International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

15 September 2023

PRE-PUBLISHED VERSION

DEL10.9

FG-AI4H Topic Description Document for the Topic Group on AI for ophthalmology (TG-Ophthalmo)

Summary

This topic description document (TDD) specifies a standardized benchmarking for AI in ophthalmology. It covers scientific, technical, and administrative aspects relevant for setting up this benchmarking.

Keywords

Artificial intelligence; benchmarking; health; topic groups; overview; ethics; regulations; data quality; data audit; clinical relevance; topic description; ophthalmology; diabetic retinopathy; age-related macular degeneration; glaucoma; pathological myopia; red eye

Change Log

This document contains Version 1 of the Deliverable DEL10.9 on "*FG-AI4H Topic Description Document for the Topic Group on AI for ophthalmology (TG-Ophthalmo)*" approved on 15 September 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editor:

Arun Shroff
Medindia/Xtend.ai
India / USA

Tel: +1 908-208-1100
Email: arunshroff@gmail.com

Contributors:

Yanwu (Frank) Xu
Intelligent Healthcare Unit, Baidu, China

Tel: +86 13918541815
Fax: +86 10 59922186
Email: 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

Dr Suneetha N
St John's Medical College
India

Tel: +91 8197673923
Email: suneetha.n.lobo@gmail.com

Rajaraman Subramanian
Calligo Technologies
India

Tel: +919845239446
Email: Rajaraman.subramanian@calligotech.com

Sriganesh Rao
Calligo Technologies
India

Tel: +919845155800
Email: Sriganesh.rao@calligotech.com

Sushil Kumar
TEC, New Delhi
India

Tel: +911123323471
Email: sushil.k.123@gmail.com

CONTENTS

Page

Page

List of Tables

List of Figures

ITU-T FG-AI4H Deliverable 10.9

FG-AI4H Topic Description Document for the Topic Group on AI for ophthalmology (TG-Ophthalmo)

1 Introduction

This topic description document specifies the standardized benchmarking for AI for Ophthalmology systems. It serves as deliverable No. DEL10.09 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

The specific conditions and diseases in this topic group include Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), Glaucoma (GC), Pathological Myopia (PM) and Red Eye (RE). Additional diseases and conditions that are relevant to this Topic Group may be added in the future

Several hundred million people worldwide suffer from these ophthalmological conditions and diseases can lead to vision loss and blindness. As a result, it is estimated that approximately 21% of the global population experiences vision loss and 36 million people have blindness.

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 ophthalmologists and healthcare professionals available to screen or diagnose 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 healthcare professionals, many of the affected people live in remote areas with little or no access to an eye care clinic or a screening centre.

The ophthalmological conditions above are usually diagnosed by capturing and examining images of the eye and/or retina. For example, fundus images are generally captured to detect and diagnose DR, AMD and GC. Recent advances in Artificial Intelligence algorithms including neural networks and deep learning for image recognition and classification have shown to be effective in the detection and diagnosis of these conditions and diseases. The use of AI for early detection and diagnosis of these diseases has the potential to bridge the gap in healthcare professionals worldwide and prevent vision loss and blindness for millions.

The input to these AI systems is generally the fundus images of the retina and/or other relevant images of the retina / eye. The AI system is trained to output the diagnosis of the particular condition or disease. In some cases it may also provide segmentation or annotation of the diseased parts of the image.

While systems for AI-based ophthalmological diagnosis have great potential to improve health care, the lack of consistent standardisation makes it difficult for organizations like the WHO, governments, and other key players to adopt such applications as part of their solutions to address global health challenges.

The implementation of a standardized benchmarking for AI based ophthalmology by the ITU/WHO AI4H Focus Group will therefore be an important step towards closing this gap. Paving the way for the safe and transparent application of AI technology will help improve access to eye care and prevent vision loss and blindness for millions globally.

2 About the FG-AI4H topic group on Ophthalmology

The introduction highlights the potential of a standardized benchmarking of AI systems for Ophthalmology to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-Ophthalmo at meeting C in Lausanne, Switzerland, 22-25 January 2019. This was based on the following use case, submitted by Arun Shroff of Medindia/Xtend.AI, which was accepted at the November 2018 meeting B in New York:

FGAI4H-B-028-R1: Proposal: Using AI for early detection of Diabetic Retinopathy to prevent vision loss

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD. During FG-AI4H meeting C in Lausanne, Switzerland, 22-25 January 2019, Arun Shroff from Medindia/Xtend.AI was nominated as topic driver for the TG-Ophthalmo.

2.1 Documentation

This document is the TDD for the TG-Ophthalmo. It introduces the health topic including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for Ophthalmology. It describes the existing approaches for assessing the quality of AI for Ophthalmology systems and provides the details that are likely relevant for setting up a new standardized benchmarking. It specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subsections for all versions of the benchmarking. Finally, it summarizes the results of the topic group's benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as deliverable "DEL 10.09 Ophthalmology (TG-Ophthalmo)." The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

Table 1 – Topic Group output documents

Number	Title
FGAI4H-x-017-A01	Latest update of the Topic Description Document of the TG-Ophthalmo
FGAI4H-x-017-A02	Latest update of the Call for Topic Group Participation (CfTGP)
FGAI4H-x-017-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-Ophthalmo

The working version of this document can be found in the official topic group SharePoint directory.

– https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Ophthalmo.aspx

(a free ITU account is required to access these documents)

2.2 Status of this topic group

The following subsections describe the update of the collaboration within the TG-Ophthalmo for the official focus group meetings.

2.2.1 Status update for meeting D (Shanghai)

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

2.2.2 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.
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)
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:
7. Pearse A Keane, MD MSc FRC Ophth MRCSI, Consultant Ophthalmologist, Moorfields Eye Hospital, U.K.
8. 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
9. Ashley Kras, M.D. M. S., Ophthalmologist & Bioinformatician (Harvard Medical School)

2.2.3 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:
3. Dr Covadonga Bascaran, PHEC MSc Programme Director, International Centre for Eye Health (ICEH), London School of Hygiene & Tropical Medicine
4. Inês Sousa, Head of Intelligent Systems, Fraunhofer.
5. We had several online Meetings / calls and discussion with:
6. Prof Leo Celi, Clinical Research Director, Harvard MIT Division of Health Science and Technology and Ash Krasley: (online meeting on June 22, 2019)

7. Details about MIT Open Access Project

8. Potential collaboration with FGAI4H / Contribution of Data

9. Dr. Jorge Cuadros, O.D., PhD. (CEO), Founder & CEO of EyePACS LLC, USA (online meeting on July 22, 2019).

10. 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.

11. 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.

2.2.4 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
2. 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
3. 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

2.2.5 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
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.
5. Other TDD updates:
   - Topic Group Thematic Classification updated.
   - Added Quadratic Kappa Metric for multi-label classification.
   - Added Kaggle DR challenge datasets and results
   - Miscellaneous edits/corrections

2.2.6 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 :
   - Dr. Jorge Cuadros, EyePACS, CA, to discuss collaboration for obtaining datasets for testing.
   - 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
   - Rajaraman Subramanian and Sriganesh Rao, Calligo Technologies to discuss TDD updates, subtopic creation.

2.2.7 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.
4. Four new members have joined the TG since last meeting:

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

2.2.8 Status update for Meeting J (E-meeting)

1. Collaboration with the Data and AI solution assessment methods workgroup (WG-DAISAM). Worked with the group in providing Ophthalmology data and models to assess bias, robustness, explainability and uncertainty. Results were included in a paper "ML4H Auditing: From Paper to Practice", which was selected for the NeurIPS 2020 workshop on Machine Learning for Health. It has been published in the Proceedings of Machine Learning Research and is available at http://proceedings.mlr.press/v136/oala20a/oala20a.pdf
2. Outreach via email and social media. Also presented the work of the Focus Group at the 16th International Conference of Telemedicine Society of India on Dec 20, 2020 as part of a talk on “How AI is Transforming Healthcare”.
3. Started working on converting the TDD to the new TDD template (J-105).

2.2.9 Status update for Meeting K (E-meeting)

1. Collaboration with the Data and AI solution assessment methods workgroup (WG-DAISAM). Worked with the group in providing Ophthalmology data and models to assess bias, robustness, explainability and uncertainty. Results were included in a paper "ML4H Auditing: From Paper to Practice", which was selected for the NeurIPS 2020 workshop on Machine Learning for Health. It has been published in the Proceedings of Machine Learning Research and is available at http://proceedings.mlr.press/v136/oala20a/oala20a.pdf
2. Outreach via email and social media. Also presented the work of the Focus Group at the 16th International Conference of Telemedicine Society of India on Dec 20, 2020 as part of a talk on “How AI is Transforming Healthcare”.
3. Started working on converting the TDD to the new TDD template (J-105).

2.2.10 Status update for Meeting L (E-meeting)

1. Work is ongoing on converting the TDD to the new TDD template (J-105).
2. Meetings with WG-DAISAM group and collaboration to help with setting up challenge for TG-Ophthalmo use case for DR.
3. Outreach to organizations to obtain test data for the challenge. Email exchanges with :
    - Dr. Pearse A Keane, Consultant Ophthalmologist, Moorfields Eye Hospital, U.K
    - Dr. Alastair Denniston, Director of INSIGHT - the Health Data Research Hub for Eye Health, who have agreed in principle to work with us to provide data.

