1. ABSTRACT

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 colour 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 multi-centre annotations by multiple experts, specific to clinical tasks, and capable of running submitted software automatically on the data stored, with clear and widely agreed performance criteria, to provide a fair comparison.
2. INTRODUCTION

This paper was born of discussions involving several international research groups on automatic retinal image analysis (ARIA; see [22] for a recent review) at IEEE 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. 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 applicable 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.

(a) **Screening/monitoring**, e.g., 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 clinical attention. False negatives (missing 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 [31][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 healthcare providers and the government, and improve the general eye health.

(b) **Computer-assisted diagnosis and risk stratification**, e.g., diagnosis of 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 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).

(c) **Biomarkers discovery**, 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, e.g., 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 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 [1,2].

In addition, three further areas would benefit from reliable ARIA systems:
(d) **Longitudinal studies**, whereby ARIA provides a means to study quantitatively the evolution and characterization of a disease, to assist treatment planning or gauging patient response to a treatment.

(e) **Computer-aided or image-guided surgery**, an emerging application of ARIA algorithms [93][94][95][96], e.g., vitreo-retinal microsurgery, for which ARIA allows registration of intra-operative imagery with pre-operative imagery for image-guided surgical interventions [4].

(f) **Telehealth.** ARIA disease screening and monitoring could play a very important role here, e.g., in less developed countries where the incidence of diabetes is rising and screening made difficult by the lack of resources and clinicians [5]. A WHO 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 US, and applied on a societal scale under various national healthcare programs, e.g., in the UK [99,100].

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 the appendix a list of public data sets currently available for testing ARIA algorithms.

### 3. VALIDATION IN RETINAL IMAGE ANALYSIS

#### 3.1. 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, e.g., regions traced manually around landmarks or lesions, image quality level, or scores attached to DR screening images.

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

#### 3.2. 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):

(a) **outcome oriented**, e.g., disease/no disease;
(b) **disease grading**, e.g., severity of DR, ROP plus or pre-plus;
(c) **feature grading**, e.g., tortuosity level of vessels or eye vasculature, width of retinal vessels;
(d) **image/pixel/measurement oriented**, e.g., 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 advocated strongly by some authors [97].

Two major issues in the generation of reference standard data are the variability of expert judgement and the need of generating annotations directly comparable to the algorithm’s output. The former is addressed by having several experts annotate the same data set; however it is not ultimately clear how to obtain a single reference value from multiple ones (e.g., by averaging values, by discussion and consensus, inter-rater reliability metrics such as AC1 or Kappa, or just keeping histograms/distributions of multiple values). This is because some annotations 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 screening programme, 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 consists 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 summarising annotations from multiple experts, e.g., STAPLE for image segmentation [6][7] and other tasks [8]. Imperfect reference standards are the motivation behind weak learning methods in pattern recognition and machine learning [9][10]; these methods are currently used only rarely in ARIA [73] but might provide very useful modelling tools (see also Section 5.1). An interesting viewpoint is offered by Quellec et al. [105] who found, briefly, 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.

4. Images and tasks

We discuss the characteristics of digital images and of the instruments generating them, and the clinical tasks for which images are ultimately created and analysed. Both play a substantial role in validation. For reasons of space, we omit other imaging modalities like OCT and FA angiography.

4.1. Instruments and images

**Fundus cameras**

The majority of ARIA systems reported to date consider single colour images from digital fundus cameras, although acquiring two images per eye for both eyes (posterior pole, optic disc (OD) centred) is increasingly common in screening programmes. Images are acquired with or without dilating the patient's pupil through eye drops, respectively mydriatic, non- mydriatic. 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 (Germany), Topcon (USA), Nidek (Japan), Canon (Japan) and Kowa (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 degrees with a magnification of 2.5x. Some modifications are possible through zoom or auxiliary lenses, from 15 degrees, which provides 5x magnification, to 140 degrees with a wide-angle lens, which reduces the image by half. The actual image resolution, i.e., how many millimetres are captured by a pixel, depends on various factors, mainly the properties of the optics and the resolution of the CMOS/CCD image sensor employed; about 3,000x3,000 pixels are now common. Higher-resolution sensors could
be used, but the optics pose a limit to the resolution achievable, i.e., to the size of the smallest distance which can be imaged in focus. As the optical quality of the eye prevents resolving features smaller than 20μm, 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; existing fundus cameras are bulky, expensive (about $25,000) and require special skills to operate. User-friendly, cost-effective, hand-held retinal camera have been developed to tackle these issues, but work in this area is still limited [10][11].

