|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-J-017-A01 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | E-meeting, 30 September – 2 October 2020 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Ophthalmo Topic Driver | | |
| **Title:** | | Att.1 – TDD update (TG-Ophthalmo) | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Arun Shroff Medindia/Xtend.ai India / USA | | Tel: +1 908-208-1100 Email: [arunshroff@gmail.com](mailto:arunshroff@gmail.com) |

|  |  |
| --- | --- |
| **Abstract:** | This topic description document (TDD) specifies a standardized benchmarking for the use of AI for Ophthalmology. It follows the structure defined in [FGAI4H-C-105](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-C-105.docx?d=w50606d7d9bf340198b6423e4d5babbe6) relevant for setting up this benchmarking. This document is a draft and a work-in-progress until it is finally approved by the Focus Group. |

|  |  |
| --- | --- |
| **Change Notes:** | ***Version 7.0 (submitted as FG-AI4H-DEL10.9 for e-meeting J)***   * *Updated sections on DR classification to include DME and referrable /non-referrable DR.* * *Miscellaneous edits and corrections* * *Updated progress since meeting I and list of new members.*   ***Version 6.0 (submitted as FGAI4H-I-017-A01 for e-meeting I)***   * *Incorporated TDD updates received from Xingxing Cao, Baidu. Rajaraman Subramanian, Calligo Technologies and Parvathi Ram*, *St. John's Medical College, India* * *Miscellaneous edits and corrections.*   ***Version 5.0 (submitted as FGAI4H-H-017-A01 for meeting H in Brasilia)***   * *Incorporated topic group contributions received in Meeting G into relevant sections of TDD: FGAI4H-G-030-R01 (St, John's Medical College) on Red Eye and FG-AI4H-G-028 (Calligo Technologies) on Leveraging Edge Analytics* * *Added 1.1.1 Topic Group Thematic Classification* * *Separated authors and contributors according to ITU rules* * *Added Change Notes to the TDD* * *Separated Binary/Multi-label classification scores/metrics; added Kappa – Quadratic Kappa Metric.* * *Miscellaneous edits / corrections*   ***Version 4.0 (submitted as FGAI4H-G-012 for meeting G in New-Delhi)***   * *Added Status Update for Meeting G* * *Added new topic group members* * *Minor edits / corrections*   ***Version 3.0 (submitted as FGAI4H-F-012 for meeting F in Tanzania)***   * *Added Status Update for Meeting F* * *Added new topic group members* * *Minor edits / corrections\*   ***Version 2.0 (submitted as FGAI4H-E-014-R01 for meeting E in Geneva)***   * *Updated to include Pathological Myopia (PM)* * *Added Status Update for Meeting E* * *Added new topic group members*   ***Version 1.0 (submitted as FGAI4H-D-038 for meeting D in Shanghai)***  *This is the initial draft version of the TDD.* |

**Contributors**

|  |  |  |
| --- | --- | --- |
| **Contact:** | Yanwu (Frank) XU Intelligent Healthcare Unit, Baidu, China | Tel: +86 13918541815 Fax: +86 10 59922186 Email: [xuyanwu@baidu.com](mailto:xuyanwu@baidu.com) |
|  | Xingxing Cao, Intelligent Healthcare Unit, Baidu, China | Tel: + 86 18500936701 Fax: + 86 10 59922189 Email: caoxingxing@baidu.com |
|  | Parvathi Ram St John's Medical College India | Tel: +91 9972234011 Email: [pramo282@gmail.com](mailto:pramo282@gmail.com) |
|  | Dr Suneetha N St John's Medical College India | Tel: +91 8197673923 Email: [suneetha.n.lobo@gmail.com](mailto:suneetha.n.lobo@gmail.com) |
|  | Rajaraman Subramanian  Calligo Technologies  India | Tel: +919845239446 Email: [Rajaraman.subramanian@calligotech.com](mailto:Rajaraman.subramanian@calligotech.com) |
|  | Sriganesh Rao  Calligo Technologies  India | Tel: +919845155800  Email: [Sriganesh.rao@calligotech.com](mailto:Sriganesh.rao@calligotech.com) |
|  | Sushil Kumar  TEC, New Delhi India | Tel: +911123323471 Email: [sushil.k.123@gmail.com](mailto:sushil.k.123@gmail.com) |

**CONTENTS**

| Page |
| --- |
| [1 Introduction 5](#_Toc51950675)  [1.1 Topic Description 5](#_Toc51950676)  [1.1.1 Topic Thematic Classification: 5](#_Toc51950677)  [1.1.2 Relevance 6](#_Toc51950678)  [1.1.3 Current approaches and gold standards for detection 7](#_Toc51950679)  [1.1.4 Impact of AI 9](#_Toc51950680)  [1.1.5 Impact of benchmarking AI Solutions 10](#_Toc51950681)  [1.2 Ethical considerations 11](#_Toc51950682)  [1.3 Existing AI solutions (includes datasets, systems and benchmarks) 11](#_Toc51950683)  [1.3.1 DR: 11](#_Toc51950684)  [1.3.2 GC 12](#_Toc51950685)  [1.3.3 AMD 12](#_Toc51950686)  [1.3.4 PM: 12](#_Toc51950687)  [1.3.5 Red eye 13](#_Toc51950688)  [2 AI4H Topic Group 16](#_Toc51950689)  [2.1 General mandate of the Topic Group 17](#_Toc51950690)  [2.2 Topic description document 17](#_Toc51950691)  [2.3 Subtopics 17](#_Toc51950692)  [2.4 Topic group participation 17](#_Toc51950693)  [2.5 Status of this Topic Group 17](#_Toc51950694)  [2.5.1 Status update for meeting D (Shanghai): 18](#_Toc51950695)  [2.5.2 Status Update for Meeting E (Geneva) 18](#_Toc51950696)  [2.5.3 Status Update for Meeting F (Zanzibar): 18](#_Toc51950697)  [2.5.4 Status Update for Meeting G (New Delhi): 19](#_Toc51950698)  [2.5.5 Status Update for Meeting H (Brasilia): 19](#_Toc51950699)  [2.5.6 Status Update for Meeting I (E-meeting): 20](#_Toc51950700)  [2.5.7 Status Update for Meeting I (E-meeting): 20](#_Toc51950701)  [2.5.8 Current members of the topic group: 21](#_Toc51950702)  [2.6 Next meetings 21](#_Toc51950703)  [3 Method 22](#_Toc51950704)  [3.1 Overview of the benchmarking 22](#_Toc51950705)  [3.2 AI input data structure 22](#_Toc51950706)  [3.3 AI output data structure 22](#_Toc51950707)  [3.4 Test data labels 23](#_Toc51950708)  [3.5 Scores and Metrics: 24](#_Toc51950709)  [3.5.1 Classification Tasks: 24](#_Toc51950710)  [3.5.2 Segmentation Tasks: 25](#_Toc51950711)  [3.5.3 Localization Tasks: 25](#_Toc51950712)  [3.6 Undisclosed test data set collection 25](#_Toc51950713)  [3.7 Benchmarking methodology and architecture 25](#_Toc51950714)  [4 Reporting methodology 25](#_Toc51950715)  [5 Results 26](#_Toc51950716)  [6 Discussion 26](#_Toc51950717)  [7 Declaration of conflict of interest 26](#_Toc51950718)  [Appendix A: Glossary 28](#_Toc51950719)  [References 29](#_Toc51950720) |

**List of Tables**

| Page |
| --- |
| [Table 1 – FG-AI4H thematic classification scheme 5](#_Toc52219837)  [Table 2 – ETDRS and International Retinopathy classification scales 8](#_Toc52219838) |

**List of Figures**

| Page |
| --- |
| [Figure 1 – Red eye diagnostic algorithm 14](#_Toc51950723)  [Figure 2 – New algorithm for eye disease diagnostic 15](#_Toc51950724)  [Figure 3 – Gold standard: Iritis 16](#_Toc51950725)  [Figure 4 – Gold standard: Infective conjunctivitis 16](#_Toc51950726) |

# Introduction

As part of the work of the WHO/ITU Focus Group (FG) AI for health (AI4H), this document specifies a standardized benchmarking approach for AI-based applications for Ophthalmology

## Topic Description

This topic group is devoted to standardized benchmarking of artificial intelligence for Ophthalmology. The specific conditions and diseases include Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), Glaucoma (GC), Pathological Myopia (PM) and Red Eye.

Additional diseases and conditions that are relevant to this Topic Group may be added in the future.

DR is a serious eye-disease caused by diabetes that affects blood vessels in the light-sensitive tissue called the retina that lines the back of the eye. It is the most common cause of vision loss among people with diabetes and the leading cause of vision impairment and blindness among working-age adults worldwide.

AMD causes damage to the macula and is a leading cause of vision loss among people age 50 and older. The macula is a small spot near the center of the retina and the part of the eye needed for sharp, central vision, which lets us see objects that are straight ahead. While AMD by itself does not lead to complete blindness but loss of central vision in it can interfere with simple everyday activities.

[GC](https://nei.nih.gov/health/glaucoma/glaucoma_facts) is a group of diseases that damage the eye's optic nerve—the bundle of nerve fibers that connects the eye to the brain and leads to vision loss and blindness. In adults, diabetes nearly doubles the risk of glaucoma.