They have certain requirements including an appication and approval process before data can be provided. Dialog has been initiated and discussions are ongoing.

- Dr. Jorge Cuadros of EYEPACS, USA who has also responded. Discussions are ongoing.

2.2.11 Status update for Meeting M (E-meeting)

1. Major revisions to the TDD made to adapt it to the new template (J-105).
2. The TDD now has split the topic into the various subtopics for ophthalmology – DR, AMD, GC, PM, and RE. The new template contains many new sections that need to be completed by the respective contributors to the subtopics.
3. Initiated an application with the INSIGHT Health Data Research Hub - a NHS led partnership in U.K. to obtain access to their data sets of labelled imaged for DR and AMD for the DR challenge and testing.
4. Working with the ML4H team (Luis Oala. Pradeep Balachandran, et al) on Trial Audits Iteration 2 for TG-Ophthalmo - team formation completed.

2.2.12 Status update for Meeting N (E-meeting)

1. Minor edits to TDD
2. Collaborated with ML4H Audit Team (team (Luis Oala. Pradeep Balachandran, et al) on Trial Audits Iteration 2.0 for TG-Ophthalmo
3. Completed the first draft of the Audit Verification checklist for TG-Ophthalmo.
4. Completed setup of benchmarking practice tasks on MLAudit platform including:
   a. Registration on Mlaudit platform
   b. Updating challenge configuration files
   c. Creating & hosting a challenge (text prediction based)
   d. Participating in a challenge
   e. Creating an annotations & submission file for text predictions

2.2.13 Status update for Meeting O

1. Continued collaboration with ML4H Audit team to host a challenge for benchmarking for TG-Ophthalmology. The challenge can be viewed at https://health.aiaudit.org/web/challenges/challenge-page/371/overview
2. Submitted responses to the Audit Verification checklist for the hosted benchmark and responding to the feedback received.

2.2.14 Status update for Meeting P

1. Continued collaboration with ML4H Audit team to host a challenge for benchmarking for TG-Ophthalmology. The challenge can be viewed at https://health.aiaudit.org/web/challenges/challenge-page/371/overview
2. Completed the Model Summary section of the TG-Ophthalmology Challenge ML Audit Trials 2.0R - Audit Report. All sections of the audit report are now complete and ready for review and feedback. The report can be viewed at:

   https://health.aiaudit.org/web/challenges/challenge-page/371/my-report

2.3 Topic Group participation

The participation in both, the Focus Group on AI for Health and in a TG is generally open to anyone (with a free ITU account). For this TG, the corresponding 'Call for TG participation' (CfTGP) can be found here:

– https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/tg/CfP-TG-Ophthalmo.pdf

Each topic group also has a corresponding subpage on the ITU collaboration site. The subpage for this topic group can be found here:

– https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-Ophthalmo.aspx

For participation in this topic group, interested parties can also join the regular online meetings. For all TGs, the link will be the standard ITU-TG 'zoom' link:

– https://itu.zoom.us/my/fgai4h

All relevant administrative information about FG-AI4H—like upcoming meetings or document deadlines—will be announced via the general FG-AI4H mailing list fgai4h@lists.itu.int.

All TG members should subscribe to this mailing list as part of the registration process for their ITU user account by following the instructions in the 'Call for Topic Group participation' and this link:

– https://itu.int/go/fgai4h/join

In addition to the general FG-AI4H mailing list, TG-Ophthalmo has a separate mailing list:

– fgai4htgophthalmo@lists.itu.int

Regular FG-AI4H workshops and meetings proceed about every two months at changing locations around the globe or remotely. More information can be found on the official FG-AI4H website:

– https://itu.int/go/fgai4h

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.

2.3.1 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

23. Aaron Y. Lee, MD, MSCI, Associate Professor, Department of Ophthalmology, University of Washington, USA

24. Sheena Macpherson, ObjectivAI

3 Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI in Ophthalmology and how this can help to solve a relevant ‘real-world’ problem.

Topic Groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. The TG-Ophthalmo currently has the following subtopics: Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), Glaucoma (GC), Pathological Myopia (PM) and Red Eye (RE).

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 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."

RE is a condition where the sclera has become reddened or “bloodshot”. 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.

3.1 Subtopic Diabetic Retinopathy (DR)

3.1.1 Definition of the AI task

This section provides a detailed description of the specific task the AI systems for DR are expected to solve. It is *not* about the benchmarking process (this will be discussed more detailed in chapter 4). This section corresponds to DEL03 *"AI requirements specifications*," which describes the functional, behavioural, and operational aspects of an AI system.

AI systems for DR typically provide a diagnosis of the absence or presence of DR, and/or the stage of the DR, by processing one or more fundus images of the retina. This is generally considered a classification task for the AI model.

The input data for the AI model for DR is generally as follows:

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

- Image File Format: JPG or PNG format (or any other acceptable image 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.
- Meta Data : This may include anonymised patient data such as: Patient's Gender, Age, Diabetes (Y/N), Number of Years with Diabetes, Any other relevant medical data.

The output of the AI DR model should the following:

- The diagnosis of the retinal image as per the AI algorithm. The labels will depend upon the specific condition and the type of classification that is being benchmarked which could be one of the following classifications:

  i) Classification: All DR severity levels:
    - 0 (Non-gradable Image)
    - 1(No DR)
    - 2 (Mild)
    - 3 (Moderate NPDR)
    - 4 (Severe NPDR)
    - 5 (PDR)

  - Classification: Presence or absence of diabetic macular edema:
    - 0 (Non-gradable Image)
    - 1 (No Diabetic Macular Edema)
    - 2 (Diabetic Macular Edema present)

  ii) Classification: Referable or Non-referable DR:
    - 0 (Nongradable Image)
    - 1 (Non-referable Retinopathy – No DR or Mild DR without Diabetic Macular Edema)

- 2 (Referable Retinopathy - Moderate, Severe, PDR, or Diabetic macular edema present)

3.1.2 Current gold standard

This section provides a description of the established gold standard of the addressed health topic.

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 DR, moderate DR or vision-threatening DR, 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	Any of the following (4-2-1 rule) and no PDR – Severe H in each quadrant – VB in 2 or more quadrants

Diabetic Retinopathy Grade	Simplified ETDRS Scale	International Scale
	– IRMA ≥ standard photo 8A in at least 1 quadrant	– 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%.[1].

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.[2]

3.1.3 Relevance and impact of an AI solution

This section addresses the relevance and impact of the AI solution (e.g., on the health system or the patient outcome) and describes how solving the task with AI improves a health issue.

The WHO estimates that there are over 422 million people with diabetes worldwide[3].

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[4].

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.

An accurate way of benchmarking the performance of AI solutions to detect and diagnose DR can have a major impact on selecting and implementing the best solution to address the global healthcare challenge posed by these diseases especially 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.

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.[5]

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, the amount of data they generate will grow exponentially.

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.

3.1.4 Existing AI solutions

This section provides an overview of existing AI solutions for the same health topic that are already in operation. It contains details of the operations, limitations, robustness, and the scope of the available AI solutions. The details on performance and existing benchmarking procedures will be covered in chapter 6.

The following areas are for further study:

- Description of the general status and the maturity of AI systems for the health topic of your TG (e.g., exclusively prototypes, applications, and validated medical devices)
- Which are the currently known AI systems and their inputs, outputs, key features, target user groups, and intended use (if not discussed before)? This can also be provided as a table.
- What are the common features found in most AI solutions that might be benchmarked?
- What are the relevant metadata dimensions characterizing the AI systems in this field and with relevance for reporting (e.g., systems supporting offline functions, availability in certain languages, and the capability to process data in a specific format)?
- Description of existing AI systems and their scope, robustness, and other dimensions.

3.1.4.1 DR datasets

- Publicly available datasets include the EyePACS dataset (around 90,000 fundus images, 5 levels of severity), [6]
- MESSIDOR dataset (1,200 images, 4 levels of severity), [7]
- The DIARETDB dataset (around 200 images marked with lesions), etc.[8]
- High-Resolution Fundus (HRF) Image Database.[9]
- 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).

3.1.4.2 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). [10]

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%.[11]

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% .[12]

This work combined features both from deep ResNet and from meta-data 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 [13] [14] [15], 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.

Another multi-center, noninterventional device validation study was conducted evaluating a total of 311,604 retinal images from 23,724 veterans who presented for teleretinal DR screening at the Veterans Affairs (VA) Puget Sound Health Care System (HCS) or Atlanta VA HCS from 2006 to 2018. Five companies provided seven algorithms, including one with FDA approval, that independently analysed all scans, regardless of image quality. The sensitivity/specificity of each algorithm when classifying images as referable DR or not were compared with original VA tele-retinal grades and a regraded arbitrated data set.

Although high negative predictive values (82.72–93.69%) were observed, sensitivities varied widely (50.98–85.90%). Most algorithms performed no better than humans against the arbitrated data set, but two achieved higher sensitivities, and one yielded comparable sensitivity (80.47%, P = 0.441) and specificity (81.28%, P = 0.195). Notably, one had lower sensitivity (74.42%) for proliferative DR ($P = 9.77 \times 10{-}4$) than the VA tele-retinal graders.

