jmf}@uninova.pt, pmv@fct.unl.pt

Abstract. Age-related macular degeneration (ARMD) is a common disease in the elderly and is currently the main cause of blindness in developed countries. Drusen are one of its risks factors, which are visible in a retinal examination. Its quantitative analysis is important in the follow up of the ARMD. The authors have previously developed two tools for semi-automatic and automatic drusen quantification. In this paper five tests performed on these tools are presented and discussed in order to evaluate their accuracy, variability and repeatability. The statistical results show that the automatic tool is as accurate as the experts in Drusen quantification, with the advantage of being a reproducible method. The semi-automatic quantification method, which was used by the experts for Drusen quantification, proved statistically to produce a high intra-observer agreement.

Keywords: Drusen quantification; image processing; ARMD; Inter-observer agreement

1 Introduction

ARMD (Age-Related Macular Degeneration) is the most common cause for irreversible blindness in the developed countries [1, 2]. According to the World Health Organization it is the third cause of irreversible blindness. Drusen are considered as one of the ARMD most significant risk factors [3]. They are visible in retina images as yellow round deposits which can be located anywhere in the retina (see Figure 1). However, they have more severe consequences when located in the macula. Diagnosis and follow up of drusen are usually done with the use of fundus imaging and its evaluation is obtained following the guidelines of the protocol defined by the International ARM Epidemiological Study Group [4], which provides a qualitative evaluation. This qualitative analysis suffers from accuracy and, more important, from reproducibility [5].

Previously we have published two methodologies for drusen quantification, one semi-automatic (MD3RI) [6] and another automatic (AD3RI) [7]. The MD3RI was developed for computer assisted drawing of drusen contours, to be used by retinal

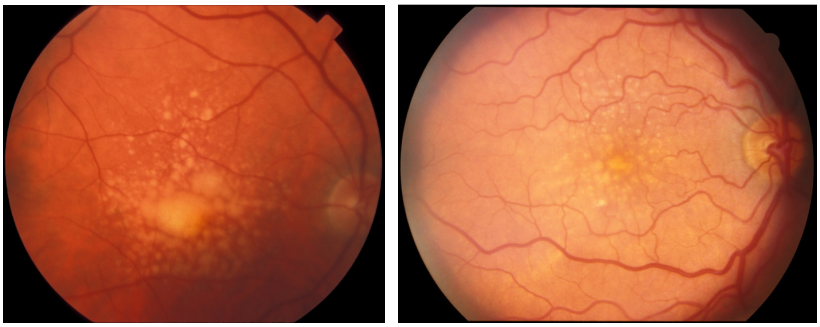

Fig. 1. Retinal images containing drusen. Drusen are the yellow brighter spots on center of the image, having the optic disk on their right side.

graders. The objective was to produce a reference dataset of retinal images graded by experts, which could be used as the ground truth of any automatic quantification technique. In our case, it was used by the AD3RI methodology.

On previous works the accuracy and repeatability of semi-automatic or automatic drusen quantification methods are usually restricted to one study with a small amount of graders and retinal images. In this article we present five different studies that evaluated the variability and repeatability performance of the MD3RI and the AD3RI for almost 100 images and 9 graders.

2 Materials and Methods

2.1 Materials

The retinal images which were used in these studies were collected in two collaborating research centers. The original dataset contained a total of 200 retinal images, from which 99 were acquired using a red-free filter and are in gray-scale, 24 were acquired in color-scale, and 77 were from the ECAM study [8] acquired in color scale. It was verified that several of these images were not gradable due to deficient image quality and were not selected for the following subsets.

A first subset of 7 images containing small, medium and large confluent drusen was selected for grading drusen using the MD3RI and assessing intra-observer agreement. A second subset of 22 images containing no drusen, small, medium and large confluent drusen were selected and analyzed by four ophthalmologists (OP1, OP2, OP3, OP4) and four trained graders (TG1, TG2, TG3, TG4) using the MD3RI tool, and also by AD3RI, in order to evaluate the tools' accuracy and the inter-observer agreement. A second subset containing 39 images was later prepared to reevaluate AD3RI's accuracy. This subset was analyzed by one ophthalmologist (OP1) and one trained grader (TG1).

