Comparative Performance of an AI Tool and First-Year Residents in Retinal Disease Detection and Glaucoma Assessment: A Study in a Mexican Tertiary Care Setting

Dalia Camacho-García-Formentí[1], **Gabriela Baylón-Vázquez**[2], **Karen Arriozola-Rodríguez**[2], **Enrique Avalos-Ramirez**[2], **Curt Hartleben-Matkin**[2], **Hugo Valdez-Flores**[3], **Damaris Hodelin-Fuentes**[4], and **Alejandro Noriega**[1]

[1]Prosperia Salud S.A. de C.V., Mexico City, Mexico.
[2]Glaucoma Department, Instituto de Oftalmología Fundación Conde de Valenciana IAP, Mexico City, Mexico.
[3]Retina Department, Hospital de Nuestra Señora de la Luz, Mexico City, Mexico
[4]Ophthalmology Department, Hospital General Dr. Juan Bruno Zayas Alfonso, Santiago de Cuba, Cuba

Abstract

Background: Artificial intelligence (AI) shows promise in ophthalmology, but its performance in diverse healthcare settings remains understudied. We evaluated AI, an AI-powered screening tool developed with Mexican data, against first-year ophthalmology residents in a tertiary care setting in Mexico City.

Methods: We analyzed 435 adult patients undergoing their first ophthalmic evaluation. AI and residents' assessments were compared against expert annotations for retinal lesions, cup-to-disk ratio (CDR) measurements, and glaucoma suspect detection. We also evaluated a synergistic approach combining AI and resident assessments.

Results: For glaucoma suspect detection, AI outperformed residents in accuracy (88.6% vs 82.9%, $p = 0.016$), sensitivity (63.0% vs 50.0%, $p = 0.116$), and specificity (94.5% vs 90.5%, $p = 0.062$). While, the synergistic approach deemed a higher sensitivity (80.4%) than ophthalmic residents alone or AI alone

($p < 0.001$). AI's CDR estimates showed lower mean absolute error (0.056 vs 0.105, $p < 0.001$) and higher correlation with expert measurements ($r = 0.728$ vs $r = 0.538$). In retinal lesion detection, AI demonstrated superior sensitivity (90.1% vs 63.0% for medium/high-risk lesions, $p < 0.001$) and specificity (95.8% vs 90.4%, $p < 0.001$). Furthermore, differences between AI and residents were statistically significant across all metrics. The synergistic approach achieved the highest sensitivity for retinal lesions (92.6% for medium/high-risk, 100% for high-risk) while maintaining good specificity (87.4%).

Conclusion: AI outperforms first-year residents in key ophthalmic assessments. The synergistic use of AI and resident assessments shows potential for optimizing diagnostic accuracy, highlighting the value of AI as a supportive tool in ophthalmic practice, especially for early-career clinicians.

Keywords: Artificial intelligence; glaucoma; cup-disk ratio; retinal disease; ophthalmology residents.

1. Introduction

The need for ophthalmic screenings in Mexico has increased significantly due to high prevalence of risk factors associated with ophthalmic diseases. Diabetes and hypertension, which affect approximately 15 million and 40 million Mexicans over 20 years old, respectively,[1–3] are risk factors for several conditions. These include glaucoma, diabetic retinopathy (DR), diabetic macular edema (ME), hypertensive retinopathy, and cataracts.[4–7] Furthermore, the elderly population, approximately 18 million individuals,[8] also faces higher risks of glaucoma, age-related macular degeneration (AMD) and cataracts.[4,7,9]

Glaucoma presents additional risk factors, including family history and increasing age,[4] affecting primarily individuals over 40 years old.[10] This age group comprises over 44 million people in Mexico.[11] Additionally, it is estimated that over 1.5 million Mexicans have glaucoma and 50% remain undiagnosed.[12] Moreover, glaucoma is the second leading cause of irreversible blindness in Mexico and worldwide.[12]

While periodic ophthalmic evaluations are recommended for these at-risk populations, Mexico faces a significant shortage of ophthalmologists, with only 4,213 registered on July 2024. Of whom 31.5% are concentrated in Mexico City.[13] This scarcity makes comprehensive screening of all at-risk individuals unfeasible.