3.2 Subtopic Age-related Macular Degeneration (AMD)

3.2.1 Definition of the AI task

This section provides a detailed description of the specific task the AI systems for AMD are expected to solve. It is *not* about the benchmarking process (this will be discussed more detailed in chapter 4). This section corresponds to DEL03 *"AI requirements specifications*," which describes the functional, behavioural, and operational aspects of an AI system.

AI systems for AMD typically provide a diagnosis of the absence or presence of AMD, and/or the stage of the AMD, by processing one or more fundus images of the retina. This is generally considered a classification task for the AI model. In some cases, it may also be considered as an ordinal regression task, in which the AI model is trained to output a number indicating the stage of the disease.

The input data for the AI model for AMD is generally as follows:

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

- Image File Format: JPG or PNG format (or any other acceptable image 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.
- Meta Data : This may include anonymised patient data such as : Patient's Gender, Age, Diabetes (Y/N), Number of Years with Diabetes, Any other relevant medical data.

The output of the AI AMD model should the following:

- The diagnosis of the retinal image as per the AI algorithm. The labels will depend upon the specific condition and the type of classification that is being benchmarked which could be one of the following classifications:
  - 0 (Non-gradable image)
  - 1 (No/early-stage AMD)
  - 2 (Intermediate/advanced stage AMD)

3.2.2 Current gold standard

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 oedema, 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 aetiology 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.

We will introduce the macular degeneration sign annotation, which focuses on the whole macular region. Take 2 times of the maximum diameter of optic disc, and set as *a*. The minimum distance between macular fovea and optic disc edge is *b*. The minimum distance from macular fovea to the superior and inferior arcuate vessels (main vein vessels) is *c*. The final radius of the macular region is the minimum value of *a*, *b* and *c*. Candidates for macular degeneration sign annotation include but are not limited to the following:

a) Unable to determine: the macular region is unreadable so the presence of referral lesion in the region cannot be determined,

b) Without referral (Low risk): no abnormal signs associated with macular degeneration are observed,

c) Suggest referral (Medium risk): at least one abnormal sign is suspected in macular region,

d) Identified referral (High risk): at least one abnormal sign is found in macular region.

Candidates for macular degeneration signs annotation are shown in Table 3.

Table 3 – Criteria for macular degeneration sign annotation

Discriminant Candidate	Criteria
Unable to determine	Meet at least one of the following conditions: More than a third of the macular region is invisible, More than a third of the macular region cannot be read due to underexposure or overexposure.
Without referral (Low risk)	No signs of the discriminant candidate ‘identified referral (high risk)’ are found.
Suggest referral (Medium risk)	Suspected occurrence of any of the signs of the discriminant candidate ‘identified referral (high risk)’. (Note: if it is not clear to confirm the sign as the discriminant candidate ‘identified referral (high risk)’ through the fundus color photograph only, further examination is preferred based on clinical experience.)
Identified referral (High risk)	At least one of the following signs is found in the macular region: Drusen1: there is at least one drusen with a diameter greater than 125μm (equivalent to the diameter of the vein at the inferior temporal margin of the optic disc) Geographic atrophy Neovascularization (accompanied by haemorrhage or exudation, with at least one lesion of haemorrhage or exudation larger than 125μm in diameter) Exudation (at least one exudation lesion with diameter greater than 125μm) Hemorrhage (at least one haemorrhage lesion with diameter greater than 125μm) Scar Pigment proliferation (the signs of pigment proliferation which involve the macular region and may affect vision) Macular hole (stage II and above) Macular epiretinal membrane (phase II and above) Macular oedema (moderate and above) Diffuse choroid atrophy or macular atrophy lesion, pigment (black) Fuchs spots, scar, lacquer crack2, macular epiretinal membrane,

Discriminant Candidate	Criteria
	macular hole and retinal detachment (global or local) in the macular region caused by myopia
1 Drusen is colloidal or transparent body, and is a kind of degeneration disease that happens in choroid retina. It is caused by the abnormal deposit of the abnormal metabolite in pigment epithelial cells in the retina. 2 Lacquer crack is the common change of posterior pole of degenerative myopia fundus. Yellow white or white stripes can be seen in the macular or posterior pole, which are reticulated or branched and resemble cracks on lacquerware.	

3.2.3 Relevance and impact of an AI solution

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 [16] And AMD is the third leading cause of vision loss worldwide, by 2010, it has been responsible for approximately 5% of all blindness globally [17]. 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%[18].

An accurate way of benchmarking the performance of AI solutions to detect and diagnose AMD 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.

3.2.4 Existing AI solutions

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 aAMD 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. [19] [20]. 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 [21] [22] [23]. 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 [24] [25]The best reported accuracy is 93.4%.

3.3 Subtopic Glaucoma (GC)

3.3.1 Definition of the AI task:

This section provides a detailed description of the specific task the AI systems for GC are expected to solve. It is *not* about the benchmarking process (this will be discussed more detailed in chapter 4). This section corresponds to DEL03 *"AI requirements specifications*," which describes the functional, behavioural, and operational aspects of an AI system.

AI systems for GC typically provide a diagnosis of the absence or presence of GC, and/or the stage of the AMD, by processing one or more fundus images of the optic disc and retina. This is generally considered a classification task for the AI model.

The input data for the AI model for AMD is generally as follows:

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

- Image File Format: JPG or PNG format (or any other acceptable image 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.
- Metadata: This may include anonymised patient data such as : Patient's Gender, Age, Diabetes (Y/N), Number of Years with Diabetes, Any other relevant medical data.

The output of the AI GC model should the following:

- The diagnosis of the retinal image as per the AI algorithm. The labels will depend upon the specific condition and the type of classification that is being benchmarked which could be one of the following classifications:
  - 0 (Non-gradable image)
  - 1 (No GC)
  - 2 (GC)

3.3.2 Current gold standard

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.

Glaucoma annotation is mainly based on the fundus morphology of the optic disc region, and the observation range is a local circular region with a diameter of about two optic disc diameters on the fundus colour photography. Candidates for glaucoma annotation include but are not limited to the following:

a) Unable to determine: the optic disc is unreadable and the presence of glaucoma cannot be determined,

b) Without referral (Low risk): no abnormal signs associated with glaucoma are found,

c) Suggest referral (Medium risk): one abnormal sign associated with glaucoma is found,

d) Identified referral (High risk): at least two abnormal signs associated with glaucoma are found.

Candidates for glaucoma annotation are shown in Table 4.

Table 4 – Criteria for glaucoma annotation

Discriminant Candidate	Criteria
Unable to determine	Meet at least one of the following conditions: There is a defect in the optic disc region that affects the image reading; There is a problem of overexposure or underexposure in the optic disc region that affects the image reading; There is a problem of poor image sharpness that affects the image reading; Changes in optic disc structure due to high myopia that affects image reading.
Without referral (Low risk)	No signs of the discriminant candidate ‘identified referral (high risk)’ are found.
Suggest referral (Medium risk)	Meet at least one of the following conditions: Only one of the signs of the discriminant candidate ‘identified referral (high risk)’ is found; Vertical cup to disk ratio is greater than 0.8.
Identified referral (High risk)	Meet at least two of the following signs: Non-physiological expansion of the optic cup: the expansion of the optic cup is generally manifested as an increase in cup to disc ratio. During glaucoma annotation, the vertical cup to disc ratio is the main reference. If vessels inside the optic cup are obviously squeezed to the edge, it can be annotated as suspected glaucoma. Otherwise, it should be considered as the physiological expansion of the optic cup. Disc rim missing or disc rim notch: mainly refer to the disc rim missing in the vertical direction, especially in the inferior part of optic disc. Optic disc haemorrhage: the optic disc haemorrhage associated with glaucoma is generally linear or flame-shaped. Bayoneting of blood vessels: blood vessels extending from the rim of the optic cup show obviously ascent. Glaucoma annotation should give priority to the bayoneting sign of the inferior side of the optic disc. Do not conform to the ISNT rule of the normal optic nerve: the rule of disc rim thickness distribution which the normal fundus satisfying is I (inferior side of the optic disc rim) ≥ S (superior side) ≥ N (nasal side) ≥ T (tempel side).

3.3.3 Relevance and impact of an AI solution

There are nearly 40 million blind people in the world today, according to World Health Organization [26]. Another 285 million have visual impairment. Globally, 8% of all blindness is attributable to glaucoma, making it the leading cause of global irreversible blindness [27]. 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 [28]. 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.

An accurate way of benchmarking the performance of AI solutions to detect and diagnose GC 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.

3.3.4 Existing AI solutions

3.3.4.1 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 (REFUGE2/iChallenge-GON, 2000 fundus images, 3 tasks: classification of clinical glaucoma; segmentation of optic disc and cup; localization of fovea)
  - 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 [29] [30], or a classification based on optic disk and cup information.[31] 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. [32] 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.

3.4 Subtopic Pathological myopia (PM)

3.4.1 Definition of the AI task

This is for further study.

3.4.2 Current gold standard

This is for further study.

3.4.3 Relevance and impact of an AI solution

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.

An accurate way of benchmarking the performance of AI solutions to detect and diagnose PM 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.

3.4.4 Existing AI solutions

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.

3.5 Subtopic Red eye (RE)

3.5.1 Definition of the AI task

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.

3.5.2 Current gold standard

This is for further study.

