Validating Retinal Fundus Image Analysis Algorithms: Issues and a Proposal

Emanuele Trucco,1 Alfredo Ruggeri,2 Thomas Karnowski,3 Luca Giancardo,4 Edward Chaum,5 Jean Pierre Hubschman,6 Bashir al-Diri,7 Carol Y. Cheung,8 Damon Wong,9 Michael Abràmoff,10–13 Gilbert Lim,14 Dinesh Kumar,15 Philippe Burlina,16 Neil M. Bressler,17 Herbert F. Jelinek,18 Fabrice Meriaudeau,19 Gwénéolé Quellec,20 Tom MacGillivray,21 and Bal Dhillon22

1VAMPIRE project, School of Computing, University of Dundee, Dundee, United Kingdom
2Department of Information Engineering, University of Padova, Padova, Italy
3Oak Ridge National Laboratory, Oak Ridge, Tennessee
4Istituto Italiano di Tecnologia, Genoa, Italy
5University of Tennessee Health Science Center, Memphis, Tennessee
6Jules Stein Eye Institute, Los Angeles, California
7REVIEW Group, University of Lincoln, Lincoln, United Kingdom
8Singapore Eye Research Institute, Singapore
9Institute for Infocomm Research, A*STAR, Singapore
10Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa City, Iowa
11Department of Electrical and Computer Engineering, University of Iowa, Iowa City, Iowa
12Department of Biomedical Engineering, University of Iowa, Iowa City, Iowa
13Department of Veterans Affairs, Iowa City VA Medical Center, Iowa City, Iowa
14School of Computing, National University of Singapore, Singapore
15RMIT University, Melbourne, Australia
16Applied Physics Laboratory, Johns Hopkins University, Laurel, Maryland
17Wilmer Eye Institute, Johns Hopkins University, Baltimore, Maryland
18Charles Sturt University, Albury, Australia
19University of Burgundy, Le2i, Le Creusot, France
20INSERM, U1101, SFR ScInBioS, Brest, France
21VAMPIRE project, Clinical Research Imaging Centre, University of Edinburgh, Edinburgh, United Kingdom
22VAMPIRE project, NHS Princess Alexandra Eye Pavilion, University of Edinburgh, Edinburgh, United Kingdom

Correspondence: Emanuele Trucco, VAMPIRE/CVIP, School of Computing, University of Dundee, Dundee, UK; emanuelttrucco@computing.dundee.ac.uk.
Submitted: June 6, 2012
Accepted: December 2, 2012

This paper concerns the validation of automatic retinal image analysis (ARIA) algorithms. For reasons of space and consistency, we concentrate on the validation of algorithms processing color fundus camera images, currently the largest section of the ARIA literature. We sketch the context (imaging instruments and target tasks) of ARIA validation, summarizing the main image analysis and validation techniques. We then present a list of recommendations focusing on the creation of large repositories of test data created by international consortia, easily accessible via moderated Web sites, including multicenter annotations by multiple experts, specific to clinical tasks, and capable of running submitted software automatically on the data stored, with clear and widely agreed-on performance criteria, to provide a fair comparison.

Keywords: validation, fundus image analysis, reference standards

This paper was born of discussions involving several international research groups on automatic retinal image analysis (ARIA; see Ref. 1 for a recent review) at IEEE Engineering in Medicine and Biology Conference (EMBC) 2011, Boston. There was unanimous recognition of two key facts.

1. Efforts in the community are shifting from generating algorithms to detect, localize, or measure retinal features and properties, validated with small sets of test data, to generating measurements of clinical and public health significance for clinicians, eye care providers, and biomedical scientists and researchers, requiring larger and “real-life” sets of test data.
2. The current methods for validating ARIA algorithms are neither uniform nor widely agreed upon. Issues include how to deal with the variability of expert annotations; the availability of public, large, structured “real-life” datasets for testing; and the accepted definition of reference (gold) standards in different applicative contexts.
It was felt that the discussion was sufficiently important to seek the opinion of further groups, and to publish the result of the discussion. This paper is that result.

Purely for reasons of space and consistency, we concentrate on the validation of algorithms processing fundus camera images, currently the largest section of the ARIA literature. ARIA algorithms are currently used for the following main purposes:

1. **Screening/monitoring**, for example of diabetic retinopathy (DR), glaucoma, or age-related macular degeneration. The goal is to identify images showing signs of a target condition in large sets (tens of thousands to millions). The images (patients) selected are referred for further investigation. False negatives (missed patients with disease) must be minimized; limited amounts of false positives (false alarms) are acceptable, but should also be considered as a factor to avoid unnecessary use of resources or side effects of unnecessary treatments. It has been shown that appropriate screening of DR is cost-effective.\(^2,3\) DR screening facilitates early detection of patients with mild stages of DR, and thus early intervention (e.g., by targeting a patient’s blood glucose and blood pressure levels, or by laser treatment) and ultimately the prevention of vision loss (outcome of interest). ARIA screening promises to eliminate inefficiencies within the current DR screening workflow by providing a faster, more cost-effective and accurate disease diagnosis. It will also eventually improve economics of eye disease management and cost savings for patients, public health care providers, and the government, and improve general eye health.

2. **Computer-assisted** diagnosis and risk stratification, for example diagnosis of retinopathy of prematurity (ROP)/Plus disease given measurements of tortuosity and width not detected readily by clinical examination alone. The purpose is to detect the presence or likelihood of a disease from specific signs. ARIA performance must be demonstrated to be more precise than diagnosis in the absence of computer assistance or generate richer data improving a clinician’s diagnosis. Unlike that with screening and monitoring, the outcome is not necessarily binary (refer/do not refer), and the diagnosis usually depends on a combination of factors beyond ARIA measurements (e.g., age, symptoms, clinical features).

3. **Biomarkers** aimed to determine whether the occurrence of measurable features in retinal images is linked significantly (in a statistical sense) with specific outcomes or conditions that impact treatment decisions, prognosis, or diagnosis, for example retinal vessel width with lacunar stroke and coronary heart disease. ARIA features may also be useful for testing effects of new drugs and therapies (e.g., changes in retinal vascular parameters with a novel drug for hypertension) or for discovery of novel pathways in the natural history of diseases (e.g., microvascular disease pathways in heart attacks). Links to cognitive performance and gene expression have also been reported.\(^4,5\)

In addition, three further areas would benefit from reliable ARIA systems:

4. **Longitudinal** studies, whereby ARIA provides a means to study quantitatively the evolution and characterization of a disease to assist treatment planning or gauge patient response to a treatment.