Two other subsets were created to evaluate AD3RI's repeatability, using repeated images of the same eye. The first subset contained 9 pairs of images taken from the same eye in the same acquiring conditions. The second subset, contained images from

the ECAM study, where retinal images of patients suffering from cataracts were taken before and after cataract surgery and are expected to have approximately the same drusen areas. In this subset a total of 7 pairs of retinal images of at least acceptable quality were selected. These pairs of images were linearly registered using Matlab (MathWorks), in order to obtain images with equally resolution and ROI location.

In all the five studies, the analysis protocol adopted some of the Wisconsin Grading System recommendations [9], namely, the inner-macula was defined as the region of interest (circular region of 3000μm diameter centered in the Macula), and only drusen wider that 63μm were considered. The ROI was defined manually by one of the trained graders.

2.2 Manual Quantification Method

The Manual Drusen Deposits Detection in Retinal Images (MD3RI) tool [6] was used by graders to quantify drusen from retinal images using a semi-automatic and user-friendly method. In this tool before analyzing the image, its resolution is measured in comparison with the optic disk diameter in pixels and considering its diameter to be approximately 1500μm [4]. Then, the ROI is manually centered in the fovea and the image preprocessed. To mark one druse the grader selects with the mouse pointer its location and slides the mouse pointer to the left or right to increase or decrease the druse contour. The area inside the contour is added to the total affect area.

The algorithm that supports the semi-automatic grading method begins by converting the image to gray scale using the green color information, chosen due to the good contrast between drusen and retina surface of this color channel. It is followed by the gradient path labeling algorithm, a morphological and labeling algorithm that detects drusen considering image intensity as topographical information and computing the image two-dimensional gradient using the Sobel operator, which at pixels' level returns a path to the higher intensity pixels. The final result is an auxiliary image where brighter spots such as drusen are represented by their area of influence.

The drawing of the contours is obtained applying a threshold on the selected area of influence and where the threshold value is obtained by the mouse displacement. The result is a semi-automatic contour detection where the contour grows or shrinks controlled by the mouse (Figure 2).

2.3 Automatic Quantification Method

The Automatic Detection of Drusen Deposits on Retinal Images (AD3RI) tool [7], is an application specifically developed for this purpose, which requires little or no user intervention for the image analysis. The algorithms that support the AD3RI tool are the same as in the MD3RI, except the output from gradient path labeling algorithm that is used to obtain the location of the main drusen spots.

In order to segment and quantify drusen, and motivated by the intensity elevations shown on drusen areas on the tri-dimensional representation, these intensity spots are then modeled using modified Gaussian functions. This class of analytical functions is

adjustable in translation, rotation, amplitude, scaling and shape modifications, allowing it to represent the typical drusen shape. The Levenberg-Marquardt Least-Squares optimization algorithm [10] was used to fit the multiple elementary functions to the image adjusting the functions parameters in order to minimize the mean square error between the model and the image. This approach of modeling the drusen spots improved the drusen segmentation and characterization algorithms already published [11-16] by providing a shape consistent segmentation and by being more reproducible.

The contour of drusen spots and their area are calculated by thresholding the analytical model (Figure 2). The threshold value that produces more accurate contours was determined by comparing the false-positives and the false-negative pixels between the automated method and all the manually graded images.

