Clinical Ophthalmology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Evaluating the Influence of Clinical Data on Inter-Observer Variability in Optic Disc Analysis for AI-Assisted Glaucoma Screening

Sayeh Pourjavan[1,2], Gen-Hua Bourguignon[1], Cristina Marinescu[2], Loic Otjacques[2], Antonella Boschi[1]

[1]Department of Ophthalmology, Cliniques Universitaires Saint Luc, UCL, Brussels, Belgium; [2]Department of Ophthalmology, Chirec Hospital Groups, Delta Hospital, Brussels, Belgium

Correspondence: Sayeh Pourjavan, Email Sayeh.pourjavan@saintluc.uclouvain.be

Clinical Ophthalmology downloaded from https://www.dovepress.com/
For personal use only.

Purpose: This study aims to evaluate the inter-observer variability in assessing the optic disc in fundus photographs and its implications for establishing ground truth in AI research.

Methods: Seventy subjects were screened during a screening campaign. Fundus photographs were classified into normal (NL) or abnormal (GS: glaucoma and glaucoma suspects) by two masked glaucoma specialists. Referrals were based on these classifications, followed by intraocular pressure (IOP) measurements, with rapid decisions simulating busy outpatient clinics.In the second stage, four glaucoma specialists independently categorized images as normal, suspect, or glaucomatous. Reassessments were conducted with access to IOP and contralateral eye data.

Results: In the first stage, the agreement between senior and junior specialists in categorizing patients as normal or abnormal was moderately high. Knowledge of IOP emerged as an independent factor influencing the decision to refer more patients. In the second stage, agreement among the four specialists varied, with greater concordance observed when additional clinical information was available. Notably, there was a statistically significant variability in the assessment of optic disc excavation.

Conclusion: The inclusion of various risk factors significantly influences the classification accuracy of specialists. Risk factors like IOP and bilateral data influence diagnostic consistency among specialists. Reliance solely on fundus photographs for AI training can be misleading due to inter-observer variability. Comprehensive datasets integrating multimodal clinical information are essential for developing robust AI models for glaucoma screening.

Plain Language Summary: Glaucoma is a leading cause of irreversible blindness, and early detection is critical in preventing blindness. Screening for glaucoma using fundus photographs is one approach, but there is significant variability in how specialists interpret these images. This study evaluated how consistently different eye specialists assess these photographs and what this variability means for developing artificial intelligence (AI) tools to detect glaucoma.

The study involved 70 individuals screened for glaucoma using fundus photographs. Two specialists initially classified the images as either normal or abnormal (including glaucoma suspects). The agreement between the specialists was moderate, showing that different clinicians sometimes reach different conclusions based on the same images. The study also tested how additional information, like intraocular pressure (IOP), affects these classifications. Surprisingly, including IOP data introduced more variability, making agreement between the specialists even lower.

The research highlights that relying solely on fundus photos without considering other clinical factors, like IOP or data from both eyes, could be misleading when developing AI tools. For AI to effectively assist in glaucoma detection, it must be trained on comprehensive datasets that include more than just fundus images.

These findings emphasize the importance of using a broad range of clinical data when training AI models for glaucoma screening to improve accuracy and reliability in real-world settings.

Keywords: artificial intelligence, diagnostic imaging, glaucoma screening, clinical decision support, multimodal diagnostic

Received: 24 October 2024
Accepted: 20 December 2024
Published: 27 December 2024

© 2024 Pourjavan et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v3.0) License (http://creativecommons.org/licenses/by-nc/3.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

Introduction

Glaucoma is the leading cause of irreversible blindness worldwide, affecting 3.5% of individuals aged 40 to 80 years.[1] Over half of cases remain undiagnosed due to the asymptomatic nature of early disease,[2] often leading to advanced stages with significant visual impairment, impacting the quality of the patient's life. Additionally, the economic burden associated with late-stage disease is substantial; the direct care costs for advanced glaucoma (stages 4–5, Bascom Palmer system) are 1.7 times higher than for early-stage disease.[3] These challenges highlight an urgent need for improved strategies for early detection to prevent blindness.

However, mass screening faces logistical and clinical challenges, including reliance on skilled personnel, limited diagnostic tools, and high patient volumes in underserved regions. Inter-observer variability in optic disc assessment further complicates glaucoma diagnosis, emphasizing the need for innovative tools to enhance accuracy.

Recent advancements in artificial intelligence (AI), particularly deep learning, have demonstrated diagnostic performance exceeding 95% AUC in internal datasets.[4–6] AI has the potential to reduce variability and provide consistent evaluations, making it invaluable in busy clinics and resource-limited settings.

This study originated from a glaucoma screening campaign conducted on World Sight Day. Seventy individuals were screened, and fundus photographs were rapidly classified as normal or abnormal, reflecting the time constraints typical of a busy outpatient clinic condition. The observed unexpected variability during this rapid decision-making process among trained professionals motivated further investigation with masked evaluations by experienced colleagues, transforming this campaign into a research opportunity.