To address the growing need of ophthalmic screenings, researchers have developed AI models for automatic detection of retinal diseases in fundus images,[14–21] as well as models that analyse the optic disk for glaucoma suspect detection.[22–27]

Even though most AI research in ophthalmology has been done in Asia, North America and Europe,[28] several AI systems and models have been validated on real-world settings in Latin America. Arenas-Cavalli *et al* have validated DART, a DR screening tool, on the Chilean health system.[20] Rogers *et al* evaluated the Pegasus AI system for DR detection on a Mexican cohort,[16] and González-Briceño *et al* evaluated their models on primary care data from the Mexican Institute of Social Security.[15]

The latter validations are focused mainly on DR screening on primary care settings. However, to our knowledge, AI systems that perform both retinal disease detection and glaucoma suspect detection have not been evaluated, nor its potential on tertiary care settings in Latin America.

AI can be a valuable tool in clinical practice in ophthalmology hospitals, as even experienced ophthalmologists exhibit significant variability in CDR estimation, DR severity grading and retinal disease detection.[29–32] Even when ophthalmologists manually annotate optic cup and optic disk, CDR estimates vary.[29,30] Also, on ophthalmological referrals there is moderate agreement between a single retina specialist compared to a consensus of three retina specialists. In contrast, general ophthalmologists have minimal agreement in the same comparison.[32] This suggests that first-year ophthalmology residents would likely exhibit lower agreement for ophthalmological referral assessment.

Previous studies have demonstrated that AI can enhance ophthalmologists' sensitivity in detecting diabetic retinopathy (DR).[33] Moreover, AI-based CDR measurements for glaucoma assessment have been shown to surpass the accuracy of the average expert.[30] Thus, using AI as an assistive tool can help standardize and improve key diagnostic measurements such as CDR.

Moreover, AI tools may offer particular benefits for early-career clinicians.[34,35] Some researchers have explored how AI can enhance the learning process of ophthalmology residents for DR and pathological myopia (PM) detection.[34,35] This potential for AI to support early-career clinicians adds another dimension to its value in ophthalmic practice.

Our study evaluates an AI-powered ophthalmic screening tool, in a tertiary care setting. We consider it may prove valuable in ophthalmology hospitals, particularly during initial consultations often conducted by first-year residents, who are still developing their diagnostic skills.

We extend current knowledge by comparing AI performance not only to expert annotations but also to first-year ophthalmology residents, providing insights into how AI might support early-career clinicians. We also explore how AI could aid first-year residents in diagnosis and subspecialty referrals as part of the triage process. Additionally, we assess the agreement between AI and first-year residents in cataract

detection to evaluate the software's efficacy in identifying media opacities.

By evaluating AI performance across multiple ophthalmic conditions simultaneously, we also contribute to understanding how integrated AI systems might function in real-world clinical scenarios.

2. Materials and methods

We performed ophthalmic screenings using an ophthalmic screening tool on adult patients over 18 years old, that underwent their first ophthalmic evaluation by a first-year ophthalmic resident. These studies were carried out from February 12 to March 14, 2024, at Conde de Valenciana Centro, an ophthalmic institute in Mexico City.

This project was approved by Fundación de Asistencia Privada Conde de Valenciana IAP Ethics in Research Committee (CEI-2023/12/01), the Fundación de Asistencia Privada Conde de Valenciana IAP Biosecurity Committee (CB-0053-2023), and the Fundación de Asistencia Privada Conde de Valenciana IAP Research Committee (CI-053-2023). Patients included in this study signed an informed consent, previous to their participation. This study was conducted according to the Declaration of Helsinki guidelines.

For each patient, we collected the hospital medical record ID, personal information, information related to risk factors, and ophthalmic symptoms. Some of the variables were sex, age, self-reported diabetes and hypertension, blurry vision, increased sensitivity to light, and difficulty to see at night among other variables. The complete list of variables is presented on Appendix A.

The latter information is required by the AI platform, to perform part of the screening process in which logical rules are applied to obtain some of the outcomes from the platform, such as media opacity.

Retinal fundus images were also required for screening, in order for the AI tool to evaluate presence of retinal disease and to calculate CDR. For each patient, retinal fundus images were taken with a Horus 45° autofocus portable fundus non-mydriatic camera from Jedmed.

2.1. Data

A total of 468 screenings were done on 464 patients that underwent their first ophthalmic evaluation by a first-year ophthalmic resident at Conde de Valenciana Centro.

Nine screenings were excluded from the analysis due to an error in registration that resulted in ages less than 18 years old. Furthermore, six studies were excluded

because of missing images resulting from camera problems. In addition, 15 patients were excluded, due to having empty medical records. Therefore, screenings for 435 patients were considered for analysis.

The average age was 59.1 (15.7 SD) years. Of the evaluated patients, 34.0% were men and 66.0% were women, 32.2% reported having been diagnosed with diabetes and 39.3% with hypertension.