4.2. Image quality

Image quality depends on acquisition procedures, operators and their training, blur, occlusions (e.g., cataract, eyelashes), widespread lesions, artefacts introduced by the instrument, and conditions; for instance, fundus images of babies 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 programmes.

As quality influences the performance of ARIA systems, it seems recommendable to divide a set of test images into quality classes, e.g., 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 programme, for instance, defines three image quality grades (inadequate, minimum, achievable) [19].

Automated systems for assessing retinal image quality exist [15,16,17,18,20,21], but quality is not always considered in the wider ARIA literature when preparing data set. 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 using 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 OR-colour characteristics like the shape of the colour histogram.

4.3. 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 which the reader is referred to the recent review by Abràmoff et al.[22].

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 centres, 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 centres is generally sufficient to establish a retinal coordinate system; estimating the ellipticity of the OD requires accurate contour detection including parapapillary atrophy segmentation [101].

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 and neovascularization. Measurements can be local (e.g., width, branching angles) and global (e.g., fractal dimension of whole network) [129,130]. Subtle vessel changes may occur during the early stages of disease development, consequent on changes to blood flow dynamics, and associations of these changes have been found with age, or with risk of stroke [128, 129]. There is therefore substantial interest in automatically segmenting the vasculature and measuring its properties. This requires, typically, locating the vessels, establishing overall branching patterns and connectivity, and computing target measurements.

The majority of datasets used for validating vasculature-related ARIA algorithms (see Appendix) concentrate on the first step. They consist of images with corresponding reference standard in the form of vascular masks (binary pixel images) generated by clinicians using a software drawing tool. Two well-known examples are DRIVE and STARE.

These datasets suffer from two limitations. First, there is currently no absolute, objective definition of the location of the edge of a retinal blood vessel. The observed vessel in a standard retinal image corresponds to the blood column within the vessel. However, as the column depth reduces towards the vessel edge, the intensity drops off, blurring the edge's appearance. Second, generating vessel masks is one of the most labour-intensive annotation tasks. The anti-aliasing effect along vessel edges make it difficult to determine exactly whether an individual pixel belongs to a vessel properly.

Both DRIVE and STARE use multiple observers, who provide sometimes significantly different reference standards. As algorithms approach human levels of performance, it becomes difficult to use such variable standards to assess performance. On the other hand, the datasets provide an easily accessible reference standard. The REVIEW dataset uses a different approach. A limited number of vessels have their edges marked using a contour tool with sub-pixel accuracy. This allows more accurate assessment of algorithms for vessel width determination, but does not provide sufficient detail to analyse the overall segmentation performance.

Work addressing vascular connectivity exists and some standard definitions of angles have been suggested [23]. However there appears to be relatively little published work on vessel branching angles, and we know no publicly available reference standard dataset addressing this aspect [24][25].

**Diabetic retinopathy and diabetes-related retinal lesions**

DR has attracted a large part of ARIA research, and we dedicate a longer background section to it. Prevalence is expected to grow exponentially [26][27], affecting 300 million people worldwide by 2025 [28]. DR, a specific microvascular complication of diabetes, is a major cause of vision loss in people aged 25-60. Of the 246 million people with diabetes, about a third have DR, and a third of these have vision-threatening retinopathy, the majority caused by diabetic macular edema (DME) [29]. DR poses a huge economic burden on patients, healthcare systems and society, estimated at US$500 million annually in the US alone.

This background poses a demand for reliable automated early-screening procedures. The challenge for ARIA is to find cost-effective techniques with sufficient sensitivity and specificity to identify those at risk of vision loss reliably. Clinically, the primary validation method of interest is outcome-oriented: how well can the presence of disease be detected (refer/no refer), and how sensitive can the systems be made (near-zero false negatives) with a manageable level of specificity.