Purpose

To evaluate inter-observer variability in optic disc classification using colour fundus photographs, with and without intraocular pressure (IOP) data, and to assess the reliability of ground truth labels based solely on fundus images for the external validation of AI algorithms in glaucoma detection.

Methods

This cross-sectional observational study was conducted at our hospital in accordance with the principles of the Declaration of Helsinki. The Institutional Ethical Commission of the Cliniques Universitaires Saint-Luc in Brussels approved the study protocol (OVO-AI Study: Observer Variability in Optic disc and AI Implications). This research adheres to the principles outlined in the Declaration of Helsinki for studies involving human subjects. It complies with all relevant national ethical guidelines, including the General Data Protection Regulation (GDPR) for data privacy and confidentiality within Europe. While HIPAA is specific to the United States, the GDPR ensures similar standards in the European context. Written informed consent was obtained from each participant before inclusion in the study.

Fundus images from both eyes were obtained from a cohort of 70 individuals presenting randomly during a glaucoma screening campaign in the hospital. The only two exclusion criteria were age and a confirmed diagnosis of glaucoma. The minimum age for screening was set at 30 years due to the racial heterogeneity (Caucasians, Asians, Latinos, and African descendants) in Brussels. The clinical parameters were age, bilateral fundus pictures, taken with a Crystalvue NFC600 camera and intraocular pressure measured with air tonometry, NIDEK Tonoref III.

The Crystalvue NFC600 camera is equipped with MONA AI-driven software capable of providing a probability score for glaucoma based on fundus images. MONA, a Belgian healthcare AI startup, is a spin-off of KULeuven and VITO (the Flemish technology agency). Initially developed for detecting diabetic retinopathy (DR) and diabetic macular edema (DME), MONA AI was trained on an extensive dataset of over 273,000 retinal images. The system achieved a sensitivity of 90% and specificity of 95% for referable DR and similarly high accuracy for DME detection. Its convolutional neural network (CNN) architecture leverages these vast datasets to identify subtle pathological changes with high consistency. Although specific details about MONA's glaucoma training dataset are unavailable, it likely employs a comparable AI framework, assigning optic disc scores from 0.0 to 1.0, with higher scores indicating a greater likelihood of glaucoma.

While the MONA AI system is not yet commercially available, its use in this study was dictated by camera availability. AI-generated probability scores were accessed post hoc for supplementary analysis and were not part of the initial evaluations. The inclusion of MONA was not intended as validation or a comparative assessment of the software.

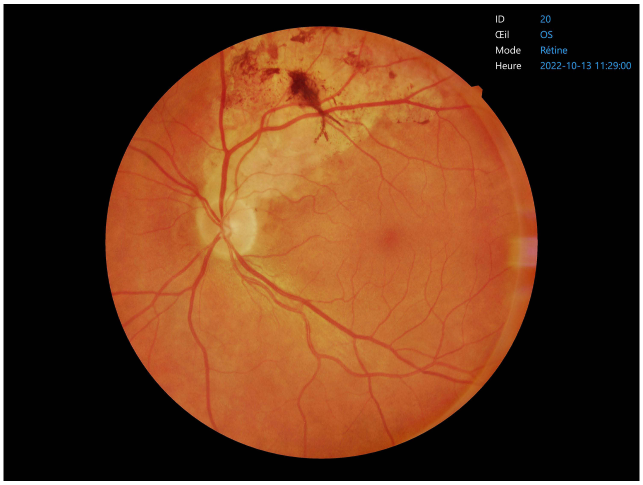

Figure 1 Asymptomatic patient during screening with branch occlusion and vasculitis caused by recent toxoplasmosis infection.

Fundus photographs were visually evaluated by two masked glaucoma specialists (SP and GB) who categorized the images as normal (NL) or abnormal (GS) including glaucoma and glaucoma suspects, in real-time, simulating the time constraints of a busy outpatient clinic. The classification adhered to the European Glaucoma Society (EGS) guidelines,[7] using parameters such as optic disc size and rim, vessel bearing, peripapillary haemorrhages, and cup-to-disc ratio.

Intraocular pressure (IOP) was subsequently measured using an air tonometer (NIDEK Tonoref III), which underwent routine calibration per manufacturer recommendations to ensure accuracy. Elevated IOP values were defined as >21 mmHg. Patients with abnormal fundus classifications (GS) and/or elevated IOP were referred for further ophthalmological evaluation. During the screening, one asymptomatic patient showed an important branch occlusion (Figure 1) and was directly sent to the Ophthalmology emergency for further examination and was diagnosed with a toxoplasmosis-caused occlusion. He was discarded from the study.

In the second phase, the fundus photographs of 138 optic discs were reanalyzed by four ophthalmologists (three glaucoma specialists and one resident). The images were categorized into three groups: normal (N), glaucoma-suspect (S), and glaucomatous (G) per EGS guidelines. Vertical Cup-to-Disc Ratio (VCDR) values were recorded for each optic nerve. The classifications were conducted in a masked manner without knowledge of IOP or the condition of the contralateral disc. Following unblinding, IOP and contralateral disc appearance were revealed sequentially, allowing observers to update their classifications.