PM represents a subgroup of myopia and affects up to 3% of the world population. Vision loss related to pathologic myopia is of great clinical significance as it can be progressive, irreversible and affects individuals during their most productive years. High myopia is defined as refractive error of at least -6.00D or an axial length of 26.5mm or more. Pathological or degenerative myopia is defined as "high myopia with any posterior myopia-specific pathology from axial elongation."

Conjunctivitis is the most common cause of red eye. Other causes include blepharitis, corneal abrasion, foreign body, subconjunctival haemorrhage, keratitis, iritis, glaucoma, chemical burn, and scleritis. Although most causes are usually benign and can be managed by primary care physicians, certain uncommon conditions with red eye like keratitis, iritis and glaucoma require early recognition, initiation of treatment and quick referral to a higher centre for appropriate management. There is a high likelihood of complications including irreversible loss of vision if referral is delayed.

### Topic thematic classification

According to the current version of the thematic classification scheme document C-104 of the FG, the categorization of this topic "AI for Ophthalmology " is applicable as described in Table 1.

Table 1 – FG-AI4H thematic classification scheme

| Level | | Thematic classification |
| --- | --- | --- |
| Level 1 | Public Health (Level-1A) | 1.1. Health service  1.2. Health systems  1.10. Non-communicable diseases  sub-classes applicable:  4. health services delivery  6. community health  9. informatics  10. public health interventions  11. public policy |
| Clinical Health  (Level-1B) | 1.1. Prevention  1.2. Diagnosis  sub-classes applicable:  23. Ophthalmology |
| Level-2 Thematic Classification (Artificial Intelligence)  AI-benchmarking class type | | 1. Machine Learning  1.1. Classification  1.2. Regression  1.7. Anomaly detection  5. Perception  5.1. Visual recognition (photo/video) |
| Level-3 Thematic Classification  (nature of data types) | | 3.1. Anonymized Electronic Health Record data  3.2. Medical Images, photographs |

### Relevance

#### Diabetic retinopathy (DR)

The WHO estimates that there are over 422 million people with diabetes worldwide.[[1]](#endnote-1) Of these 35% or over 148 million are estimated to have DR with potential for vision impairment and 11% or 48 million are estimated to have Vision Threating DR (VTDR) that can lead to blindness.[[2]](#endnote-2) Both the number of people with diabetes and those affected by DR are growing at alarming rates – and projected by 2040 to be 642 million with diabetes, 225 million with DR and 64 million with VTDR.

#### Age-related macular degeneration (AMD)

According to Lancet research, the number of people living with macular degeneration is expected to reach 196 million worldwide by 2020 and increase to 288 million by 2040 [[3]](#endnote-3) And AMD is a leading cause (3rd) of vision loss worldwide, by 2010, it has been responsible for approximately 5% of all blindness globally [[4]](#endnote-4). Age is a prominent risk factor for AMD. The risk of getting advanced AMD increases from 2% for those ages 50-59, to nearly 30% for those over the age of 75. Studies suggest in China the prevalence of early AMD in Chinese persons aged 50 years or older was 9.5% and that of late AMD was 1.0%[[5]](#endnote-5).

#### Glaucoma (GC)

There are nearly 40 million blind people in the world today, according to World Health Organization [[6]](#endnote-6). Another 285 million have visual impairment. Globally, 8% of all blindness is attributable to glaucoma, making it the leading cause of global irreversible blindness [[7]](#endnote-7). There were 60 million people with glaucoma in the world in 2010 and will be nearly 80 million by 2020. Of these 60 million, 7.4 million were bilaterally blind from glaucoma in 2010 and 11.2 million (14%) will be bilaterally blind in 2020.

In China, according to a study, it was estimated that 9.4 million (2.6%) people aged 40 years and older have glaucomatous optic neuropathy [[8]](#endnote-8). Of this number, 5.2 million (55%) are blind in at least one eye and 1.7 million (18.1%) are blind in both eyes.

#### Pathological myopia (PM)

PM has become a global burden of public health. Among myopic patients, about 35% have high myopia. Myopia leads to elongation of axial length, potentially causing pathological changes in retina and choroid. With an increase in myopic refraction, high myopia will develop into pathologic myopia, which is characterized by formation of pathologic changes at: (1) posterior pole, including tessellated fundus, posterior staphyloma, retino-choroidal degeneration, etc; (2) optic disc, including parapapillary atrophy, tilting, etc; (3) myopic maculopathy, including lacquer crack, Fuchs spot, CNV, etc. Pathologic myopia causes irreversible visual impairment to patients. Therefore, it is important to have early diagnosis and regular follow-up.

The overall global prevalence is estimated to be 0.9-3.1% with regional variability. The prevalence of pathological myopia-related visual impairment has been reported as 0.1%-0.5% in European studies and 0.2% to 1.4% in Asian studies.

#### Red eye (RE)

Eye problems are the reason for 2-3% visits to primary health centres and emergency facilities, the majority of which are cases of red eye.[[9]](#endnote-9) Red eye is one of the most common problems seen in eye clinics in developing countries. The majority of red eye cases are seen at community clinics, primary health centres and health sub centres, where diagnosis and management are done by primary care physician, community health nurses, midwives and health workers.

Recognising the need for emergent referrals to an ophthalmologist for some causes is the key in the primary care management of red eye. If primary health care workers can accurately diagnose the cause of red eye and provide primary level treatment, then patients can be managed quicker and closer to where they live. Furthermore, secondary centres will be relieved of treating simple conditions, allowing more time and resources for eye conditions that need the attention of specialists.

### Current approaches and gold standards for detection

#### DR

DR detection/diagnosis requires capturing a photograph of the retina using specialized equipment such as a slit-lamp and fundus camera. The image is then examined by an ophthalmologist, optometrist or a trained professional to detect abnormalities such as microaneurysms, exudates, haemorrhages, diabetic macular edema (DME), etc. to determine if DR and/or DME is present and its severity and stage of progression.

In general DR can be classified as mild, moderate or vision-threatening, which includes severe non-proliferative DR (NPDR), proliferative DR (PDR) and diabetic macular edema (DME). Accurate diagnosis of DR from fundus camera images and grading its severity requires professional expertise and training.

Diabetic macular edema (DME) is an additional condition that can occur independently of whether or not DR is present and at any DR severity level. DME therefore has a separate diagnosis and usually has 3 classification levels: no DME, noncentral-involved DME, or central-involved DME. The presence or absence of DME and its severity level, along with presence or absence of DR and its severity level, are taken together to determine the recommended course of action, treatment and whether or not a referral is required to an ophthalmologist.

The outcome of screening for DR is sometimes termed as either non-referrable DR (which includes no DR, as well as mild NPDR, and no DME), or referrable DR (which includes moderate NPDR, severe PDR, PDR, or presence of DME). This classification approach is useful in screening and is currently used by some AI systems, rather than the more granular grading of DR severity levels.

The two most commonly used classification systems for grading severity of diabetic retinopathy are the simplified Early Treatment Diabetic Retinopathy Study (ETDRS) scale and the International Clinical Diabetic Retinopathy Disease Severity Scale. The classification systems distinguish different levels of non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR).

The International Scale is easier to use because it is a simpler system and does not rely on reference images from the Airlie House classification system. Though the two grading scales are similar, they are not interchangeable. Both scales are summarized in Table 1.

Table 2 – ETDRS and International Retinopathy classification scales

| Diabetic Retinopathy Grade | Simplified ETDRS Scale | International Scale |
| --- | --- | --- |
| No apparent DR |  | No abnormalities |
| Mild NPDR | At least one MA but no H/MA ≥ standard photo 2A | MA only |
| Moderate NPDR | H/MA ≥ standard photo 2A, and/or CWS, VB, IRMA but NOT satisfying criteria for severe NPDR | More than just MA but less than severe NPDR |
| Severe NPDR | One or more of the following:   * H/MA ≥ standard photo 2A in all 4 quadrants * VB in at least 2 quadrants * IRMA ≥ standard photo 8A in at least 1 quadrant | Any of the following (4-2-1 rule) and no PDR   * Severe H in each quadrant * VB in 2 or more quadrants * IRMA in 1 or more quadrants |
| PDR |  | One or more of the following:   * Neovascularization * VH or PRH |
| Early PDR | New vessels and definition not met for high risk PDR |  |
| High risk PDR | One or more of the following:   * NVD >1/3 DD * NVD with VH or PRH * NVE >1/4 DD and VH or PRH |  |

**Abbreviations:** CWS: cotton-wool spots; H: Hemorrhages; IRMA: Intraretinal microvascular abnormalities; MA :Microaneurysms; NVD: New vessels at the optic disc; NVE: new vessels elsewhere; PRH: preretinal hemorrhage; VB: Venous Beading ; VH: vitreous hemorrhage ;  
ETDRS – Early Treatment diabetic retinopathy Study; NPDR – Non-proliferative diabetic retinopathy; PDR – Proliferative diabetic retinopathy; DME – Diabetic Macular Edema; CSME – Clinically significant macular edema; FDP – Fibrous proliferations disc; FPE – Fibrous proliferations elsewhere; DD – Disc diameter;