3.5.3 Relevance and impact of an AI solution

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.[33] 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.

3.5.4 Existing AI solutions

(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.

3.5.4.1 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.

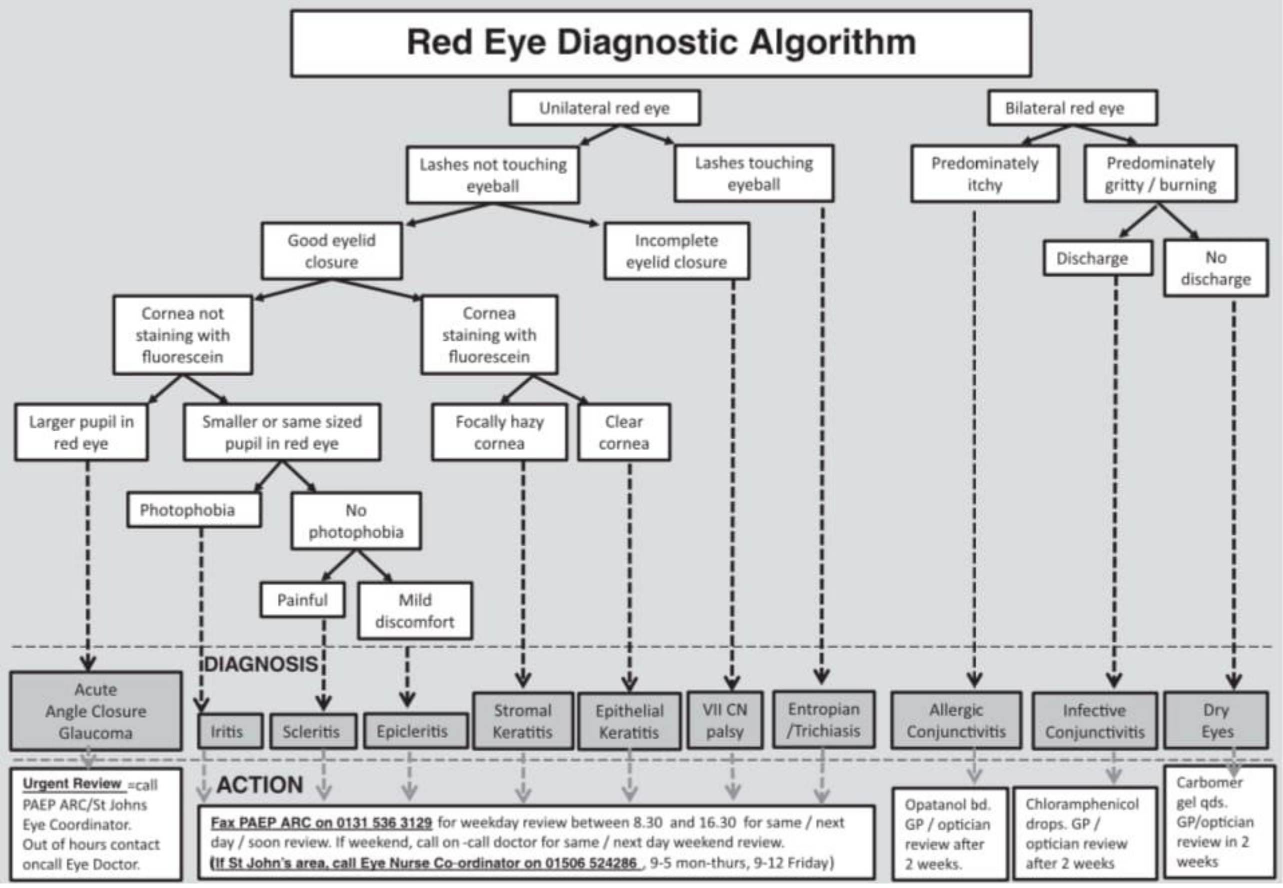

Red Eye Diagnostic Algorithm
Unilateral red eye
Bilateral red eye
Lashes not touching eyeball
Lashes touching eyeball
Predominately itchy
Predominately gritty / burning
Good eyelid closure
Incomplete eyelid closure
Discharge
No discharge
Cornea not staining with fluorescein
Cornea staining with fluorescein
Larger pupil in red eye
Smaller or same sized pupil in red eye
Focally hazy cornea
Clear cornea
Photophobia
No photophobia
Painful
Mild discomfort
DIAGNOSIS
Acute Angle Closure Glaucoma
Iritis
Scleritis
Epicleritis
Stromal Keratitis
Epithelial Keratitis
VII CN palsy
Entropian /Trichiasis
Allergic Conjunctivitis
Infective Conjunctivitis
Dry Eyes
ACTION
Urgent Review =call PAEP ARC/St Johns Eye Coordinator. Out of hours contact oncall Eye Doctor.
Fax PAEP ARC on 0131 536 3129 for weekday review between 8.30 and 16.30 for same / next day / soon review. If weekend, call on -call doctor for same / next day weekend review.
(If St John's area, call Eye Nurse Co-ordinator on 01506 524286 , 9-5 mon-thurs, 9-12 Friday)
Opatanol bd. GP / optician review after 2 weeks.
Chloramphenicol drops. GP / optician review after 2 weeks
Carbomer gel qds. GP/optician review in 2 weeks

Figure 1 – Edinburgh red eye diagnostic algorithm

3.5.4.2 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."

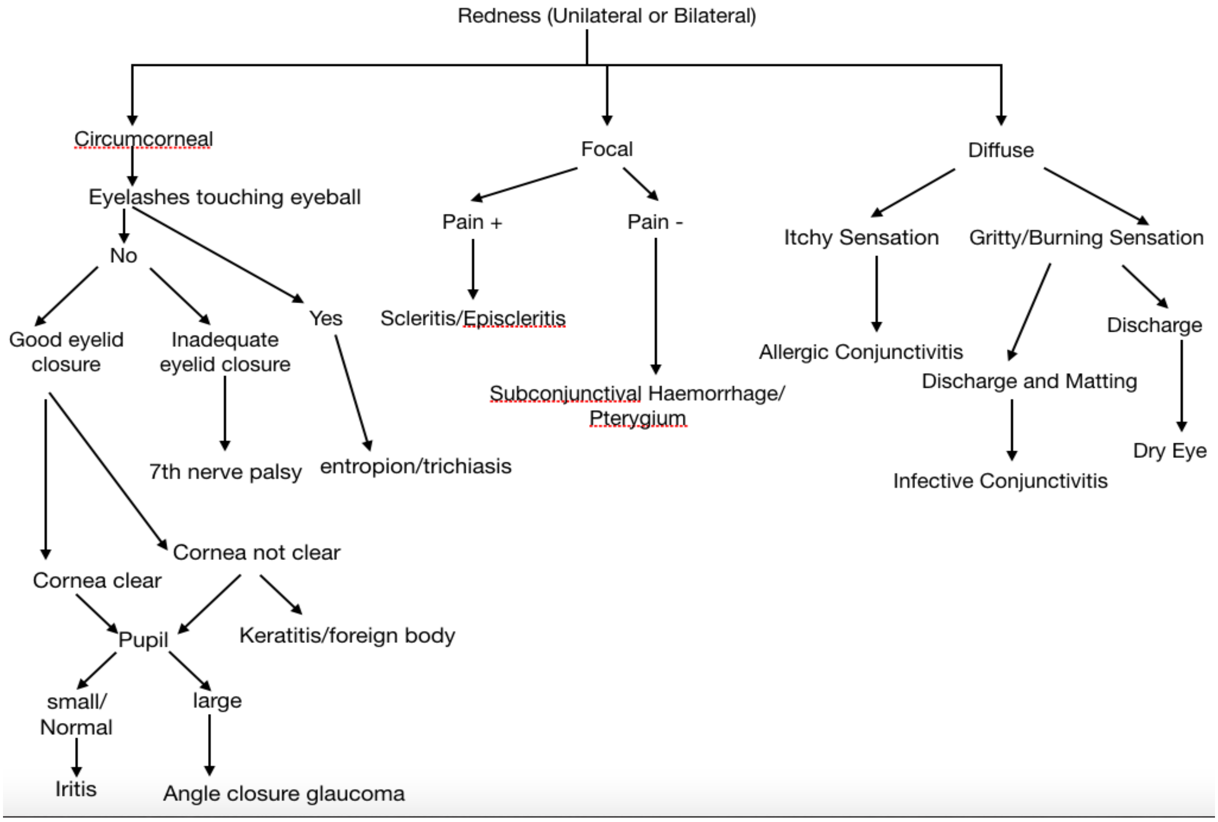


Figure 2 – New algorithm for red eye disease diagnostic

3.5.4.3 Data quality

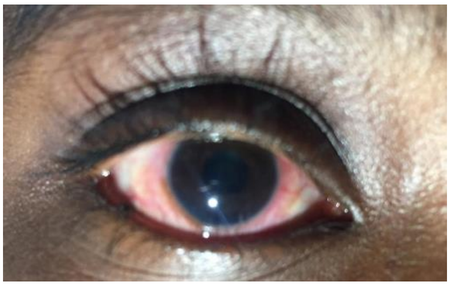
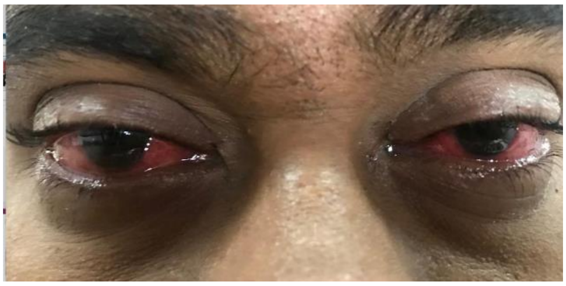
The images are captured and annotated by expert ophthalmologists at SJMCH. We are interested in collaborating with other institutes to expand our dataset.