Statistical Analysis

The Kappa coefficients were used to assess the agreement among different specialists for categorical variables and the Weighted Kappa coefficients for categorical ordinal variables. Kappas with their standard error are reported.

Intraclass correlation coefficients (ICC) were used to evaluate agreement in VCDR scoring on the image level for continuous variables.

Results

138 images of 69 subjects were taken. The statistical descriptive data of the continuous variables are shown in Table 1.

Results from the Hospital Screening

During the screening in the hospital, the flow of incoming patients limited the decision and scoring time. It was nearly instantaneous, less than one minute per eye mimicking a normal clinical setting.

Of the total 138 right and left eye pictures, 120 were labelled NL (61/ 69 for the RE, 61 and 59/ 69 for the LE). Eighteen subjects were referred for further examination based solely on their fundus photographs when either observer classified one or both eyes as GS. The number of referred subjects increased to 27, based on their initial allocation to the

Table 1 Descriptive Analysis of the Participants

F/M	47.7/52.3%	
Variable	**Mean ± SD**	**Min-max**
Age (years)	57 ± 12	33–82
IOP RE	15.5 ± 4.5	10–28
IOP LE	15.8 ± 4.4	10–30

Table 2 Number of Referred Patients with or Without Knowledge of the IOP

	Referral Based on Pictures	Referral Based on Pictures and IOP
SP	5	4
GB	8	6
Both	5	17
total	18	27

Table 3 Reproducibility Between SP & GB Based on Fundus Pictures

	Cohen's Weighted Kappas	*p*
RE	0.53 ± 0.17	0.001
LE	0.57 ± 0.15	< 0.001
BE	0.55 ± 0.11	<0.001

Table 4 Reproducibility Between SP & GB Based on Fundus Pictures with or Without IOP

	Cohen's Weighted Kappas		*p*	
	Without IOP	With IOP	Without IOP	With IOP
RE	0.38 ± 0.12	0.24 ± 0.13	0.001	0.015
LE	0.35 ± 0.19	0.29 ± 0.14	< 0.001	0.007

GS group and, or high IOP, after revealing the eye pressure (Table 2). It indicates that IOP, for clinicians, is considered a crucial and independent risk factor for further investigation.

The concordance between the 2 masked observers (Table 3) during the screening, where the observers took the pictures and scored them only as NL or GS without knowing the IOP, is quantified by the kappa's values for both eyes (kappa ± SE = 0.55 ± 0 0.11, $p < 0.001$). It means an acceptable reproducibility between the glaucoma specialist (SP) and her junior resident (GB) score.[8]

The Kappa values for the initial assessment (without IOP) indicate "moderate" agreement (Table 4).

The values of the weighted Kappa's decline when the IOP is revealed. The lower Kappa values after including IOP fall into the "fair" agreement category, suggesting less consistency when IOP data is factored in. This reduction in Kappa values suggests that the inclusion of IOP data may have introduced some variability in the assessment, leading to a lower level of agreement between the observers. This variability could arise from differing interpretations of the significance of IOP values in the context of fundus images.

Results from the Masked Picture Analysis

In the second phase of the study, three glaucoma specialists (SP, LO, CM) and one glaucoma resident (GB) independently reviewed the fundus photographs under various conditions to assess classification consistency. The evaluations were performed in the following scenarios:

1. Separate Classification: Each eye was assessed individually without additional information or time constraints.
2. Incorporating Contralateral Eye Information: Observers compared the images of both eyes from the same subject.
3. With and Without IOP Consideration: Observers classified the images without IOP data initially, followed by a second round where IOP values were revealed, allowing for potential reclassification.

The images were categorized into three groups: Normal (N), Suspect (S), and Glaucoma (G). Additionally, the vertical cup-to-disc ratio (VCDR) was recorded for each optic disc. Table 5 and Table 6 summarize the descriptive analysis of classifications and inter-observer agreement, respectively.

The comparison of classification performance across the different conditions is summarized below:

Table 5 Descriptive Analysis of Patient Classifications Across Different Conditions and Comparisons, with Data Provided by Multiple Observers. N= Normal, S= Suspect, G= Glaucoma

	RE			**RE + IOP**			**RE Compared to LE**			**RE+IOP Compared to LE**			**LE**			**LE + IOP**			**LE Compared to RE**			**LE + IOP Compared to RE**		
	N	**S**	**G**	**N**	**S**	**G**	**N**	**S**	**G**	**N**	**S**	**G**	**N**	**S**	**G**	**N**	**S**	**G**	**N**	**S**	**G**	**N**	**S**	**G**
SP	60	8	1	60	8	1	60	7	2	60	7	2	59	8	2	59	8	2	59	8	2	59	8	2
GB	59	8	2	60	7	2	60	6	3	60	9	3	59	9	1	61	8	0	57	10	2	58	10	1
LO	56	10	1	54	13	1	53	13	1	53	14	1	57	8	2	55	13	0	54	11	2	55	12	1
CM	56	10	3	49	16	3	56	9	4	55	9	4	57	9	3	47	19	3	54	12	3	53	12	3