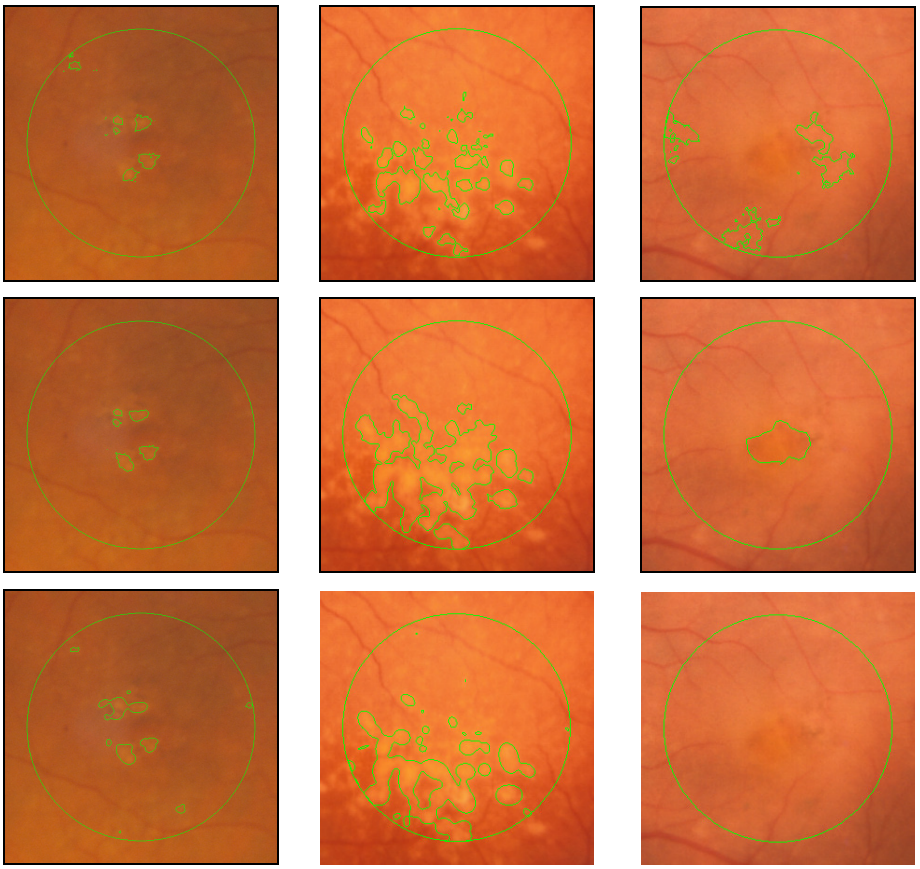

Fig. 2. Grading examples. This figure shows, from left to right, images graded by different experts using MD3RI (OP1 and TE3) and by AD3RI.

2.4 Statistical analysis

To validate and assess the accuracy of the manual (MD3RI) and the automated (AD3RI) methods, both overall and local agreement indicators were used. The overall indicators evaluate variability and agreement calculating the total or regional areas affected by the disease. However, this measurement only evaluates the agreement in the area value of the lesions, which can come from different locations. In the local agreement the comparison is made at pixels' level, where the agreement is measured in number of false positive, false negative, true positive and true negative pixels. This is a much accurate measure of agreement, although is more sensitive to noise and to a perfect alignment of sequences of images.

For estimating the overall agreement indicator it was calculated, for every image, the total area affected in the ROI and evaluated the *coefficient of variation* (CV) and the *intra-class correlation coefficient* (ICC) between the graders being analyzed (experts and AD3RI). The local agreement was obtained comparing pairs of images from different graders (experts and AD3RI) and computing the *sensitivity*, the *specificity* and the *kappa* coefficient as agreement indicators. To evaluate the manual method, four graders repeated the same measurement three times over a subset of seven images and the CV and ICC were calculated.

3 Results & Discussion

In this section, the results obtained in the five studies undertaken to estimate the variability and repeatability of the drusen quantification tools will be presented.

3.1 MD3RI Repeatability study

To assess the intra-observer agreement among the specialists, a group of four trained graders was invited to repeat for three times their markings in a selection of seven images (Table 1). The mean intra-observer variability observed was 10%, being in accordance with the expected value. Some of these results were influenced by the fact that some of the repetitions were made in different computers, which was the case for TG1 and TG3 who obtained the higher variability. The ICC showed a high agreement between the repeated measurements of all the specialists.

Table 1. Intra-observer variability among the four trained graders.

	CV / Image #							Summary		
Spec	*1*	*4*	*7*	*13*	*15*	*19*	*21*	*Mean(CV)*	*Std(CV)*	*ICC*
TG1	12%	10%	14%	16%	10%	4%	7%	10%	4%	0,994
TG2	22%	5%	8%	11%	5%	2%	11%	9%	6%	0,996
TG3	3%	5%	13%	35%	13%	7%	5%	12%	11%	0,988
TG4	9%	6%	6%	7%	16%	3%	6%	8%	4%	0,997
Mean(CV)	11%	6%	10%	17%	11%	4%	7%	10%	6%	0,965
Std(CV)	8%	2%	4%	12%	5%	2%	3%			

3.2 MD3RI and AD3RI Accuracy studies

In these studies the AD3RI accuracy was evaluated by comparison with a ground truth obtained from the average grading of a panel of eight experts in study I (Table 2.a) and a panel of two experts in study II (Table 2.b). The same experts were also evaluated among themselves, in order to produce an efficiency score for each of them and evaluate inter-observer variability using MD3RI.

Regarding the MD3RI inter-observer agreement (Table 2.a and Table 2.b), was obtained at the pixel level a *kappa* agreement of 0.60 in both studies and at area's level a CV of 22.5% and 18.7%, showing a reasonably high variability. These values are similar to the ones reported in [17] for the qualitative evaluation of drusen using the International Classification and Grading System of ARMD (ICGS), where the accuracy is also moderate (*kappa* = 0.27-0.69). We believe that even with similar *kappa* values, due to its qualitative nature, the ICGS is not as precise as the MD3RI quantitative measurement.

