Clinical Ophthalmology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

Artificial Intelligence and Glaucoma: Going Back to Basics

Saif Aldeen AlRyalat[1], Praveer Singh[2], Jayashree Kalpathy-Cramer[2], Malik Y Kahook[2]

[1]Department of Ophthalmology, The University of Jordan, Amman, 11942, Jordan; [2]Department of Ophthalmology, University of Colorado School of Medicine, Sue Anschutz-Rodgers Eye Center, Aurora, CO, USA

Correspondence: Malik Y Kahook, Department of Ophthalmology, University of Colorado School of Medicine, Sue Anschutz-Rodgers Eye Center, Aurora, CO, USA, Email Malik.Kahook@cuanschutz.edu

Abstract: There has been a recent surge in the number of publications centered on the use of artificial intelligence (AI) to diagnose various systemic diseases. The Food and Drug Administration has approved several algorithms for use in clinical practice. In ophthalmology, most advances in AI relate to diabetic retinopathy, which is a disease process with agreed upon diagnostic and classification criteria. However, this is not the case for glaucoma, which is a relatively complex disease without agreed-upon diagnostic criteria. Moreover, currently available public datasets that focus on glaucoma have inconstant label quality, further complicating attempts at training AI algorithms efficiently. In this perspective paper, we discuss specific details related to developing AI models for glaucoma and suggest potential steps to overcome current limitations.

Keywords: artificial intelligence, glaucoma, deep learning, optic disc, segmentation

Overview

Diagnosis of glaucoma is complex, often requiring the analysis of multiple data points from clinical exams and diagnostic testing to arrive at a decision for the presence of disease, lack of disease, or need for further longitudinal evaluations to reach a definitive conclusion.[1] Traditionally, there are two main applications for Artificial Intelligence (AI) projects in glaucoma. The first is to determine the level of suspicion for the presence of disease, in which applications can be deployed in non-ophthalmic or non-glaucoma focused settings to guide in patient referral. A specific example is the use of AI tools developed by Cybersight, a not-for-profit telemedicine platform, in low- and middle-income countries leveraging fundus photos that are uploaded and analyzed with rapid feedback to the clinician.[2,3] The second, and perhaps more difficult, application of AI in glaucoma is diagnosing disease, in which applications are typically deployed directly in ophthalmology clinics. A specific example of this would be diagnosing glaucoma from optical coherence tomography (OCT) images.[4] A recent review of utilizing AI for glaucoma found 23 deep learning projects were under exploration for assisting clinicians in diagnosing and treating glaucoma.[4] Given the widely recognized complexity of diagnosing glaucoma, how can an AI model provide useful and actionable information when ophthalmologists themselves often do not agree on the basics of disease presence and progression?[1,5,6] To understand and simplify the complexity of developing useful glaucoma-specific AI applications, we discuss possible steps to achieve successful implementation in the clinical setting.

The Need for a Gold Standard

Creating algorithms to recognize presence or absence of disease requires a clear and concrete definition of the disease with high certainty. An example would be teaching an AI platform to identify the presence or absence of brain hemorrhage on computed tomography.[7] The need for a firm "consensus definition" of disease and the ability to sort out a clear presence or absence of disease sets the bar for model accuracy. This might not be a major issue for other ophthalmic diseases like diabetic retinopathy, where a single fundus image can be used reliably to diagnose the presence

Received: 22 March 2023
Accepted: 24 May 2023
Published: 31 May 2023

of bleeding or vascular abnormalities.[8] However, glaucoma is a disease with no consensus “gold-standard” definition to rely on making the development of a robust AI-based diagnostic model very problematic.