Table 6 Comparison of Classification Performance Across Different Conditions and Among Different Specialists. The Bolded Values Reflect the Highest Agreement for Each Specialist

Weighted Kappas ± SE								
	Condition 1		**Condition 2**		**Condition 3**		**Condition 4**	
	Separate Classification for Each Eye (With & Without IOP Consideration)		**Classification with Contralateral Eye Consideration (With & Without IOP)**		**Classification with and Without Consideration of the Other Eye (IOP unknown)**		**Classification with and Without Consideration of the Other Eye (IOP known)**	
	RE	**LE**	**RE**	**LE**	**RE**	**LE**	**RE**	**LE**
SP	0,42±0,19	0,56±0,15	**1,00±0,00**	**1,00±0,00**	0,84±0,09	0,71±0,10	0,39±0,19	0,45±0,16
GB	**0,95±0,05**	**0,82±0,09**	0,91±0,07	0,82±0,08	0,91±0,06	0,77±0,08	0,95±0,05	0,77±0,10
LO	0,72±0,11	0,61±0,11	0,42±0,12	0,57±0,10	**0,87±0,08**	**0,87±0,08**	0,72±0,09	0,77±0,09
CM	0,80±0,08	0,69±0,10	0,74±0,09	0,71±0,09	**0,96±0,03**	**0,89±0,06**	0,89±0,06	0,92±0,05

- SP (Senior Specialist) achieved perfect agreement (Kappa = 1.00) when using contralateral eye information (Condition 2). However, performance declined when relying solely on IOP or without any additional information.
- GB (Resident) demonstrated consistently high performance across all conditions, indicating robust diagnostic capability irrespective of data inputs.
- LO (Specialist) showed significant improvement when contralateral eye information was available (Conditions 2 and 3) but lower agreement when relying only on IOP.
- CM (Specialist) consistently performed well in all conditions, with the highest agreement observed when both contralateral eye information and IOP were included (Condition 3).

Table 6 shows that SP performs perfectly when the other eye's information is available (Condition 2), but its performance is lower when relying solely on IOP or without any additional information. GB performs consistently well across all conditions, indicating its robustness to various data inputs. LO performance benefits most when the other eye's information is available (Conditions 2 and 3) but shows lower performance when relying solely on IOP. CM performs strongly in all conditions, with the highest agreement when other eye information is available (Conditions 2 and 3).

In Condition 3, where classifications were based on contralateral eye information but excluded IOP, emerged as the optimal scenario, showing high inter-observer agreement. This condition provides a balance where all raters perform well, suggesting that using the other eye's information without IOP gives the most reliable classification across the board and makes it the most robust scenario for ground truth classification. In conclusion, including contralateral eye information significantly improved classification reliability across all observers. In contrast, including IOP often introduced variability, leading to lower agreement in several conditions. Condition 3—leveraging contralateral eye data without IOP—proved to be the most reliable for establishing ground truth. The consistently high performance of GB and CM further highlights the potential for enhanced diagnostic accuracy when incorporating bilateral information.

Results from VCDR Scoring

Estimation of the optic disc excavation by all the raters is shown in Table 7.

This table shows that the more "experimented" ophthalmologists rate the excavations smaller than the "younger" colleagues.

Figure 2 shows the Vertical Cup to Disc scoring (VCDR) for each rater.

Table 7 Estimation of the Optic Disc Excavation by All the Raters

	Nr	Minimum C/D	Maximum C/D	Average C/D ±SD	
SP OD	69	0.00	0.60	0.27	0.13
SP OG	69	0.10	0.60	0.27	0.12
GB OD	69	0.00	0.75	0.30	0.19
GB OG	69	0.00	0.75	0.32	0.18
LO OD	67	0.00	0.70	0.33	0.16
LO OG	67	0.00	0.70	0.33	0.18
CM OD	69	0.10	0.65	0.27	0.16
CM OG	69	0.10	0.65	0.29	0.16
Valid Nr	67				