The UK National Institute for Clinical Excellence (NICE) guideline states that a DR screening test should have sensitivity and specificity of at least 80% and 95% respectively, with a technical failure rate of less than 5%.[[10]](#endnote-10)

The gold standard photography method for the detection of DR is stereoscopic color fundus photography in 7 standard fields (30°) as defined by the Early Treatment Diabetic Retinopathy Study (ETDRS) group. [[11]](#endnote-11)

#### AMD

The initial evaluation of a patient with signs and symptoms suggestive of AMD includes all features of the comprehensive adult medical eye evaluation, with particular attention to those aspects relevant to AMD. In the physical examination, stereoscopic biomicroscopic examination of the macula is usually needed. Binocular slit-lamp bio microscopy of the ocular fundus is often necessary to detect subtle signs of choroidal neovascularization. These includes small areas of haemorrhage, hard exudates, subretinal fluid, macular edema, subretinal fibrosis, or pigment epithelial elevation. Optical coherence tomography is important in diagnosis and managing AMD, particularly in determining the presence of subretinal fluid and in documenting the degree of retinal thickening. Fundus photography may be obtained when fluorescein angiography is performed, because they are useful in finding landmarks, evaluating serous detachments and determining the etiology of blocked fluorescence. Fundus photographs may also be used as a baseline reference for selected patients with advanced non-neovascular AMD and for follow-up of treated patients.

There are lots of classifications of AMD in the literature. The classification from the Age-related Eye Disease Study (AREDS) is as follows: no AMD, early AMD, intermediate AMD and advanced AMD.

#### GC

Glaucoma diagnosis mainly contains the following aspects: 1) measuring intraocular pressure (tonometry); 2) Testing for optic nerve damage with a dilated eye examination and imaging tests; 3) checking for areas of vision loss (visual field test); 4) measuring corneal thickness (pachymetry); 5) inspecting the drainage angle (gonioscopy). Fundus photography and optical coherence tomography are often used in this course.

Glaucoma can be generally classified as primary glaucoma, secondary glaucoma and congenital glaucoma. Among them, primary glaucoma contains primary angle-closure glaucoma and primary open angle glaucoma.

(Todo: current approaches for other conditions to be added: PM/Red Eye)

### Impact of AI

#### DR, AMD, GC, PM

For all the diseases described above, vision loss and blindness can be delayed or prevented by early detection and treatment of the condition. This requires an examination and screening by a trained ophthalmologist or an eye care professional.

However, given the large numbers of people affected worldwide by these conditions, there are not sufficient specialists globally to screen everyone at risk. The shortfall is particularly acute in developing countries, including India, China and many countries in Asia and Africa. In addition to the dire shortage of trained professionals, many of the affected people live in remote areas with little or no access to an eye care clinic or a screening centre.

In India, for example, there are over 72 million people with diabetes and an estimated 25 million have some stage of DR and about 7 million have VTDR. However, India only has 15,000 trained ophthalmologists, which in a nation with 1.3 billion people amounts to a mere 9 specialists per million. Kenya, with a population of 48 million has less than 100 ophthalmologists, and Angola, less than 20 for 29 million people.[[12]](#endnote-12)

*(The following is from Calligo Technologies contribution FG-AI4H-G-028 received during meeting G.)*

With the advent of Edge Computing, health care industry has transformed itself considerably, while hospitals and clinics are gearing up to take better and faster care of their patients. In fact, Edge Computing has permeated the industry in such a powerful manner that clinicians and doctors heavily rely on them to treat patients. As more and more devices get connected in the health care industry, networking among them all has really become huge because the data that keeps comes in is never going to slow down.

A frequent problem in mass eyecare check-ups is that the quality of images captured might not always be usable for an ophthalmologist to grade for diabetic retinopathy. In such situations, the patients are asked to come back and undergo the process again. Now with AI, the system checks the image as soon as it is clicked and prompts the technician to click another image in case it is not good enough. Now, even a minimally skilled technician can take usable images of the eye fundus.

Once usable images are captured, the system grades the images, again in real-time, and identifies if the images have diabetic retinopathy. In case a patient is found to be diabetic retinopathy positive, they are advised to consult an ophthalmologist to determine the next course of action.

Checking on patients with high risk problems and ensuring a more effective, customized treatment approach can thus be facilitated. Lack of data makes the creation of patient-centric care programs more difficult, so one can clearly understand why utilizing big data can be so highly important in the industry.

#### Red eye

In the case of red eye detection, an AI solution could be aimed at providing better eye care to those living in rural areas. In most rural health centres, the healthcare provider may be a nurse, midwife or a non-ophthalmologist doctor. This holds true in most developing countries.

An AI solution would be able to accurately diagnose the cause of red eye and recommend treatment or referral to an expert ophthalmologist. Such a solution would also increase the efficiency of treatment of red eye cases in rural centres.

AI and deep learning-based systems offer the following benefits:

* Bridge the shortage of healthcare professionals and provide access to screening where none exists.
* Increase overall efficiency and scalability of current screening methods.
* Provide earlier detection of many eye diseases thereby preventing vision loss for millions.
* Decrease overall health-care costs via earlier interventions when it is easier and less expensive to treat these diseases.

### Impact of benchmarking AI Solutions

An accurate way of benchmarking the performance of AI solutions to detect and diagnose DR, AMD, GC, PM and Red Eye can have a major impact on selecting and implementing the best solution to address the global healthcare challenge posed by these diseases specially in the LMICs. This can in turn improve the lives of millions at risk for vison impairment and vision loss globally because they do not have access to human experts and infrastructure to get screened. This also fulfils the important objective of achieving the UN's SDGs in health.

## Ethical considerations

(to be completed)

* Data ownership
* Data security and privacy
* Related regulations and laws
* Responsibilities
* Algorithm bias

## Existing AI solutions (includes datasets, systems and benchmarks)

### DR

#### DR datasets

* Publicly available datasets include the EyePACS dataset (around 90,000 fundus images, 5 levels of severity), [[13]](#endnote-13)
* MESSIDOR dataset (1,200 images, 4 levels of severity), [[14]](#endnote-14)
* The DIARETDB dataset (around 200 images marked with lesions), etc.[[15]](#endnote-15)
* High-Resolution Fundus (HRF) Image Database.[[16]](#endnote-16)
* Diabetic Retinopathy datasets from Kaggle:
* Kaggle DR Challenge 2015: 35,000 images of DR classified into 5 levels of severity (No DR, Mild, Moderate, Severe, Proliferative DR).
* APTOS 2019 Blindness Detection Challenge: 3664 Images classified into 5 levels of severity ((No DR, Mild, Moderate, Severe, Proliferative DR).

#### DR systems and benchmarks

A team at Google published results in 2016 of a study for detecting DR working with doctors in India and the US. The results show that their AI model's performance for DR detection and grading its severity was on-par with that of ophthalmologists. Their model had a combined accuracy score of 0.95, which was slightly better than the median of the 8 ophthalmologists consulted (measured at 0.91). [[17]](#endnote-17)

Currently, IDx-DR is the first FDA approved device for AI DR screening. Based on a customized CNN architecture and lesion characteristics, this device can achieve a sensitivity of 96.8% and a specificity of 87%.[[18]](#endnote-18)

The best reported performance on binary classification of no DR/non-referable DR vs. referable DR is a sensitivity of 94% and specificity of 98% .[[19]](#endnote-19)

This work combined features both from deep ResNet and from metadata, and classified the features with a gradient boosting decision tree.

For five level classification of no DR, mild, moderate, severe non-proliferative DR, and proliferative DR [[20]](#endnote-20) [[21]](#endnote-21) [[22]](#endnote-22), the best accuracy reported is 96% by a combination of GoogleNet and ResNet model.

In the APTOS 2019 Blindness Detection challenge organized by Kaggle in Sep 2019, 2931 teams competed, and the top solution achieved a Quadratic Kappa weighted score of 0.9361 on an undisclosed test data set.

"Calligo Health Engine" is an Edge Analytics solution which is easy to use, industry gradable, low cost & low resource and is capable of identifying Diabetic Retinopathy using Artificial Intelligence with an accuracy **over 96%** and within **Seconds**.

### GC

#### GC datasets

* Online retinal fundus image dataset for glaucoma Analysis (ORIGA, 650 fundus images)
* Retinal fundus images for glaucoma analysis (RIGA, 760 fundus images)
* ACHIKO-K (258 fundus images)
* DRISHTI-GS (100 images mainly for optic disk and cup segmentation)
* Glaucoma Dataset from iChallenge:
* Retinal Fundus Glaucoma Challenge dataset (REFUGE/iChallenge-GON, 1200 fundus images, 2 tasks: classification of clinical glaucoma; segmentation of optic disc and cup)
* Angle closure Glaucoma Evaluation Challenge dataset (AGE/iChallenge-PACG, 4800 AS-OCT images, 2 tasks: angle closure classification; scleral spur localization)

AI practice on suspected glaucoma classification generally follow two approaches, i.e. an end-to-end whole image classification [[23]](#endnote-23) [[24]](#endnote-24), or a classification based on optic disk and cup information.[[25]](#endnote-25) For the end-to-end approach, a resulting AUC of 0.986 by training an inception-v3 network on their private dataset of 48000+ images was reported. [[26]](#endnote-26) A multitask deep CNN model based on a U-net sharing features for the glaucoma classification task was set up and the disc and cup segmentation task, achieving an AUC of 0.95 while providing some medical interpretability.