There are currently multiple professional societies and associations that try to provide definitions for what qualifies as “glaucoma”, including the American Academy of Ophthalmology (AAO),[9] The National Institute for Health and Care Excellence (NICE),[10] World Glaucoma Association (WGA),[11] European Glaucoma Society (EGS),[12] Glaucoma Research Foundation (GRF),[13] among others. Almost all of these societies and associations agree on the word “progressive” in the definition of glaucoma. However, such progression also needs to be differentiated from the age-related changes that occur in the optic nerve. Some significant points to be considered for identifying progression of glaucoma in order of importance:

- Rate of progression: Glaucomatous progression in optic nerve fiber layer loss is faster than expected from age-related changes alone, despite no agreed cut-off point for the rate to be considered glaucomatous.[14,15]
- Pattern of damage: Glaucomatous progressive optic neuropathy usually occur as selective loss of nerve fibers, typically starting in the parapapillary area and progressing from the peripheral to the central visual field.[16,17]
- Risk factors: Presence of risk factors classically associated with glaucoma is also an indicator for the diagnosis in patients with progressive optic neuropathy. Elevated intraocular pressure is a major risk factor for the development and progression of glaucoma, while age-related changes in the optic nerve head are less likely to be related to intraocular pressure.[18] Other risk factors include aging, thin corneas, and family history of glaucoma.[15,18]

Several published articles have focused on the establishment of a gold standard. For instance, one study proposed a definition of glaucomatous optic neuropathy (GON) using specific parameters from OCT and standard automated perimetry.[19] This definition was used as a reference standard to develop a deep learning algorithm for detecting GON in fundus photos, demonstrating high performance in differentiating GON from normal. Similarly, another publication identified objective criteria from OCT and perimetry that could denote a specific definition of GON in eyes with open-angle glaucoma. This study suggested that objective criteria could be practical and useful for comparisons among clinical studies, supplementing subjective clinical assessments.[20] Such attempts underline the importance of establishing objective criteria and pave the way for advancements in AI applications in glaucoma diagnosis. Accordingly, a high-quality dataset needs to show evidence of progression, preferably providing longitudinal images (rather than just mentioning that there was progression), along with clinical data to be assessed by prospective dataset users prior to its utilization in developing AI algorithms.

The Need for a Large Reliable Dataset

AI projects generally require large datasets, which can be relatively easy for diabetic retinopathy or age-related macular degeneration due to the large volume of patients and the ease of diagnosis.[21] However, for openly accessible glaucoma datasets, the patient numbers are generally small, on the mild to the moderate spectrum of disease severity, and of questionable reliability.[22] The issue becomes more complicated when we consider the quality of the datasets. In a study published in Eye, none of the reviewed datasets reached a definitive diagnosis of glaucoma.[22] In fact, when authors describe the methodology used for glaucoma diagnosis, there was a lack of sufficient detail to judge the progressive nature of the disease.[22] Furthermore, the available large and well-designed glaucoma clinical trials do not have a unified diagnostic definition for the disease.[23] Accordingly, we should be cautious while using openly accessible datasets that have patients labeled as “glaucoma”, as the methodology of such labeling should be double-checked by glaucoma experts within the team. In this regard, it is important to consider the importance of having a clinical expert in almost any AI-related glaucoma project.[24]

Longitudinal data from untreated patients represents a critical aspect of understanding the natural progression of glaucoma. For instance, untreated patient data could help train AI models to distinguish between the affects of the disease progression and the affects of the treatment, enhancing the robustness and validity of AI prediction models. It is worth noting that the generation of such datasets may pose ethical and practical challenges, given that standard of care would necessitate treatment. Despite this, some existing observational studies and clinical trials, such as the Early Manifest

Glaucoma Trial (EMGT) and the Ocular Hypertension Treatment Study (OHTS), do provide longitudinal data from untreated patients and have significantly contributed to our understanding of glaucoma progression.[25,26] Incorporating and collecting such data would undoubtedly improve the performance and clinical utility of AI systems in glaucoma.

Accurate Assessment of Ground Truth Optic Disc Segmentation

Datasets that have annotated fundus images are also used to train AI models for the classification of glaucoma. This is done through the optic disc and optic cup assessment, which is an important indicator for the presence or absence of disease as well as gauging severity. Among the most used datasets in this regard is the DRISHTI dataset.[27] As of October 2022, the DRISHTI dataset has been downloaded 275 times and cited 360 times.[27,28] We performed an assessment of the accuracy of ground truth optic disc/cup segmentation provided with the openly accessible DRISHTI dataset.

