The Definition of Glaucomatous Optic Neuropathy in Artificial Intelligence Research and Clinical Applications

Felipe A. Medeiros,[1,2] Terry Lee,[1] Alessandro A. Jammal,[1] Lama A. Al-Aswad[3,4], Malvina B. Eydelman[5] and Joel S. Schuman[3,6-8] for the Collaborative Community for Ophthalmic Imaging Executive Committee and Glaucoma Workgroup

[1]Department of Ophthalmology, Duke University School of Medicine, Durham, NC, USA
[2]Department of Electrical and Computer Engineering, Pratt School of Engineering, Duke University, Durham, NC, USA
[3]Department of Ophthalmology, NYU Langone Health, NYU Grossman School of Medicine, New York, NY, USA.
[4]Department of Population Health, NYU Langone Health, NYU Grossman School of Medicine, New York, NY, USA.
[5]US Food and Drug Administration, Silver Spring, M.D., USA
[6]Departments of Biomedical Engineering and Electrical and Computer Engineering, New York University Tandon School of Engineering, Brooklyn, NY, USA
[7]Center for Neural Science, NYU, New York, NY, USA
[8] Neuroscience Institute, NYU Langone Health, New York, NY, USA.

Presented in part at the Collaborative Community for Ophthalmic Imaging United States Food and Drug Administration Virtual Workshop, September 3rd and 4th, 2020 and January 28, 2022.

Corresponding Author: Felipe A. Medeiros, MD, PhD, Duke Eye Center, Department of Ophthalmology, Duke University, 2351 Erwin Rd, Durham, North Carolina, 27705. E-mail: felipe.medeiros@duke.edu

FDA participates as a member of the Collaborative Community on Ophthalmic Imaging. This article reflects the views of the authors and should not be construed to represent FDA's views or policies.

Financial Support: Funding from the National Institutes of Health (Bethesda, MD, USA) R01EY021818 (FAM), and R01-EY013178 and U01EY033001 (JSS). An unrestricted grant from Research to Prevent Blindness (New York, NY) to the Department of Ophthalmology, NYU Langone Health, NYU Grossman School of Medicine, New York, NY.

Disclosures:
Lama A. Al-Aswad, MD, MPH
Aerie Pharmaceuticals, Inc.: Consultant/Advisor
GlobeChek: Equity Owner
AI Optics: Advisor
New World Medical Inc: Grant Support
Save Vision Foundation: Grant Support
Topcon Medical Systems Inc.: Research support and consultant
Verily: Consultant
Zeiss: Adviser

Michael D. Abramoff, MD, PhD
Digital Diagnostics: ICP
Novago AG, ICP

Bhavna J. Antony, PhD
IBM Research, Employee

Michael Boland, MD, PhD
Carl Zeiss Meditec – consulting
Topcon Healthcare – consulting
Janssen – consulting
Allergan – consulting

Balwantray C. Chauhan, Ph.D.
Canadian Institutes of Health Reserch (Research Support)
Alcon Reserch Instutute (Award and Research Suport)
Dalhousie Medical Research Foundation (Research Support)
CenterVue (Equipment support)
Heidelberg Engineering (Equipment and Reserch Support)
Topcon (Equipment Support)

Michael Chiang, MD
NIH: Grant support
NSF: Grant support
Genentech: Grant support
Novartis: Consultant
InTeleretina, LLC: Equity owner

Jeffrey L Goldberg, MD, PhD
Carl Zeiss Meditec – consulting

Naama Hammel, MD
Google Health, Employee

Felipe A. Medeiros, MD, PhD
Aeri Pharmaceuticals (Consultant)
Allergan (Consultant, Financial support)
Annexon (Consultant)
Biogen (Consultant)
Carl ZeissMeditec (Consultant, Financial support)
Galimedix (Consultant)
Google Inc (Financial support)
Heidelberg Engineering (Financial support)
nGoggle Inc (Patent)
Novartis (Financial support)
Stealth Biotherapeutics (Consultant)
Stuart Therapeutics (Consultant)
Reichert (Consultant, Financial support)

Louis R Pasquale, MD
Consultant for each:
Eyenovia
Syke Biosciences
Twenty twenty

Harry A. Quigley, MD
Consultant for, Injectsense, IDx, Gore; Research Support: Heidelberg, Topcon.
National Eye Institute: Grant Support