### AMD

Recently Automatic Detection challenge on Age-related Macular degeneration (ADAM) has been held. The ADAM challenge focuses on the investigation and development of algorithms associated with diagnosis of AMD and segmentation of lesions in fundus images. The challenge has 4 tasks: classification of AMD and non-AMD fundus images; detection and segmentation of optic disc; localization of fovea; detection and segmentation of lesions from fundus images. ADAM dataset contains 1200 fundus images.

Currently, most existing work of detecting AMD in fundus images addresses the problem as a binary classification between no/early stage AMD and intermediate/advanced stage AMD. The two commonly used datasets are the Age-Related Eye Disease Study (AREDS) dataset, which consists of fundus images from around 4,700 participants, and the Cooperative Health Research in the Region of Augsburg (KORA) dataset, which consists of fundus images from 2,840 patients. Most state-of-the-art methods for AMD binary classification are in one of the three following categories:

1. Using CNNs of existing architectures such as GoogleNet, VGG, etc. [[27]](#endnote-27) [[28]](#endnote-28). The best reported performance of this type of method is 94.3% accuracy, using an ensemble of several CNNs.
2. Using customized deep CNN models [[29]](#endnote-29) [[30]](#endnote-30) [[31]](#endnote-31). The best reported result is an AUC of 0.96 and an accuracy of 91.6% on AREDS dataset.
3. Using deep image features from a pretrained CNN model and then classifying with a Support Vector Machine or Random Forest based model [[32]](#endnote-32) [[33]](#endnote-33)The best reported accuracy is 93.4%.

### PM

Now there is only the PALM challenge which focuses on the investigation and development of algorithms associated with the diagnosis of Pathological Myopia (PM) and segmentation of lesions in fundus photos from PM patients. The goal of the challenge is to evaluate and compare automated algorithms for the detection of pathological myopia on a common dataset of retinal fundus images. The medical image analysis community were invited to participate for developing and testing existing and novel automated fundus classification and segmentation methods. This challenge has 4 tasks: classification of PM and non-PM (including high myopia and normal) fundus images; detection and segmentation of disc; localization of fovea; detection and segmentation of retinal lesions (atrophy and detachment) from fundus images. PALM dataset contains 1200 fundus images.

### Red eye

(the document sections on Red Eye have been contributed by Parvathi Ram and Dr. Suneetha N, St. John's Medical College, India based on their submission FG-AI4H-G-030-R19)

From a preliminary review, no previous studies concerned with AI-based etiological diagnosis of anterior segment conditions of the eye were found. However, the following algorithm may serve as a starting point for development of an AI based system for Red Eye detection and diagnosis.

#### The Edinburgh red eye diagnostic algorithm

The Edinburgh Red Eye diagnostic algorithm was designed by Timlin et. al. to assist clinicians referring patients to the acute ophthalmology service within Edinburgh. This algorithm aims to aid primary care physicians to diagnose anterior segment conditions resulting in red eye in the same way an experienced ophthalmologist approaches such a patient- analysing the symptoms and signs and using a combination of pattern recognition and deductive reasoning to arrive at a diagnosis.

The accuracy of this algorithm was tested by analysing the concordance between the algorithm-assisted diagnosis (made by primary care physicians) and the 'gold-standard' diagnosis (made by expert ophthalmologists).

The results showed a 72% diagnostic accuracy for the Edinburgh red eye diagnostic algorithm, which rises to 76% when only severe eye conditions are included (iritis, keratitis and AACG). The algorithm is depicted in Figure 1.

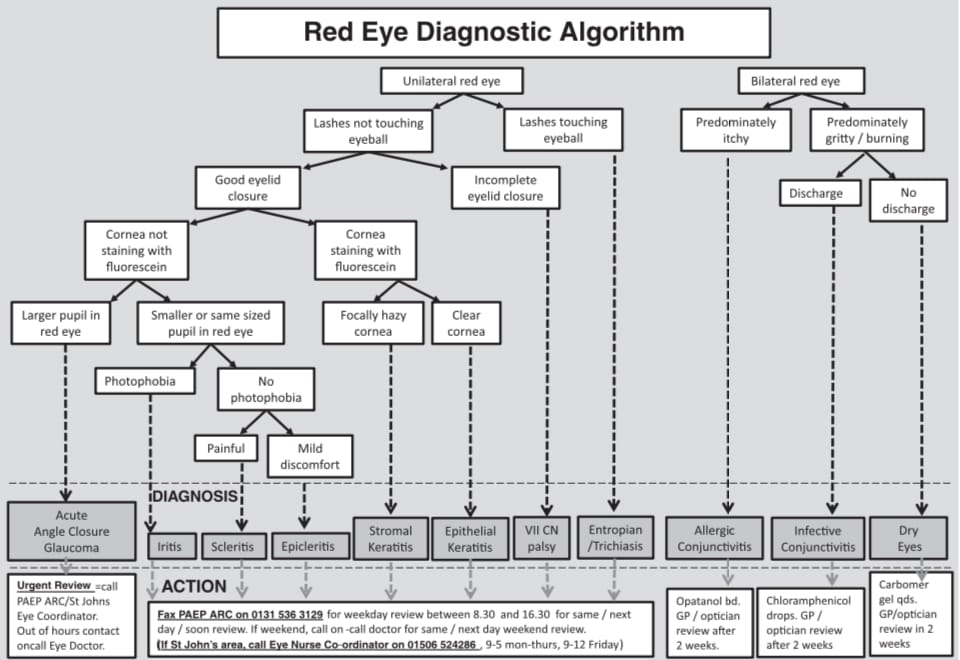


Figure 1 – Edinburgh red eye diagnostic algorithm

#### Red eye datasets

Currently, no known public datasets are available for Red Eye diagnosis and detection. In their contribution Parvathi Ram and Dr Suneetha N. state:

"The images are to be obtained from the St John's Medical College Hospital Ophthalmology Department. We are making and annotating a dataset of red eye images for this study. Data collection has started, and we have annotated and labelled 30 images. A new algorithm has been developed and has been applied this to the images in the dataset (Figure 2). Currently there are no available datasets of red eye images, as concluded from a preliminary literature survey. Currently the test data cannot be made available to individuals outside SJNAHS, because of ethical concerns and policies. However, we may be open to contributing to an open database in the future."

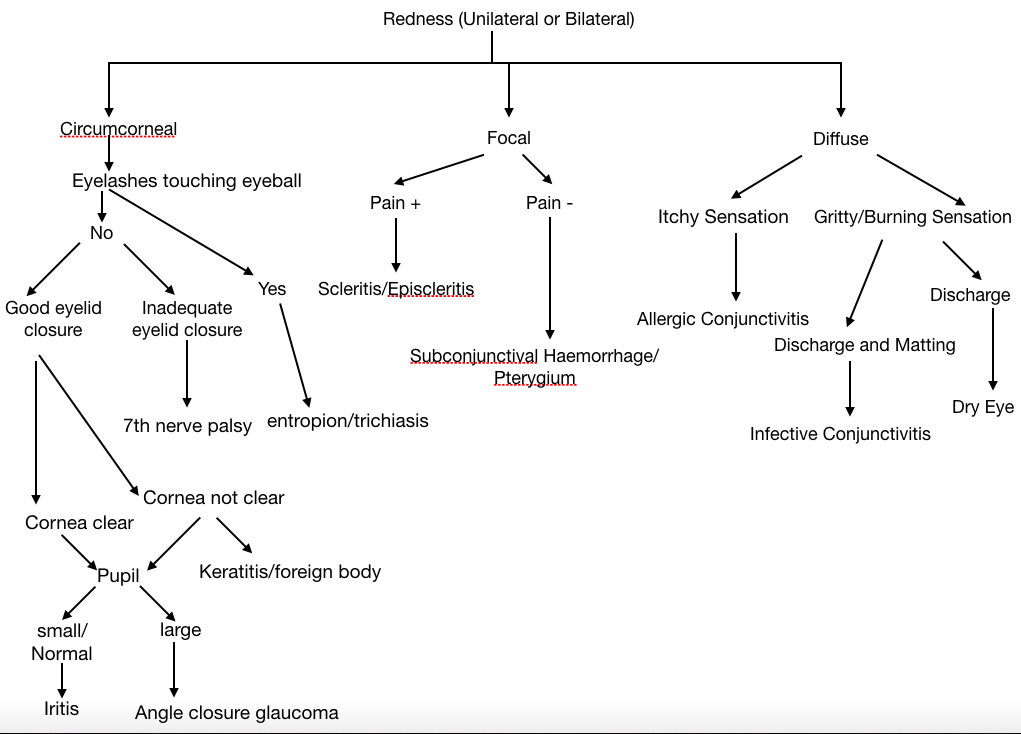


Figure 2 – New algorithm for red eye disease diagnostic

#### Data quality

The images are captured and annotated by expert ophthalmologists at SJMCH. We are interested in collaborating with other institutes to expand our dataset.