5. **Computer-aided** or image-guided surgery, an emerging application of ARIA algorithms,\(^5,9\) for example, vitreoretinal microsurgery, for which ARIA allows registration of intraoperative imagery with preoperative imagery for image-guided surgical interventions.\(^10\)

6. **Telehealth.** ARIA disease screening and monitoring could play a very important role here, for example in less developed countries where the incidence of diabetes is rising and screening made difficult by the lack of resources and clinicians.\(^11\) A World Health Organization consultation report reviewing principles for the care of diabetic retinopathy noted that retinal imaging systems and image analysis methods perform at least as well as human providers. Computer-assisted telehealth programs can therefore become a scalable method for providing expert care. The cost-effectiveness of telehealth screening of DR is accepted as an alternative to eye specialist exams in the United States, and is applied on a societal scale under various national health care programs, for example in the United Kingdom.\(^12,13\)

This paper combines input from 14 international research groups on the validation of ARIA algorithms. We first define “validation,” hence the scope of our discussion, and sketch the main techniques reported to date. We then give some compact background on the images and tasks for which ARIA algorithms are developed. We then discuss the issues making validation of ARIA algorithms difficult, and conclude with recommendations. We include in Appendix A a list of public datasets currently available for testing ARIA algorithms.

### Validation in Retinal Image Analysis

#### Validation: A Definition

For our purposes, validation indicates the process of showing that an algorithm performs correctly by comparing its output with a reference standard. In ARIA, target performance levels are normally represented by a reference (“gold”) standard defined by expert performance, for example, regions traced manually around landmarks or lesions, image quality level, or scores attached to DR screening images.

Validating ARIA algorithms implies therefore (1) selecting a data (image) sample representative for the specific validation purposes, (2) collecting reference standard annotations on the sample images, (3) running algorithms on the sample images, and (4) comparing the output with the reference standard by performing statistical analysis to assess the agreement of the algorithm’s output and reference standard, for example sensitivity, specificity, positive and negative predictive value, and area under ROC (receiver operating characteristic) curve.

#### Techniques

We identify four main types of validation processes in the ARIA literature, each involving its own reference standards. From the most general (defined in terms of clinical concepts) to the most detailed (defined in terms of image elements), these are as follows:

1. **Outcome** oriented, for example disease/no disease;
2. **Disease** grading, for example severity of DR, ROP Plus or Pre-plus;
3. **Feature** grading, for example tortuosity level of vessels or eye vasculature, width of retinal vessels;
4. **Image/pixel/measurement** oriented, for example locating microaneurysms, measuring area or perimeter of target regions, locating vessel bifurcations.
The key validation task is to assess quantitatively the agreement of automatic and manual measurements. How exactly to declare agreement or disagreement between sets of data is the object of discussion in the literature. Techniques vary with the type of the validation process (see above), but in general they are drawn from statistics. The ones reported most frequently in ARIA papers include graphs (e.g., scattergrams, Bland–Altman plots), integral indexes like correlation coefficients (e.g., Pearson), and statistical tests. ROC curves and associated coefficients (e.g., specificity, sensitivity, area under the curve), imported from signal processing, are frequently used to quantify the accuracy of detection and classification and are advocated strongly by some authors.\textsuperscript{14}

Two major issues in the generation of reference standard data are the variability of expert judgment and the need for generating annotations directly comparable to the algorithm's output. The former is addressed by having several experts annotate the same dataset; however, it is not ultimately clear how to obtain a single reference value from multiple ones (e.g., by averaging of values, discussion and consensus, interrater reliability metrics such as AC1 or Kappa, or just keeping histograms/distributions of multiple values). This is because some annotation tasks are not part of normal clinical practice and clinicians are not used to them, or do not see their relevance (e.g., tracing accurate contours around lesions).

A related question is where to set the “outcome” for validation. In a screening program, a refer/no refer decision with an associated uncertainty level seems the obvious choice; other cases are not so clear. A related point is that ARIA algorithms often consist of a pipeline of modules; and while testing the outcome of the algorithm is the main goal, it may be interesting and useful to test each individual module.

For these reasons some authors have begun to explore alternative validation paradigms. One is estimating simultaneously the quality of ARIA results and reference standards summarizing annotations from multiple experts, for example STAPLE for image segmentation\textsuperscript{15,16} and other tasks.\textsuperscript{17} Imperfect reference standards are the motivation behind weak learning methods in pattern recognition and machine learning\textsuperscript{18,19}; these methods are currently used only rarely in ARIA\textsuperscript{20} but might provide very useful modeling tools (see also subsection “Variation of Expert Judgment”). An interesting viewpoint is offered by Quellec et al.,\textsuperscript{21} who found, in brief, that a disagreement on DR severity between the algorithm and one expert would predict disagreement between the expert and a more experienced one. The ability to model expert disagreement would be a powerful tool for validation.

Images and Tasks

We discuss the characteristics of digital images and of the instruments generating them, as well as the clinical tasks for which images are ultimately created and analyzed. Both play a substantial role in validation. For reasons of space, we omit other imaging modalities like optical coherence tomography (OCT) and fluorescein angiography (FA) angiography.

Instruments and Images

The majority of ARIA systems reported to date consider single color images from digital fundus cameras, although acquiring two images per eye for both eyes (posterior pole, optic disc [OD] centered) is increasingly common in screening programs. Images are acquired with or without dilating the patient’s pupil through eye drops, respectively mydriatic and nonmydriatic. In the latter case, cameras require a high-power flash to allow enough light to enter the pupil and be reflected back by the retina. Concurrent illumination and imaging, performed through the same optical path, is the main engineering challenge faced by fundus cameras.

The main manufacturers of fundus cameras are currently Zeiss (Wetzlar, Germany), CenterVue (Padova, Italy), Topcon (Oakland, NJ), Nidek (Gamagori, Japan), Canon (Tokyo, Japan), and Kowa (Nagoya, Japan). Depending on the clinical application, various optical filters are available, the most common being green filters for red-free photography and barrier filters for fluorescein angiography. A typical camera has a field of view (FOV) of 30° to 50° with ×2.5 magnification. Some modifications are possible through zoom or auxiliary lenses, from 15°, which provides ×5 magnification, to 140° with a wide-angle lens, which reduces the image by half. The actual image resolution, that is, how many millimeters are captured by a pixel, depends on various factors, mainly the properties of the optics and the resolution of the complementary metal-oxide-semiconductor/charge-coupled device (CMOS/CCD) image sensor employed; approximately 3000 × 3000 pixels is now common. Higher-resolution sensors could be used, but the optics pose a limit to the resolution achievable, that is, to the size of the smallest distance that can be imaged in focus. As the optical quality of the eye prevents resolving features smaller than 20 μm, very high resolutions without adaptive optics may not be useful. We notice that the diffusion of current imaging equipment and technology is somewhat limited by infrastructure requirements; most existing are bulky, are expensive (approximately $25,000), and require special skills to operate. User-friendly, cost-effective handheld retinal cameras have been developed to tackle these issues, but work in this area is still limited.\textsuperscript{19,22}

Image Quality

Image quality depends on acquisition procedures, operators and their training, blur, occlusions (e.g., cataract, eyelashes), widespread lesions, artifacts introduced by the instrument, and conditions; for instance, fundus images of infants for ROP assessment tend to be lower quality than those of adults. Quality considerations are essential for proper validation, as image quality is at the basis of exclusion criteria applied in screening programs.