3.5.4.4 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.

3.5.4.5 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.

3.5.4.6 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.

4 Ethical considerations

The rapidly evolving field of AI and digital technology in the fields of medicine and public health raises a number of ethical, legal, and social concerns that have to be considered in this context. They are discussed in deliverable DEL01 "*AI4H ethics considerations,"* which was developed by the working group on "Ethical considerations on AI4H" (WG-Ethics). This section refers to DEL01 and should reflect the ethical considerations of the TG-Ophthalmo. These areas are for further study:

- Data ownership
- Data security and privacy
- Related regulations and laws
- Responsibilities
- Algorithm bias
  - What are the ethical implications of applying the AI model in real-world scenarios?
  - What are the ethical implications of introducing benchmarking (having the benchmarking in place itself has some ethical risks; e.g., if the test data are not representative for a use case, the data might create the illusion of safety and put people at risk)?

- What are the ethical implications of collecting the data for benchmarking (e.g., how is misuse of data addressed, is there the need for an ethics board approval for clinical data, is there the need for consent management for sharing patient data, and what are the considerations about data ownership/data custodianship)?
- What risks face individuals and society if the benchmarking is wrong, biased, or inconsistent with reality on the ground?
- How is the privacy of personal health information protected (e.g., in light of longer data retention for documentation, data deletion requests from users, and the need for an informed consent of the patients to use data)?
- How is ensured that benchmarking data are representative and that an AI offers the same performance and fairness (e.g., can the same performance in high, low-, and middle-income countries be guaranteed; are differences in race, sex, and minority ethnic populations captured; are considerations about biases, when implementing the same AI application in a different context included; is there a review and clearance of 'inclusion and exclusion criteria' for test data)?
- What are your experiences and learnings from addressing ethics in your TG?

5 Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and Ophthalmology for quality assessment. It addresses different aspects of the existing work on benchmarking of AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics, and clinical evaluation attempts). The goal is to collect all relevant learnings from previous benchmarking that could help to implement the benchmarking process in this topic group.

5.1 Subtopic DR

5.1.1 Publications on benchmarking systems

While a representative comparable benchmarking for Ophthalmology does not yet exist, some work has been done in the scientific community assessing the performance of such systems. This section summarizes insights from the most relevant publications on this topic. It covers parts of the deliverable DEL7 *"AI for health evaluation considerations,"* DEL7.1 *"AI4H evaluation process description,"* DEL7.2 *"AI technical test specification,"* DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM),"* and DEL7.4 *"Clinical Evaluation of AI for health"*.

These areas are for further study:

- What is the most relevant peer-reviewed scientific publications on benchmarking or objectively measuring the performance of systems in your topic?
- State what are the most relevant approaches used in literature?
- Which scores and metrics have been used?
- How were test data collected?
- How did the AI system perform and how did it compare the current gold standard? Is the performance of the AI system equal across less represented groups? Can it be compared to other systems with a similar benchmarking performance and the same clinically meaningful endpoint (addressing comparative efficacy)?
- How can the utility of the AI system be evaluated in a real-life clinical environment (also considering specific requirements, e.g., in a low- and middle-income country setting)?

- Have there been clinical evaluation attempts (e.g., internal and external validation processes) and considerations about the use in trial settings?
- What are the most relevant gaps in the literature (what is missing concerning AI benchmarking)?

5.1.2 Benchmarking by AI developers

All developers of AI solutions for Ophthalmology implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group.

These areas are for further study:

- Which scores and metrics have been used?
- How did they approach the acquisition of test data?

Current systems that have implemented benchmarking for AI solutions for DR have usually done so based on the performance of the algorithm on undisclosed test data-sets. The scores and metrics used for benchmarking DR depend upon the type of task performed by the AI, which for DR would generally be classification.

5.1.2.1 Classification tasks for DR

Classification of the conditions being considered may be either binary (2 classes) – for example DR or no DR 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.

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 correctly classified.

– 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 DR algorithms:

5.1.2.1.1 Binary classification tasks for DR

– **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.

5.1.2.1.2 Multi-label classification tasks for DR

In this case the most commonly used metrics 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.
- **Macro $\boldsymbol{F_1}$-score:** this metric is defined as the mean of class-wise/label-wise F_1-scores:

  $$Macro\ F_1\ = \frac{1}{N}\sum_{i=0}^{N} F_{1_i},$$

  where I is the class /label index and N is the number of classes/labels. Macro F_1-score = 1 is the best, and the worst value is 0.
- **Micro $\boldsymbol{F_1}$-score:** this metric measures the F_1-score of the aggregated contributions of all classes. It defined as the harmonic mean of the precision and recall:

  $$Micro\ F_1\ = 2 * Micro\ precision * Micro\ recall / Micro\ precision + Micro\ recall.$$

  Micro F_1-score =1 is the best value, and the worst value is 0.

5.1.3 Relevant existing benchmarking frameworks

Triggered by the hype around AI, recent years have seen the development of a variety of benchmarking platforms where AIs can compete for the best performance on a determined dataset. Given the high complexity of implementing a new benchmarking platform, the preferred solution is to use an established one. This section reflects on the different existing options that are relevant for this topic group and includes considerations of using the assessment platform that is currently developed by FG-AI4H and presented by deliverable DEL7.5 *"FG-AI4H assessment platform"* (the deliverable explores options for implementing an assessment platform that can be used to evaluate AI for health for the different topic groups).

Document C031 provides a list of the available platforms. While not specific for ophthalmology it provides important examples for many aspects of benchmarking ranging from operational details, over scores & metrics, leader boards, reports to the overall architecture. Due to high numbers of participants and the prestige associated with a top rank, the platforms have also substantial experience in designing the benchmarking in a way that is hard or impossible to manipulate.

These areas are for further study:

- Which benchmarking platforms could be used for this topic group (e.g., EvalAI, AIcrowd, Kaggle, and CodaLab)?
- Are the benchmarking assessment platforms discussed, used, or endorsed by FG-AI4H an option?
- Are there important features in this topic group that require special attention?
- Is the reporting flexible enough to answer the questions stakeholders want to get answered by the benchmarking?
- What are the relative advantages and disadvantages of these diverse solutions?

5.2 Subtopic AMD

Current systems that have implemented benchmarking for AI solutions for AMD have usually done so based on the performance of the algorithm on undisclosed test data-sets. The scores and metrics used for benchmarking DR depend upon the type of task performed by the AI, which for AMD would generally be classification.

5.2.1.1 Classification tasks for AMD

Classification of the conditions being considered may be either binary (2 classes) – for example AMD or no AMD or multi-class - for example, in the case of AMD an image may be classified as having No AMD, early AMD, intermediate AMD, advanced AMD (4 classes).

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 correctly classified.

– 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 DR algorithms:

5.2.1.1.1 Binary classification tasks for AMD

– **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.

5.2.1.1.2 Multi-label classification tasks for AMD

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.
- **Macro $\boldsymbol{F_1}$-score:** this metric is defined as the mean of class-wise/label-wise F_1-scores:

  $$Macro\ F_1 \ = \frac{1}{N}\sum_{i=0}^{N} F_{1_i},$$

  where *I* is the class /label index and *N* is the number of classes/labels. Macro F_1-score = 1 is the best, and the worst value is 0.
- **Micro $\boldsymbol{F_1}$-score:** this metric measures the F_1-score of the aggregated contributions of all classes. It defined as the harmonic mean of the precision and recall:

  $$Micro\ F_1 = 2 * Micro\ precision * Micro\ recall / Micro\ precision + Micro\ recall.$$

  Micro F_1-score =1 is the best value and the worst value is 0.

5.3 Subtopic GC

Current systems that have implemented benchmarking for AI solutions for GC have usually done so based on the performance of the algorithm on undisclosed test data-sets. The scores and metrics used for benchmarking DR depend upon the type of task performed by the AI, which for GC would generally be classification. In some cases, the AI model may also be evaluated on Segmentation of the optic disc.

5.3.1.1 Classification tasks for GC:

Classification of the conditions being considered may be either binary (2 classes) – for example GC or no GC or multi-class - for example, in the case of GC an image may be classified as having low risk, medium risk, or high risk of GC, (3 classes).

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 correctly classified.
- 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 DR algorithms:

5.3.1.1.1 Binary classification tasks for GC:

– **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.

5.3.1.1.2 Multi-label classification tasks for GC:

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.

– **Macro $\boldsymbol{F_1}$-score:** this metric is defined as the mean of class-wise/label-wise F_1-scores:

$$Macro\ F_1 = \frac{1}{N}\sum_{i=0}^{N} F_{1_i},$$

where *I* is the class /label index and *N* is the number of classes/labels. Macro F_1-score = 1 is the best, and the worst value is 0.