2.2. Medical records

The medical records of the patients involved in this study were identified by the medical record ID. We extracted the following information: CDR by eye, initial diagnosis defined by the first year residents including whether it was associated with glaucoma, retina, or cataract, and reference to specialists or additional tests corresponding to glaucoma subspecialty, retina subspecialty, or cataract surgery.

2.3. AI screening tool

For this project, the screenings were performed with an ophthalmic screening tool named retinIA, previously developed using Mexican data.[36] It analyzes retinal images, as well as patient information, such as risk factors and ophthalmic symptoms. The latter to give a recommendation whether to schedule an appointment with an ophthalmologist, to update their glasses, or to repeat the study the following year.

Along with the main recommendation, the screening tool gives a possible prediagnosis, which may include: DR, ME, AMD, PM, other retinal lesions, media opacity, possible ophthalmic damage due to chronic conditions, and possible presence of other ophthalmic conditions. It also estimates the CDR to help detect suspects of glaucoma.

For retinal disease analysis, the AI tool incorporates a multi-output convolutional neural network. The convolutional base corresponds to that of the Inception V3 architecture.[37] This is followed by a dense layer of 2,048 units, that further divides into two branches: one for image quality and laterality, and another for disease detection and associated retinal lesion severity. The complete architecture can be seen on Appendix B. Prior to implementing the model, images are resized to 512x512 and preprocessed using a method based on.[38]

CDR estimation involves several steps. First, images are resized to 256x256, and a U-Net architecture[39] is used to detect the region of interest (ROI) where the optic disk is present. The optic disk ROI is extracted and resized to 256x256. Image quality is then verified using a convolutional neural network based on Inception V3 with 256x256 input. If quality meets the criteria, a second U-Net is used to extract the optic cup. The segmented image is re-scaled to the original ROI size, and the heights of the optic cup and optic disk are extracted to calculate CDR.

This AI tool also includes explainability features for the AI analysis performed on retinal images. These features include a colormap for retinal anomalies, obtained by combining GradCam[40] and SmoothGrad[41] methods. Additionally, a close-up of the optic disk is provided, with markings of the optic disk and optic cup heights and the CDR estimate.

2.4. Ground-truth annotation

A total of 1,013 fundus images were obtained during data collection, 918 belonged to the 435 patients that were considered for analysis.

To determine the groundtruth value for CDR analysis, three ophthalmologists annotated bounding boxes around the optic disk and the optic cup. This was performed on the 861 images where the optic disk was present in the image. Bounding box annotation was performed on LinkedAI annotation platform.[42]

The cup disk ratio (CDR) was calculated using the height of the optic disk and the height of the optic cup, determined by the bounding boxes. The groundtruth CDR was calculated as the average of the CDR of the three experts.

We determined the groundtruth for glaucoma suspect as follows. If the CDR was 0.6 or more, or if the CDR of one eye exceeded the CDR of the other eye by 0.2 or more, the patient was then considered as a glaucoma suspect. This criteria were based on Harizman *et al*'s definition on absence of glaucomatous optic neuropathy, considering measurements associated to CDR.[43]

To determine the groundtruth for presence of retinal lesions, all 918 images were annotated by a retinal expert and an ophthalmic expert. For images in which there was a discordance between the retina expert and the ophthalmologist, the groundtruth was obtained from the agreement between the retina expert and the ophthalmologist through negotiation.

Annotations were performed on the Televal platform,[44] where experts selected different retinal findings and lesions that are associated with different ophthalmic diseases, including DR, ME, AMD, and PM, among others. A complete list of all annotated findings can be found in Appendix C and the logical rules to determine a possible prediagnosis and risk of visual loss are presented in Appendix D.

3. Results

For both glaucoma suspects and presence of retinal lesions, we calculated accuracy, specificity, sensitivity, positive predictive value (PPV) and F1-score. To assess the statistical significance of the performance differences between AI and residents, we used a bootstrap estimation method to calculate p-values. This method estimated

the proportion of cases where residents obtained better metrics than AI across multiple resamples of the data. Additionally, for CDR we evaluated the absolute and relative errors, and calculated the Pearson correlation coefficient (r). Furthermore, we calculated the receiver operating characteristic curve (ROC) considering glaucoma suspect detection as the outcome and the maximum estimated CDR by patient as a model. Finally, we present the percentage of agreement between residents diagnosis of cataract and AI's detection of media opacities.

3.1. Glaucoma suspect and cup disk ratio

The AI software analysed CDR for 61.6% of the patients, for the remaining 38.4% of patients the software considered patients' images had insufficient quality for CDR evaluation. Meanwhile, groundtruth CDR was obtained for 78.6% of patients, where all three ophthalmologists considered that images had sufficient quality for grading.