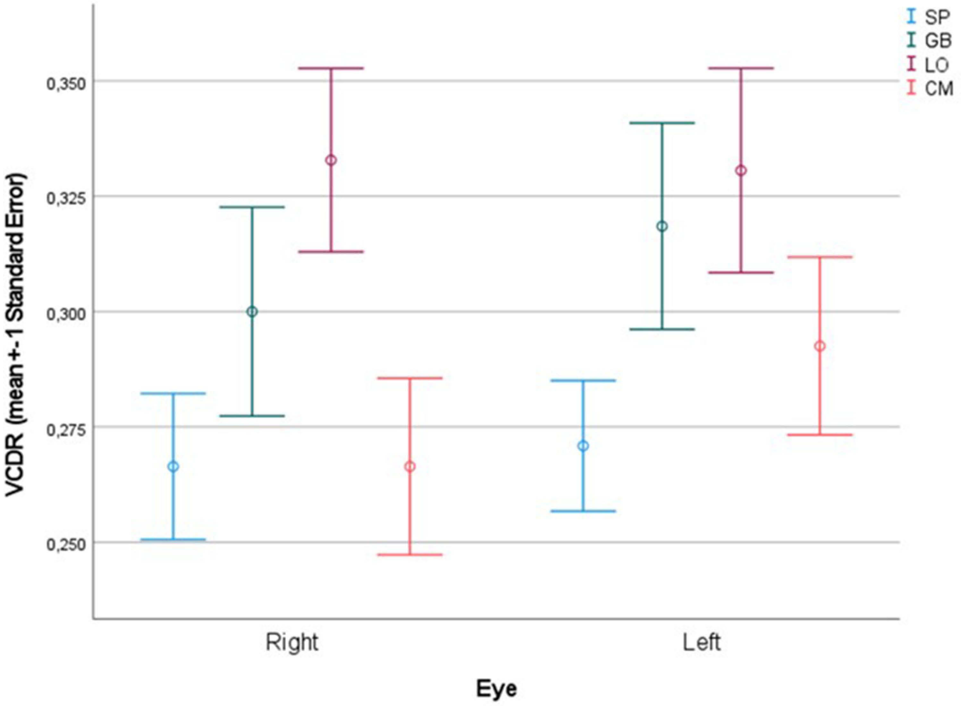

Figure 2 Shows the Vertical Cup to Disc scoring (VCDR) for each rater.

Result from AI-MONA

The MONA algorithm analyzes fundus images to assess the likelihood of glaucoma, as shown in Figure 3. While the system does not provide a numeric image-quality score, it automatically rejects poor-quality images where the optic nerve head is not visible or is improperly centred. For acceptable images, MONA generates a probability score between 0 and 1, where higher scores indicate a greater likelihood of glaucoma.

- The algorithm operates with a defined threshold: Scores below 0.73: Classified as healthy and Scores of 0.73 or higher: Indicate a high probability of glaucoma.

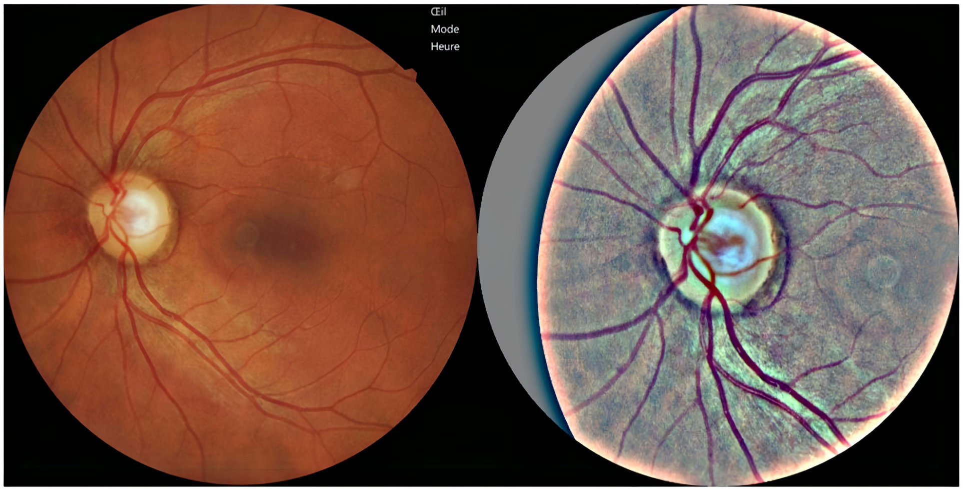

Figure 3 The original picture and the processed picture by MONA.

https://doi.org/10.2147/OPTH.S492872

Table 8 The Descriptive Analysis by MONA, Based on Risk Scores. The Higher the Risk Scores, the Higher the Possibility of Having Glaucoma

Risk Score	Mean ± SD	Min	Max	Rejected
RE	0,587,494,512 ± 0,11,121,722	0,282,902,298	0,282,902,298	7/69
LE	0,586,386,255 ± 0,12,020,194	0,319,392,204	0,831,739,731	1/69

Notably, the algorithm does not classify images as "Suspect", focusing only on binary outcomes (healthy or glaucomatous). Table 8 provides a descriptive analysis of the risk scores generated by MONA, illustrating the distribution of scores and their correlation with the likelihood of glaucoma (Table 8). Figure 4 shows the Scatter Plot of MONA- AI risk factors for both eyes separately.

The Pearson correlation coefficient between the RE and LE is 0.23 indicating a weak positive correlation.

Figure 5 shows the different histograms and kernel density estimates (KDEs) for both eyes to represent the distribution differences.