Joel S. Schuman, MD
Aerie Pharmaceuticals, Inc.: Consultant/Advisor, Equity Owner
BrightFocus Foundation: Grant Support
Boehringer Ingelheim: Consultant/Advisor
Carl Zeiss Meditec: Patents/Royalty/Consultant/Advisor
Massachusetts Eye and Ear Infirmary and Massachusetts Institute of Technology: Intellectual Property
National Eye Institute: Grant Support
New York University: Intellectual Property
Ocugenix: Equity Owner, Patents/Royalty
Ocular Therapeutix, Inc.: Consultant/Advisor, Equity Owner
Opticient: Consultant/Advisor, Equity Owner
Perfuse, Inc.: Consultant/Advisor
Regeneron, Inc.: Consultant/Advisor
SLACK Incorporated: Consultant/Advisor
Tufts University: Intellectual property
University of Pittsburgh: Intellectual property

Remo Susanna, MD
Adapt: Patents/Royalty

Jayme Vianna, MD
EadieTech: Consulting

Linda Zangwill, PhD
Grant support: National Eye Institute
Equipment and research support: Carl Zeiss Meditec Inc., Heidelberg Engineering GmbH, Optovue Inc., Topcon Medical Systems Inc.
Patent licensed to: Zeiss Meditec
Consultant: Abbvie

ABSTRACT

Objective: Although Artificial intelligence (AI) models may offer innovative and powerful ways to use the wealth of data generated by diagnostic tools, there are important challenges related to their development and validation. Most notably is the lack of a perfect reference standard for glaucomatous optic neuropathy (GON). As AI models are trained to predict presence of glaucoma or its progression, they generally rely on a reference standard that is used to train the model and assess its validity. If an improper reference standard is used, the model may be trained to detect or predict something that has little or no clinical value. This article summarizes the issues and discussions related to the definition of GON in AI applications as presented by the Glaucoma Workgroup from the Collaborative Community for Ophthalmic Imaging (CCOI) United States Food and Drug Administration (FDA) Virtual Workshop, on September 3 and 4, 2020 and on January 28, 2022.

Study Design: Review and Conference Proceedings

Subjects: No human or animal subjects or data therefrom were used in the production of this article.

Methods: A summary of the Workshop was produced with input and/or approval from all participants.

Main Outcome Measures: Consensus position of the CCOI Workgroup on the challenges in defining GON and possible solutions.

Results: The Workshop reviewed existing challenges that arise from the use of subjective definitions of GON and highlighted the need for a more objective approach to characterize GON that could facilitate replication and comparability of AI studies, and allow for better clinical validation of proposed AI tools. Different tests and combination of parameters for defining a reference standard for GON have been proposed. Different reference standards may need to be considered depending on the scenario in which the AI models are going to be applied, such as community-based or opportunistic screening versus detection or monitoring of glaucoma in tertiary care.

Conclusions: The development and validation of new AI-based diagnostic tests should be based on rigorous methodology with clear determination of how the reference standards for glaucomatous damage are constructed and the settings where the tests are going to be applied.

Glaucoma is an optic neuropathy characterized by progressive degeneration of retinal ganglion cells (RGCs) that may lead to irreversible loss of visual function. Despite the availability of effective treatments, glaucoma remains one of the leading causes of blindness in the world.[1] The number of patients with glaucoma is predicted to increase substantially as the result of an ageing population, with estimates of over 110 million people affected by 2040.[2]

The loss of RGCs in glaucoma tends to follow an insidious course, with the majority of patients being asymptomatic and unaware they have the disease until late stages. In fact, it is estimated that approximately 1 in 3 patients may have advanced visual field loss in at least one eye at the time of presentation.[3,4] In developing countries, population-based studies show that over 90% of patients with glaucoma are unaware they have the disease. Besides its asymptomatic nature in early stages, presentation with advanced disease may also occur because of factors such as economic cost, access to services, or health perceptions.

Currently, there are no effective screening strategies to identify all patients with glaucoma and a diagnosis of glaucoma or suspicious glaucoma typically occurs opportunistically during routine visits to ophthalmologists or community optometrists. Even for patients already diagnosed with glaucoma, monitoring progression over time can be challenging due to the insidious nature of the disease and the large variability often seen in tests to detect change. Thus, there is a pressing need for more effective strategies for detecting and monitoring glaucoma.

Artificial intelligence (AI) and, in particular, deep learning, has risen to the forefront of innovative approaches for screening, diagnosis and detection of glaucoma progression. Deep learning models have been applied to a variety of tests such as fundus photography, optical coherence tomography, and standard automated perimetry. However, there are challenges related to the development and validation of such models. Most notable is the lack of a perfect reference standard, or “gold standard,” in glaucoma. As these models are generally trained to predict presence of glaucoma or its progression, these models usually rely on a reference standard that is used to train the model and assess its validity. If an improper reference standard is used, the model may be trained to detect or predict something that has little or no clinical utility. In fact, the selection of the proper reference standard for validating new AI-based tests ultimately depends on the purpose of the application, i.e., whether for population-based or opportunistic screening, clinic-based diagnostics or detection of progression.