Some researchers [86][87][38][31][104] have reported systems with performance deemed acceptable for clinical deployment, but reports of large-scale studies remain rare and may not hold true for particular datasets with unique idiosyncrasies. Lesions targeted by ARIA systems include microaneurysms, cotton wool spots, soft exudates and small hemorrhages for non-proliferative retinopathy, and ischemic areas in the retina, loss of vessels and vessel proliferation for proliferative retinopathy [88][51][81][37].

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 (colour 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 data sets for validation. For instance, the winning team of the Retinopathy Online Challenge
(http://roc.healthcare.uiowa.edu/, Niemeijer et al. [30]), led by Quellec, achieved only 60% sensitivity
with 8 false positives per image; but earlier work by Niemeijer (2005) [43] reported 100% sensitivity
and 87% specificity on image-level screening, despite the per-lesion sensitivity being only about 30%.

<table>
<thead>
<tr>
<th>Reference</th>
<th>DR Task</th>
<th>Method</th>
<th>Public dataset?</th>
<th># of images</th>
<th>Image type</th>
<th>Validation level</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abràmoff 2010 [31]</td>
<td>MA, Dot HM</td>
<td>Pixel clustering, kNN</td>
<td>No</td>
<td>16670</td>
<td>Color fundus</td>
<td>Image</td>
<td>47.7%</td>
<td>90%</td>
</tr>
<tr>
<td>Giancardo 2011 [32]</td>
<td>MA</td>
<td>Radon transform, wavelet pre-processing</td>
<td>Yes</td>
<td>100</td>
<td>Color fundus</td>
<td>Lesion</td>
<td>50%</td>
<td>&gt;10 false positive per image</td>
</tr>
<tr>
<td>Niemeijer 2010 [30]</td>
<td>MA</td>
<td>Wavelet transform</td>
<td>Yes</td>
<td>100</td>
<td>Color fundus</td>
<td>Lesion</td>
<td>60%</td>
<td>8 false positive per image</td>
</tr>
<tr>
<td>Mizutani 2009 [33]</td>
<td>MA</td>
<td>Double-ring filter</td>
<td>Yes</td>
<td>100</td>
<td>Color fundus</td>
<td>Lesion</td>
<td>65%</td>
<td>27 false positive per image</td>
</tr>
<tr>
<td>Quellec 2008 [34]</td>
<td>MA</td>
<td>Wavelet, genetic algorithm</td>
<td>No</td>
<td>1115</td>
<td>Color fundus + Angiogram</td>
<td>Lesion</td>
<td>90.24%</td>
<td>89.75%</td>
</tr>
<tr>
<td>Walter 2007 [35]</td>
<td>MA</td>
<td>Diameter closing, feature extraction, classification</td>
<td>No</td>
<td>115</td>
<td>Color fundus</td>
<td>Lesion</td>
<td>88.50%</td>
<td>2.13 false positive per image</td>
</tr>
<tr>
<td>Huang 2007 [36]</td>
<td>MA</td>
<td>Edge Inference</td>
<td>No</td>
<td>49</td>
<td>Color fundus</td>
<td>Lesion</td>
<td>68%</td>
<td>&gt;40 false positive per image</td>
</tr>
<tr>
<td>Niemeijer 2007 [37]</td>
<td>Exudates</td>
<td>Machine Learning</td>
<td>No</td>
<td>430</td>
<td>Color fundus</td>
<td>Lesion</td>
<td>95%</td>
<td>88%</td>
</tr>
<tr>
<td>Philip 2007 [38]</td>
<td>MA, Dot HM</td>
<td>Combination of methods</td>
<td>No</td>
<td>15473</td>
<td>Color fundus</td>
<td>Image</td>
<td>97.90%</td>
<td>67.40%</td>
</tr>
<tr>
<td>Fleming 2006 [39]</td>
<td>MA</td>
<td>Watershed contrast normalization</td>
<td>No</td>
<td>1677</td>
<td>Color fundus</td>
<td>Image</td>
<td>85.40%</td>
<td>83.10%</td>
</tr>
<tr>
<td>Quellec 2006 [40]</td>
<td>MA</td>