We randomly chose 10 images from the DRISHTI dataset to be assessed. We recruited five ophthalmologists and senior trainees from a tertiary referral hospital in Jordan. Each ophthalmologist annotated the optic disc and cup from each image using the cloud-based Computer Vision Annotation Tool (CVAT), a free, online, interactive video and image annotation tool for computer vision.[29] Each participant underwent training on how to use the CVAT platform. We created the rater's ground truth using the STAPLE algorithm. The STAPLE algorithm is a popular technique for estimating a consensus segmentation of a structure from multiple human or machine-derived (imperfect) segmentations.[30] Finally, an expert glaucoma specialist with more than 20 years of experience (MYK) was presented with the original optic disc image alongside each ground truth to decide which one was more accurate. Both ground truth masks utilized the same colors, as shown in Figure 1, and the rater masked to the source of ground truth (ie, DRISHTI or new ground truth) was masked from the rater.

In our analysis of 10 images, we found that the DRISHTI dataset's ground truth segmentation was deemed superior in 8 cases when compared to the segmentations generated by our group of five ophthalmologists. This indeed supports the high quality of the DRISHTI dataset. In two instances, however, the quality of the DRISHTI segmentation was deemed less accurate. It should be noted that, in one case, both DRISHTI and our raters incorrectly classified a region of peripapillary atrophy as optic disc neuroretinal rim (Figure 1). Such suboptimal accuracy in the ground truth used to train AI models may result in lower accuracy than reported after model training. These findings suggest the need for caution upon using openly accessible datasets to train optic disc segmentation models. We suggest reviewing masks calculated from consensus segmentation, even if originally developed using multiple experts.

Development of Accurate AI Models

The diagnosis of glaucoma may require the assessment of anatomical and functional data for the same patient, in addition to the clinical and demographic data.[23] The majority of available AI models depend on a single modality as an input to diagnose glaucoma,[31] which is not how physicians approach diagnosing the presence or progression of disease in clinical practice. Recently, the use of multimodal AI models as a way of harvesting data from multiple modalities as an input to provide a more reliable assessment has been introduced in medicine.[32] More recently, Xiong et al published a study on

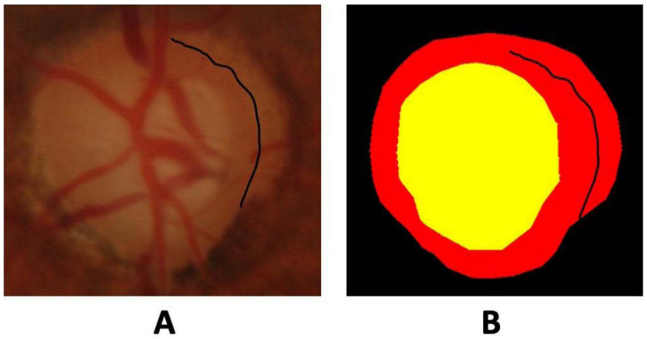

Figure 1 Fundus image of an optic disc (**A**) showing peripapillary atrophy wrongly annotated as being part of the neuroretinal rim in the original ground truth (**B**).

the use of a multimodal algorithm to have visual fields and peripapillary OCT scans to detect glaucomatous optic neuropathy, where they showed how such a multimodal algorithm was superior to models that depended on either visual fields or peripapillary OCT scans alone.[33] Whether such advances in AI models can improve accuracy in clinical settings is yet to be determined. However, multimodal algorithms are closer to what physicians do in disease diagnoses.

Clinical Settings Implementation

Models developed and tested showing high accuracies do not usually show similar accuracy upon real-life implementation. This has been shown in different ophthalmic subspecialties[34,35] including glaucoma.[36] As a result, AI models should be tested extensively in the actual setting in which they will be implemented.