The Key Role of the Reference Standard for Training and Validating AI Models

Deep learning neural networks are computer algorithms made of several layers of interconnected artificial “neurons,” whose development was inspired by biological brain cells, but which do not really reflect their biological complexity. In a deep learning model, each artificial neuron receives input from other neurons and then performs computations in order to produce an output. Data are fed to the neural network and processed by the many (usually thousands or millions) of interconnected artificial

neurons with the goal of producing a certain desired outcome. However, before such deep learning networks can be used for specific tasks, they need to be trained so that the specific computations performed at each artificial neuron and their pattern of interconnections can be determined. This training process involves feeding the network with data, observing the results, making modifications to the model, and repeating the process iteratively, until a certain desired outcome is achieved. After the network has been trained, it can then be used to obtain predictions on previously unseen data.

There are essentially 3 ways to train a deep learning model: supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning involves training the network using a completely labeled dataset. For example, if an algorithm is aimed at identifying glaucoma on fundus photographs, it can be trained by feeding the network with labeled photos of glaucoma and normal eyes. The network then "learns" the optimal features that will lead to the best discrimination of glaucoma from a normal photo. This learning process is done by comparing the algorithm's predictions to the actual labels and readjusting the weights of the artificial neurons, in a process known as backpropagation.[5] Unsupervised learning, on the other hand, involves training the algorithm with unlabeled data with the goal of discovering hidden patterns in the data, without providing any information to the network regarding what the final outcome should be. This approach has been used, for example, to classify patterns of visual field loss in glaucoma, as well as to detect progressive change over time.[6-10] Finally, semi-supervised learning uses a combination of the two approaches.[11]

Supervised learning has been the most widely used method for developing deep learning models for detection of glaucoma.[12-22] As these models are trained to predict a certain label (e.g., glaucoma versus normal or progression versus stability), the process of labeling the data (i.e., the reference standard used), is essential. Ultimately, the deep learning model can only be as good as the labeled data. If a poor, biased or imprecise reference standard is used to label the data, this will result in a deep learning model that will essentially replicate those imperfections. Even if unsupervised training is used, the deep learning model ultimately has to be tested against some valid reference standard to assess its clinical validity. Therefore, the reference standard is key to the process of training and validation of deep learning models for diagnosis, screening and detection of glaucoma progression.

Reference Standards for AI Applications in Glaucoma: A Summary of the Collaborative Community on Ophthalmic Imaging (CCOI) Discussions

This article summarizes the issues and discussions related to the definition of GON in AI applications as presented by the Glaucoma Workgroup from the Collaborative Community for Ophthalmic Imaging (CCOI) United States Food and Drug Administration (FDA) Virtual Workshop, on September 3 and 4, 2020 and on January 28, 2022. As this work does not involve any patient or animal data nor collection or analyses of research data, institutional review board and patient informed consents

were not required. The presentation of this article further adheres to the tenets of the Declaration of Helsinki.

When establishing reference standards for training and evaluating deep learning models, it is essential to consider the goal at hand. **David Garway-Heath, MD, Moorfields Eye Hospital,** noted in the CCOI meeting that the goal of screening for a disease is very much different from that of diagnostics in a clinical setting, which in turn is different from detecting progression in known disease. Different clinical settings have different pre-test probabilities for disease and different tolerance for false-positive and false-negative rates from a cost-effectiveness perspective. For example, a test used to screen the population at large cannot have subpar specificity: that would result in massive referrals of false-positive patients who do not actually have glaucoma, thereby overwhelming specialists and draining public resources. On the other hand, a test used to monitor for progression for an already-established clinical population may be designed to tolerate more false-positives in order to ensure that the false-negative rate is minimized (i.e., not missing any patients who progress).

An important misconception concerns what constitutes early glaucoma diagnosis from a screening standpoint, which is often meant to imply diagnosis at a very early disease stage, before any significant visual field loss is detectable by perimetry or sometimes even before the appearance of clear signs of optic nerve damage. However, focusing on such early stages for screening may lead to significant problems related to uncertainty in diagnosis, besides being largely unnecessary. From a public health standpoint, an early diagnosis means diagnosis at a stage earlier than when the patient would have presented symptomatically. As symptomatic presentation of glaucoma generally occurs at a late stage, almost any stage of glaucoma can in fact be considered early detection from the point of view of screening. Given the relatively low prevalence of glaucoma and the difficulties related to discriminating early glaucoma from normal variation, attempting to focus screening programs on detection of very early disease will likely lead to failure. Moving the focus to well-established cases of glaucoma, but who would still be asymptomatic, will lead to much improved diagnostic accuracy and effectiveness. This has key implications on determining suitable reference standards to be used for development and validation of deep learning models for glaucoma screening.