– **Micro $\boldsymbol{F_1}$-score:** this metric measures the F_1-score of the aggregated contributions of all classes. It defined as the harmonic mean of the precision and recall:

$$Micro\ F_1 = 2 * Micro\ precision * Micro\ recall / Micro\ precision + Micro\ recall.$$

Micro F_1-score =1 is the best value and the worst value is 0.

5.3.1.2 Segmentation tasks for GC

– **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:** This metric 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)
- **Dice coefficient:** This metric can be used in segmentation tasks. Given two sets, X and Y, Dice coefficient is defined as $\text{Dice} = 2|X \cap Y|/|X| + |Y|$.

In all the above cases – higher values are better and algorithms would be ranked in descending

5.4 Subtopic PM

Current systems that have implemented benchmarking for AI solutions for PM have usually done so based on the performance of the algorithm on undisclosed test data-sets. The scores and metrics used for benchmarking PM depend upon the type of task performed by the AI, which for PM would generally be classification.

5.4.1.1 Classification tasks for PM

Classification of the conditions being considered may be either binary (2 classes) – for example P< or no PM or multi-class - for example, in the case of PM an image may be classified as having No PM, High Myopia (HM), or PM (3 classes).

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 correctly classified.
- 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 DR algorithms:

5.4.1.1.1 Binary classification tasks for PM

- **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.

5.4.1.1.2 Multi-label classification tasks for PM

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.

– **Macro $\boldsymbol{F_1}$-score:** this metric is defined as the mean of class-wise/label-wise F_1-scores:

$$Macro\ F_1 \ = \frac{1}{N}\sum_{i=0}^{N} F_{1_i},$$

where *I* is the class /label index and *N* is the number of classes/labels. Macro F_1-score = 1 is the best, and the worst value is 0.

– **Micro $\boldsymbol{F_1}$-score:** this metric measures the F_1-score of the aggregated contributions of all classes. It defined as the harmonic mean of the precision and recall:

$$Micro\ F_1 = 2 * Micro\ precision * Micro\ recall / Micro\ precision + Micro\ recall.$$

Micro F_1-score =1 is the best value and the worst value is 0.

5.5 Subtopic RE

This is for further study.

6 Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the Ophthalmology AI task including subsections for each version of the benchmarking that is iteratively improved over time.

It reflects the considerations of various deliverables: DEL5 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL5.1 *"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL5.2 *"Data acquisition"*, DEL5.3 *"Data annotation specification"*, DEL5.4 *"Training and test data specification"* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), DEL5.5 *"Data handling"* (which outlines how data will be handled once they are accepted), DEL5.6 *"Data sharing practices"* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), DEL06 *"AI training best practices specification"* (which reviews best practices for proper AI model training and guidelines for model reporting), DEL7 *"AI for health evaluation considerations"* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), DEL7.1 *"AI4H evaluation process description"* (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), DEL7.2 *"AI* *technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), DEL7.4*"Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), DEL7.5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL9 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL9.1 *"Mobile based AI applications,"* and DEL9.2 *"Cloud-based AI applications"* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

6.1 Subtopic DR

The benchmarking of Ophthalmology is going to be developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This section outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to the subsequent sections, where the actual benchmarking methodology for each version will be described.

TG-Ophthalmo in collaboration with the FGAI4H MLAudit initiative has completed the following tasks:

- Audit Verification Checklist for TG-Ophthalmo use case (first draft completed)
- Setup of benchmarking for DR on the MLAudit platform has been started and the following practice tasks have been completed:
  - Registration on Mlaudit platform
  - Updating challenge configuration files
  - Creating & hosting a challenge (text prediction based)
  - Participating in a challenge
  - Creating an annotations & submission file for text predictions

The following tasks need to be completed:

- Dockerized model testing with images (when ready)
- Get undisclosed test data set of images

6.1.1 Benchmarking version [Y]

This section includes all technological and operational details of the benchmarking process for the benchmarking version [Y] (latest version, chronologically reversed order).

6.1.1.1 Overview

This section provides an overview of the key aspects of this benchmarking iteration, version [Y]. The following is for further study:

- What is the overall scope of this benchmarking iteration (e.g., performing a first benchmarking, adding benchmarking for multi-morbidity, or introducing synthetic-data-based robustness scoring)?
- What features have been added to the benchmarking in this iteration?

6.1.1.2 Benchmarking methods

This section provides details about the methods of the benchmarking version [Y]. It contains detailed information about the benchmarking system architecture, the dataflow and the software for the benchmarking process (e.g., test scenarios, data sources, and legalities).

6.1.1.2.1 Benchmarking system architecture

This section covers the architecture of the benchmarking system. For well-known systems, an overview and reference to the manufacturer of the platform is sufficient. If the platform was developed by the topic group, a more detailed description of the system architecture is required. The following is for further study:

- How does the architecture look?
- What are the most relevant components and what are they doing?
- How do the components interact on a high level?
- What underlying technologies and frameworks have been used?
- How does the hosted AI model get the required environment to execute correctly? What is the technology used (e.g., Docker/Kubernetes)?

6.1.1.2.2 Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture. The following is for further study:

- How do benchmarking data access the system?
- Where and how (data format) are the data, the responses, and reports of the system stored?
- How are the inputs and the expected outputs separated?
- How are the data sent to the AI systems?
- Are the data entries versioned?
- How does the lifecycle for the data look?

6.1.1.2.3 Safe and secure system operation and hosting

This section addresses security considerations about the storage and hosting of data (benchmarking results and reports) and safety precautions for data manipulation, data leakage, or data loss.

In the case of a manufactured data source (vs. self-generated data), it is possible to refer to the manufacturer’s prescriptions. The following is for further study:

- Based on the architecture, where is the benchmarking vulnerable to risk and how have these risks been mitigated (e.g., did you use a threat modelling approach)? A discussion could include:
    - Could someone access the benchmarking data before the actual benchmarking process to gain an advantage?
    - What safety control measures were taken to manage risks to the operating environment?
    - Could someone have changed the AI results stored in the database (your own and/or that of competitors)?
    - Could someone attack the connection between the benchmarking and the AI (e.g., to make the benchmarking result look worse)?

  - How is the hosting system itself protected against attacks?
- How are the data protected against data loss (e.g., what is the backup strategy)?
- What mechanisms are in place to ensure that proprietary AI models, algorithms and trade-secrets of benchmarking participants are fully protected?
- How is it ensured that the correct version of the benchmarking software and the AIs are tested?
- How are automatic updates conducted (e.g., of the operating system)?
- How and where is the benchmarking hosted and who has access to the system and the data (e.g., virtual machines, storage, and computing resources, configurational settings)?
- How is the system's stability monitored during benchmarking and how are attacks or issues detected?
- How are issues (e.g., with a certain AI) documented or logged?
- In case of offline benchmarking, how are the submitted AIs protected against leakage of intellectual property?

6.1.1.2.4 Benchmarking process

This section describes how the benchmarking looks from the registration of participants, through the execution and resolution of conflicts, to the final publication of the results. The following is for further study:

- How are new benchmarking iterations scheduled (e.g., on demand or quarterly)?
- How do possible participants learn about an upcoming benchmarking?
- How can one apply for participation?
- What information and metadata do participants have to provide (e.g., AI autonomy level assignment (IMDRF), certifications, AI/machine learning technology used, company size, company location)?
- Are there any contracts or legal documents to be signed?
- Are there inclusion or exclusion criteria to be considered?
- How do participants learn about the interface they will implement for the benchmarking (e.g., input and output format specification and application program interface endpoint specification)?
- How can participants test their interface (e.g., is there a test dataset in case of file-based offline benchmarking or are there tools for dry runs with synthetic data cloud-hosted application program interface endpoints)?
- Who is going to execute the benchmarking and how is it ensured that there are no conflicts of interest?
- If there are problems with an AI, how are problems resolved (e.g., are participants informed offline that their AI fails to allow them to update their AI until it works? Or, for online benchmarking, is the benchmarking paused? Are there timeouts?)?
- How and when will the results be published (e.g., always or anonymized unless there is consent)? With or without seeing the results first? Is there an interactive drill-down tool or a static leader board? Is there a mechanism to only share the results with stakeholders approved by the AI provider as in a credit check scenario?

- In case of online benchmarking, are the benchmarking data published after the benchmarking? Is there a mechanism for collecting feedback or complaints about the data? Is there a mechanism of how the results are updated if an error was found in the benchmarking data?

6.1.1.3 AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of Ophthalmology. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking. This is the only TDD section addressing this topic. Therefore, the description needs to be complete and precise. This section does *not* contain the encoding of the labels for the expected outcomes. It is only about the data the AI system will see as part of the benchmarking. The following is for further study:

- What are the general data types that are fed in the AI model?
- How exactly are they encoded? For instance, discuss:
  - The exact data format with all fields and metadata (including examples or links to examples)
  - Ontologies and terminologies
  - Resolution and data value ranges (e.g., sizes, resolutions, and compressions)
  - Data size and data dimensionality

6.1.1.4 AI output data structure

Similar to the input data structure for the benchmarking, this section describes the output data the AI systems are expected to generate in response to the input data. It covers the details of the data format, coding, and error handling at the level of detail needed for an AI to participate in the benchmarking. The following is for further study:

- What are the general data output types returned by the AI and what is the nature of the output (e.g., classification, detection, segmentation, or prediction)?
  - How exactly are they encoded? Discuss points like:
    - The exact data format with all fields and metadata (including examples or links to examples)
    - Ontologies and terminologies
- What types of errors should the AI generate if something is defective?