#### Annotation/label quality

The annotations are relevant, of high quality and are made by expert ophthalmologists at SJMCH.

We have developed a new algorithm instead of using the original Edinburgh red eye algorithm to take into consideration:

* Eye conditions more commonly seen in tropical countries such as India
* Parameters that would be easier to incorporate into an AI algorithm

Examples of application of the algorithm are shown in Figure 3 and Figure 4.

|  |  |
| --- | --- |
| Figure 3 – Gold standard: Iritis | Figure 4 – Gold standard: Infective conjunctivitis |

In Figure 3, the examination shows that the patient's eyelashes are not touching the eyeball and the patient has good eyelid closure. From the image it is noted that the patient has circumcorneal redness and a clear cornea. Therefore, the flowchart reveals the correct diagnosis of Iritis.

In Figure 4, data collected at the time of examination shows that the patient does not experience itchiness. From the image it is noted that the patient has diffuse redness. Therefore, the flowchart reveals the correct diagnosis of Infective Conjunctivitis.

#### Data provenance

The data is being collected in a professional and ethical way as per the guidelines set down by the Institutional Ethics Committee at SJMC. The data comes from both urban and rural clinical backgrounds as we are affiliated with various rural healthcare centres, from where we will be obtaining data.

#### Benchmarking

As a pilot study for the new algorithm (described in Clause 8), we have reviewed 20 images of patients seen in the SJMCH outpatient department. Comparing with gold standard diagnoses of expert ophthalmologists, 15 of those 20 images were found to be accurately diagnosed by a medical student using the algorithm provided.

# AI4H Topic Group

Over the past decade, considerable resources have been allocated to exploring the use of AI for health, which has revealed an immense potential. Yet, due to the complexity of AI models, it is difficult to understand their strengths, weaknesses, and limitations. If the technology is poorly designed or the underlying training data are biased or incomplete, errors or problematic results can occur. AI technology can only be used with complete confidence if it has been quality controlled through a rigorous evaluation in a standardized way. Towards developing this standard assessment framework of AI for health, the ITU has established FG-AI4H in partnership with the WHO.

Thus far, FG-AI4H has established thirteen (20) topic groups. These are concerned with: AI and cardiovascular disease risk prediction, dermatology, falls among the elderly, histopathology, neuro-cognitive disorders, outbreak detection, ophthalmology, psychiatry, radiotherapy, snakebite and snake identification, symptom assessment, tuberculosis and Volumetric chest computed tomography. A current list can be found at the FGAI4H website at <https://itu.int/go/fgai4h>.

As the work by the Focus Group continues, new Topic Groups will be created. To organize the Topic Groups, for each topic the Focus Group chose a topic driver. The exact responsibilities of the topic driver are still to be defined and are likely to change over time. The preliminary and yet-to-confirm list of the responsibilities includes:

* Creating the initial draft version(s) of the topic description document.
* Reviewing the input documents for the topic and moderating the integration in a dedicated session at each Focus Group meeting.
* Organizing regular phone calls to coordinate work on the topic description document between meetings.

During meeting C in Lausanne, Arun Shroff from Medindia/Xtend.AI was selected as topic driver for the "Topic Group - AI for Ophthalmology

## General mandate of the Topic Group

The Topic Group is a concept specific to the AI4H-FG. The preliminary responsibilities of the Topic Groups are:

1. Provide a forum for open communication among various stakeholders
2. Agree upon the benchmarking tasks of this topic and scoring metrics
3. Facilitate the collection of high-quality labelled test data from different sources
4. Clarify the input and output format of the test data
5. Define and set-up the technical benchmarking infrastructure
6. Coordinate the benchmarking process in collaboration with the Focus Group management and working groups

## Topic description document

The primary output of each Topic Group is the topic description document (TDD) specifying all relevant aspects of the benchmarking for the individual topics. **This document is the TDD for the Topic Group on "AI for Ophthalmology)" (TG-Ophthalmo)**. The document has been developed over several FG-AI4H meetings starting from meeting D in Shanghai. Any contributions from members have been incorporated in the relevant sections. Suggested changes to the document will be submitted as input documents for each meeting. The relevant changes will then be discussed and integrated into an official output document until the TDD ready for the first official benchmarking.

## Subtopics

Topic groups summarize similar AI benchmarking use cases to limit the number of use case specific meetings at the Focus Group meetings and to share similar parts of the benchmarking. However, in some cases, it is expected that inside a Topic Group different subtopic groups can be established to pursue different topic-specific specializations. TG-Ophthalmo will start without separate subtopic Groups. They will be introduced once participants with different benchmarking requirements join the Topic Group. It is expected to introduce subtopic Groups for "DR", "AMD", "Glaucoma", Pathological Myopia and "Red Eye" which may be added as soon as partners that are interested in benchmarking systems for these conditions.

## Topic group participation

The participation in both the focus and Topic Group is generally open and free of charge. Anyone who is from a member country of the ITU may participate. On the 14. of March 2019 the ITU published an official "call for participation" document outlining the process for joining the Focus Group and the Topic Group. For this topic, the corresponding call can be found [here](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/ophthalmo.aspx).

## Status of this Topic Group

During meeting D it was discussed that the TDD should contain an explicit section describing the progress since the last meeting for the upcoming meeting. The following subsections serve this purpose:

### Status update for meeting D (Shanghai)

1. The first draft of the Topic Group Document was created and shared with members of the topic group.
2. We promoted the topic group via email and our networks to invite anyone interested to join and contribute.
3. Topic Group Member, Baidu.com provided some suggestions on including AMD, Glaucoma and other imaging methods that were incorporated into the TDD
4. The first version of the TDD was published (as [FGAI4H-D-038](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-D-038.docx)).

### Status update for Meeting E (Geneva)

1. The updated Call for Topic Group participation for TG-Ophthalmo was published on the ITU website and can be [downloaded here](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Ophthalmo.pdf).
2. We had several email exchanges with the topic group members to request inputs and updates to the TDD. Yanwu XU indicated that Dr Xingxing Cao from Baidu would provide topic group updates on their behalf.
3. Dr. Xingxing Cao, Baidu added the condition Pathological Myopia (PM) to the TDD.
4. An updated and revised TDD was published (as [FGAI4H-E-014-R01](https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/docs/FGAI4H-E-014-R01.docx))
5. We reached out to our networks via email and social media (LinkedIn, Twitter), sharing the call for topic group participation and to spread the word.
6. We had email exchanges, calls and discussions with several groups and individuals interested in contributing to the topic group including the following:

* Pearse A Keane, MD MSc FRC Ophth MRCSI, Consultant Ophthalmologist, Moorfields Eye Hospital, U.K.
* Prof Leo Celi M.D. M.S. M.P.H., Clinical Research Director, Laboratory for Computational Physiology, Harvard-MIT Division of Health Science and Technology, Open Access (MIT) Project
* Ashley Kras, M.D. M. S., Ophthalmologist & Bioinformatician (Harvard Medical School)

### Status update for Meeting F (Zanzibar)

1. We continued to promote the topic group via email to our networks and via social media.
2. As a result, we received several inbound emails and interest in joining/contributing to the group. We had two new Topic group members who joined the group:

* Dr Covadonga Bascaran, PHEC MSc Programme Director, International Centre for Eye Health (ICEH), London School of Hygiene & Tropical Medicine
* Inês Sousa, Head of Intelligent Systems, Fraunhofer.

1. We had several online Meetings / calls and discussion with:

* Prof Leo Celi, Clinical Research Director, Harvard MIT Division of Health Science and Technology and Ash Krasley: (online meeting on June 22, 2019)
* Details about MIT Open Access Project
* Potential collaboration with FGAI4H / Contribution of Data
* Dr. Jorge Cuadros, O.D., PhD. (CEO), Founder & CEO of EyePACS LLC, USA (online meeting on July 22, 2019).
* EyePACS LLC is a US company with a telemedicine platform for ophthalmology deployed in more than 600 community health and primary care centres across the US and internationally.
* They have a database of over 5 million diagnostic retinal images. We discussed possibility of collaborating with the topic group and contributing data. They expressed interest and discussions are ongoing.