Discussion

Initially designed as an awareness campaign, this study uncovered significant inter-observer variability in clinical glaucoma diagnostics. Despite shared training and adherence to standardized methodologies, the findings highlighted the inherently subjective nature of optic disc evaluation. Notably, each specialist's diagnostic reasoning, shaped by individual experience and cognitive processes, operated like a distinct "black box", even within a structured framework. For example, while intraocular pressure (IOP) often introduced complexity, assessing both optic discs simultaneously improved diagnostic accuracy in some cases, further illustrating the variability in clinical approaches.

The small sample size of 70 subjects reflects the practical limitations of data collection during the campaign. However, the results provide valuable insights into diagnostic variability and decision-making processes under real-world conditions. The increase in referrals from 18 to 27 when intraocular pressure (IOP) data was included underscores IOP's role as an independent glaucoma risk factor, though its inclusion introduced more significant variability and lowered agreement among observers. Notably, Condition 3, where bilateral fundus information was available, but IOP was excluded, showed the highest inter-observer agreement, indicating its potential as a reliable framework for establishing ground truth.

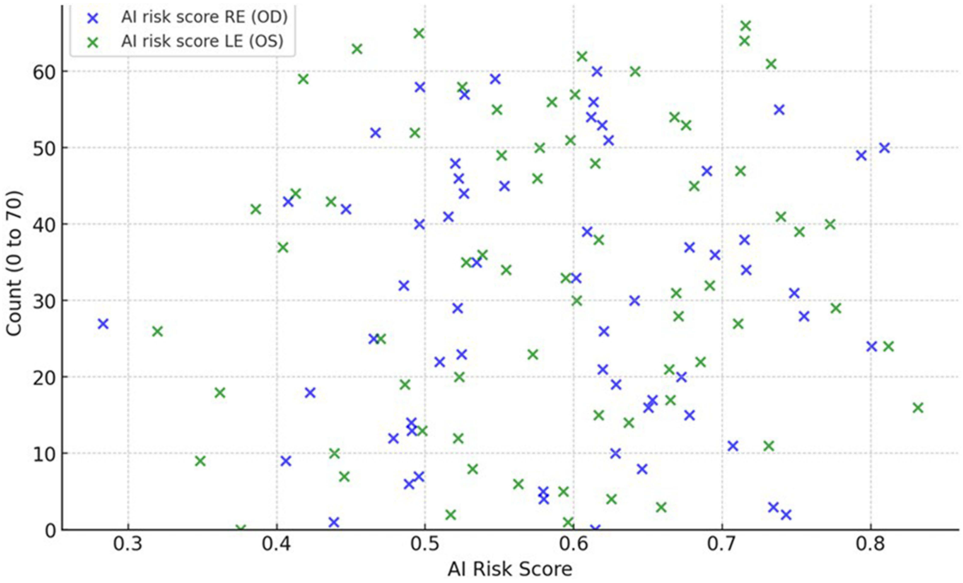

Figure 4 The Scatter Plot of MONA- AI risk factors for both eyes separately.

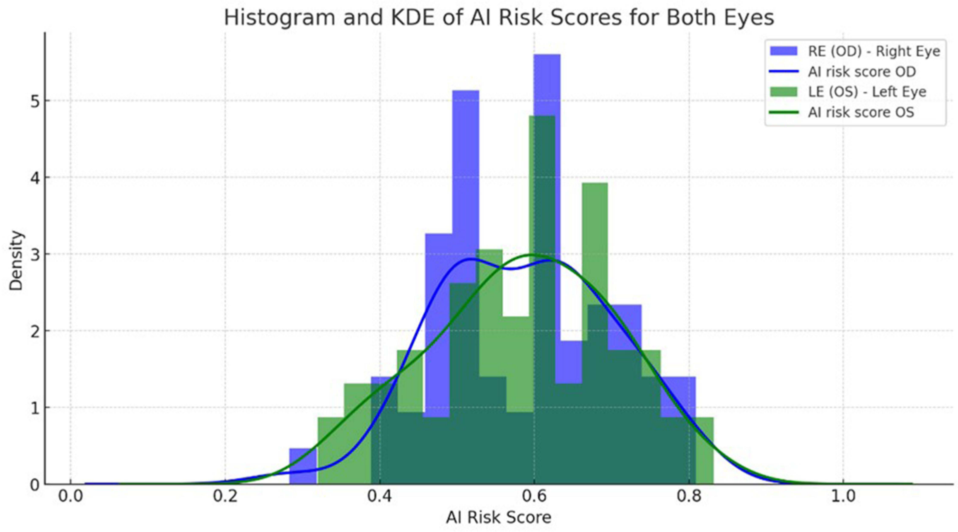

Figure 5 Histograms and kernel density estimates (KDEs) for both eyes to represent the.

Clinical and AI Implications

The moderate agreement observed in this study highlights limitations in using fundus photographs as the sole basis for training and validating AI models. Despite employing the European Glaucoma Society (EGS) Guidelines, subjective interpretations of these criteria led to inconsistencies, complicating comparisons between AI and human evaluations. These findings align with prior research showing that AI models trained on datasets with comprehensive clinical parameters demonstrate superior generalizability compared to those reliant solely on fundus photo-based labels.