6.1.1.5 Test data label/annotation structure

While the AI systems can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called ‘labels’) for each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately. The following is for further study:

- What are the general label types (e.g., expected results, acceptable results, correct results, and impossible results)?
- How exactly are they encoded? Discuss points like:
  - The exact data format with all fields and metadata (including examples or links to examples)

  - Ontologies and terminologies

- How are additional metadata about labelling encoded (e.g., author, data, pre-reviewing details, dates, and tools)?
- How and where are the labels embedded in the input data set (including an example; e.g., are there separate files or is it an embedded section in the input data that is removed before sending to the AI)?

6.1.1.6 Scores and metrics

Scores and metrics are at the core of the benchmarking. This section describes the scores and metrics used to measure the performance, robustness, and general characteristics of the submitted AI systems. The following is for further study:

- Who are the stakeholders and what decisions should be supported by the scores and metrics of the benchmarking?
- What general criteria have been applied for selecting scores and metrics?
- What scores and metrics have been chosen/defined for robustness?
- What scores and metrics have been chosen/defined for medical performance?
- What scores and metrics have been chosen/defined for non-medical performance?
  - Metrics for technical performance tracking (e.g., monitoring and reporting when the performance accuracy of the model drops below a predefined threshold level as a function of time; computational efficiency rating, response times, memory consumption)
- What scores and metrics have been chosen/defined for model explainability?
- Describe for each aspect
  - The exact definition/formula of the score based on the labels and the AI output data structures defined in the previous sections and how they are aggregated/accumulated over the whole dataset (e.g., for a single test set entry, the result might be the probability of the expected correct class which is then aggregated to the average probability of the correct class)
  - Does it use some kind of approach for correcting dataset bias (e.g., the test dataset usually has a different distribution compared to the distribution of a condition in a real-world scenario. For estimating the real-world performance, metrics need to compensate this difference.)
  - What are the origins of these scores and metrics?
  - Why were they chosen?
  - What are the known advantages and disadvantages?
  - How easily can the results be compared between or among AI solutions?
  - Can the results from benchmarking iterations be easily compared or does it depend too much on the dataset (e.g., how reproducible are the results)?
- How does this consider the general guidance of WG-DAISAM in DEL7.3 “Data and artificial intelligence assessment methods (DAISAM)”?
- Have there been any relevant changes compared to previous benchmarking iterations? If so, why?

6.1.1.7 Test dataset acquisition

Test dataset acquisition includes a detailed description of the test dataset for the AI model and, in particular, its benchmarking procedure including quality control of the dataset, control mechanisms, data sources, and storage. The following is for further study:

- How does the overall dataset acquisition and annotation process look?
- How have the data been collected/generated (e.g., external sources vs. a process organized by the TG)?
- Have the design goals for the benchmarking dataset been reached (e.g., please provide a discussion of the necessary size of the test dataset for relevant benchmarking results, statistical significance, and representativeness)?
- How was the dataset documented and which metadata were collected?
    - Where were the data acquired?
    - Were they collected in an ethical-conform way?
    - Which legal status exists (e.g., intellectual property, licenses, copyright, privacy laws, patient consent, and confidentiality)?
    - Do the data contain 'sensitive information' (e.g., socially, politically, or culturally sensitive information; personal identifiable information)? Are the data sufficiently anonymized?
    - What kind of data anonymization or deidentification has been applied?
    - Are the data self-contained (i.e., independent from externally linked datasets)?
    - How is the bias of the dataset documented (e.g., sampling or measurement bias, representation bias, or practitioner/labelling bias)?
    - What addition metadata were collected (e.g., for a subsequent detailed analysis that compares the performance on old cases with new cases)? How was the risk of benchmarking participants accessing the data?
- Have any scores, metrics, or tests been used to assess the quality of the dataset (e.g., quality control mechanisms in terms of data integrity, data completeness, and data bias)?
- Which inclusion and exclusion criteria for a given dataset have been applied (e.g., comprehensiveness, coverage of target demographic setting, or size of the dataset)?
- How was the data submission, collection, and handling organized from the technical and operational point of view (e.g., folder structures, file formats, technical metadata encoding, compression, encryption, and password exchange)?
- Specific data governance derived by the general data governance document (currently F-103 and the deliverables beginning with DEL5)
- How was the overall quality, coverage, and bias of the accumulated dataset assessed (e.g., if several datasets from several hospitals were merged with the goal to have better coverage of all regions and ethnicities)?
- Was any kind of post-processing applied to the data (e.g., data transformations, repackaging, or merging)?
- How was the annotation organized?
    - How many annotators/peer reviewers were engaged?

    - Which scores, metrics, and thresholds were used to assess the label quality and the need for an arbitration process?
    - How have inter-annotator disagreements been resolved (i.e., what was the arbitration process)?
    - If annotations were part of the submitted dataset, how was the quality of the annotations controlled?
    - How was the annotation of each case documented?
    - Were metadata on the annotation process included in the data (e.g., is it possible to compare the benchmarking performance based on the annotator agreement)?
- Were data/label update/amendment policies and/or criteria in place?
- How was access to test data controlled (e.g., to ensure that no one could access, manipulate, and/or leak data and data labels)? Please address authentication, authorization, monitoring, logging, and auditing
- How was data loss avoided (e.g., backups, recovery, and possibility for later reproduction of the results)?
- Is there assurance that the test dataset is undisclosed and was never previously used for training or testing of any AI model?
- What mechanisms are in place to ensure that test datasets are used only once for benchmarking? (Each benchmarking session will need to run with a new and previously undisclosed test dataset to ensure fairness and no data leakage to subsequent sessions)

6.1.1.8 Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also DEL5.5 on *data handling* and DEL5.6 on *data sharing practices*). The following is for further study:

- Which legal framework was used for data sharing?
- Was a data sharing contract signed and what was the content? Did it contain:
    - Purpose and intended use of data
    - Period of agreement
    - Description of data
    - Metadata registry
    - Data harmonization
    - Data update procedure
    - Data sharing scenarios
        - Data can be shared in public repositories
        - Data are stored in local private databases (e.g., hospitals)
    - Rules and regulation for patients’ consent
    - Data anonymization and de-identification procedure
    - Roles and responsibilities
        - Data provider

    - Data protection officer
    - Data controllers
    - Data processors
    - Data receivers
- Which legal framework was used for sharing the AI?
- Was a contract signed and what was the content?

6.1.1.9 Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed. The following is for further study:

- Does this topic require comparison of the AI model with a baseline (gold standard) so that stakeholders can make decisions?
- Is the baseline known for all relevant application contexts (e.g., region, subtask, sex, age group, and ethnicity)?
- Was a baseline assessed as part of the benchmarking?
- How was the process of collecting the baseline organized? If the data acquisition process was also used to assess the baseline, please describe additions made to the process described in the previous section.
- What are the actual numbers (e.g., for the performance of the different types of health workers doing the task)?

6.1.1.10 Reporting methodology

This section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and general public. The following is for further study:

- What is the general approach for reporting results (e.g., leader board vs. drill down)?
- How can participants analyse their results (e.g., are there tools or are detailed results shared with them)?
- How are the participants and their AI models (e.g., versions of model, code, and configuration) identified?
- What additional metadata describing the AI models have been selected for reporting?
- How is the relationship between AI results, baselines, previous benchmarking iterations, and/or other benchmarking iterations communicated?
- What is the policy for sharing participant results (e.g., opt in or opt out)? Can participants share their results privately with their clients (e.g., as in a credit check scenario)?
- What is the publication strategy for the results (e.g., website, paper, and conferences)?
- Is there an online version of the results?

- Are there feedback channels through which participants can flag technical or medical issues (especially if the benchmarking data was published afterwards)?
- Are there any known limitations to the value, expressiveness, or interpretability of the reports?

6.1.1.11 Result

This section gives an overview of the results from runs of this benchmarking version of your topic. Even if your topic group prefers an interactive drill-down rather than a leader board, pick some context of common interest to give some examples. The following is for further study:

- When was the benchmarking executed?
- Who participated in the benchmarking?
- What overall performance of the AI systems concerning medical accuracy, robustness, and technical performance (minimum, maximum, average etc.) has been achieved?
- What are the results of this benchmarking iteration for the participants (who opted in to share their results)?

6.1.1.12 Discussion of the benchmarking

This section discusses insights of this benchmarking iterations and provides details about the ‘outcome’ of the benchmarking process (e.g., giving an overview of the benchmark results and process). The following is for further study:

- What was the general outcome of this benchmarking iteration?
- How does this compare to the goals for this benchmarking iteration (e.g., was there a focus on a new aspect to benchmark)?
- Are there real benchmarking results and interesting insights from this data?
  - How was the performance of the AI system compared to the baseline?
  - How was the performance of the AI system compared to other benchmarking initiatives (e.g., are the numbers plausible and consistent with clinical experience)?
  - How did the results change in comparison to the last benchmarking iteration?
- Are there any technical lessons?
  - Did the architecture, implementation, configuration, and hosting of the benchmarking system fulfil its objectives?
  - How was the performance and operational efficiency of the benchmarking itself (e.g., how long did it take to run the benchmarking for all AI models vs. one AI model; was the hardware sufficient)?
- Are there any lessons concerning data acquisition?
  - Was it possible to collect enough data?
  - Were the data as representative as needed and expected?
  - How good was the quality of the benchmarking data (e.g., how much work went into conflict resolution)?
  - Was it possible to find annotators?
  - Was there any relevant feedback from the annotators?
  - How long did it take to create the dataset?