### Status update for Meeting G (New Delhi)

1. We have received two new submissions to the Topic Group as potential sub-topics:

* "Proposal for sub-topic - AI based Aetiological Classification of Red Eye" from Ms. Parvathi Ram and Dr. Suneetha N, St. John's Medical College, India
* "Leveraging Edge analytics and Artificial Intelligence for the rapid assessment of avoidable blindness" from Rajaraman Subramanian, Sriganesh Rao, Calligo Technologies and Sushil Kumar TEC, New Delhi, India

1. Topic Group Meeting (Nov 7, 2019)

* Participants:
* Parvathi Ram, Medical Student, St. John's Medical College, India
* Rami Verbin, IT Professional, Israel,
* Discussion:
* Overview and status of the TDD and sections requiring updates.
* Discussion with Parvathi Ram about her submission to the topic group and possibility of making it into a subtopic, based on obtaining labeled datasets for redeye.
* Need for undisclosed datasets for topic group

1. New members to the topic group:

* Parvathi Ram, St. John's Medical College, India
* Dr. Suneetha N, St John's Medical College, India
* Dr. Sheila John, Sankara Netralaya, Chennai, India
* Rajaraman Subramanian, Calligo Technologies, India
* Sriganesh Rao, Calligo Technologies, India
* Sushil Kumar TEC, New Delhi India

### Status update for Meeting H (Brasilia)

1. We continued to promote the topic group via email to our networks and via social media.
2. We have a new dedicated mailing list for the topic group: [fgai4htgophthalmo@lists.itu.int](mailto:fgai4htgophthalmo@lists.itu.int)
3. An email was sent to all topic group members and an online meeting scheduled. However, we received no inputs or contributions from current members, no members attended the online meeting, and no new members have joined the group since Meeting G.
4. We updated the TDD to incorporate the submissions received during Meeting G:

* FGAI4H-G-030-R01 (St, John's Medical College) on Red Eye incorporated into relevant sections.
* FG-AI4H-G-028 (Calligo Technologies) on Leveraging Edge Analytics incorporated into current AI systems overview.

1. Other TDD updates:
2. Topic Group Thematic Classification updated.
3. Added Quadratic Kappa Metric for multi-label classification.
4. Added Kaggle DR challenge datasets and results
5. Miscellaneous edits/corrections

### Status update for Meeting I (E-meeting)

1. Outreach via email, social media.
2. Emails to all topic group members requesting inputs and contributions.
3. Updates to TDD received from Xingxing Cao, Baidu, Rajaraman Subramanian, Calligo Technologies and Parvathi Ram of St. John's Medical College, India
4. Calls with :
5. Dr. Jorge Cuadros, EyePACS, CA, to discuss collaboration for obtaining datasets for testing.
6. José Tomás Arenas C., Co-Founder & CEO of TeleDx.Org, to bring them on board to the topic group and explore opportunities for collaboration in South Americas
7. Rajaraman Subramanian and Sriganesh Rao, Calligo Technologies to discuss TDD updates, subtopic creation.

### Status update for Meeting J (E-meeting)

1. Outreach via email, social media.
2. Collaboration with the Data and AI solution assessment methods workgroup (WG-DAISAM) : Working to provide inputs for evaluation that go beyond performance of the TG-Ophthalmo AI models, to assess factors such as bias, robustness, explainability and uncertainty. The goal is to better understand the process, and submit the results of the analysis to the ML4H workshop at the NeurIPS conference.
3. Received two contributions during Meeting I from Tencent Healthcare (China) on

* FGAI4H-I-041 - Evaluation method and index of artificial intelligence glaucoma assisted screening system based on fundus image
* FGAI4H-I-040 - Data set construction and annotation of artificial intelligence assisted screening system based on fundus image

We are working with them to incorporate the content of both the proposals above into relevant sections of this TDD.

1. Seven new members have joined the TG since last meeting:
2. Daniel Ting MD (1st Hons) PhD, Consultant, Vitreo-retinal Service, Singapore National Eye Center, Head, AI and Digital Innovation, Singapore Eye Research Institute
3. Dr. Karthik Srinivasan, Medical Officer, Vitreo retinal Services, Aravind Eye Hospital, Chennai.
4. João Victor Dias, Lead Data Scientist, NTT Data Brazil, Artificial Intelligence for HealthTech and Financial Machine Learning
5. Jianrong Wu, Yanchun Zhu, Man Tat Alexander Ng and Yajun Zhang, Tencent Healthcare (Shenzhen), China

### Current members of the topic group

1. Arun Shroff, CEO, Xtend.AI (USA)& CTO, Medindia.net (India) Topic Driver
2. Yanwu XU, Intilligent Healthcare Unit, Chief Scientist, Baidu, China
3. Xingxing Cao, Intilligent Healthcare Unit, Baidu, China
4. Jingyu WANG, Artificial Intelligence Group, Baidu, China
5. Shan Xu, CAICT, China
6. Ashley Kras, M.D. M. S., Ophthalmologist & Bioinformatician (Harvard Medical School)
7. Dr Covadonga Bascaran, PHEC MSc Programme Director, International Centre for Eye Health (ICEH), London School of Hygiene & Tropical Medicine
8. Inês Sousa, Head of Intelligent Systems, Fraunhofer.
9. Parvathi Ram, St. John's Medical College, India
10. Dr. Suneetha N, St John's Medical College, India
11. Dr. Sheila John, Sankara Netralaya, Chennai, India
12. Rajaraman Subramanian, Calligo Technologies, India
13. Sriganesh Rao, Calligo Technologies, India
14. Sushil Kumar TEC, New Delhi India
15. José Tomás Arenas C., Ricoleta, Chile
16. Daniel Ting MD, PhD, Consultant, Vitreo-retinal Service, Singapore National Eye Center, Head, AI and Digital Innovation, Singapore Eye Research Institute
17. Dr. Karthik Srinivasan, Medical Officer, Vitreo retinal Services, Aravind Eye Hospital, Chennai.
18. João Victor Dias, Lead Data Scientist, NTT Data Brazil, GeekVision (São Paulo), Brazil.
19. Jianrong Wu, Tencent Healthcare (Shenzhen), China
20. Yanchun Zhu, Tencent Healthcare (Shenzhen), China
21. Man Tat Alexander Ng, Tencent Healthcare (Shenzhen), China
22. Yajun Zhang, Tencent Healthcare (Shenzhen), China

## Next meetings

The Focus Groups meets about every two months at changing locations. The upcoming meetings are:

* K: November 2020 – TBC. An up to date list can be found at the official [ITU FG AI4H website](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx).

Tools/process of TG cooperation - to be filled out according to FG regulations

TG interaction with WG, FG- to be filled out according to FG regulations

# Method

## Overview of the benchmarking

The benchmarking of the algorithms for detecting DR, GC, AMD will be done on a sufficiently large and previously undisclosed test data set. All data will be provided as per the data acceptance guidelines published by the focus group. All data will be labelled by licensed ophthalmologists or eye-care professionals.

(TBC: The availability of data and benchmarking for Red Eye needs to be determined and added)

## AI input data structure

The following input data structure is being proposed for the following eye conditions - DR, AMD, GC. The input data structures for PM and Red Eye have not yet been defined.

Images of each retina captured with fundus cameras should be submitted as separate files in the following format:

* Image File Format: JPG or PNG format
* Image File Names: Images names will be anonymised to exclude any patient identifying information.
* Image Resolution: the images will be supplied in their original resolution as captured from the fundus cameras.

## AI output data structure

The output of the algorithm should be a CSV file in text format with the following columns:

* Name of the image file processed (for example: I705656.JPG OR L566768.PNG)
* The diagnosis of the retinal image as per the algorithm. The labels will depend upon the specific condition and the type of classification that is being benchmarked:

1. **DR:**
2. Classification: All DR severity levels:

* 0 (Non-gradable Image) (TBD: should this be one of the classifications.)
* 1(No DR)
* 2 (Mild)
* 3 (Moderate NPDR)
* 4 (Severe NPDR)
* 5 (PDR)

1. Binary Classification: Referable or Non-referable DR:

* 0 (Ungradable Image. TBD)
* 1 (Non-referable Retinopathy – No DR or Mild DR)
* 2 (Referable Retinopathy - Moderate, Severe, PDR)

1. **AMD:**

* 0 (Image Ungradable. TBD)
* 1 (No/early stage AMD
* 2 (Intermediate/advanced stage AMD)

1. **GC:**

* 0 (Image Ungradable. TBD)
* 1 (No GC)
* 2 (GC)

1. **PM:**

* 0 (Image Ungradable. TBD)
* 1 (No PM/HM)
* 2 (HM: high myopia)
* 3 (PM)

The output data structures for PM and Red Eye have not yet been defined.

## Test data labels

A separate CSV file in text format will be provided containing the following columns:

* Name of the Image File (example: R705656.JPG OR L566768.PNG)
* Label or Annotation of the Image that contains the diagnosis of the retinal image. The labels will depend upon the specific condition that is being benchmarked and also the type of classification. Currently, the following are being proposed:

1. **DR:**
2. Classification: All DR severity levels:

* 0 (Non-gradable Image) (TBD: should this be one of the classifications.)
* 1(No DR)
* 2 (Mild)
* 3 (Moderate NPDR)
* 4 (Severe NPDR)
* 5 (PDR)

1. Classification: Referable or Non-referable DR:

* 0 (Ungradable Image. TBD)
* 1 (Non-referable Retinopathy – No DR or Mild DR)
* 2 (Referable Retinopathy - Moderate, Severe, PDR)

1. **AMD:**

* 0 (Image Ungradable. TBD)
* 1 (No/early stage AMD
* 2 (Intermediate/advanced stage AMD)

1. **GC:**

* 0 (Image Ungradable. TBD)
* 1 (No GC)
* 2 (GC)

1. **PM:**

* 0 (Image Ungradable. TBD)
* 1 (No PM/HM)
* 2 (HM: high myopia)
* 3 (PM)

## Scores and metrics

All metrics will be computed based on the performance of the algorithm on the undisclosed test data-set. The scores and metrics used for benchmarking AI will depend upon the type of task performed by the AI, which for the conditions in this topic group would generally be either classification or segmentation.