On the other hand, residents of ophthalmology that evaluate patients at an in-person consultation registered CDR for 95.4% of patients. Additionally, 97.7% of those detected by the resident with glaucoma were referred to glaucoma subspecialty or to further tests.

We compared CDR annotated by the residents with referral to further glaucoma assessment, considering the 0.6 threshold for CDR. Among the referred patients, 89.4% had a CDR $\geq$ 0.6. Meanwhile, 73.6% of patients with a CDR $\geq$ 0.6 were referred to a glaucoma subspecialist or for further tests.

The analysis to asses metric performance was done on 245 patients (56.3%), corresponding to those who had a CDR evaluation by residents of ophthalmology, the AI tool, and all three ophthalmologists.

For performance assessment of AI, we classified patients as glaucoma suspects if their CDR was 0.55 or higher in at least one eye. This threshold was set by maximizing the F1-score between the highest CDR per patient and the ground truth for glaucoma suspect. This threshold also results on a high specificity point (94.5%).

Metrics for glaucoma suspect detection are shown in Table 1. They include the evaluation of the AI software and the ophthalmology residents' diagnosis. Additionally, we present the metrics that would arise from a possible synergy between ophthalmology residents and AI, where a positive outcome for glaucoma is determined if either the residents of ophthalmology or AI indicated possible presence of glaucoma.

When comparing AI's performance to that of the ophthalmology residents, AI consistently outperformed the residents across all metrics. The AI tool achieved an accuracy of 88.6% compared to 82.9%, a sensitivity of 63.0% compared to 50.0%, a

specificity of 94.5% compared to 90.5%, a PPV of 72.5% compared to 54.8%, and an F1-score of 67.4% compared to 52.3%.

Additionally, the highest sensitivity arises from the synergy between the residents' diagnosis and AI's detection, achieving a sensitivity of 80.4%. Moreover, this combination yields an F1-score of 67.3%, maintaining a similar balance between sensitivity and PPV as AI on its own.

	Accuracy	**Sensitivity**	**Specificity**	**PPV**	**F1 score**
Resident diagnosis	82.9% (78.0%, 87.3%)	50.0% (35.3%, 64.8%)	90.5% (86.2%, 94.1%)	54.8% (39.1%, 69.7%)	52.3% (38.0%, 64.5%)
AI	**88.6%** (84.5%, 92.2%)	63.0% (50.0%, 77.3%)	**94.5%** (91.1%, 97.6%)	**72.5%** (57.1%, 86.1%)	**67.4%** (55.2%, 77.8%)
Resident or AI	85.3% (80.4%, 89.4%)	**80.4%** (68.3%, 91.5%)	86.4% (81.2%, 90.7%)	57.8% (44.9%, 69.6%)	67.3% (55.4%, 76.0%)

Table 1: Performance metrics for glaucoma suspect detection. The values considered for AI correspond to a high specificity point, where the best F1-sccore is met.

Statistical analysis using bootstrap estimation was performed to assess the differences in performance between AI and residents. The analysis revealed statistically significant differences in accuracy ($p = 0.016$), positive predictive value ($p = 0.02$), and F1-score ($p = 0.026$). The differences in sensitivity ($p = 0.116$) and specificity ($p = 0.062$) were not statistically significant at the 0.05 level.

We also evaluated significance of the increase in sensitivity obtained from the synergistic approach. This combined approach differed significantly from both AI's performance ($p < 0.001$) and the resident's performance ($p < 0.001$).

In Figure 1 we show the ROC curves for AI and residents of ophthalmology that would arise from considering the maximum CDR by patient as the model for both AI and the residents. The area under the ROC (AUC-ROC) for AI is 0.848, while the AUC-ROC for residents of ophthalmology is 0.801.

We marked specific values for sensitivity and specificity based on AI's results (high specificity point and best F1-score), the residents' diagnoses, and the synergistic point between ophthalmology residents and AI. Additionally, we include markings for AI's performance at a specificity matched to that of the residents' diagnoses, as well as a point corresponding to an 80% sensitivity for AI.