The AD3RI performance was classified as comparable to the experts, since it performed similarly to them in all indicators. At pixel level, *sensitivity* and *specificity* reveled that it has a slight tendency to overestimate drusen areas, since in the ACC-I study the specificity (0.96) was lower than the experts (0.97), while the sensitivity was higher (AD3RI = 0.68 and Avg Grader = 0.67). This observation was confirmed with the visual analysis of the graded images and is justified by a more detailed and systematic analysis that detects all spots and is not dependent on the user attention.

The areas' comparison (CV and ICC) showed that, although the CV obtained by the AD3RI (28.8%) was above the average among the experts (22.5%), the ICC (0.92) revealed a strong correlation between AD3RI and the experts. The images containing few drusen spots were the main cause for a higher CV, where an over or under estimation of drusen can cause a significant relative variation on the total area, increasing its CV.

Table 2. Summary of agreement indicators on the accuracy studies I (a) and II (b), and the same indicators for the repeatability studies I (c) and II (d).

		CV	ICC	*Sensitivity*	*Specificity*	*Kappa*
(a) ACC - I	**AD3RI**	28,8%	0,922	0,68	0,96	0,58
	OP1	28,3%	0,860	0,58	0,98	0,55
	OP2	23,6%	0,794	0,66	0,98	0,61
	OP3	15,0%	0,918	0,69	0,97	0,64
	OP4	19,6%	0,917	0,61	0,96	0,53
	TG1	16,3%	0,925	0,66	0,98	0,64
	TG2	14,2%	0,891	0,65	0,98	0,62
	TG3	21,9%	0,903	0,76	0,96	0,64
	TG4	41,2%	0,822	0,77	0,93	0,57
	Avg Grader	22,5%	0,879	0,67	0,97	0,60
(b) ACC - II	**AD3RI**	36,1%	0,935	0,76	0,98	0,61
	OP1	17,9%	0,912	0,63	0,98	0,61
	TG1	19,5%	0,912	0,75	0,99	0,60
	Avg Grader	18,7%	0,912	0,69	0,98	0,60
(c) REP - I	**AD3RI**	2,6%	0,998	0,78	0,97	0,76
(d) REP - II	**AD3RI**	3,0%	0,984	0,75	0,97	0,71

3.3 AD3RI Repeatability studies

For studying the repeatability of AD3RI, its results on pairs of images taken from the same eye with no significant changes in drusen areas were compared. The repeatability study was separated in two (Table 2.c and Table 2.d), due to the nature of their images subsets, in particular within the second study REP II that can eventually contain unwanted changes in drusen areas, because there is a 6 months gap between images. At areas' level the average variability was approximately 3%, which we considered a low value. The ICC obtained almost absolute agreement between the paired analyses, reinforcing the repeatability of the algorithm. The agreement at pixel level was also high. It obtained a sensitivity of 0.78 (REP I) and 0.75 (REP II), and a Kappa coefficient of 0.76 (REP I) and 0.71 (REP II), which are considered as *substantial* agreement [18].

4 Conclusions

The development of methods to quantitatively measure drusen in a reproducible and accurate procedure will certainly improve the quality of the follow up of this disease and potentiate epidemiologic studies and clinical trials. These studies, that collect thousands of images throughout several years, must be graded using a reproducible method to allow comparison during all the study period. Currently, this is manually done by trained experts with a fastidious procedure, lacking accuracy and reproducibility.

This article presented five studies in which the variability and accuracy of a semi-automatic and an automatic quantification of drusen methodologies were assessed. Since there is no standard assessment technique to be applied in this type of studies, most of the published works use different performance indicators making comparison between studies inaccurate or even impossible. In our work, performance was assessed using several indicators allowing direct comparison with other studies. This comparison showed that the results produced by the AD3RI were similar or better than the others.

From the above, we considered that AD3RI demonstrated promising results. It compares positively with the panel of human experts and since is a determinist method it is not dependent on factors such as attention or accuracy.

Acknowledgments. The authors acknowledge the University of Aberdeen, Rudolfstiftung Hospital, and Hospital Santa Maria for supplying the retinal images used in this work and for their valuable feedback while testing the software and grading the retinal images.