<td>Wavelet template matching</td>
<td>No</td>
<td>995</td>
<td>Green channel fundus</td>
<td>Lesion</td>
<td>87.90%</td>
<td>96.20%</td>
</tr>
<tr>
<td>Pallawala 2005 [41]</td>
<td>MA</td>
<td>Generalized eigenvectors</td>
<td>No</td>
<td>70</td>
<td>Color fundus</td>
<td>Region</td>
<td>93%</td>
<td>NA</td>
</tr>
<tr>
<td>Serrano 2004 [42]</td>
<td>MA</td>
<td>2D adaptive filtering, region growing</td>
<td>No</td>
<td>11</td>
<td>Angiogram</td>
<td>Region</td>
<td>90.72%</td>
<td>82.35%</td>
</tr>
<tr>
<td>Niemeijer 2005 [43]</td>
<td>MA, Dot HM</td>
<td>Pixel classification</td>
<td>No</td>
<td>240</td>
<td>Color fundus</td>
<td>Region</td>
<td>100%</td>
<td>87%</td>
</tr>
<tr>
<td>Larsen 2003 [44]</td>
<td>MA, Dot HM</td>
<td>RetinalLyze (proprietary)</td>
<td>No</td>
<td>400</td>
<td>Color fundus</td>
<td>Image</td>
<td>96.70%</td>
<td>71.40%</td>
</tr>
<tr>
<td>Rapantzi-kos 2003 [45]</td>
<td>Drusen</td>
<td>Histogram adaptive local thresholding</td>
<td>No</td>
<td>23</td>
<td>Color fundus</td>
<td>Region</td>
<td>98.80%</td>
<td>99.31%</td>
</tr>
<tr>
<td>Sinthanayothin 2002</td>
<td>MA, HM</td>
<td>Morphological, region growing</td>
<td>No</td>
<td>142</td>
<td>Color fundus</td>
<td>Region</td>
<td>77.50%</td>
<td>88.70%</td>
</tr>
<tr>
<td>Reference</td>
<td>Method</td>
<td>Disease</td>
<td>Technique Details</td>
<td>Result 1</td>
<td>Result 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------</td>
<td>---------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>----------</td>
<td>-----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsu 2001</td>
<td>neural network</td>
<td>Exudates</td>
<td>Dynamic clustering with domain knowledge</td>
<td>(543) Color fundus</td>
<td>Image 100% 74%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yang 2001</td>
<td>MA</td>
<td>Morphological, region growing</td>
<td>(46) Color fundus</td>
<td>Image 90% 80%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wang 2000</td>
<td>Exudates</td>
<td>Statistical classification with local window-based verification</td>
<td>(200) Color fundus</td>
<td>Image 100% 70%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hipwell 2000</td>
<td>MA</td>
<td>Manual rule-based classifier</td>
<td>(3885) Red-free</td>
<td>Image 85% 76%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ege 2000</td>
<td>MA, HM, Exudate</td>
<td>Cotton Wool</td>
<td>Estimate background intensity, extract candidate regions for classification</td>
<td>(268) Color fundus</td>
<td>Lesion 94% 69%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee 1999</td>
<td>MA, Dot HM</td>
<td>Pattern recognition</td>
<td>(400) 35mm color slides</td>
<td>Lesion NA NA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cree 1997</td>
<td>MA</td>
<td>Manual rule-based classifier</td>
<td>(88) Angiogram</td>
<td>Lesion 82% 84%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gardner 1996</td>
<td>HM, Exudates</td>
<td>Neural network</td>
<td></td>
<td>(480) Red-free</td>
<td>Region 73.80% 73.80%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spencer 1992</td>
<td>MA</td>
<td>Matched filter</td>
<td>(6) Angiogram</td>
<td>Lesion 45% &gt;150 false positive per image</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.