As quality influences the performance of ARIA systems, it seems advisable to divide a set of test images into quality classes, for example good, acceptable, and unusable. In the interest of applicability, quality classes should be defined by practitioners, using national standards for specific tasks whenever present. The UK diabetes screening program, for instance, defines three image quality grades (inadequate, minimum, achievable).\textsuperscript{23}

Automated systems for assessing retinal image quality exist,\textsuperscript{24–29} but quality is not always considered in the wider ARIA literature with respect to preparing datasets. Capturing quality definitions applied by experts in an ARIA algorithm is difficult, as clinicians learn from examples and practice; images considered viable for clinical analysis may not always produce good results with ARIA systems. ARIA quality estimation has often relied on heuristic measures (e.g., contrast level of vessels in specific retinal regions) and the use of approximate classes for quality classifications such as “adequate” or “inadequate.” Such labels are used in supervised learning systems, combined with pattern recognition and image-processing methods based on image features like sharpness of vessel regions, quantity of the blood vessel, and color characteristics like the shape of the color histogram.
Validating Retinal Fundus Image Analysis Algorithms

Tasks

To complete the background picture for our discussion of validation, we identify the main ARIA tasks addressed in the literature and the main measures used for validation. We do not aim to review ARIA techniques for each task; for these the reader is referred to the recent review by Abràmoff et al.1

Anatomical Landmarks: Location and Measurement.

In fundus images, the main landmarks are the OD, the macular region, and the vasculature. Most detection methods reported combine anatomical knowledge (relative locations of retinal landmarks, vasculature geometry) with image features (brightness levels, vessel density, orientation and thickness) to locate OD and macula, thus identifying a reference coordinate system for the location of retinal lesions.

Reference standard sets must specify therefore OD and macula locations. Measures used to compare reference standard annotations and ARIA estimates are normally the distance between annotated and estimated OD centers, or integral measures of contour agreement based on area (e.g., the Dice coefficient), or point-to-point distances along the contours.

The target clinical task determines the accuracy of the “location of the anatomical landmark.” For example, locating OD (with approximate size) and macula centers is generally sufficient to establish a retinal coordinate system; estimating the ellipticity of the OD requires accurate contour detection including parapapillary atrophy segmentation.30

Vasculature Segmentation and Measurement.

The retinal vasculature is an important indicator of various diseases; its changes underlie the development of other signs such as retinal lesions. Indicators of disease in the vasculature include width and tortuosity changes, venous beading, focal arterial retinal lesions are an important indicator of various diseases; and vessel proliferation for proliferative retinopathy. 45–48

Glaucoma.

Glaucoma is a disease of the optic nerve, resulting in a gradual and progressive loss of vision. The main indicator in ARIA is the cup-to-disc ratio (CDR), that is, the ratio of the size of the optic cup to that of the OD. Various imaging techniques are used in relation to glaucoma. Currently, fundus photography remains the only modality in which the characteristic colors of the retina and pathologies are preserved. Stereo imaging approaches exist in which a depth map of the OD region is computed from two retinal photographs acquired from displaced viewpoints.71–72

Table 1 summarizes ARIA work on DR detection, in particular microaneurysms and exudates, over the past 20 years. Many algorithms have been designed for the detection of DR in various types of retina images (color fundus, angiogram, red-free). The majority of these algorithms are validated on modest numbers of retina images annotated by experts, usually not available publicly.

Performance is evaluated in terms of sensitivity and specificity at either the lesion, region, or image level.

In general, results depend on methodology and dataset, stressing the need for large, internationally agreed-upon datasets for validation. For instance, the winning team of the Retinopathy Online Challenge (provided in the public domain at http://roc.healthcare.uiowa.edu/, Niemeijer et al.46), led by Quellec, achieved only 60% sensitivity with eight false positives per image; but earlier work by Niemeijer (2005)40 reported 100% sensitivity and 87% specificity on image-level screening, even though the per lesion sensitivity was only approximately 30%.

Glucoma.

Glucoma is a disease of the optic nerve, resulting in a gradual and progressive loss of vision. The main indicator in ARIA is the cup-to-disc ratio (CDR), that is, the ratio of the size of the optic cup to that of the OD. Various imaging techniques are used in relation to glucoma. Currently, fundus photography remains the only modality in which the characteristic colors of the retina and pathologies are preserved. Stereo imaging approaches exist in which a depth map of the OD region is computed from two retinal photographs acquired from displaced viewpoints.71–72

Table 1 summarizes ARIA work on DR detection, in particular microaneurysms and exudates, over the past 20 years. Many algorithms have been designed for the detection of DR in various types of retina images (color fundus, angiogram, red-free). The majority of these algorithms are validated on modest numbers of retina images annotated by experts, usually not available publicly. Performance is evaluated in terms of sensitivity and specificity at either the lesion, region, or image level.

In general, results depend on methodology and dataset, stressing the need for large, internationally agreed-upon datasets for validation. For instance, the winning team of the Retinopathy Online Challenge (provided in the public domain at http://roc.healthcare.uiowa.edu/, Niemeijer et al.46), led by Quellec, achieved only 60% sensitivity with eight false positives per image; but earlier work by Niemeijer (2005)40 reported 100% sensitivity and 87% specificity on image-level screening, even though the per lesion sensitivity was only approximately 30%.

Vascular Segmentation and Measurement.

The retinal vasculature is an important indicator of various diseases; its changes underlie the development of other signs such as retinal lesions. Indicators of disease in the vasculature include width and tortuosity changes, venous beading, focal arterial retinal lesions are an important indicator of various diseases; and vessel proliferation for proliferative retinopathy. 45–48

Glaucoma.