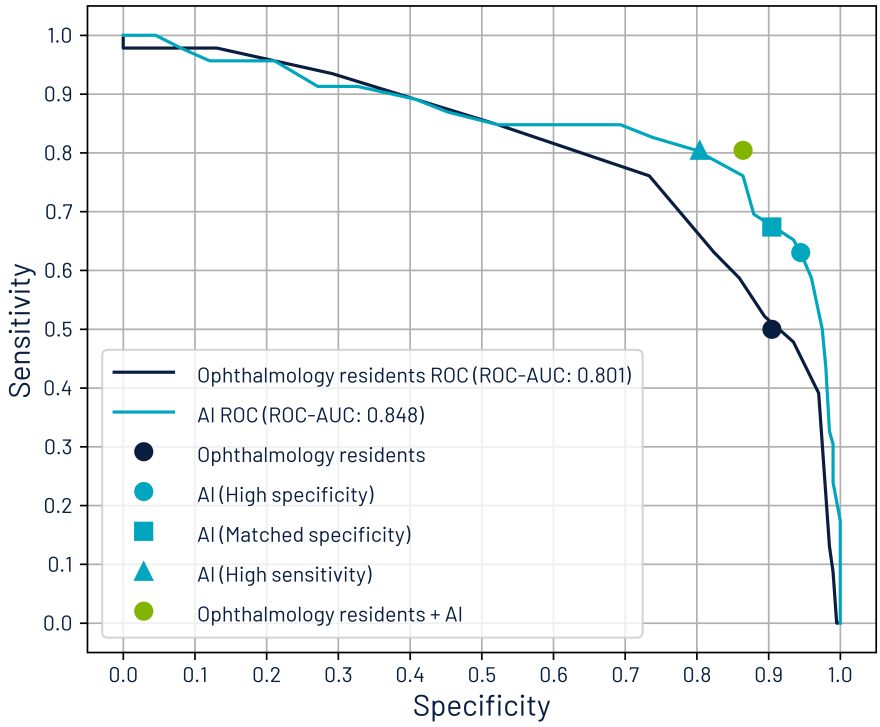

Figure 1: Receiver Operating Characteristic Curves for Glaucoma Detection Based on Maximum CDR by patient.

The matched specificity point considering the data from AI would result in a sensitivity of 67.4% and a specificity of 90.5%, that matches that of the residents. Even at this point, AI's sensitivity is higher than that of the residents of ophthalmology. This point would arise when a CDR of 0.527 or higher is considered as glaucoma suspect.

The high sensitivity point that corresponds to a sensitivity of 80.4% was obtained when considering a CDR of 0.506 or higher as glaucoma suspect. This point, also matches the sensitivity of the synergistic point between AI and ophthalmology residents. However, specificity is lower (80.4%) than that attained at the synergistic point (86.4%).

We further analysed the CDR estimates by eye, by comparing them to the groundtruth CDR. This was evaluated on 362 eyes that were evaluated by all 3 ophthalmologists, by the residents of ophthalmology and by the AI tool. The main results are shown in Table 2.

	Absolute error	Relative error	r	p-value
Resident	0.105 (SD: 0.074)	20.8% (SD: 14.4%)	0.538	< 0.001
AI	**0.056** (SD: 0.042)	**10.9%** (SD: 7.9%)	**0.728**	< 0.001

Table 2: Cup disk ratio evaluation

The mean absolute error for the CDR estimate for AI was 0.056 (SD: 0.042), and the relative error was 10.9% (SD: 7.9%). Meanwhile, the mean absolute error of residents of ophthalmology was 0.105 (SD: 0.074), and the relative error was 20.8% (SD: 14.4%). Moreover, the difference between AI and residents was significative with $p < 0.001$.

In Figure 2, we show the comparison between CDR groundtruth values and CDR estimates for both AI and ophthalmology residents. For both estimates, there is a significant relationship between estimation and groundtruth ($p < 0.001$). However, the Pearson correlation coefficient for AI was higher ($r = 0.728$) than for ophthalmology residents ($r = 0.538$).