**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), i.e., 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 colours 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 [56][57]. Tomographic imaging of the retina in 3D has been made possible by confocal laser scanning, used in the Heidelberg retinal tomograph, and optical coherence tomography (OCT), which exploits interferometry to achieve tomographic micrometre-resolution imaging of the retinal layers.

Table 2 summarizes glaucoma-related ARIA reports (although numerous papers have described individual optic disk and cup detection, only papers leading to a CDR or glaucoma detection outcome have been included). ARGALI is a recent ARIA system for glaucoma assessment [58][59][60]. It uses active contour methods based on level sets to segment the cup and the disc and calculate the cup-to-disc ratio. A similar approach was adopted by Joshi et al. [61]. Some authors reported stereo techniques recovering depth information [56][57][62][63]. An alternative approach is the use of machine learning to assign a predictive score for the risk of glaucoma directly to images [64]. AGALIA (Automatic GLaucoma Diagnosis and Its Genetic Association Study through Medical Image InformAtics) [65][66][67] 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, e.g., ORIGA-light [65][66][67].

<table>
<thead>
<tr>
<th>Reference</th>
<th>Method</th>
<th>Public dataset?</th>
<th># of images</th>
<th>Image type</th>
<th>Validation level</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xu 2012 [107]</td>
<td>Intra-image learning</td>
<td>ORIGA [108], on request</td>
<td>650</td>
<td>Non-stereo Fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.081</td>
</tr>
<tr>
<td>Yin 2012 [109]</td>
<td>Statistical model-based</td>
<td>ORIGA [108], on request</td>
<td>650</td>
<td>Non-stereo Fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.100</td>
</tr>
<tr>
<td>Joshi 2012 [110]</td>
<td>Depth-discontinuity model</td>
<td>No</td>
<td>138</td>
<td>Stereo Fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.09</td>
</tr>
<tr>
<td>Xu 2011 [111]</td>
<td>Sliding window, regression</td>
<td>ORIGA [108], on request</td>
<td>650</td>
<td>Non-stereo Fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.091</td>
</tr>
<tr>
<td>Liu 2011 [112]</td>
<td>AGLAIA framework</td>
<td>No</td>
<td>291</td>
<td>Non-stereo Fundus</td>
<td>Glaucoma</td>
<td>AUC: 0.73</td>
</tr>
<tr>
<td>Joshi 2011 [113]</td>
<td>Active contour model, r-bends</td>
<td>No</td>
<td>138</td>
<td>Non-stereo Fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.09</td>
</tr>
<tr>
<td>Acharya 2011 [114]</td>
<td>Higher order spectra, texture features</td>
<td>No</td>
<td>60</td>
<td>Non-stereo Fundus</td>
<td>Glaucoma</td>
<td>Detection: 61%</td>
</tr>
<tr>
<td>Muramatsu 2011 [115]</td>
<td>Active contour, depth reconstruction</td>
<td>No</td>
<td>80</td>
<td>Stereo Fundus, glaucoma</td>
<td>Mean CDR error: 0.110 AUC: 0.90</td>
<td></td>
</tr>
<tr>
<td>Bock 2010 [116]</td>
<td>Appearance-based analysis</td>
<td>No</td>
<td>575</td>
<td>Non-Stereo Fundus</td>
<td>Glaucoma</td>
<td>AUC: 0.88</td>
</tr>
<tr>
<td>Hatanaka 2010 [117]</td>
<td>Intensity profiling</td>
<td>No</td>
<td>50</td>
<td>Non-stereo Fundus</td>
<td>CDR, Glaucoma</td>
<td>Mean CDR error: 0.14 AUC: 0.87</td>
</tr>
<tr>
<td>Joshi 2010 [118]</td>
<td>Regional information</td>
<td>No</td>
<td>170</td>
<td>Non-stereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.100</td>
</tr>
<tr>
<td>Wong 2009 [119]</td>
<td>Level-set</td>
<td>No</td>
<td>104</td>
<td>Non-stereo fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.089</td>
</tr>
<tr>
<td>Muramatsu 2009 [120]</td>
<td>Thresholding, 3D reconstruction</td>
<td>No</td>
<td>80</td>
<td>Stereo Fundus</td>
<td>Glaucoma</td>
<td>AUC: 0.83</td>
</tr>
<tr>
<td>Wong 2009 [121]</td>
<td>Hybrid wavelet-edges, kinking</td>
<td>No</td>
<td>27</td>
<td>Non-Stereo Fundus</td>
<td>CDR</td>
<td>Mean CDR error: 0.093</td>
</tr>
<tr>
<td>Liu 2008 [122]</td>
<td>ARGALI</td>
<td>No</td>
<td>23</td>
<td>Non-Stereo Fundus</td>
<td>CDR</td>
<td>Correlation w/ground truth = 0.89</td>
</tr>
<tr>
<td>Xu 2007 [123]</td>
<td>Deformable model</td>
<td>No</td>
<td>25</td>
<td>Stereo Fundus</td>
<td>CDR</td>
<td>Correlation w/ground truth: 0.71</td>
</tr>
<tr>
<td>Abramoff 2007 [124]</td>
<td>Pixel features, kNN</td>
<td>No</td>
<td>58</td>
<td>Stereo Fundus</td>
<td>CDR</td>
<td>Correlation w/ground truth: 0.93</td>
</tr>
<tr>
<td>Inoue 2005 [125]</td>
<td>Discriminatory analysis</td>
<td>No</td>
<td>N/A</td>
<td>Non-stereo Fundus</td>
<td>N/A</td>
<td>Not provided</td>
</tr>
</tbody>
</table>