An example of the challenges and importance of the reference standard in AI applications, comes from deep learning models applied to fundus photographs. Fundus photography represents a relatively low-cost option for screening for certain eye diseases, such as diabetic retinopathy.[23] There are several inexpensive, portable nonmydriatic fundus cameras that can be used in low-resource settings, making this method attractive for community-based or opportunistic screening.[24] Once a deep learning model is successfully trained to recognize the presence of disease on fundus photographs, it can then be deployed to provide gradings on previously unseen photos in real-time. Ting and colleagues[17] proposed that a deep learning algorithm could be developed to screen for glaucoma in existing teleretinal imaging. Using a large database of 494,661 teleretinal photographs, they developed an algorithm capable of

detecting images that were considered “referable” for glaucoma. The reference standard was based on subjective grading of the photographs by ophthalmologists or professional graders. In the test dataset, their algorithm detected “referable” glaucoma on photographs with an area under the receiver operating characteristic (ROC) curve of 0.942, sensitivity of 96.4%, and specificity of 87.2%. It is important to note that a specificity of 87.2% would translate into 13% of those without disease being labeled as false positives. When applied in the context of screening, this would likely result in a large number of healthy individuals being unnecessarily referred for evaluation. Therefore, to minimize the number of false positives, targeting well-established cases of glaucoma rather than all suspicious “referable” ones may be warranted in the context of population-based screening.

In contrast to diabetic retinopathy, the approach of training deep learning models to replicate human gradings of fundus photographs with the goal of screening for glaucoma may have significant limitations. Subjective gradings tend to have limited reproducibility[25-27] and poor interrater reliability.[26-28] Also, ophthalmologists tend to over diagnose glaucoma in eyes with physiologically enlarged discs and miss damage in eyes with small discs.[13] Overall, subjective gradings tend to have low specificity. If such subjective gradings are used as the reference standard to train a deep learning model, the trained model can only perform as well as those gradings and will carry all their imperfections. If used in the context of screening for the disease targeting high specificity, graders trained to detect well-established, unequivocal nerve damage, not dubious, potentially “referable” or suspect cases will be critical.

The limitations of subjective reference standards for training and validating deep learning models for glaucoma diagnosis has led to the quest for a consensus toward an objective reference standard that could be used for this purpose. **Joel S Schuman, MD, New York University-Langone,** pointed out during the CCOI meeting that since the vast majority of glaucoma research in the present day utilizes optical coherence tomography (OCT) and standard automated perimetry, these would be potential tools to be used for establishing such a reference standard. **Harry A Quigley, MD, Johns Hopkins University,** described at the CCOI meeting a recent attempt to define glaucomatous optic neuropathy (GON), based on a recent consensus process carried out with 110 glaucoma experts throughout the world.[29] The specialists were asked to agree upon several features that should be considered in defining GON including, among other factors: that the clinical examination of the retina and optic nerve would be necessary to rule out conditions simulating GON, that IOP should not be a criterion for diagnosis, that an OCT defect must be in the corresponding opposite hemifield from the visual field defect, and that OCT retinal nerve fiber layer (RNFL) assessment should be included either alone or with segmented macular thickness.[17] To find a set of OCT and perimetry criteria to define GON, they recruited participating clinicians across 13 international centers who entered 2 reliable OCT and 2 perimetric tests from eyes seen in their clinics, along with the clinician’s classification of *definite GON*, *probably GON*, or *not GON*, taking into consideration the history, clinical exam, perimetry and OCT. Classifications for a total of 2580 eyes from 1531 patients were collected. The investigators then derived objective criteria derived from OCT and Standard Automated

Perimetry (SAP) measures to predict the glaucoma status of each subject. OCTs were graded using software classifications of *normal*, *borderline*, or *abnormal* in the superior or inferior quadrants; perimetry was graded as abnormal if a glaucoma hemifield test (GHT) 'outside normal limits' with 3 points in the pattern deviation plot at a $P < 5\%$ or worse in the abnormal hemifield was present. Using this data, combinations of OCT and VF measures were used to create 4 criteria by which to determine the presence of GON, and sensitivity and specificity of each criterion was tested: sensitivity ranged from 65 to 77%, and the specificity ranged from 98 to 99%.[18] The best performing criterion achieved a sensitivity of 77% and specificity of 98% by defining GON on the basis of abnormal OCT in the superior or inferior RNFL quadrants with matching opposite, abnormal GHT in at least 1 of the 2 most recent pairs of tests (**Table 1**).[30]