- Is there any feedback from stakeholders about how the benchmarking helped them with decision-making?
    - Are metrics missing?
    - Do the stakeholders need different reports or additional metadata (e.g., do they need the "offline capability" included in the AI metadata so that they can have a report on the best offline system for a certain task)?
- Are there insights on the benchmarking process?
    - How was the interest in participation?
    - Are there reasons that someone could not join the benchmarking?
    - What was the feedback of participants on the benchmarking processes?
    - How did the participants learn about the benchmarking?

6.1.1.13 Retirement

This section addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this section DEL04 "*AI software lifecycle specification*" (identification of standards and best practices that are relevant for the AI for health software life cycle). The following is for further study:

- What happens with the data after the benchmarking (e.g., will they be deleted, stored for transparency, or published)?
- What happens to the submitted AI models after the benchmarking?
- Could the results be reproduced?
- Are there legal or compliance requirements to respond to data deletion requests?

6.1.2 Benchmarking version

Future version of this section will include all technological and operational details of the benchmarking process for a benchmarking version.

7 Overall discussion of the benchmarking

This section discusses the overall insights gained from benchmarking work in this topic group. This should not be confused with the discussion of the results of a concrete benchmarking run (e.g., in section 6.1.1.12). The following is for further study:

- What is the overall outcome of the benchmarking thus far?
- Have there been important lessons?
- Are there any field implementation success stories?
- Are there any insights showing how the benchmarking results correspond to, for instance, clinical evaluation?
- Are there any insights showing the impact (e.g., health economic effects) of using AI systems that were selected based on the benchmarking?
- Was there any feedback from users of the AI system that provides insights on the effectiveness of benchmarking?
    - Did the AI system perform as predicted relative to the baselines?

  - Did other important factors prevent the use of the AI system despite a good benchmarking performance (e.g., usability, access, explainability, trust, and quality of service)?
- Were there instances of the benchmarking not meeting the expectations (or helping) the stakeholders? What was learned (and changed) as a result?
- What was learned from executing the benchmarking process and methodology (e.g., technical architecture, data acquisition, benchmarking process, benchmarking results, and legal/contractual framing)?

8 Regulatory considerations

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *“Regulatory considerations on AI for health” (WG-RC)* compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are DEL2 *“AI4H regulatory considerations”* (which provides an educational overview of some key regulatory considerations), DEL2.1 *“Mapping of IMDRF essential principles to AI for health software”,* and DEL2.2 *“Guidelines for AI based medical device (AI-MD): Regulatory requirements”* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). DEL04 identifies standards and best practices that are relevant for the “*AI software lifecycle specification.”* The following sections discuss how the different regulatory aspects relate to the TG-Ophthalmo.

This section requires further study.

8.1 Existing applicable regulatory frameworks

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for Ophthalmology. The following is for further study:

- What existing regulatory frameworks cover the type of AI in this TDD (e.g., MDR, FDA, GDPR, and ISO; maybe the systems in this topic group always require at least “MDR class 2b” or maybe they are not considered a medical device)?
- Are there any aspects to this AI system that require additional specific regulatory considerations?

8.2 Regulatory features to be reported by benchmarking participants

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This section outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates a screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements). The following is for further study:

- Which certifications and regulatory framework components of the previous section should be part of the metadata (e.g., as a table with structured selection of the points described in the previous section)?

8.3 Regulatory requirements for the benchmarking systems

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This section outlines the regulatory requirements for software used for benchmarking in this topic group. The following is for further study:

- Which regulatory frameworks apply to the benchmarking system itself?
- Are viable solutions with the necessary certifications already available?
- Could the TG implement such a solution?

8.4 Regulatory approach for the topic group

Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the DEL2 *"AI4H regulatory considerations."* The following is for further study:

- Documentation & Transparency
  - How will the development process of the benchmarking be documented in an effective, transparent, and traceable way?
- Risk management & Lifecycle approach
  - How will the risk management be implemented?
  - How is a life cycle approach throughout development and deployment of the benchmarking system structured?
- Data quality
  - How is the test data quality ensured (e.g., the process of harmonizing data of different sources, standards, and formats into a single dataset may cause bias, missing values, outliers, and errors)?
  - How are the corresponding processes document?
- Intended Use & Analytical and Clinical Validation
  - How are technical and clinical validation steps (as part of the lifecycle) ensured (e.g., as proposed in the IMDRF clinical evaluation framework)?
- Data Protection & Information Privacy
  - How is data privacy in the context of data protection regulations ensured, considering regional differences (e.g., securing large data sets against unauthorized access, collection, storage, management, transport, analysis, and destruction)? This is especially relevant if real patient data is used for the benchmarking.
- Engagement & Collaboration
  - How is stakeholder (regulators, developers, healthcare policymakers) feedback on the benchmarking collected, documented, and implemented?

Annex A:
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

Acronym/Term	Expansion	Comment
AI	Artificial intelligence	
AI4H	Artificial intelligence for health	
AI-MD	AI based medical device	
API	Application programming interface	
CfTGP	Call for topic group participation	
CSME	Clinically significant macular oedema	
CWS	Cotton-wool spots	
DD	Disc diameter	
DEL	Deliverable	
DME	Diabetic Macular Edema	
ETDRS:	Early Treatment diabetic retinopathy Study	
FDA	Food and Drug administration	
FDP	Fibrous proliferations disc	
FG	Focus Group	An instrument created by ITU-T providing an alternative working environment for the quick development of specifications in their chosen areas.
FGAI4H	Focus Group on AI for Health	
FPE	Fibrous proliferations elsewhere	
GDP	Gross domestic product	
GDPR	General Data Protection Regulation	
IIC	International Computing Centre	The United Nations data centre that will host the benchmarking infrastructure
IMDRF	International Medical Device Regulators Forum	
IP	Intellectual property	
IRMA	Intraretinal microvascular abnormalities	
ISO	International Standardization Organization	
ITU	International Telecommunication Union	
LMIC	Low-and middle-income countries	
MA	Microaneurysms	
MDR	Medical Device Regulation	
MVB	minimal viable benchmarking	
NGO	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)
NPDR	Non-proliferative diabetic retinopathy	
NVD	New vessels at the optic disc	
NVE	new vessels elsewhere	
PDR	Proliferative diabetic retinopathy	
PII	Personal identifiable information	
PRH	Preretinal haemorrhage	
SaMD	Software as a medical device	
SDG	Sustainable Development Goals	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 on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group Ophthalmo
TG	Topic Group	
WG	Working Group	
WHO	World Health Organization	

Annex B: Declaration of conflict of interests

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

Not provided by the participants.

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

References

1. National Institute for Clinical Excellence. Management of Type 2 Diabetes: Retinopathy—Screening and Early Management. London, UK: Inherited Clinical Guideline; 2002

2. https://www.ncbi.nlm.nih.gov/pubmed/2062513/

3. https://www.who.int/en/news-room/fact-sheets/detail/diabetes

4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4657234/

5. International Council of Ophthalmology. Number of Ophthalmologists in Practice and Training Worldwide. http://www.icoph.org/ophthalmologists-worldwide.html

6. http://www.eyepacs.com/data-analysis

7. http://www.adcis.net/en/third-party/messidor/

8. http://www.it.lut.fi/project/imageret/diaretdb1/

9. http://www.it.lut.fi/project/imageret/diaretdb1/

10. https://www5.cs.fau.de/research/data/fundus-images

11. 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

12. 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.

13. R. Gargeya, T. Leng, Automated identification of diabetic retinopathy using deep learning, Ophthalmology 124 (7) (2017) 962–969.

14. 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.

15. 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.

16. 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.

17. https://diabetesjournals.org/care/article/44/5/1168/138752/Multicenter-Head-to-Head-Real-World-Validation

18. https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(13)70145-1/fulltext (global)

19. https://www.brightfocus.org/macular/article/age-related-macular-facts-figures (global)

20. https://www.ncbi.nlm.nih.gov/pubmed/29302323 (China)

21. 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 colour fundus photography, Ophthalmology

22. 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.

23. 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.

24. 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

25. 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.

26. 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.

27. 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.

28. https://www.aaojournal.org/article/S0161-6420(14)00433-3/abstract (global)

29. https://www.glaucoma.org/glaucoma/glaucoma-facts-and-stats.php (global)

30. https://bjo.bmj.com/content/85/11/1277 (China)

31. 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 colour fundus photographs, Ophthalmology 125 (8) (2018) 1199–1206. doi:10.1016/j.ophtha.2018.01.023.

32. 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/

33. 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.

34. A. Chakravarty, J. Sivswamy, A deep learning based joint segmentation and classification framework for glaucoma assessment in retinal colour fundus images, arXiv preprint arXiv:1808.01355

35. Pflipsen, M., Massaquoi, M. and Wolf, S., 2016. Evaluation of the Painful Eye. American family physician, 93(12).