Glaucoma is a disease of the optic nerve, resulting in a gradual and progressive loss of vision. The main indicator in ARIA is the cup-to-disc ratio (CDR), that is, the ratio of the size of the optic cup to that of the OD. Various imaging techniques are used in relation to glucoma. Currently, fundus photography remains the only modality in which the characteristic colors of the retina and pathologies are preserved. Stereo imaging approaches exist in which a depth map of the OD region is computed from two retinal photographs acquired from displaced viewpoints.71–72

Table 1 summarizes ARIA work on DR detection, in particular microaneurysms and exudates, over the past 20 years. Many algorithms have been designed for the detection of DR in various types of retina images (color fundus, angiogram, red-free). The majority of these algorithms are validated on modest numbers of retina images annotated by experts, usually not available publicly. Performance is evaluated in terms of sensitivity and specificity at either the lesion, region, or image level.

In general, results depend on methodology and dataset, stressing the need for large, internationally agreed-upon datasets for validation. For instance, the winning team of the Retinopathy Online Challenge (provided in the public domain at http://roc.healthcare.uiowa.edu/, Niemeijer et al.46), led by Quellec, achieved only 60% sensitivity with eight false positives per image; but earlier work by Niemeijer (2005)40 reported 100% sensitivity and 87% specificity on image-level screening, even though the per lesion sensitivity was only approximately 30%.
Table 2 summarizes glaucoma-related ARIA reports (although numerous papers have described individual optic disc and cup detection, only papers leading to a CDR or glaucoma detection outcome have been included). ARGALI is a recent ARIA system.73–75 It uses active contour methods based on level sets to segment the cup and the disc and calculate the CDR. A similar approach was adopted by Joshi et al.76 Some authors have reported stereo techniques recovering depth information.71,72,77,78 An alternative approach is the use of machine learning to assign a predictive score for the risk of glaucoma directly to images. 79 AGLAIA (Automatic GLaucoma Diagnosis and Its Genetic Association Study through Medical Image InformAtics)67,80,81 computes 13 image cues for glaucoma assessment, and aims to integrate clinical and genome data in a holistic glaucoma analysis.

No publicly available datasets for validation are known to us, although some may be available on request from the authors under specific agreements, for example ORIGA-light.80–82

Retinopathy of Prematurity. Retinopathy of prematurity (ROP) is a disease involving abnormal development of retinal vasculature in premature infants, which can lead to retinal detachment and visual loss. The main indicators of ROP severity are the Plus disease, the stage when treatment is required, and the Pre-plus disease, a predictor of sight-threatening Plus disease development. Plus and Pre-plus diseases can be diagnosed by recognizing their specific signs, namely abnormal vascular dilation and tortuosity. The number of infants requiring ROP examinations has recently increased, thanks to improved survival of very low-birth-weight infants.
Therefore, computer-assisted solutions that can either increase the productivity of ophthalmologists’ screening or allow trained paramedical personnel to carry out part of the screening themselves will be of significant clinical benefit. Many ROP-related systems for computer-aided diagnosis have been reported recently, and Table 3 attempts a summary. Retinal Image multiScale Analysis (RISA) provides a semiautomatic tool for the labeling of the skeleton trees, followed by an automatic procedure to measure vessel width and tortuosity and from these derive Plus or Pre-plus diagnosis. 102 The Computer-Aided Image Analysis of the Retina (CAIAR) system semiautomatically identifies the retinal vessels, with provision for manual pixel editing if any vessel is inappropriately represented, and then automatically measures width and tortuosity of each identified vessel. 103 ROPtool semiautomatically traces retinal blood vessels; its reliability in measuring tortuosity and dilation of vessels was assessed in two distinct studies. 104,106 VesselMap (Imedos, Jena, Germany) is a commercial semiautomatic software program developed to analyze vessels in an adult retinal image. It performs the tracking of the main retinal vessels, providing information only about the vessel diameter, and was used also in ROP images. 105 No public annotated datasets on ROP seem to be available to date for ARIA system training and validation.

### Table 2. ARIA Literature on Glaucoma at a Glance

<table>
<thead>
<tr>
<th>Reference No.</th>
<th>Method</th>
<th>Public Dataset?</th>
<th>No. of Images</th>
<th>Image Type</th>
<th>Validation Level</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>Intra-image learning</td>
<td>ORIGA, 84</td>
<td>650</td>
<td>Nonstereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.081</td>
</tr>
<tr>
<td>85</td>
<td>Statistical model based</td>
<td>ORIGA, 84</td>
<td>650</td>
<td>Nonstereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.100</td>
</tr>
<tr>
<td>86</td>
<td>Depth-discontinuity model</td>
<td>No</td>
<td>138</td>
<td>Nonstereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.09</td>
</tr>
<tr>
<td>87</td>
<td>Sliding window, regression</td>
<td>ORIGA, 84</td>
<td>650</td>
<td>Nonstereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.091</td>
</tr>
<tr>
<td>88</td>
<td>AGLAIA framework</td>
<td>No</td>
<td>291</td>
<td>Nonstereo fundus</td>
<td>Glaucoma</td>
<td>AUC: 0.73</td>
</tr>
<tr>
<td>89</td>
<td>Active contour model, r-bends</td>
<td>No</td>
<td>138</td>
<td>Nonstereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.09</td>
</tr>
<tr>
<td>90</td>
<td>Higher-order spectra, texture features</td>
<td>No</td>
<td>60</td>
<td>Nonstereo fundus</td>
<td>Glaucoma</td>
<td>Detection: 61%</td>
</tr>
<tr>
<td>91</td>
<td>Active contour, depth reconstruction</td>
<td>No</td>
<td>80</td>
<td>Stereo fundus</td>
<td>CDR, glaucoma</td>
<td>Mean CDR error: 0.110, AUC: 0.90</td>
</tr>
<tr>
<td>92</td>
<td>Appearance-based analysis</td>
<td>No</td>
<td>575</td>
<td>Nonstereo fundus</td>
<td>Glaucoma</td>
<td>AUC: 0.88</td>
</tr>
<tr>
<td>93</td>
<td>Intensity profiling</td>
<td>No</td>
<td>50</td>
<td>Nonstereo fundus</td>
<td>CDR, Glaucoma</td>
<td>Mean CDR error: 0.14, AUC: 0.87</td>
</tr>
<tr>
<td>94</td>
<td>Regional information</td>
<td>No</td>
<td>170</td>
<td>Nonstereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.100</td>
</tr>
<tr>
<td>95</td>
<td>Level set</td>
<td>No</td>
<td>104</td>
<td>Nonstereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.089</td>
</tr>
<tr>
<td>96</td>
<td>Thresholding, 3-D reconstruction</td>
<td>No</td>
<td>80</td>
<td>Stereo fundus</td>
<td>Glaucoma</td>
<td>AUC: 0.83</td>
</tr>
<tr>
<td>97</td>
<td>Hybrid wavelet edges, kinking</td>
<td>No</td>
<td>27</td>
<td>Nonstereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.093</td>
</tr>
<tr>
<td>98</td>
<td>ARGALI</td>
<td>No</td>
<td>23</td>
<td>Nonstereo fundus</td>
<td>CDR</td>
<td>Correlation with ground truth: 0.89</td>
</tr>
<tr>
<td>99</td>
<td>Deformable model</td>
<td>No</td>
<td>25</td>
<td>Stereo fundus</td>
<td>CDR</td>
<td>Correlation with ground truth: 0.71</td>
</tr>
<tr>
<td>100</td>
<td>Pixel features, kNN</td>
<td>No</td>
<td>58</td>
<td>Stereo fundus</td>
<td>CDR</td>
<td>Correlation with ground truth: 0.94</td>
</tr>
<tr>
<td>101</td>
<td>Discriminatory analysis</td>
<td>No</td>
<td>NA</td>
<td>Nonstereo fundus</td>
<td>NA</td>
<td>Not provided</td>
</tr>
</tbody>
</table>