Felipe A Medeiros, Duke University, presented at the CCOI meeting the results of another proposed objective definition for GON.[31] The criteria proposed that a diagnosis of GON should involve corresponding structural and functional damage, based on RNFL assessment by spectral-domain OCT (SDOCT) and visual field assessment by SAP. The set of criteria are summarized in **Table 2** and uses both global and localized parameters with the requirement that there be topographic correspondence between structural and functional damage which will enhance specificity. The investigators assessed the proposed objective reference standard against a subjective classification by glaucoma experts and found a 95.2% overall agreement, with a weighted kappa of 0.87, indicating excellent agreement. They then developed a deep learning model that used fundus photographs to discriminate glaucoma from normal eyes, which had been classified based on the objective reference standard on a dataset comprised of 9830 fundus photos from 2927 eyes of 2025 individuals. The deep learning model achieved an overall area under the ROC curve of 0.92 to discriminate between objectively defined GON and normal. Interestingly, when the same deep learning model was tested against subjectively (by glaucoma experts) defined GON and normal, the same performance was achieved. These results illustrate the potential to develop deep learning models based on objective criteria for GON.

It should be noted that the proposed approaches above for defining an objective reference standard for GON both require clinical examination to exclude other potential confounding conditions that could lead to OCT abnormality and visual field damage, such as, for example, diabetic retinopathy. Therefore, an AI algorithm trained against such reference standards may not be used solely to diagnose glaucoma, but rather as a tool to assist in referral or as an ancillary test to help in making a final diagnosis. Of course, AI algorithms could also be trained to evaluate for the presence of other conditions such as diabetic retinopathy, besides glaucoma, in more comprehensive approaches targeted at screening.

In another approach to specifying an objective definition of a definition for GON, **Jayme R Vianna, MD, Dalhousie University,** and colleagues presented at the CCOI the methods of a Crowd-Sourced Glaucoma Study, in which they created an online database of 1270 subjects with or without diagnosis of glaucoma provided by clinicians

around the around. This database included an optic disc photograph, a Humphrey 24-2 or Octopus G1 perimetry result, and OCT imaging of the optic nerve for 1 eye from each subject. Glaucoma specialists worldwide were then invited to assess eyes for likelihood of glaucoma on a scale from 0 to 100 using only the presented exam findings, with the goal that each eye receive evaluations from 20 clinicians. While data collection is ongoing the primary analysis of this study will be to assess which objective characteristics from perimetry and OCT—or a combination thereof—best discriminate between patients with high and low glaucoma likelihood. This and other approaches that utilize crowd-sourcing of glaucoma experts' opinions offer the advantage that they may help mitigate biases in glaucoma assessment that may be unique to a specific institution or study group, since this reference standard would reflect the combined opinions of experts worldwide. In doing so they also bring the collective expertise of glaucoma specialists to groups that may lack that expertise. However, because they are still based on subjective assessments, albeit those of experts, they are still subject to potential human errors and biases. In addition, crowd-sourcing may sometimes be expensive, time-consuming and difficult to achieve under a variety of scenarios.

The aforementioned approaches utilize the most widely used structural and functional metrics for glaucoma assessment in clinical practice: mean deviation (MD), pattern standard deviation (PSD), and GHT of the 24-2 or 30-2 SAP; and the OCT peripapillary scan that acquires global and sectoral RNFL thicknesses. While these are the most commonly used, **Donald C Hood, PhD, Columbia University,** pointed out at the CCOI meeting that there is some evidence that using 24-2 visual field and OCT disc/RNFL scan alone may miss some eyes with glaucoma in early disease stages. Hood and De Moraes argue that these tests may miss damage to the macular retinal ganglion cells, which can occur even in early stages of glaucoma. To address this, they proposed a new automated method[32] that uses topographical agreement between structure (OCT RNFL and retinal ganglion cell complex [RGC+] probability map) and function (10-2 and 24-2 visual field) for detecting abnormal glaucomatous changes. Hood and colleagues have also recently published on an approach where deep learning models were trained based on OCT probability maps.[33] Such models were successful in replicating expert gradings of OCTs, suggesting that they could be eventually used to assist in the diagnosis of glaucomatous damage while decreasing the reliance on expert gradings.