For example, an AI model developed for screening might be implemented in general ophthalmic practice rather than glaucoma-specialized settings. Another point that needs to be addressed is the feasibility of such implementation. In an analysis of AI-glaucoma screening in rural China, the authors found an excess cost of such implementation that could not be offset by the reduction in disease progression.[37] Potential reasons why an AI model for glaucoma diagnosis might not show accurate results in clinical practice include:

- Lack of diversity in training data: AI models rely on large datasets to learn how to accurately classify images. If the training dataset is biased or lacks diversity, the AI model may not perform as well on new data.[38] For example, if an AI model is developed using mostly images of glaucoma in Caucasian populations, it may not perform as well on images of glaucoma in other populations, such as Asian or African populations.
- Variability in imaging quality: AI models are sensitive to variability in image quality, and may not perform as well on images that are of lower quality.[39] Variability in imaging quality can arise from factors such as differences in equipment, lighting, or patient positioning, and can lead to variability in the features that the AI model uses to classify images.
- Limited generalizability: AI models may perform well on the specific type of images that they were trained on, but may not generalize well to other types of images.[40] For example, an AI model that was developed using fundus photographs may not perform as well on other types of images, such as OCT scans.

Next Steps in Glaucoma AI

While the role of glaucoma-focused AI platforms in non-specialized settings is clearer, its role in ophthalmic and glaucoma-specific settings requires more intensive contemplation. Going back to the basics in this regard might be beneficial to set a path toward enhanced utility in real-world settings. First, defining what a “gold standard” is for diagnosing glaucoma that is also acceptable for training AI models is imperative. The diagnosis of glaucoma is now mostly based on clinical assessment rather than defined criteria, which would create high variability between experts with agreement only on advanced cases.[41] Depending on clinical assessment, it can also create glaucoma diagnosis variability due to demographic variation between physicians and their available resources. To minimize such diagnostic variability, we believe that we can initially agree on cases that can be definitively diagnosed as glaucoma (eg, cases of documented accelerated progression), and then move toward less certain cases that represent the “gray zone” between definitive glaucoma and definitive non-glaucoma cases.[42] After that, a cloud-based registry can be established that includes definitively diagnosed glaucoma patients according to the agreed-upon definitions. Finally, developing a sophisticated AI model would be of greater value, considering the accuracy of its ground truth data (Figure 2).

Rather than adding more deep learning models that might have poor performance, going back to the basics by optimizing the diagnosis of glaucoma is likely the better option. Otherwise, developing an AI that has poor performance would further complicate clinical setting adoption by physicians who are seeking tools that assist rather than confuse decision-making. On the other hand, the benefits of developing an AI model that performs well or helps diagnose glaucoma patients are unquestionable. Finally, reliable digital tools are most needed in low-resource settings, where diagnostic equipment is often absent, skilled clinicians are often in short supply, and where time to make clinical decisions is limited.

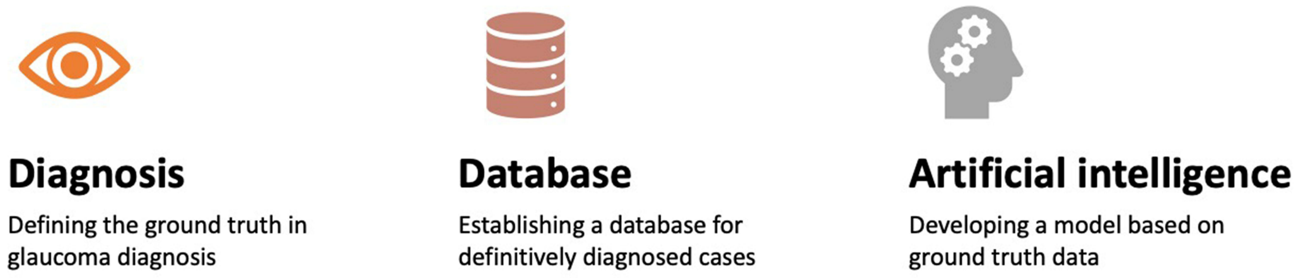
Diagnosis
Defining the ground truth in glaucoma diagnosis
Database
Establishing a database for definitively diagnosed cases
Artificial intelligence
Developing a model based on ground truth data

Figure 2 Proposed framework for enhancing artificial intelligence research in glaucoma.

Disclosure

Professor Jayashree Kalpathy-Cramer reports grants from NIH, Genentech, and GE healthcare; consultant for Silioam Vision LLC, outside the submitted work. The authors report no other conflicts of interest in this work.