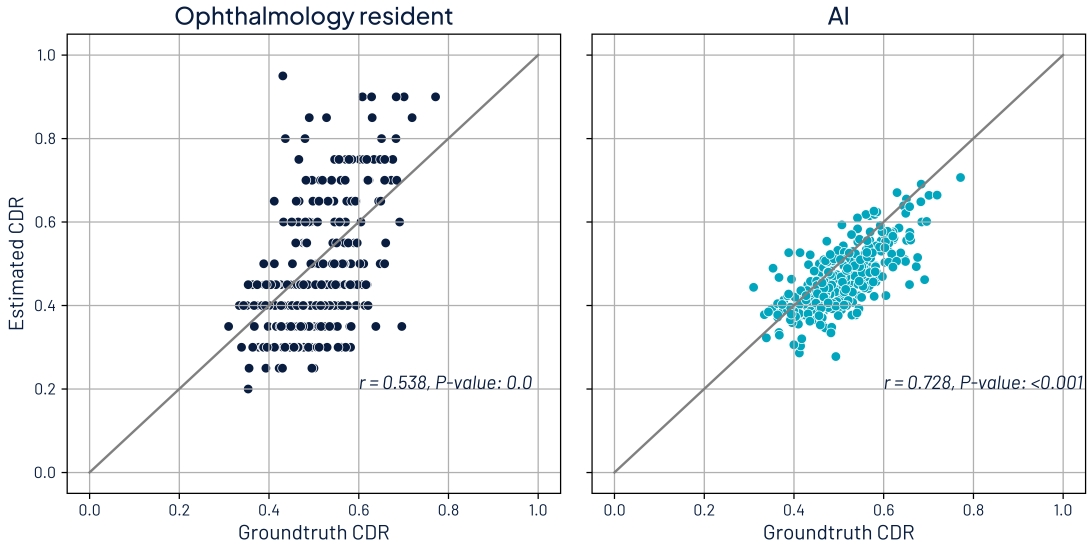

Figure 2: Comparison of CDR estimates with groundtruth CDR. The left image shows the comparison with the residents of ophthalmology and the right image shows the comparison with AI.

From this analysis, we can conclude that the AI tool helps determine CDR values more accurately and may guide residents in estimating CDR and identifying glau-

coma suspects. Moreover, the best sensitivity for glaucoma suspects was obtained when considering both the residents' diagnoses and AI's results.

Nevertheless, residents' diagnoses are valuable on their own, since only 61.6% of the patients had an estimated CDR by the AI tool due to image quality constraints. Meanwhile, residents can evaluate a broader range of patients (95.4%) due to the advantages of performing in-person evaluations.

Additionally, to better understand possible reasons related to insufficient image quality, we compared the prevalence of initial diagnosis (assigned by residents) between the subset with groundtruth CDR (78.6%) and the subset without (21.4%). Initial diagnosis where prevalence on the subset without groundtruth exceeded in at least 25% relative difference to that of the subset with groundtruth were PM (6.7% vs 0.5%), retinal detachment (10.0% vs 1.2%), vitreous hemorrhage (7.5% vs 1.5%), uveitis (0.8% vs 0.2%), non-functional eye or prosthesis (5.8% vs 1.5%), AMD (1.7% vs 0.5%), DR (5.8% vs 4.4%), and cataracts (9.2% vs 7.2%). From the latter, we may notice that opacities arising from either retinal detachment, vitreous hemorrhage, and cataracts may be related to a decreased image quality for OD annotation. Moreover, high myopia (related to PM) may also affect image quality resulting in blurry optic disk images. It's also straight forward that images cannot be acquired on the prosthetic eyes and therefore annotations can't be performed.

3.2. Retinal lesions

The analysis was performed on patients where at least one eye was evaluated by AI to determine presence of retinal lesions. This subset consisted on 395 patients (90.8%).

If the AI tool indicated the presence of possible DR, ME, AMD, PM, or other retinal lesions, it was considered positive for presence of retinal lesions.

Of those patients that where diagnosed by residents of ophthalmology with retinal diseases, only 78.8% were referred to either retina subspecialty or to further tests.

Table 3 presents different performance metrics. We distinguished three categories for sensitivity analysis. The first category includes the presence of all retinal lesions, including those associated with mild stages of retinal disease (excluding tessellated fundus).

The second category includes retinal lesions associated with medium or high risk of vision loss, such as large drusen, intraretinal hemorrhages, as well as more severe findings like vitreous hemorrhage or retinal detachment.

The third category includes lesions associated with a high risk of vision loss and encompasses vitreous hemorrhages, retinal detachment, neovascularization, macular holes, and others.

A complete list of what is included in each category is provided in Appendix D.

	Accuracy	**Sensitivity**	**Sensitivity medium risk and over**	**Sensitivity high risk**	**Specificity**	**PPV**	**F1 score**
Resident diagnosis	75.2% (70.9%, 79.7%)	51.9% (44.0%, 59.9%)	63.0% (52.2%, 73.3%)	80.5% (67.9%, 92.1%)	90.4% (86.4%, 93.9%)	77.9% (70.0%, 85.4%)	62.3% (54.9%, 69.0%)
AI	**88.1%** (84.8%, 91.1%)	76.3% (69.9%, 82.6%)	90.1% (83.5%, 96.1%)	**100.0%** (100.0%, 100.0%)	**95.8%** (93.1%, 98.3%)	**92.2%** (87.3%, 96.5%)	**83.5%** (78.7%, 87.6%)
Resident or AI	86.1% (82.8%, 89.4%)	**84.0%** (78.4%, 89.9%)	**92.6%** (86.5%, 97.7%)	**100.0%** (100.0%, 100.0%)	87.4% (82.9%, 91.2%)	81.4% (75.3%, 86.7%)	82.6% (78.1%, 86.9%)

Table 3: Performance metrics for presence of retinal lesions detection.