The decision on which specific tests and parameters to use for defining a reference standard in AI studies may depend largely on the purpose of the application. For example, if an AI algorithm is being developed for population-based or opportunistic screening for glaucoma, the reference standard that will serve as the basis for its development and validation should exhibit high specificity. Such a reference standard should be capable of detecting well-established glaucoma cases at a level of disease severity that would avoid the large diagnostic uncertainty that is seen in very early glaucoma. It seems unlikely that inclusion of macular OCT and 10-2 tests would be necessary to compose such reference standard in this context, as most unequivocal glaucoma cases can be promptly diagnosed by a combination of conventional 24-2 visual field test and OCT RNFL assessment. In contrast, if an AI algorithm is being

developed to assist clinicians in detecting the earliest signs of disease in glaucoma suspect patients being evaluated at tertiary hospitals, it may make sense for the reference standard to also make use of other tests to increase the sensitivity for early damage, such as macular OCT, for example.

Other approaches have been proposed to overcome the subjective reference standards used to train deep learning models in glaucoma. In an approach named machine-to-machine (M2M) proposed by Medeiros et al.[14] a deep learning algorithm was trained on color fundus photographs that were labeled with an objective quantitative reference standard, the corresponding global RNFL thickness measurement from SDOCT. By training the M2M deep learning algorithm to predict the RNFL thickness value when assessing a color fundus photograph, the degree of glaucomatous damage could be quantified rather than just "qualified". A strong correlation was demonstrated between the predicted RNFL value from the photo-based deep learning algorithm and the actual RNFL thickness value from the corresponding SDOCT ($r=0.832$, $p<0.001$), with a mean absolute error of approximately 7 microns. In a subsequent work[16], the authors showed that the Bruch's membrane opening-minimum rim width (BMO-MRW) parameter could also be used as a reference standard for labeling optic disc photographs. The deep learning predictions were also highly correlated with the actual BMO-MRW values (Pearson's $r=0.88$, $p<0.001$). Compared to training using subjective human labeling as reference standard or objective binary definitions of GON, the M2M approach may offer a distinct advantage, since the output is quantitative rather than qualitative, of allowing cut-offs to be established in order to optimize its application to achieve desired specificity levels.[34] In more recent longitudinal studies, the authors have also shown that the M2M model was able to successfully detect progressive glaucoma over time[35] as well as predict development of visual field loss among glaucoma suspects.[36]

It should be noted that there may be scenarios where a subjective reference standard may be a feasible option for development and validation of AI models for aiding detection of disease progression. For example, suppose that one wishes to develop a deep learning model that can replicate in a clinical setting the performance of glaucoma experts in detecting disease progression. It is then reasonable to set up a study where experts will produce the reference standard by grading tests for progression, i.e., a series of OCTs or visual fields, or both, perhaps accompanied by other clinical information, and a deep learning model will be trained to attempt to replicate such standard. Such an AI model could then potentially assist in bringing general practitioners to a level comparable to those of experts when assessing for progression in a clinical setting. When creating such reference standard, however, it is important to make sure that it represents a valid clinically relevant outcome that is also reproducible.

Reaching a consensus on an objective definition of GON has been an elusive task to the clinical and scientific community for years. However, as **Balwantray Chauhan MD, Dalhousie University,** discussed at the CCOI meeting, perhaps this is a result of genuine differences among clinicians and researchers as to what exactly

glaucoma is. And yet, as the CCOI participants emphasized, there is a dire need for such a definition, both to improve the quality and consistency of clinical practice and to facilitate research using AI in glaucoma. More importantly, most clinicians agree on obvious cases of GON. So, as **Chauhan** notes, perhaps the goal should be to arrive at a consensus on a working definition of GON first. Albeit not a perfect all-encompassing definition that includes all stages of glaucoma from its very earliest changes, it would still serve as an objective and standardized definition by which the burgeoning new AI algorithms could be compared and evaluated.

In conclusion, AI approaches offer enormous potential to develop tools for glaucoma diagnosis and assessment of progression. However, it is critically important that the development and validation of new AI-based diagnostic tests be based on rigorous methodology with clear determination of how the reference standards were constructed and the settings where the tests are ultimately going to be applied. The suitable reference standards may differ significantly depending on the proposed application. Similarly, the requirements for diagnostic accuracy may vary considerably, depending on whether the test is being considered for community-based or opportunistic screening versus detection or monitoring of disease in tertiary care. The use of objective approaches to define reference standards for GON and its progression may help improve the comparability of AI studies and allow better clinical validation of proposed tests.