References

1. 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*. 2009;147:39–44.e1. doi:10.1016/j.ajo.2008.07.023
2. Glaucoma Today. AI for glaucoma care. Available from: https://glaucomatoday.com/articles/2022-july-aug/ai-for-glaucoma-care. Accessed May 25, 2023.
3. Cybersight. AI-driven automated interpretation for fundus images. Available from: https://cybersight.org/ai-consult/. Accessed May 25, 2023.
4. Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Transl Vis Sci Technol*. 2020;9:42. doi:10.1167/tvst.9.2.42
5. Yokoyama Y, Tanito M, Nitta K, et al. Stereoscopic analysis of optic nerve head parameters in primary open angle glaucoma: the glaucoma stereo analysis study. *PLoS One*. 2014;9:e99138. doi:10.1371/journal.pone.0099138
6. 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*. 2019;126:1627–1639. doi:10.1016/j.ophtha.2019.07.024
7. Kuo W, Häne C, Mukherjee P, Malik J, Yuh EL. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. *Proc Natl Acad Sci*. 2019;116:22737–22745. doi:10.1073/pnas.1908021116
8. AlRyalat SA, Muhtaseb R, Alshammari T. Simulating a colour-blind ophthalmologist for diagnosing and staging diabetic retinopathy. *Eye*. 2020;1–4. doi:10.1038/s41433-020-01232-z
9. American Academy of Ophthalmology (AAO). American Academy of Ophthalmology [Homepage]. Available from: https://www.aao.org/. Accessed May 25, 2023.
10. The National Institute for Health and Care Excellence. NICE [Homepage]. Available from: https://www.nice.org.uk/. Accessed May 25, 2023.
11. World Glaucoma Association (WGA). World Glaucoma Association [Homepage]. Available from: https://wga.one/. Accessed May 25, 2023.
12. European Glaucoma Society (EGS). Available from: https://www.eugs.org/eng/default.asp. Accessed May 25, 2023.
13. Glaucoma Research Foundation is dedicated to finding a cure; 2021. Available from: https://glaucoma.org/. Accessed May 25, 2023.
14. Musch DC, Gillespie BW, Lichter PR, et al. Visual field progression in the Collaborative Initial Glaucoma Treatment Study the impact of treatment and other baseline factors. *Ophthalmology*. 2009;116:200–207. doi:10.1016/j.ophtha.2008.08.051
15. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. *JAMA*. 2014;311:1901–1911. doi:10.1001/jama.2014.3192
16. Ha A, Park KH. Optical coherence tomography for the diagnosis and monitoring of glaucoma. *Asia-Pac J Ophthalmol*. 2019;8:135.
17. Lloyd MJ, Mansberger SL, Fortune BA, et al. Features of optic disc progression in patients with ocular hypertension and early glaucoma. *J Glaucoma*. 2013;22:343–348. doi:10.1097/IJG.0b013e31824c9251
18. Jammal AA, Berchuck SI, Thompson AC, Costa VP, Medeiros FA. The effect of age on increasing susceptibility to retinal nerve fiber layer loss in glaucoma. *Invest Ophthalmol Vis Sci*. 2020;61:8. doi:10.1167/iovs.61.13.8
19. Mariottoni EB, Jammal AA, Berchuck SI, et al. An objective structural and functional reference standard in glaucoma. *Sci Rep*. 2021;11:1752. doi:10.1038/s41598-021-80993-3
20. Iyer JV, Boland MV, Jefferys J, Quigley H. Defining glaucomatous optic neuropathy using objective criteria from structural and functional testing. *Br J Ophthalmol*. 2021;105:789–793. doi:10.1136/bjophthalmol-2020-316237
21. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health*. 2021;3:e51–e66. doi:10.1016/S2589-7500(20)30240-5
22. AlRyalat SA. Machine learning on glaucoma: the missing point. *Eye*. 2021;1–2. doi:10.1038/s41433-021-01561-7
23. AlRyalat SA, Ertel MK, Seibold LK, Kahook MY. Designs and methodologies used in landmark clinical trials of glaucoma: implications for future big data mining and actionable disease treatment. *Front Med*. 2022;9:818568. doi:10.3389/fmed.2022.818568
24. AlRyalat SA, Al-Ryalat N, Ryalat S. Machine learning in glaucoma: a bibliometric analysis comparing computer science and medical fields' research. *Expert Rev Ophthalmol*. 2021;16:511–515. doi:10.1080/17469899.2021.1964956
25. Heijl A, Leske MC, Bengtsson B, Hyman L, Bengtsson B, Hussein M. Reduction of intraocular pressure and glaucoma progression: results from the early manifest glaucoma trial. *Arch Ophthalmol*. 2002;120:1268–1279. doi:10.1001/archopht.120.10.1268