References

1. AMD Alliance International: What's Happening in AMD? An answer based on presentations at AAO's "Retina 2006: Emerging New Concepts" - Internal Report, www.amdalliance.org/documents/FINAL_approved_report_from_AAO_Nov06.pdf, (2006).

2. Cook, H.L., Patel, P.J., Tufail, A.: Age-related macular degeneration: diagnosis and management. Br Med Bull. 85, 127-149 (2008).
3. Hageman, G.S., Luthert, P.J., Victor Chong, N.H., Johnson, L.V., Anderson, D.H., Mullins, R.F.: An integrated hypothesis that considers drusen as biomarkers of immune-mediated processes at the RPE-Bruch's membrane interface in aging and age-related macular degeneration. Prog Retin Eye Res. 20, 705-732 (2001).
4. Bird, A.C., Bressler, N.M., Bressler, S.B., Chisholm, I.H., Coscas, G., Davis, M.D., de Jong, P.T., Klaver, C.C., Klein, B.E., Klein, R., et al.: An international classification and grading system for age-related maculopathy and age-related macular degeneration. The International ARM Epidemiological Study Group. Surv Ophthalmol. 39, 367-374 (1995).
5. Sparrow, J.M.L., Dickinson, A.J., Duke, A.M.: The Wisconsin Age-related Macular Degeneration grading system: Performance in an independent centre. Ophthalmic Epidemiology. 4, 49-55 (1997).
6. Mora, A., Vieira, P., Fonseca, J.: MD3RI a Tool for Computer-Aided Drusens Contour Drawing. Fourth IASTED International Conference on Biomedical Engineering - BIOMED2006. ACTA Press, Innsbruck, Austria (2006).
7. Mora, A.D., Vieira, P.M., Manivannan, A., Fonseca, J.M.: Automated drusen detection in retinal images using analytical modelling algorithms. Biomedical Engineering Online. 10, (2011).
8. Brunner, S., Krebs, I., Stolba, U., Falkner, C.I., Binder, S., Bauer, P.: Cataract Surgery in Nonexsudative Age-Related Macular -First Results of a Prospective, Randomized, Multicenter Trial (ECAM-1). Association for Research in Vision and Ophthalmology. p. poster 195/B169 (2005).
9. Klein, R., Davis, M.D., Magli, Y.L., Segal, P., Klein, B.E., Hubbard, L.: The Wisconsin age-related maculopathy grading system. Ophthalmology. 98, 1128-1134 (1991).
10. Marquardt, D.W.: An algorithm for least-squares estimation of non-linear parameters. Journal of the Society for Industrial and Applied Mathematics. 11, 431-441 (1963).
11. Sebag, M., Peli, E., Lahav, M.: Image analysis of changes in drusen area. Acta Ophtalmologica. 69, 603-610 (1991).
12. Phillips, R., Forrester, J., Sharp, P.: Automated detection and quantification of retinal exudates. Graefes Arch Clin Exp Ophthalmol. 231, 90-94 (1993).
13. Morgan, W.H., Cooper, R.L., Constable, I.J., Eikelboom, R.H.: Automated extraction and quantification of macular drusen from fundal photographs. Australian and New Zealand Journal of Ophthalmology. 22, 7-12 (1994).
14. Shin, D.S., Javornik, N.B., Berger, J.W.: Computer-assisted, interactive fundus image processing for macular drusen quantitation. Ophthalmology. 106, 1119-1125 (1999).
15. Soliz, P., Wilson, M.P., Nemeth, S.C., Nguyen, P.: Computer-aided methods for quantitative assessment of longitudinal changes in retinal images presenting with maculopathy. Medical Imaging 2002: Visualization, Image-Guided Procedures, and Display. pp. 159-170. SPIE, San Diego, CA, USA (2002).
16. Smith, R.T., Chan, J.K., Nagasaki, T., Ahmad, U.F., Barbazetto, I., Sparrow, J., Figueroa, M., Merriam, J.: Automated detection of macular drusen using geometric background leveling and threshold selection. Arch Ophthalmol. 123, 200-206 (2005).
17. Scholl, H.P.N., Peto, T., Dandekar, S., Bunce, C., Xing, W., Jenkins, S., Bird, A.C.: Inter- and intra-observer variability in grading lesions of age-related maculopathy and macular degeneration. Graefe's Archive for Clinical and Experimental Ophthalmology. 241, 39-47 (2003).
18. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics. 159-174 (1977).