References

1. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. *JAMA*. May 14 2014;311(18):1901-11. doi:10.1001/jama.2014.3192
2. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. Nov 2014;121(11):2081-90. doi:10.1016/j.ophtha.2014.05.013
3. Heijl A, Bengtsson B, Oskarsdottir SE. Prevalence and severity of undetected manifest glaucoma: results from the early manifest glaucoma trial screening. *Ophthalmology*. Aug 2013;120(8):1541-5. doi:10.1016/j.ophtha.2013.01.043
4. Boodhna T, Crabb DP. Disease severity in newly diagnosed glaucoma patients with visual field loss: trends from more than a decade of data. *Ophthalmic Physiol Opt*. Mar 2015;35(2):225-30. doi:10.1111/opo.12187
5. Chollet F. *Deep Learning with Python*. Manning Publications Co.; 2018:361.
6. Sample PA, Boden C, Zhang Z, et al. Unsupervised machine learning with independent component analysis to identify areas of progression in glaucomatous visual fields. *Invest Ophthalmol Vis Sci*. Oct 2005;46(10):3684-92. doi:10.1167/iovs.04-1168
7. Wang M, Tichelaar J, Pasquale LR, et al. Characterization of Central Visual Field Loss in End-stage Glaucoma by Unsupervised Artificial Intelligence. *JAMA Ophthalmol*. Jan 2 2020;doi:10.1001/jamaophthalmol.2019.5413
8. Yousefi S, Balasubramanian M, Goldbaum MH, et al. Unsupervised Gaussian Mixture-Model With Expectation Maximization for Detecting Glaucomatous Progression in Standard Automated Perimetry Visual Fields. *Transl Vis Sci Technol*. May 2016;5(3):2. doi:10.1167/tvst.5.3.2
9. Goldbaum MH, Sample PA, Zhang Z, et al. Using unsupervised learning with independent component analysis to identify patterns of glaucomatous visual field defects. *Invest Ophthalmol Vis Sci*. Oct 2005;46(10):3676-83. doi:10.1167/iovs.04-1167
10. Huang X, Saki F, Wang M, et al. An Objective and Easy-to-Use Glaucoma Functional Severity Staging System Based on Artificial Intelligence. *J Glaucoma*. Aug 1 2022;31(8):626-633. doi:10.1097/IJG.0000000000002059
11. Zhao R, Chen X, Xiyao L, Zailiang C, Guo F, Li S. Direct Cup-to-Disc Ratio Estimation for Glaucoma Screening via Semi-supervised Learning. *IEEE journal of biomedical and health informatics*. Aug 12 2019;doi:10.1109/jbhi.2019.2934477
12. Asaoka R, Murata H, Hirasawa K, et al. Using Deep Learning and transform learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *Am J Ophthalmol*. Oct 11 2018;doi:10.1016/j.ajo.2018.10.007
13. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology*. Aug 2018;125(8):1199-1206. doi:10.1016/j.ophtha.2018.01.023
14. Medeiros FA, Jammal AA, Thompson AC. From Machine to Machine: An OCT-Trained Deep Learning Algorithm for Objective Quantification of Glaucomatous Damage in Fundus Photographs. *Ophthalmology*. Apr 2019;126(4):513-521. doi:10.1016/j.ophtha.2018.12.033
15. Thompson AC, Jammal AA, Berchuck SI, Mariottoni EB, Medeiros FA. Assessment of a Segmentation-Free Deep Learning Algorithm for Diagnosing Glaucoma From Optical