| 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 themselves part of the screening will be of significant clinical benefit.

<table>
<thead>
<tr>
<th>Reference</th>
<th>DR Task</th>
<th>Method</th>
<th>Public dataset?</th>
<th># of images</th>
<th>Image type</th>
<th>Validation level</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Koreen 2007 [68]</td>
<td>ROP diagnosis</td>
<td>semi-automated</td>
<td>No</td>
<td>20</td>
<td>Color fundus</td>
<td>Image</td>
<td>50%-100%</td>
<td>46%-93%</td>
</tr>
<tr>
<td>Wilson 2008 [69]</td>
<td>vessel tortuosities</td>
<td>semi-automated</td>
<td>No</td>
<td>10</td>
<td>Color fundus</td>
<td>Image feature</td>
<td>ground truth correlation: 0.49-0.67 (tortuosities), 0.42 (width)</td>
<td></td>
</tr>
<tr>
<td>Wallace 2007 [70]</td>
<td>ROP diagnosis</td>
<td>semi-automated</td>
<td>No</td>
<td>185</td>
<td>Color fundus</td>
<td>Image</td>
<td>97%</td>
<td>94%</td>
</tr>
<tr>
<td>Wallace 2009 [106]</td>
<td>vessel width measure</td>
<td>semi-automated</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>Johnson 2007 [71]</td>
<td>vessel width measure</td>
<td>semi-automated</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 3.

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 semi-automatic tool for the labelling of the skeleton trees, followed by an automatic procedure to measure vessel width, tortuosity and from these derive Plus or Pre-plus diagnosis [68]. The Computer-Aided Image Analysis of the Retina (CAIAR) system semi-automatically 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 [69]. ROPtool traces semi-automatically retinal blood vessels; its reliability in measuring tortuosity and dilation of vessels was assessed in two distinct studies [70, 106]. VesselMap (Imedos, Jena, Germany) is a commercial semi-automatic software, developed to analyse vessels in 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 [71].

No public annotated datasets on ROP seem available, to date, for ARIA system training and validation.

Age-related macular degeneration

Age-Related Macular Degeneration (AMD) is a condition of great interest because of its relevance and prevalence. It has however attracted less ARIA research than DR and we do not discuss it in this paper. See [9][10][45] [22] for examples of studies.

5. VALIDATION ISSUES AND CHALLENGES

We now discuss the factors introducing uncertainties in the reference (Section 3). Some are shared with other areas of medical image processing. Indeed, the very definition of the reference standard varies with a number of factors, and may not reflect the true state of a disease. All this results in assessment variations. As it is unreasonable to pursue an accuracy higher than that of the reference standard used, it seems essential to characterize quantitatively the variations of reference standards.
(see Section 3.2). 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 data sets. 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 point “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 which impact the four types of validation listed in Section 3.2: identifying the presence of DR (refer/no refer decision), and identifying specific lesions from which conclusions are then drawn by the specialist [72].

5.1. Variation of expert judgement

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 judgements vary, in general, with experts (inter-observer variations) and, to a lesser extent, over repeated judgements by the same expert (intra-observer variations).

These variations depend at least on experience, task, background, image quality (e.g. this makes often 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 and Abrâmoff [73,31] 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 data set used), whereas it could be improved significantly when comparing against a committee of experts.

Hubschman et al. [89] quantified the inter-observer 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, non-perfused, insufficient quality). The cardiac cycle can also have an influence on vessel calibre measurements (see Section 5.3).

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 is available, and expert numbers are usually small (2 to 5 in most ARIA papers). Relevant techniques have been mentioned in Section 3.2.