AUC, area under the ROC curve.

### Table 3. ARIA Literature on ROP at a Glance

<table>
<thead>
<tr>
<th>Reference No.</th>
<th>DR Task</th>
<th>Method</th>
<th>Public Dataset?</th>
<th>No. of Images</th>
<th>Image Type</th>
<th>Validation Level</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>102</td>
<td>ROP diagnosis</td>
<td>Semiautomated</td>
<td>No</td>
<td>20</td>
<td>Color fundus, CDR</td>
<td>Image feature</td>
<td>50%–100%</td>
<td>46%–93%</td>
</tr>
<tr>
<td>103</td>
<td>Vessel tortuosities and width measurement</td>
<td>Semiautomated</td>
<td>No</td>
<td>10</td>
<td>Color fundus, CDR</td>
<td>Image feature</td>
<td>0.49–0.67 (tortuosities), 0.42 (width)</td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>ROP diagnosis</td>
<td>Semiautomated</td>
<td>No</td>
<td>185</td>
<td>Color fundus</td>
<td>Image feature</td>
<td>97%</td>
<td>94%</td>
</tr>
<tr>
<td>105</td>
<td>Vessel width measurement</td>
<td>Semiautomated</td>
<td>No</td>
<td>20</td>
<td>Color fundus</td>
<td>Image feature</td>
<td>Ground truth correlation: 0.80</td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>Vessel width measurement</td>
<td>Semiautomated</td>
<td>No</td>
<td>30</td>
<td>Color fundus</td>
<td>Image feature</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

### Table 4. ARIA Literature on Glaucoma at a Glance

<table>
<thead>
<tr>
<th>Reference No.</th>
<th>DR Task</th>
<th>Method</th>
<th>Public Dataset?</th>
<th>No. of Images</th>
<th>Image Type</th>
<th>Validation Level</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>107</td>
<td>Glaucoma diagnosis</td>
<td>Semiautomated</td>
<td>No</td>
<td>150</td>
<td>Nonstereo fundus</td>
<td>Image feature</td>
<td>50%–100%</td>
<td>46%–93%</td>
</tr>
<tr>
<td>108</td>
<td>Vessel tortuosities and width measurement</td>
<td>Semiautomated</td>
<td>No</td>
<td>10</td>
<td>Color fundus, CDR</td>
<td>Image feature</td>
<td>0.49–0.67 (tortuosities), 0.42 (width)</td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>Glaucoma diagnosis</td>
<td>Semiautomated</td>
<td>No</td>
<td>185</td>
<td>Color fundus</td>
<td>Image feature</td>
<td>97%</td>
<td>94%</td>
</tr>
<tr>
<td>110</td>
<td>Vessel width measurement</td>
<td>Semiautomated</td>
<td>No</td>
<td>20</td>
<td>Color fundus</td>
<td>Image feature</td>
<td>Ground truth correlation: 0.80</td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>Vessel width measurement</td>
<td>Semiautomated</td>
<td>No</td>
<td>30</td>
<td>Color fundus</td>
<td>Image feature</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
reference standard used, it seems essential to characterize quantitatively the variations of reference standards (see “Techniques” subsection). Much work is still needed to achieve this goal, given the number and nature of the uncertainty sources involved.

Several practical challenges also exist. For instance, it is extremely time-consuming to collect sufficiently large, carefully constructed and annotated reference standard datasets. Time issues are exacerbated when it is necessary to annotate multiple images per eye, for instance with fluorescein angiography sequences or longitudinal studies. Another issue is that the point in the diagnostic process at which “outcome” should be set is not always clear, as factors beyond retinal measurements influence diagnosis. In automatic DR screening, for example, at least two possible end points exist that impact the four types of validation listed in “Techniques”: identifying the presence of DR (refer/no refer decision) and identifying specific lesions from which conclusions are then drawn by the specialist.107

### Variation of Expert Judgment

In engineering, reference standards for testing are normally objective measurements from instruments more accurate than the one being tested (see, e.g., camera calibration procedures); in medical image analysis, reference standards are instead built from statistical or explicit consensus among experts. Such judgments vary, in general, with experts (interobserver variations), and, to a lesser extent, over repeated judgments by the same expert (intraobserver variations).

These variations depend at least on experience, task, background, image quality (e.g., this often makes it difficult, even for the expert practitioner, to decide without ambiguity the presence of a druse), and interpretation of the annotation task, although a detailed protocol should minimize variations (see subsection “Annotation Protocols” below). Recent relevant work has been reported by Quellec et al. and Abramoff2,20 on the maximum meaningful performance achievable with automatic binary decision systems, given the characteristics of the reference standard obtained from clinicians. The study focused on the ROC area under the curve as the evaluation index. The authors ran tests with two ARIA systems for DR detection, 500 images, and the reference standard obtained by three experts. They concluded, interestingly, that meaningful performance measured against a single expert could not be improved (on the dataset used), whereas it could be improved significantly when compared against a committee of experts.

Hubschman et al.108 quantified the interobserver difference in an ischemia grading task with branch retinal vein occlusion. A single ultra-wide-field fluorescein angiogram image was segmented by four retina specialists into regions belonging to four different levels of retinal perfusion (normally perfused, partially perfused, nonperfused, insufficient quality). The cardiac cycle can also have an influence on vessel caliber measurements (see subsection “Physiological Short-Term Changes in Time”).