Sensitivity increases for both AI and ophthalmology residents as the risk of vision loss increases. Across all categories, AI's sensitivity was better than that of the residents and was statistically significant ($p < 0.001$). Sensitivity for all retinal lesions was 76.3% for AI compared to 51.9% for residents. For retinal lesions associated with a medium or high risk, sensitivity was 90.1% for AI and 63.0% for ophthalmology residents. Moreover, sensitivity for retinal lesions associated with high risk of vision loss was 100% for AI and 80.5% for ophthalmology residents. The latter information is shown on Figure 3.

In addition, significant differences between AI and resident were obtained across all metrics. The differences in accuracy, sensitivity, PPV, and F1-score were all statistically significant with $p < 0.001$. The difference in specificity was also significant ($p = 0.005$).

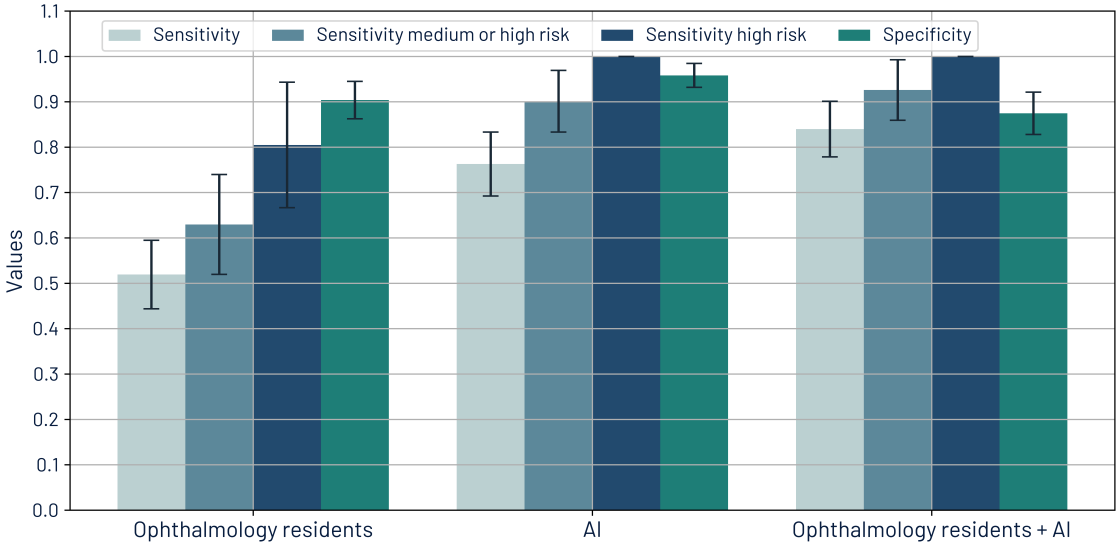

Figure 3: Sensitivities by risk of visual loss and specificity for retinal diseases for ophthalmology residents, AI, and the synergy from both.

Furthermore, AI's PPV was 92.2%, indicating that its use can help increase sensitivity with minimal false positives. Moreover, considering a possible synergy between the AI tool and ophthalmology residents —where presence of retinal lesions is considered positive when either the resident or the AI tool detected them— increases sensitivity for all retinal lesions to 84.0%, for lesions associated with medium or high risk to 92.6%, and attains 100% sensitivity for high-risk lesions. This synergistic approach achieves a PPV of 81.4%, which is even better than the baseline PPV of the residents' diagnoses.

3.3. Cataract

During the data collection process, information on cataract diagnosis and referral to cataract surgery was obtained. Furthermore, the AI tool results include possible media opacity detection. Although there was no ground truth value for cataract, and currently this AI tool does not have the capability to distinguish cataracts from other media opacities, we compared the residents' diagnoses with the results from AI.

Even though the percentage of agreement was high (86.0%), there is little agreement on detection itself, which accounts for minimal agreement in terms of Cohen's Kappa ($\kappa = 0.237$).[45] This low agreement may be due to the presence of opacities or cataract-related symptoms that correspond to different diseases, as well as the

presence of cataracts that still allow for fundus imaging with sufficient quality for analysis of the retina.

4. Discussion

In this study we compared the performance of an AI tool that detects retinal diseases, estimates CDR, and determines possible media opacities against the initial diagnosis of ophthalmology residents.

For retinal disease detection, CDR estimation and glaucoma suspect detection performance was assessed against groundtruth values obtained through fundus image annotation by experts.