5.2. 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 Section 3.2., 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 maximise 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. [73], who designed GUIs to automatically detect elements that catch the attention of clinicians in their daily clinical practice. This avoids requesting clinicians to annotate explicitly anatomical
structures, a task they have not been trained for. Usage logs (zoom, magnifying glass, etc.) are used to train weakly-supervised lesion detectors and validate their outputs.

Datasets used for high-profile clinical studies, such as AREDS (see Appendix), were annotated at national grading centres using established protocols. As the intent was to evaluate the efficacy of drugs (for AREDS, by the NIH/USDA), 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, e.g., pixel-level delineations allowing ROC analysis.

5.3. 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 [74][75][76][77][127], 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 3.1 to 3.9% for major retinal arteriololes. Venular diameter was smallest in early systole, increasing to a maximum level in early diastole and decreasing henceforth. 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 CRVE changed by 3.1% and the CRAE changed by 4.3%. However, recent work by Kumar et al [127] has shown that there is no significant change in the average of the width of six large vessels in the region often chosen for AVR estimation. As quantifying vascular changes is a primary ARIA concern [77], there is a need for more extensive studies and test sets with repeated images of the same eye in time. Gating the eye fundus camera with the electrocardiogram has also been proposed. Moret et al. [78] have measured changes to the vessel shape and size, also evidenced from the use of dynamic vessel analyzers (DVA) for disease diagnosis [79]. 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 specialised research facilities.

5.4. Different imaging instruments

Due to the nature of retinal imaging, there is usually a high level of customization for each different modality. Algorithms used for the segmentation of the optic nerve head in a confocal scanning technique may not be applicable to retinal images 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 neovascularisation. The selection of FOV and imaging field can also affect performance, particularly when relying on or assuming the visibility of retinal landmarks (OD, macula, arcades). The colour 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.

5.5. Image quality

This was discussed in Section 4.2. We re-iterate that image quality depends, among others, on instrument characteristics, acquisition procedure, and target conditions, and 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.

5.6. Data sets

The key observation is that different data sets may lead to somewhat inconsistent performance assessments as preparation protocols may differ. For instance, the best-known public retinal datasets are probably DRIVE [80] and STARE [102]. Soares et al. [81] compared binary vessels 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 [37], and a comparison of these methods has not shown large differences in accuracy with images from the STARE dataset [82].

5.7. Task, previous knowledge, condition

Expert judgement 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 task does affect annotations, it seems advisable to group reference annotations also by clinical task, and specify annotation protocols accordingly. At the moment, some public repositories group images by conditions, e.g., MESSIDOR for DR (see Appendix).

5.8. 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, i.e., clinical metadata that may be relevant for the disease incidence and for organizing consistent validation datasets. Contextual patient characterization data includes ethnicity, age, gender, medical data (hypertension, diabetes, heart diseases), lifestyle factors (e.g., smoking), systemic data (BMI, cholesterol level, blood pressure, etc.), ocular data (refractive error, lens opacity, etc.), co-morbid 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 [83]. Quellec et al. [84][85] 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 to confirm these findings is currently being performed on a dataset of 25,702 examination records from the Ophdiat screening network (http://reseau-ophdiat.aphp.fr/index.html) [3].

5.9. Human in the loop

A final consideration is that some ARIA algorithms are semi-automatic 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 de-coupling human and automatic contributions. Specialized techniques apparently, have not yet considered in the ARIA literature.

6. 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 data sets 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 [103]. 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). Most are still limited collection of images generated by individual sites, or contain limited annotations, or both. There is currently no co-ordinated consensus in the community on how to structure such data sets or what information to include, and this paper is meant to provoke thoughts towards that end.
So far, neither ARIA data sets nor epidemiological studies based on large populations have taken into account the above issues systematically. Furthermore, as images are often acquired independent of outcomes and participant characteristics, random variability may tend to underestimate the true associations found, e.g., 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, have the following properties.

- **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 ultimately useful results for clinicians.

- **Easily accessible**, ideally via websites from which data could be downloaded following a suitable registration procedure including legal disclaimers, no-redistribution clauses, acknowledgments needed for use, and so forth.

- **Regularly maintained**, ideally by a consortium of international groups monitoring distribution and guaranteeing maintenance of data and annotations.