Coherence Tomography Scans. *JAMA ophthalmology*. Feb 13 2020;doi:10.1001/jamaophthalmol.2019.5983
16. Thompson AC, Jammal AA, Medeiros FA. A Deep Learning Algorithm to Quantify Neuroretinal Rim Loss From Optic Disc Photographs. *Am J Ophthalmol*. May 2019;201:9-18. doi:10.1016/j.ajo.2019.01.011
17. Ting DSW, Cheung CY, Lim G, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*. Dec 12 2017;318(22):2211-2223. doi:10.1001/jama.2017.18152
18. Liu H, Li L, Wormstone IM, et al. Development and Validation of a Deep Learning System to Detect Glaucomatous Optic Neuropathy Using Fundus Photographs. *JAMA Ophthalmol*. Sep 12 2019;doi:10.1001/jamaophthalmol.2019.3501
19. Ahn JM, Kim S, Ahn KS, Cho SH, Lee KB, Kim US. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLoS One*. 2018;13(11):e0207982. doi:10.1371/journal.pone.0207982
20. Li F, Yan L, Wang Y, et al. Deep learning-based automated detection of glaucomatous optic neuropathy on color fundus photographs. *Graefes Arch Clin Exp Ophthalmol*. Apr 2020;258(4):851-867. doi:10.1007/s00417-020-04609-8
21. Phene S, Dunn RC, Hammel N, et al. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology*. Dec 2019;126(12):1627-1639. doi:10.1016/j.ophtha.2019.07.024
22. Shibata N, Tanito M, Mitsuhashi K, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep*. Oct 2 2018;8(1):14665. doi:10.1038/s41598-018-33013-w
23. Owsley C, McGwin G, Jr., Lee DJ, et al. Diabetes eye screening in urban settings serving minority populations: detection of diabetic retinopathy and other ocular findings using telemedicine. *JAMA Ophthalmol*. Feb 2015;133(2):174-81. doi:10.1001/jamaophthalmol.2014.4652
24. Miller SE, Thapa S, Robin AL, et al. Glaucoma Screening in Nepal: Cup-to-Disc Estimate With Standard Mydriatic Fundus Camera Compared to Portable Nonmydriatic Camera. *Am J Ophthalmol*. Oct 2017;182:99-106. doi:10.1016/j.ajo.2017.07.010
25. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. Feb 1992;99(2):215-21. doi:10.1016/s0161-6420(92)31990-6
26. Abrams LS, Scott IU, Spaeth GL, Quigley HA, Varma R. Agreement among optometrists, ophthalmologists, and residents in evaluating the optic disc for glaucoma. *Ophthalmology*. Oct 1994;101(10):1662-7. doi:10.1016/s0161-6420(94)31118-3
27. Jampel HD, Friedman D, Quigley H, et al. Agreement among glaucoma specialists in assessing progressive disc changes from photographs in open-angle glaucoma patients. *Am J Ophthalmol*. Jan 2009;147(1):39-44 e1. doi:10.1016/j.ajo.2008.07.023
28. Chan HH, Ong DN, Kong YX, et al. Glaucomatous optic neuropathy evaluation (GONE) project: the effect of monoscopic versus stereoscopic viewing conditions on optic nerve evaluation. *Am J Ophthalmol*. May 2014;157(5):936-44. doi:10.1016/j.ajo.2014.01.024
29. Iyer J, Vianna JR, Chauhan BC, Quigley HA. Toward a new definition of glaucomatous optic neuropathy for clinical research. *Curr Opin Ophthalmol*. Mar 2020;31(2):85-90. doi:10.1097/ICU.0000000000000644

30. Iyer JV, Boland MV, Jefferys J, Quigley H. Defining glaucomatous optic neuropathy using objective criteria from structural and functional testing. *Br J Ophthalmol*. Jul 22 2020;doi:10.1136/bjophthalmol-2020-316237
31. Mariottoni EB, Jammal AA, Berchuck SI, Shigueoka LS, Tavares IM, Medeiros FA. An objective structural and functional reference standard in glaucoma. *Sci Rep*. Jan 18 2021;11(1):1752. doi:10.1038/s41598-021-80993-3
32. Hood DC, Tsamis E, Bommakanti NK, et al. Structure-Function Agreement Is Better Than Commonly Thought in Eyes With Early Glaucoma. *Invest Ophthalmol Vis Sci*. Oct 1 2019;60(13):4241-4248. doi:10.1167/iovs.19-27920
33. Hood DC, La Bruna S, Tsamis E, et al. Detecting glaucoma with only OCT: Implications for the clinic, research, screening, and AI development. *Prog Retin Eye Res*. Feb 22 2022:101052. doi:10.1016/j.preteyeres.2022.101052
34. Jammal AA, Thompson AC, Mariottoni EB, et al. Human Versus Machine: Comparing a Deep Learning Algorithm to Human Gradings for Detecting Glaucoma on Fundus Photographs. *Am J Ophthalmol*. Mar 2020;211:123-131. doi:10.1016/j.ajo.2019.11.006
35. Medeiros FA, Jammal AA, Mariottoni EB. Detection of Progressive Glaucomatous Optic Nerve Damage on Fundus Photographs with Deep Learning. *Ophthalmology*. Mar 2021;128(3):383-392. doi:10.1016/j.ophtha.2020.07.045
36. Lee T, Jammal AA, Mariottoni EB, Medeiros FA. Predicting Glaucoma Development With Longitudinal Deep Learning Predictions From Fundus Photographs. *Am J Ophthalmol*. May 2021;225:86-94. doi:10.1016/j.ajo.2020.12.031