While not a validation focus in this study, the MONA AI system generated probability scores post hoc, revealing a weak positive correlation between scores for both eyes. This contrasts with the higher agreement among clinicians when bilateral images were evaluated together. MONA's inability to classify "Suspects" or identify non-glaucomatous pathologies, such as visible vasculitis, underscores the limitations of narrow task-specific training. This further emphasizes the need for broader and more diverse clinical datasets to enhance AI robustness and generalizability.

AI employment in clinical decision-making as a support system faces several challenges. Several studies indicate that high image quality is essential for accurately classifying fundus images.[9,10] Standard fundus cameras lack integrated software to assess image quality. In contrast, optical coherent tomography (OCT) and angio-OCT devices provide an image quality score and recommend discarding images that fall below a certain threshold.[11] The determination of image quality in fundus photography is often subjective. An observer may choose to classify the optic disc even if the image exhibits mild blurriness. Additionally, many patients participating in screening programs are elderly and may have cataracts or corneal ageing, leading to inherently lower image quality. The data's quality should represent the expected data level that the model will encounter in practice,[12] meaning elderly with mild to pronounced blurriness of the media. Therefore, the ground truth is fundamentally subjective by default.

Consequently, comparing the performance of AI algorithms to such variability becomes complex. These findings are in keeping with our previous work,[13,14] where the performance of a robust glaucoma-AI model trained on a dataset with ground truth labels based on deep phenotyping including all available clinical parameters had excellent generalizability in various external databases, but with the least good performance in datasets which had ground truth labels bases on only fundus photo evaluation.

In this study, other important clinical parameters, such as visual field, optic nerve head measurements with optical coherence tomography, and corneal thickness, were not performed to confirm the presence of glaucoma at the end. In a similar study by Al Aswad. et al,[15] the deep learning AI (Pegasus) could be compared to a gold standard of glaucoma versus healthy diagnosis because the testing was performed on a known database of patients: fundus images randomly selected from the Singapore Malay Eye Study (SiMES).[16]

https://doi.org/10.2147/OPTH.S492872

This study utilized a small cohort of 69 subjects obtained during a glaucoma screening campaign. While small sample sizes can be useful for initial exploratory studies, they pose significant generalizability and statistical power limitations. In glaucoma screening, the disease prevalence may vary across different populations, making it essential to have a representative sample size to draw accurate conclusions. A larger patient sample would improve the robustness and external validity of the results, allowing for better assessments of AI performance and generalizability to real-world clinical settings. In comparison, in a large systematic review of 2021 including 6 original studies comparing the performance of ophthalmologists and AI, the rate of included patients with glaucomatous optic neuropathy ranged from 41.4% to 50%. In our study, the subjects were in hospital for different reasons with a much lower rate of glaucoma, closer to the actual prevalence in the population. This difference in prevalence in our cohort was also a source of lower inter-observer reproducibility.

VCDR scoring, we observed considerable variation in how clinicians rated excavation. In this study, the longer a clinician had been practicing, the smaller their evaluation of the excavation tended to be, highlighting both inter- and possible intra-clinician variability in C/D ratings.[17,18]

Conclusion

The limitations of inter-observer variability and small cohort highlighted in this discussion, emphasize the need for cautious interpretation of studies evaluating AI systems for glaucoma screening in clinical practice solely based on fundus photos. These observations should remind AI researchers that the performance of AI algorithms relies on the robustness of the ground truth labels. We should consider a hybrid approach, for example, supplementing "Condition 3" with other clinically relevant information in this study. Future research should address these limitations by incorporating larger and more diverse patient cohorts considering multiple diagnostic parameters: specifically for glaucoma, it is recommended to determine the ground truth based on not only fundus pictures but also additional metadata such as visual field examinations or Optical Coherence Tomography scans.[19] Our findings resonate with experiences in other medical domains, where AI has been shown to complement rather than replace human expertise. Studies demonstrate that while AI and specialists independently achieve similar diagnostic accuracy for recognizing a disease, their combined efforts significantly outperform either alone. This underscores AI's strength in analyzing vast and diverse datasets, particularly for rare conditions, while clinicians bring a holistic perspective grounded in clinical experience. Similarly, in glaucoma diagnostics, AI has the potential to enhance the accuracy of human assessments by mitigating inter-observer variability and identifying subtle features that may be overlooked in busy clinical settings". By addressing these challenges, the potential of AI for glaucoma screening in clinical settings should be more effectively evaluated.

Acknowledgments

We wish to express our appreciation to MONA Health for generously granting access to the algorithm and for the prompt and insightful responses to our queries.

Funding

A grant has been awarded for the doctorate thesis of the first author of which this article is one of the publications. Glaucoma Centre Ophthalmologists Zaventem outpatient financially supported the publication of this article.

Disclosure

Gen-Hua Bourguignon is the co-author. All authors declare no competing interests.