Intuitively, uncertainty can be reduced by increasing the number of experts and the number of annotations per expert. There is, however, little consensus on how to proceed statistically when such data are available, and expert numbers are usually small (two to five in most ARIA papers). Relevant techniques have been mentioned in “Techniques.”

### Annotation Protocols

Procedures used to take photographs represent another source of variability. For example, if the eye is not positioned in the same location, the vessels may be captured at slightly different angles, resulting possibly in different measurements.

As noted in “Techniques,” annotating specific image elements, like circling regions or tracing vessels on a computer screen, is a task that doctors do not normally perform in clinical practice. To maximize annotation accuracy as well as relevance to translation, it seems desirable to align validation tasks with those that doctors specialize in. Relevant research has been reported recently by Quellec et al.,20 who designed graphical user interfaces (GUIs) to automatically detect elements that catch the attention of clinicians in their daily clinical practice. This avoids requesting clinicians to explicitly annotate anatomical structures, a task they have not been trained for. Usage logs (zoom, magnifying glass, and the like) are used to train weakly supervised lesion detectors and validate their outputs.

Datasets used for high-profile clinical studies, such as AREDS (see Appendix A), were annotated at national grading centers using established protocols. As the intent was to evaluate the efficacy of drugs (for AREDS, by the National Institutes of Health/U.S. Department of Agriculture), it should be believed that the grading protocol was highly refined and scrutinized, and should be trusted to a higher extent than those of datasets developed at individual institutions for testing a specific algorithm. However, datasets not considering ARIA validation might not provide ARIA-relevant data, for example pixel-level delineations allowing ROC analysis.

### Physiological Short-Term Changes in Time

Taking photographs at random instants in the pulse cycle may result in unrecognized variations in the measurements of retinal vessel diameters, however technically sophisticated, both among subjects and over time in the same individual. A few studies have investigated this in detail,109–113 and some conclusions appear to be conflicting. It has been reported that the maximum variation at different points in the pulse cycle ranged from 4.3% to 4.8% for major retinal venules and from 3.1% to 3.9% for major retinal arterioles. Vascular diameter was smallest in early systole, increasing to a maximum level in early diastole and decreasing thenceforth. The arteriole diameter peaked slightly earlier. In another investigation of 10 volunteers, it was shown that a summary measure of the retinal venule diameters and the arteriolar diameters change at different points in the cardiac cycle. Across the cardiac cycle, the central retinal venular equivalent (CRVE) changed by 3.1% and the central retinal arteriolar equivalent (CRAE) changed by 4.3%. However, recent work by Kumar et al.113 has shown that there is no significant change in the average of the width of six large vessels in the region often chosen for arteriole-to-venule ratio (AVR) estimation. As quantifying vascular changes is a primary ARIA concern,112 there is a need for more extensive studies and test sets with repeated images of the same eye over time. Gating the eye fundus camera with the electrocardiogram has also been proposed. Moret et al.114 have measured changes to the vessel shape and size, also evidenced from the use of dynamic vessel analyzers (DVA) for disease diagnosis.115 DVA can provide temporal resolutions above 20 frames/s but comes with limits on spatial resolution and methodology, and the instrumentation is suitable only for highly specialized research facilities.

### Different Imaging Instruments

Because of the nature of retinal imaging, there is usually a high level of customization for each modality. Algorithms used for the segmentation of the optic nerve head in a confocal scanning technique may not be applicable to retinal images.
Validating Retinal Fundus Image Analysis Algorithms

without some level of customization. Even without the same class of machines, the variation of instruments can have a large effect on algorithm processes. In retinal fundus imaging, resolution variations can have a large effect on algorithm performance, and resolution requirements depend on the task at hand. For example, a lower resolution may be acceptable for optic nerve head segmentation, but not for estimating the tortuosity of retinal neovascularization. The selection of FOV and imaging field can also affect performance, particularly when one is relying on or assuming the visibility of retinal landmarks (OD, macula, arcades). The color calibration model used in the camera CCDs for digital capture can also affect algorithm performance, possibly causing images of the same retina taken by different devices to appear significantly different.

**Image Quality**

Image quality was discussed earlier. We reiterate that image quality depends on, among other factors, instrument characteristics, acquisition procedure, and target conditions, and that capturing quality definitions applied by experts for implementation in ARIA systems is difficult; in general, images considered viable for clinical analysis may not produce good results with ARIA systems.

**Datasets**

The key observation is that different datasets may lead to somewhat inconsistent performance assessments, as preparation protocols may differ. For instance, the best-known public retinal datasets are probably DRIVE\(^{116}\) and STARE.\(^{117}\) Soares et al.\(^{17}\) compared binary vessel masks from both datasets with masks obtained from their algorithm. Accuracy differences were noted between the two datasets, associated with different segmentation methods and differences in the extent and detail of the manual tracing provided by the experts. However, other groups have reported diverse methods of vessel segmentation and identification of proliferative retinopathy,\(^{48}\) and a comparison of these methods has not shown large differences in accuracy with images from the STARE dataset.\(^{118}\)

**Task, Previous Knowledge, Condition**

Expert judgment and annotations can be different for the same image and target measure depending on the task a clinician has in mind. For instance, when asked to determine the width of a retinal vessel from a fundus images, a doctor thinking surgically might try to keep at a distance from the vessel, hence overestimating width. If the task does affect annotations, it seems advisable to group reference annotations also by clinical task and to specify annotation protocols accordingly. At the moment, some public repositories group images by conditions, for example MESSIDOR for DR (see Appendix A).

**Patient Characterization: Metadata**

The growing volume of electronic images and datasets potentially available to researchers makes it challenging to mine the rich information embedded in the data. This obviously includes image properties, but also contextual information, that is, clinical metadata that may be relevant for the disease incidence and for organizing consistent validation datasets. Contextual patient characterization data include ethnicity, age, sex, medical data (hypertension, diabetes, heart diseases), lifestyle factors (e.g., smoking), systemic data (body mass index, cholesterol level, blood pressure, and so on), ocular data (refractive error, lens opacity, and so on), comorbid diseases, and in general all data normally considered in clinical studies.

These factors are rarely discussed for data used in validating ARIA algorithms, in part because their effect on retinal images, and consequently on ARIA results, is the object of current international investigation. Content-based approaches may have particular relevance in utilizing contextual information together with image analysis to assign risk profiles to patients in a screening environment. Quellec et al.\(^{119,120}\) have reported work on the inclusion of demographic and biological data in an image-based DR severity scale, concluding that this inclusion leads to significant improvement in classification performance. A larger clinical study aimed at confirming these findings is currently being performed on a dataset of 25,702 examination records from the Ophdial screening network (provided in the public domain at http://reseau-ophdiat.aphp.fr/index.html).\(^{3}\)

**Human in the Loop**

A final consideration is that some ARIA algorithms are semiautomatic and involve deliberately human intervention, a paradigm known in robotics as “human in the loop.” This complicates the objective evaluation of performance, as it seems to require decoupling human and automatic contributions. Specialized techniques apparently have not yet been considered in the ARIA literature.