### Classification tasks

Classification of the conditions being considered may be either binary (2 classes) – for example DR or no DR, GC or no GC, or multi-class - for example, in the case of DR an image may be classified as having No DR, mild, moderate, severe or PDR (5 classes). In addition for DR, the absence of presence of DME (Diabetic Macular Edema) may be required

In the case of segmentation tasks, each pixel of an image is classified into various classes. For example, in the case of GC, usually segmentation is performed on the image by identifying pixels comprising the optic disc and optic cup areas of the retina and calculating the ratio of the areas to detect GC.

We start with a few definitions:

* An instance is either a single image (for classification tasks), or a patch or a pixel of an image (for segmentation tasks).
* True Positive (TP) is the number of positive (disease) instances which are correctlyclassified.
* True Negative (TN) is the number of negative (normal) instances which are correctly classified.
* False Positive (FP) is the number of positive (disease) instances which are incorrectly classified.
* False Negative (FN) is the number of negative (normal) instances which are incorrectly classified.

Based on the above definitions, the following are the most common metrics used to evaluate performance of algorithms:

#### Binary classification tasks

* Sensitivity or Recall or True Positive Rate is the proportion of correctly classified positive (disease) instances. It is calculated as: TP / (TP + FN)
* Specificity or True negative rate is the proportion of correctly classified negative (normal) instances. It is calculated as: TN / (TN + FP)
* Precision or Positive Predictive Value is the fraction of positive (disease) instances that are correctly classified. It is calculated as TP / (TP + FP).
* F1-Score combines Precision and Recall into a single metric. It is calculated as the harmonic mean of Precision and Recall. It is calculated as 2 x (Precision x Recall) / (Precision + Recall)
* Accuracy is the proportion of instances that are correctly classified. It is calculated as (TP + TN) / (TP + FP + TN + FN)
* AUC (Area Under Receiver Operating Curve or ROC): The ROC is a plot of True Positive Rate (Sensitivity) vs. False Positive Rate (1- Specificity)) at different predictive thresholds of the classifier. The AUC has a value between 0 and 1. The closer it is to 1 the better the performance.

#### Multi-label classification tasks

In this case the metrics most commonly used are:

* **Accuracy:** the proportion of instances that are correctly classified (the accuracy of each instance class is summed across all instance classes and divided by the number of all instance classes)
* **Cohen's Kappa and Quadratic Weighted Kappa:** This metric measures the degree of agreement between two different raters - for example between an AI model's predictions and the corresponding human verified values. This metric typically varies from 0 (in case of random agreement between raters) to 1 (complete agreement between raters). It has been used in the past in Kaggle competitions to measure the effectiveness of algorithms for detecting DR.

### Segmentation tasks

* Intersection over Union (IOU): This metric is used only for segmentation tasks. IOU is defined as follows: IOU = Area (A ∩ G) / Area ( A ∪ G)

where A indicates the segmentation from the algorithm and G indicates the manual ground truth segmentation of an image.

* F1-score

In all the above cases – higher values are better and algorithms would be ranked in descending order of these metrics.

* + 1. Localization tasks
* Average Euclidean distance between the estimations and ground truth: lower values are better.

## Undisclosed test data set collection

(TBC)

* raw data acquisition / acceptance
* test data source(s): availability, reliability,
* labelling process / acceptance
* bias documentation process
* quality control mechanisms
* discussion of the necessary size of the test data set for relevant benchmarking results
* specific data governance derived by general data governance document (currently C-004)

## Benchmarking methodology and architecture

(TBC)

* technical architecture
* hosting (IIC, etc.)
* possibility of an online benchmarking on a public test dataset
* protocol for performing the benchmarking (who does what when etc.)
* AI submission procedure including contracts, rights, IP etc. considerations

# Reporting methodology

(TBC)

* Report publication in papers or as part of ITU documents
* Online reporting
* public leaderboards vs. private leaderboards
* Credit-Check like on approved sharing with selected stakeholders
* Report structure including an example
* Frequency of benchmarking

# Results

Chapter 6 will outline the results from performing the benchmarking based on the methodology specified in this document. Since the benchmarking is still in its specification phase, there are no results available yet. Depending on the progress made on this document, first preliminary test benchmarking results on small public data sets are expected by the second quarter of 2020. The first official results form an MVB are expected after that.

# Discussion

(TBC)

# Declaration of conflict of interest

In accordance with the ITU rules in this section working on this document should define his conflicts of interest that could potentially bias his point of view and the work on this document.

Xtend.AI / Medindia

Xtenda.ai is a start-up focused on using AI for solving global challenges in health and other domains. It is backed by Medindia.net – a leading online publisher of health information, and a developer of health applications and services for consumers, doctors, healthcare professionals globally. Medindia's website and applications are visited by over 4 million visitors each month from over 230 countries. Medindia offers almost 1 million pages of trusted health and wellness information including news, special reports, articles, animations, slides, infographics, videos, health directories, drug information, calculators, personalized health record, mobile apps, interactive tools, applications and much more. All of Medindia's content is edited and authenticated by doctors and healthcare professionals. It is certified to comply with the HONCode standard for trustworthy health information. Medindia.net is headquartered in India and owned operated by Medindia4u.com Pvt. Ltd. – a private limited company based in Chennai, India. It has a marketing and support offices in USA.

Involved people: Arun Shroff, CEO of Xtend.AI and CTO of Medindia.net. Topic Driver for this topic group.

Baidu.com

Baidu is an international company with leading AI technology and platforms. Baidu's retinal algorithms focus not only on inputting an image and outputting several eye-disease risks, but also building a powerful AI retinal system that integrates all related AI capacity to provide better service and enhance the end-user experience. The AI retinal system aims to build a personal eye-health management and analysis platform for each user. Baidu's mission is to defend people's eyes and global health with AI.

Since 2016, Baidu has positioned AI as a strategic driver for the development of its business. Under the strategy of "strengthening the mobile foundation and leading in AI", Baidu has steadily improved its AI ecosystem, with productization and commercialization continuing to accelerate.

As integral components to its overall AI ecosystem, Baidu has developed two open ecosystems - the Apollo open autonomous driving platform and DuerOS, the company's conversational AI system, which operates in two important scenarios – intelligent driving and smart homes. So far, with its latest iteration – "Apollo 3.0", Baidu's autonomous driving platform has brought together over 130 partners and has been granted the first batches of licenses for autonomous driving public road tests from Beijing, Chongqing and Fujian. In the smart living field, Baidu has co-launched over 160 DuerOS-powered hardware products, covering smart speakers, children's wearables, televisions, automobiles, hotels and other vertical businesses. In September 2018, the install base of DuerOS reached 141 million devices with over 800 million voice queries. After years of commercial exploration, Baidu has formed a comprehensive AI ecosystem and is now at the forefront of the AI industry in terms of fundamental technological capability, speed of productization and commercialization, and "open" strategy. In the future, Baidu will continue to enhance user experience and accelerate the development of AI applications through the strategy of "strengthening the mobile foundation and leading in AI".

Involved people:

* Yanwu XU, Intelligent Healthcare Unit, Chief Scientist, Baidu China
* Xingxing Cao, Intelligent Healthcare Unit, Baidu, China
* Jingyu WANG Artificial Intelligence Group, Baidu, China
* Shan Xu, CAICT, China

St. John's Medical College

(To be completed by contributors from this institution)

Calligo Technologies

Calligo Technologies is a category defining Data Science and Machine Learning software company focused on helping companies seeking to realize their business potential to capture new enterprise value by leveraging the convergence of High-Performance Computing and Big Data and unleashing the potential of Dark Data using Artificial Intelligence and Machine Learning.

Calligo Technologies is

* 1 of 10 world-wide HPC Code Modernization Partner for Intel
* 1 of 4 Indian AI/ML/DL Partner for Intel
* ISV Partner – Intel AI Builders Ecosystem

Calligo Technologies flagship product CIDAP (Calligo Intelligent Data Analytics Platform) is a highly scalable platform that addresses a continuum of analytics needs from start-ups, mid-markets to large-scale enterprises and helps companies to elevate from Descriptive analytics to Prescriptive analytics. It combines the agility of Big Data processes with the scale of High-Performance Computing & Artificial Intelligence capabilities, in a converged scalable platform.

CIDAP features include:

* A Single consistent method for capturing unstructured, structured and semi-structured data
* 100% based on open source technologies
* A Highly Scalable, distributed, secure and fault tolerant architecture
* A component-based architecture that enables plug & play of new connectors
* Intel x86 optimized architecture

Calligo's Edge Analytical solution for Ophthalmology is a module of CIDAP, which can

* Assist Ophthalmologists, Diabetologists, Diagnostics centres and Insurance companies
* Focusing on providing an Edge Analytics solution that can be used easily.