References

1. 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*. 2014;121(11):2081-90. doi:10.1016/j.ophtha.2014.05.013
2. Soh Z, Yu M, Betzler BK, et al. The global extent of undetected glaucoma in adults: a systematic review and meta-analysis. *Ophthalmology*. 2021;128(10):1393-404. doi:10.1016/j.ophtha.2021.04.009
3. Lee PP, Walt JG, Doyle JJ, et al. A multicenter, retrospective pilot study of resource use and costs associated with severity of disease in glaucoma. *Arch Ophthalmol Chic Ill*. 2006;124(1):12-9.

4. Chaurasia AK, Greatbatch CJ, Hewitt AW. Diagnostic accuracy of artificial intelligence in glaucoma screening and clinical practice. *J Glaucoma*. 2022;31(5):285. doi:10.1097/IJG.0000000000002015
5. Buisson M, Navel V, Labbé A, et al. Deep learning versus ophthalmologists for screening for glaucoma on fundus examination: a systematic review and meta-analysis. *Clin Experiment Ophthalmol*. 2021;49(9):1027-38. doi:10.1111/ceo.14000
6. Islam M, Poly TN, Yang HC, Atique S, Li YCJ. Deep learning for accurate diagnosis of glaucomatous optic neuropathy using digital fundus image: a meta-analysis. *Stud Health Technol Inform*. 2020;270:153-7. doi:10.3233/SHTI200141
7. Spaeth GL. European glaucoma society terminology and guidelines for glaucoma, 5th edition. *Br J Ophthalmol*. 2021;105(Suppl 1):1-169. doi:10.1136/bjophthalmol-2021-egsguidelines
8. Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall; 1991.
9. Gonçalves MB, Nakayama LF, Ferraz D, et al. Image quality assessment of retinal fundus photographs for diabetic retinopathy in the machine learning era: a review. *Eye*. 2024;38:426–433. doi:10.1038/s41433-023-02717-3
10. Sabottke CF, Spieler BM. The effect of image resolution on deep learning in radiography. *Radiol Artif Intell*. 2020;2(1):e190015. doi:10.1148/ryai.2019190015
11. Pradhan ZS, Sreenivasaiah S, Srinivasan T, et al. The importance of signal strength index in optical coherence tomography angiography: a study of eyes with pseudoexfoliation syndrome. *Clin Ophthalmol*. 2022;16:3481–3489. doi:10.2147/OPTH.S378722
12. Dow ER, Keenan TDL, Lad EM, et al. Collaborative community for ophthalmic imaging executive committee and the working group for artificial intelligence in age-related macular degeneration. from data to deployment: the collaborative community on ophthalmic imaging roadmap for artificial intelligence in age-related macular degeneration. *Ophthalmology*. 2022;129(5):e43–e59. doi:10.1016/j.ophtha.2022.01.002
13. Tan NYQ, Friedman DS, Stalmans I, Ahmed IIK, Sng CCA. Glaucoma screening: where are we and where do we need to go? *Curr Opin Ophthalmol*. 2020;31(2):91. doi:10.1097/ICU.0000000000000649
14. Hemelings R, Elen B, Barbosa-Breda J, et al. Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. *Acta Ophthalmol*. 2020;98(1):e94-100. doi:10.1111/aos.14193
15. Al-Aswad LA, Kapoor R, Chu CK, et al. Evaluation of a deep learning system for identifying glaucomatous optic neuropathy based on color fundus photographs. *J Glaucoma*. 2019;28(12):1029-34. doi:10.1097/IJG.0000000000001319
16. Rosman M, Zheng Y, Lamoureux E, et al. Review of key findings from the Singapore Malay Eye Study (SiMES-1). *Singapore Med J*. 2012;53(2):82-7.
17. Lichter PR. Variability of expert observers in evaluating the optic disc. *Trans Am Ophthalmol Soc*. 1976;74:532–572.
18. Kwon YH, Adix M, Zimmerman MB, et al. Variance owing to observer, repeat imaging, and fundus camera type on cup-to-disc ratio estimates by stereo planimetry. *J Glaucoma*. 2009;18(4):305–310. doi:10.1097/IJG.0b013e318181545e
19. Pourjavan S, Gouverneur F, Macq B, et al. Advanced analysis of OCT/OCTA images for accurately differentiating between glaucoma and healthy eyes using deep learning techniques. *Clin Ophthalmol*. 2024;18:3493–3502. doi:10.2147/OPTH.S472231

Clinical Ophthalmology

Dovepress
Taylor & Francis Group

Publish your work in this journal

Clinical Ophthalmology is an international, peer-reviewed journal covering all subspecialties within ophthalmology. Key topics include: Optometry; Visual science; Pharmacology and drug therapy in eye diseases; Basic Sciences; Primary and Secondary eye care; Patient Safety and Quality of Care Improvements. This journal is indexed on PubMed Central and CAS, and is the official journal of The Society of Clinical Ophthalmology (SCO). The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/clinical-ophthalmology-journal