**DISCUSSION AND RECOMMENDATIONS**

The ideal way to identify effective ARIA algorithms is to compare solutions proposed in the literature on an equal footing and with datasets recognized as meaningful by a representative cross section of the clinical community. This comparison is best achieved by the creation of common, accessible, and representative datasets including automated tools to run submitted algorithms on the data provided. Other areas of image-processing research have produced such repositories, for instance stereo and multiple-image analysis.\(^{121}\) An ARIA move in this direction is the recent Diabetic Retinopathy Online Challenge.\(^{30}\) Encouragingly, a variety of public datasets have appeared in recent years (see Appendix A). Most are still limited collections of images generated by individual sites, or contain limited annotations, or both. There is currently no coordinated consensus in the community on how to structure such datasets or what information to include, and this paper is meant to provoke thoughts toward that end.

So far, neither ARIA datasets nor epidemiological studies based on large populations have systematically taken into account the above issues. Furthermore, as images are often acquired independent of outcomes and participant characteristics, random variability may tend to underestimate the true associations found, for example between retinal vascular calibers and cardiovascular diseases. Future studies need to overcome these sources of variability before retinal features could be used as a more precise biomarker.

ARIA repositories of test data (not including alternative, promising validation paradigms like indirect methods) should, ideally,

- Be created collaboratively by consortia of international groups in order to achieve large data volumes and multiple annotators, to reduce opinion bias, to guarantee international visibility and credibility, and to ultimately generate useful results for clinicians.

- Be easily accessible, ideally via Web sites from which data could be downloaded following a suitable registration proce-
an image set could be used for multiple outcomes by providing tortuosity or width measurement in feature-oriented tasks. For example, refer/do not refer in screening tasks, levels of variability in expert judgment). Uncertainty levels declared in tabular form.

- Organized by outcome, which depends on the task at hand; for example, refer/do not refer in screening tasks, levels of tortuosity or width measurement in feature-oriented tasks. An image set could be used for multiple outcomes by providing annotations for some or all the points listed under ‘Techniques.”

- Include image annotations, providing the standard reference for comparison for the outcome stated (see previous point), preferably provided by as many clinicians as possible (ideally from different sites to eliminate possible opinion bias) to estimate interobserver variability, including arbitrated annotations; each expert should ideally annotate the dataset multiple times to estimate intraobserver variability (see also last paragraph of “Techniques” section for techniques managing variations in expert judgment). Uncertainty levels declared or agreed on by the annotators should also be considered. A surrogate of direct measurement yielding “true” gold standard could be the development of phantoms of the eye, including the retinal vasculature and pumps to simulate the blood flow, which could enable the calibration of ARIA vessel measurement methods. But given the current impossibility, with ARIA tasks, of obtaining ground truth measured by independent, highly accurate instruments, the only way to estimate “how good” an algorithm is to compare its output with expert judgment.

Creating such repositories for ARIA algorithms poses important challenges. The most obvious one is the sheer complexity of the task, as presented above. Acquiring images, generating the necessary annotations, and preparing the data for public use take time and impose significant costs. Governance and ethical issues, which vary internationally, may complicate further the release of clinical data for public research. A further point is the effort required for administration and maintenance, considering for instance that the obsolescence of imaging instruments limits the life of a dataset. It is arguable that retinal scans around 600 by 600 pixels are increasingly obsolete given the availability of much higher-resolution instruments.

Ultimately, outcome-oriented measures are required. It would be of little applicative interest to develop ever more accurate ARIA techniques if they could not be used to improve clinical outcomes. Outcome measures should be considered in a public health context, taking into account health economics, risks, and the impact of changes to services using automated algorithms. However, during the development of algorithms, it is helpful to compare the isolated performance of modules that will eventually become components in a larger system so that the most effective methods can be identified. This requires the provision of comparative datasets with reference standard measurements of features that may ultimately prove diagnostic. These features must be selected based on expert clinical advice.

Acknowledgments

Supported by the National Institutes of Health Grants R01 EY018853 and R01 EY019112 (MA), and the Department of Veterans Affairs I01 CX000119 (MA).

The authors thank the additional authors listed in Appendix B.

Disclosure: E. Trucco, None; A. Ruggeri, None; T. Karnowski, None; L. Giancardo, None; E. Chaum, None; J.P. Hubschman, None; B. al-Diri, None; C.Y. Cheung, None; D. Wong, None; M. Abramoff, None; G. Lim, None; D. Kumar, None; P. Burlina, None; N.M. Bressler, None; H.F. Jelinek, None; F. Meriaudeau, None; G. Quellec, None; T. MacGillivray, None; B. Dhillon, None

References

Validating Retinal Fundus Image Analysis Algorithms


Validating Retinal Fundus Image Analysis Algorithms


APPENDIX A
A LIST OF PUBLIC DATA REPOSITORIES FOR RETINAL IMAGE ANALYSIS

This appendix lists the public retinal datasets known to us. Unless otherwise stated, all datasets listed are easily reachable by a Google search. Most descriptions are excerpts from the Web sites listed.

Structured Analysis of the Retina (STARE; available in the public domain at http://www.ces.clemson.edu/~ahooover/stare/) is one of the earliest and most often cited test sets in the ARIA literature, created for validating OD location. It consists of 31 images of healthy retinas and 50 images of retinas with disease, acquired using a Topcon TRV-50 fundus camera at 35° FOV and subsequently digitized at 605 × 700 pixels in resolution, 24 bits per pixel (standard RGB). The nerve is visible in all 81 images, although partially visible in 14 as appearing on the image border. In 5 images the nerve is completely obscured by hemorrhaging.

Digital Retinal Images for Vessel Extraction (DRIVE; available in the public domain at http://www.isi.uu.nl/Research/Databases/DRIVE/) is another much cited test set; it was created to enable comparative studies on segmentation of blood vessels in retinal images. The photographs were obtained from a DR screening program in The Netherlands. The screening population consisted of 400 diabetic subjects between 25 and 90 years of age. Forty photographs were randomly selected, 33 without and 7 with DR signs. The images were acquired using a Canon CR5 nonmydriatic 3CCD camera with a 45° FOV. Each image was captured using 8 bits per color plane at 708 by 584 pixels. The FOV of each image is circular with a diameter of approximately 540 pixels.