Involved people:

* Rajaraman Subramanian, Calligo Technologies, India
* Sriganesh Rao, Calligo Technologies, India

# Appendix A: Glossary

This section lists all the relevant abbreviations and acronyms used in the document. If there is an external source

* **AI**: [Artificial Intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) – an umbrella term that refers to one or more of the various fields of computer science including machine learning, neural networks and deep learning.
* **AI4H**: AI for health - An [ITU-T SG16 Focus Group](https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx) founded in cooperation with the WHO in July 2018.
* **API**: [Application Programming Interface](https://en.wikipedia.org/wiki/Application_programming_interface) - the software interface systems communicate through.
* **CSME**: Clinically significant macular edema
* **CWS:** cotton-wool spots
* **DD:** Disc diameter
* **DME:** Diabetic Macular Edema
* **ETDRS:** Early Treatment diabetic retinopathy Study
* **FDP:** Fibrous proliferations disc
* **FPE:** Fibrous proliferations elsewhere
* **FG:** [Focus Group](https://www.itu.int/en/ITU-T/focusgroups/Pages/default.aspx) - An instrument created by ITU-T providing an alternative working environment for the quick development of specifications in their chosen areas.
* **IIC:** International Computing Centre - The United Nations data center that will host the benchmarking infrastructure.
* **IRMA:** Intraretinal microvascular abnormalities;
* **ITU:** [International Telecommunication Union](https://www.itu.int) - The United Nations specialized agency for information and communication technologies – ICTs.
* **LMIC:** Low and Middle Income Countries
* **MA:** Microaneurysms;
* **MVB:** minimal viable benchmarking
* **NGO:** [Non Governmental Organization](https://en.wikipedia.org/wiki/Non-governmental_organization) - NGOs are usually non-profit and sometimes international organizations independent of governments and international governmental organizations that are active in humanitarian, educational, health care, public policy, social, human rights, environmental, and other areas to affect changes according to their objectives. (from Wikipedia.en)
* **NPDR:** Non-proliferative diabetic retinopathy
* **NVD:** New vessels at the optic disc
* **NVE:** new vessels elsewhere
* **PDR:** Proliferative diabetic retinopathy
* **PRH:** preretinal hemorrhage
* **SDG:** [Sustainable Development Goals](https://www.un.org/sustainabledevelopment/) - The United Nations Sustainable Development Goals are the blueprint to achieve a better and more sustainable future for all. Currently there are 17 goals defined. SDG 3 is to "Ensure healthy lives and promote well-being for all at all ages" and is therefore the goal that will benefit from the AI4H Focus Groups work the most.
* **TBC**: A topic group item to be completed.
* **TBD**: A topic group item to be discussed / determined
* **TDD**: Topic Description Document - Document specifying the standardized benchmarking for a topic FG AI4H Topic Group works on. This document is the TDD for the Topic Group "AI for Ophthalmology
* **TG**: Topic Group - Structures inside AI4H FG summarizing similar use cases and working on a TDD specifying the setup of a standardized benchmarking for the corresponding topic. The Topic Groups have been first introduced by the FG at the Meeting C, January 2019 in Lausanne. See protocol FG-AI4H-C-10x for details.
* **WHO**: [World Health Organization](https://www.who.int) - The United Nations specialized agency for international public health.

# References

1. <https://www.who.int/en/news-room/fact-sheets/detail/diabetes> [↑](#endnote-ref-1)
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4657234/> [↑](#endnote-ref-2)
3. 1. <https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(13)70145-1/fulltext> (global)

   [↑](#endnote-ref-3)
4. 1. <https://www.brightfocus.org/macular/article/age-related-macular-facts-figures> (global)

   [↑](#endnote-ref-4)
5. 1. <https://www.ncbi.nlm.nih.gov/pubmed/29302323> (China)

   [↑](#endnote-ref-5)
6. 1. <https://www.aaojournal.org/article/S0161-6420(14)00433-3/abstract> (global)

   [↑](#endnote-ref-6)
7. 1. <https://www.glaucoma.org/glaucoma/glaucoma-facts-and-stats.php> (global)

   [↑](#endnote-ref-7)
8. <https://bjo.bmj.com/content/85/11/1277> (China) [↑](#endnote-ref-8)
9. Pflipsen, M., Massaquoi, M. and Wolf, S., 2016. Evaluation of the Painful Eye. American family physician, 93(12). [↑](#endnote-ref-9)
10. National Institute for Clinical Excellence. Management of Type 2 Diabetes: Retinopathy—Screening and Early Management. London, UK: Inherited Clinical Guideline; 2002 [↑](#endnote-ref-10)
11. <https://www.ncbi.nlm.nih.gov/pubmed/2062513/> [↑](#endnote-ref-11)
12. International Council of Ophthalmology. Number of Ophthalmologists in Practice and Training Worldwide. <http://www.icoph.org/ophthalmologists-worldwide.html> [↑](#endnote-ref-12)
13. <http://www.eyepacs.com/data-analysis> [↑](#endnote-ref-13)
14. <http://www.adcis.net/en/third-party/messidor/> [↑](#endnote-ref-14)
15. <http://www.it.lut.fi/project/imageret/diaretdb1/> [↑](#endnote-ref-15)
16. <http://www.it.lut.fi/project/imageret/diaretdb1/>

    https://www5.cs.fau.de/research/data/fundus-images [↑](#endnote-ref-16)
17. Varun Gulshan, PhD1; Lily Peng, MD, PhD1; Marc Coram, PhD; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. <https://jamanetwork.com/journals/jama/fullarticle/2588763> [↑](#endnote-ref-17)
18. M. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. Folk, M. Niemeijer, Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning, Investigative ophthalmology & visual science 57 (13) (2016) 5200–5206. [↑](#endnote-ref-18)
19. 1. R. Gargeya, T. Leng, Automated identification of diabetic retinopathy using deep learning, Ophthalmology 124 (7) (2017) 962–969.

    [↑](#endnote-ref-19)
20. 1. H. Takahashi, H. Tampo, Y. Arai, Y. Inoue, H. Kawashima, Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy, PloS one 12 (6) (2017) e0179790.

    [↑](#endnote-ref-20)
21. 1. G. García, J. Gallardo, A. Mauricio, J. López, C. Del Carpio, Detection of diabetic retinopathy based on a convolutional neural network using retinal fundus images, in: International Conference on Artificial Neural Networks, Springer, 2017, pp. 635–642.

    [↑](#endnote-ref-21)
22. V. Gulshan, L. Peng, M. Coram, M. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, Jama 316 (22) (2016) 2402–2410. [↑](#endnote-ref-22)
23. L. Zhixi, Y. He, S. Keel, W. Meng, R. Chang, M. He, Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs, Ophthalmology 125 (8) (2018) 1199–1206. doi:10.1016/j.ophtha.2018.01.023. [↑](#endnote-ref-23)
24. C. Xiangyu, X. Yanwu, W. Damon, W. Tien, L. Jiang, Glaucoma detection based on deep convolutional neural network. (Aug 2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26736362/ [↑](#endnote-ref-24)
25. 1. H. Fu, J. Cheng, Y. Xu, C. Zhang, D. Wong, J. Liu, X. Cao, Disc-aware ensemble network for glaucoma screening from fundus image, IEEE Transactions on Medical Imaging.

    [↑](#endnote-ref-25)
26. A. Chakravarty, J. Sivswamy, A deep learning based joint segmentation and classification framework for glaucoma assesment in retinal color fundus images, arXiv preprint arXiv:1808.01355 [↑](#endnote-ref-26)
27. F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M. Zimmermann, B. Linkohr, A. Peters, I. Heid, C. Palm, B. Weber, A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography, Ophthalmology [↑](#endnote-ref-27)
28. A. Govindaiah, M. Hussain, R. Smith, A. Bhuiyan, Deep convolutional neural network based screening and assessment of age-related macular degeneration from fundus images, in: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, IEEE, 2018, pp. 1525– 1528. [↑](#endnote-ref-28)
29. 1. P. Burlina, N. Joshi, M. Pekala, K. Pacheco, D. Freund, N. Bressler, Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks, JAMA ophthalmology 135 (11) (2017) 1170–1176.

    [↑](#endnote-ref-29)
30. J. Tan, S. Bhandary, S. Sivaprasad, Y. Hagiwara, A. Bagchi, U. Raghavendra, A. Rao, B. Raju, N. Shetty, A. Gertych, et al., Age- related macular degeneration detection using deep convolutional neural network, Future Generation Computer Systems 87 (2018) 127–135 [↑](#endnote-ref-30)
31. S. Matsuba, H. Tabuchi, H. Ohsugi, H. Enno, N. Ishitobi, H. Masumoto, Y. Kiuchi, Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration, International ophthalmology (2018) 1–7. [↑](#endnote-ref-31)
32. 1. A. Horta, N. Joshi, M. Pekala, K. Pacheco, J. Kong, N. Bressler, D. Fre- und, P. Burlina, A hybrid approach for incorporating deep visual features and side channel information with applications to AMD detection, in: Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on, IEEE, 2017, pp. 716–720.
    2. P. Burlina, K. Pacheco, N. Joshi, D. Freund, N. Bressler, Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis, Computers in biology and medicine 82 (2017) 80–86.

    [↑](#endnote-ref-32)
33. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ [↑](#endnote-ref-33)