- **Large size**, wherever possible dimensioned statistically to maximise power; tentatively, the minimum order of magnitude should be the thousands of images. Large data repositories should be standardized, patient friendly imaging protocols allowing large populations to be image effectively; for example, single 45-degree fovea-centered, or two 45-degree fields, centered on the fovea and OD respectively [126].

- **Include metadata**, i.e., non-image data characterizing imaging instruments, patients and disease (Section 5.8).

- **Include automated tools for running software on the data**, as done by the Middlebury stereo site [103], in which executable code is loaded and run on the site, and performance assessed in terms of pre-defined measures which are displayed 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 in Section 3.2.

- **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 inter-observer variability, including arbitrated annotations; each expert should ideally annotate the data set multiple times to estimate intra-observer variability (see also last paragraph of Section 3.2, for techniques managing variations in expert judgement). Uncertainty levels declared or agreed 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, to obtain ground truth measured by independent, highly accurate instruments, the only way to estimate "how good" an algorithm is comparing its output with expert judgement.

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 takes time and 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 600x600 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.

APPENDIX: A list of public data repositories for retinal image analysis.

This appendix lists the public retinal data sets known to us. Unless otherwise stated, all data sets listed are easily reachable by a Google search. Most descriptions are excerpts from the referred websites.

STARE (http://www.ces.clemson.edu/~ahoover/stare/) is one of the earliest and most 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 field-of-view, and subsequently digitized at 605x700 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 haemorrhaging.

DRIVE (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 25-90 years of age. 40 photographs were randomly selected, 33 without and 7 with DR signs. The images were acquired using a Canon CR5 non-mydratic 3CCD camera with a 45 degree field of view (FOV). Each image was captured using 8 bits per color plane at 768 by 584 pixels. The FOV of each image is circular with a diameter of approximately 540 pixels.

DIARETDB1 (http://www2.it.lut.fi/project/imageret/diaretdb1/) consists of 89 colour fundus images. 84 contain at least mild non-proliferative 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-degree 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. 4 medical experts were asked to mark the areas related to the microaneurysms, haemorrhages, and hard and soft exudates. Ground truth confidence levels, \{<50%,~50%,~100%\}, represented the certainty of the decision that a marked finding is correct, are included.

MESSIDOR (http://messidor.crihan.fr/download-en.php) contains 1200 eye fundus color digital images of the posterior pole, acquired by 3 ophthalmologic departments using a color video 3CCD camera on a Topcon TRC NW6 non-mydratic retinograph with a 45-degree FOV, 8 bits per color plane and resolutions of 1440x960, 2240x1488 or 2304x1536 pixels. 800 images were acquired with pupil dilation (one drop of Tropicamide at 0.5%) and 400 without dilation. The 1200 images are packaged in 3 sets, one per ophthalmologic department. Each set is divided into 4 zipped sub sets containing each 100 images in TIFF format and an Excel file 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 centre, for the whole MESSIDOR set, have been made available by the Department of Electronic, Computer Systems and Automatic Engineering, University of Huelva, Spain at www.uhu.es/retinopathy/muestras/Provided_Information.zip.

REVIEW (http://reviewdb.lincoln.ac.uk/) contains several subsets, with a mix of patients with disease and no disease, and a focus on validating accurate measurements. It includes 16 images with 193 vessel segments, demonstrating a variety of pathologies and vessel types. These image sets
contain 5066 manually marked profiles. Images were assessed by three independent experts, who marked the vessel edges.

AREDS (Age-Related Eye Disease Study; https://web.emmes.com/study/areds/, http://www.areds2.org/) 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 ARMD. The dataset includes several thousands of analog and digitized fundus images showing various stages of AMD. Longitudinal studies were performed over ten 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 including 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 including an oblique view of the center of the macula near the nasal margin of the field.

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

ROC (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 website 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 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 set.

BIOIMLAB (http://bioimlab.dei.unipd.it/Data%20Sets.htm) at the Univ of Padova, Italy, maintains a number of publicly available datasets for several measurements, including vessel tortuosity (60 images from normal and hypertensive patients; 30 images of retinal arteries of similar length and calibre, 30 images of retinal veins of similar length and calibre, Matlab data structures).

HEI-MED (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.

REFERENCES