Diabetic Retinopathy Database and Evaluation Protocol (DiabRetDB1; available in the public domain at http://www2.it.lut.fi/project/imageret/diabretdb1/) consists of 89 color fundus images. Eighty-four contain at least mild nonproliferative DR signs (microaneurysms) and 5 are considered normal, not containing DR signs according to all experts who participated in the evaluation. Images were captured using the same 50° FOV digital fundus camera with varying imaging settings. The data correspond to a good (not necessarily typical) practical situation, where images are comparable and can be used to evaluate the general performance of diagnostic methods. Four medical experts were asked to mark the areas related to the microaneurysms, hemorrhages, and hard and soft exudates. Ground truth confidence levels (<50%, ~>50%, ~100%), representing the certainty of the decision that a marked finding is correct, are included.

Méthodes d’Évaluation de Systèmes de Segmentation et d’Indexation Dédiées à l’Ophthalmologie Rétinienne (MESSIDOR; available in the public domain at http://messidor.crihan.fr/download-en.php) contains 1200 eye fundus color digital images of the posterior pole, acquired by three ophthalmologic departments using a color video 3CCD camera on a Topcon TRC NW6 nonmydriatic retinograph with a 45° FOV, 8 bits per color plane, and resolutions of 1440 × 960, 2240 × 1488, or 2304 × 1536 pixels. Eight hundred images were acquired with pupil dilation (one drop of Tropicamide at 0.5%) and 400 without dilation. The 1200 images are packaged in three sets, one per ophthalmologic department. Each set is divided into four subsets, each containing 100 images in TIFF format and an Excel file (Microsoft, Redmond, WA) with medical diagnoses for each image. Currently there are no annotations (markings) on the images. Annotations by a single clinician for OD diameter and the fovea center, for the whole MESSIDOR set, have been made available by the Department of Electronic, Computer Systems and Automatic Engineering, University of Huelva, Spain (available in the public domain at www.uhu.es/retinopathy/muestras/Provided_Information.zip).

AREDS (Age-Related Eye Disease Study; available in the public domain at https://www.emmes.com/study/areds/) is a major clinical trial sponsored by the National Eye Institute (NEI) at the National Institutes of Health (http://www.nei.nih.gov/amd/), involving several US centers working on AMD. The dataset includes several thousands of analog and digitized fundus images showing various stages of AMD. Longitudinal studies were performed over 10 years showing disease progression. The images were graded by national centers for AMD as well as for lens opacity. The ground truth does not include image-level delineation of drusen. The fundus photographs consist principally of 30° images including stereo images centered on temporal margin of the disc and an oblique view of the center of the macula near the temporal margin of the field, stereo images centered on the center of the macula, and monoscopic images centered temporal to the macula and offering an oblique view of the center of the macula near the nasal margin of the field.

ARIA (available in the public domain at http://www.eyecharity.com/aria_online/) contains color fundus images collected at St Paul’s Eye Unit and the University of Liverpool, United Kingdom, as part of the ARIA project. All subjects were adults. All images were taken using a Zeiss FF450+ fundus camera, originally stored as uncompressed TIFF files and converted to compressed JPG files for World Wide Web publication. All photographs were taken at a 50° FOV. Blood vessel masks created by trained image analysis experts are available. The optic disc and fovea, where relevant, are outlined in separate file sets. The data are organized into three categories, namely, age-related macular degeneration subjects, healthy control group subjects, and diabetic subjects.

ROC (available in the public domain at http://roc.healthcare.uiowa.edu/) is a set of 100 digital color fundus photographs selected from a large dataset (150,000 images) acquired at multiple sites within the EyeCheck DR screening program (see ROC Web site references), marked as gradable by the screening program ophthalmologists and including microaneurysms. Three different types of images with different resolutions are included, acquired by a Topcon NW 100, a Topcon NW200, or a Canon CR5-45NM and resulting in two differently shaped FOVs. All images are JPEG, and compression was set in the camera. The substantial black background around the FOV present in the original type II and III images was cut off using specialized software. This complete set was
randomly split into a training and a test set, each containing 50 images. Four retinal experts, all from the Department of Ophthalmology at the University of Iowa, were asked to annotate all microaneurysms and all irrelevant lesions in all 100 images in the test and training sets.

BIOIMLAB (http://bioimlab.dei.unipd.it/Data%20Sets.htm) at the University of Padova, Italy, maintains a number of publicly available datasets for several measurements, including vessel tortuosity (60 images from normal and hypertensive patients; 50 images of retinal arteries of similar length and caliber, 50 images of retinal veins of similar length and caliber), MATLAB [Mathworks, Natick, MA] data structures).

HEI-MED (available in the public domain at http://vibot.u-bourgogne.fr/luca/heimed.php) is a collection of 169 fundus images to train and test image-processing algorithms for the detection of exudates and diabetic macular edema. The images have been collected as part of a telemedicine network for DR diagnosis. The images contain manual segmentation of exudation, and include a machine segmentation of the vascular tree and optic nerve locations. The dataset contains a mixture of ethnic groups, with roughly 60% African American, 25% Caucasian, and 11% Hispanic.

APPENDIX B

This paper is the result of an international collaboration to which a total of 41 people contributed. Editorial limits on the number of authors prevented us from listing the following contributors: Adria Perez Rovira, Kristis Zutis, Khai Sing Chin (VAMPIRE project, School of Computing, University of Dundee, Dundee, United Kingdom); Kenneth W. Tobin, who pioneered ARIA work at ORNL, and Hector Santos-Villalobos (Oak Ridge National Laboratory, Oak Ridge, TN); Enrico Grisan, Eneca Poletti (Department of Information Engineering, University of Padova, Padova, Italy); David Reed, Christopher Gee (Jules Stein Eye Institute, Los Angeles, CA); Andrew Hunter (REVIEW Group, University of Lincoln, Lincoln, United Kingdom); Tien Yin Wong, M. Kamran Ikram (SERI, Singapore); Jiang Liu, Ngaan-Meng Tan (A*STAR, Singapore); Meindert Nijemeier (Department of Biomedical Engineering, University of Iowa, Iowa City, IA); Wynne Hsu, Mong Li Lee (National University of Singapore, Singapore); Hao Hao, Behzad Alihamad, Ganesh Naik (RMIT University, Melbourne, Australia); David E. Freund (Applied Physics Laboratory, Johns Hopkins University, Laurel, MD); Juan Xu (Department of Ophthalmology, University of Pittsburgh School of Medicine, Pittsburgh, PA).