26. Gordon MO, Beiser JA, Brandt JD, et al. The ocular hypertension treatment study: baseline factors that predict the onset of primary open-angle glaucoma. *Arch Ophthalmol*. 2002;120:714–720. doi:10.1001/archopht.120.6.714
27. Sivaswamy J, Krishnadas SR, Datt Joshi G, Jain M, Syed Tabish AU, Drishti GS. Retinal image dataset for optic nerve head(ONH) segmentation. *Int Symp Biomed Imaging*. 2014;53–56. doi:10.1109/ISBI.2014.6867807
28. Drishti GS. Retina dataset for ONH segmentation. Available from: https://www.kaggle.com/datasets/lokeshsaipureddi/drishtigs-retina-dataset-for-onh-segmentation. Accessed May 25, 2023.
29. About. Computer Vision Annotation Tool (CVAT). Available from: https://openvinotoolkit.github.io/about/. Accessed May 25, 2023.
30. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*. 2004;23:903–921. doi:10.1109/TMI.2004.828354
31. Mursch-Edlmayr AS, Ng WS, Diniz-Filho A, et al. Artificial intelligence algorithms to diagnose glaucoma and detect glaucoma progression: translation to clinical practice. *Transl Vis Sci Technol*. 2020;9:55. doi:10.1167/tvst.9.2.55
32. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28:1773–1784. doi:10.1038/s41591-022-01981-2
33. Xiong J, Li F, Song D, et al. Multimodal machine learning using visual fields and peripapillary circular OCT scans in detection of glaucomatous optic neuropathy. *Ophthalmology*. 2022;129:171–180. doi:10.1016/j.ophtha.2021.07.032
34. Alryalat SA, Al-Antary M, Arafa Y, et al. Deep learning prediction of response to anti-VEGF among diabetic macular edema patients: Treatment Response Analyzer System (TRAS). *Diagnostics*. 2022;12:312. doi:10.3390/diagnostics12020312
35. Coyner AS, Oh MA, Shah PK, et al. External validation of a retinopathy of prematurity screening model using artificial intelligence in 3 low- and middle-income populations. *JAMA Ophthalmol*. 2022;140:791–798. doi:10.1001/jamaophthalmol.2022.2135
36. Chaurasia AK, Greatbatch CJ, Hewitt AW. Diagnostic accuracy of artificial intelligence in glaucoma screening and clinical practice. *J Glaucoma*. 2022;31:285. doi:10.1097/IJG.0000000000002015
37. Xiao X, Xue L, Ye L, Li H, He Y. Health care cost and benefits of artificial intelligence-assisted population-based glaucoma screening for the elderly in remote areas of China: a cost-offset analysis. *BMC Public Health*. 2021;21:1065. doi:10.1186/s12889-021-11097-w
38. Kuhlman C, Jackson L, Chunara R. No computation without representation: avoiding data and algorithm biases through diversity. *Int Conf Knowl Discov Data Min*. 2020;3593. doi:10.1145/3394486.3411074
39. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172:1122–1131.e9. doi:10.1016/j.cell.2018.02.010
40. Overfitting and generalization of AI in medical imaging. Available from: https://www.acrdsi.org/DSIBlog/2022/04/11/Overfitting-and-Generalization-of-AI-in-Medical-Imaging. Accessed May 25, 2023.
41. Quigley HA, Boland MV, Iyer JV. Evaluating an objective definition of glaucomatous optic neuropathy: an international collaborative effort. *Invest Ophthalmol Vis Sci*. 2020;61:1440.
42. Mohamed-Noriega J, Sekhar GC. Defining and diagnosing glaucoma: a focus on blindness prevention. *Community Eye Health*. 2021;34:32–35.