The AI tool demonstrated superior performance compared to ophthalmology residents in CDR estimations, obtaining a lower average error (0.056 vs 0.105, $p < 0.001$) and a higher Pearson correlation coefficient (0.728 vs 0.538). This high performance obtained from AI solutions could be expected, since CDR estimation is challenging even for experts.[29,30] Consequently, the implementation of an AI tool has the potential to standardize CDR measurements. Furthermore, we did not assess how visual outcomes from the AI tool could help residents determine CDR measurements. This can be evaluated with further research to assess the benefits of having a visual aid to determine CDR both in clinical practice and in learning processes.

Determining presence of glaucoma is highly related to OD assessments, including large CDR, notable differences in CDR from one eye to the other, evaluating whether the ISNT rule is fulfilled, as well as presence of hemorrhages in the OD.[43] Even though there are other relevant signs for glaucoma, it has been shown that CDR is a significant indicator for presence of glaucoma.[46] Moreover, 89.4% of patients referred to further tests related to glaucoma or to a glaucoma specialist had CDR estimates ≥ 0.6. Therefore, it's straight forward that a better CDR estimation results in better glaucoma suspect detection. In this work we only considered patients as glaucoma suspects, those whose CDR ≥ 0.6 or the difference of CDR between both eyes was larger than 0.2, considering CDR related aspects that determine absence of glaucomatous optic neuropathy.[43] However, under those criteria, the sensitivity was higher for AI (63.0%), that only considers CDR, than that of ophthalmology residents (50.0%), who perform an in-site evaluation and have more elements for diagnosis. While both sensitivities are relatively low when considered independently and not significantly statistically different ($p = 0.116$), when considering simultaneously those patients detected by AI and by residents, sensitivity increased to 80.4% ($p < 0.001$), while maintaining a specificity of 86.4%. Furthermore, specificity may be higher if other elements such as IOP or hemorrhages in OD were considered. This

suggests a potentially powerful synergy between AI and ophthalmology residents. Thus, implementing this assistive tool can have positive effects on patient's first consultation. Moreover, using AI offers a significant advantage in homogenizing this measurement, which is known to have high inter-observer variability.

For retinal disease detection, there is an important difference on sensitivity between AI and residents. Mild retinal lesions such as microaneurysms and small drusen may not require high sensitivity. However, retinal findings with medium or high risk of visual loss, should be acknowledged, such that the necessary considerations for referral are addressed. Sensitivity for medium or high risk findings was 90.1% for AI compared to 63.0% for residents. Even when sensitivity increased for high risk findings, residents only achieve 80.5% compared to a 100% detection by AI. Moreover, specificity is higher for AI than for residents, resulting on an overall better evaluation from AI. Furthermore, differences in accuracy, sensitivity, PPV, and F1-score were all statistically significant with $p < 0.001$. The difference in specificity was also significant ($p = 0.005$), indicating a notable performance difference between AI and first-year residents in detecting retinal lesions. Although not all patients with retinal lesions require referral to further tests or to retina specialists, high sensitivity is important in order to make a better decision in the referral process and give better recommendations to patients in terms of self-care and follow-up ophthalmic consultations.

Implementing the AI tool demonstrates significant potential in enhancing the detection of retinal lesions and guiding more accurate measurements of CDR. However, it is crucial to emphasize that the final diagnosis must always be based on the ophthalmologist's comprehensive evaluation. Ophthalmologists perform a thorough assessment that encompasses examining the peripheral retina (which may not be visible in digital images), evaluating anterior segment structures, and detecting opacities that could compromise retinal image quality. These aspects of clinical examination are critical and cannot be fully replicated by current AI technologies.

It is important to note that not all diagnoses require immediate referral to a subspecialty. Some require additional studies, and some may not be referred at all, depending on the severity and nature of their condition. The decision-making process for referrals remains a critical aspect of the resident's role. Furthermore, the role of residents in retinal disease detection should not be underestimated; notably, 35.9% of patients with images deemed insufficient for AI analysis were diagnosed with retinal disease by residents, highlighting the value of clinical experts.

Moreover, it must be considered that due to image quality, AI only estimates CDR for 61.6% of patients and does an assessment of the retina for 90.8% of patients. Especially for OD, poor image quality is a downside, this could be resolved by relaxing

quality criteria, which is possible, since all three ophthalmologists considered 78.6% of patients had images with sufficient quality for annotation. To further increase evaluation, another option is using fundus cameras with higher resolution, however, this may also result in higher implementation costs. Additionally, images of patients with opacities or PM may not have sufficient quality due to physiological differences. Which could be the underlying cause of large prevalence differences (at least 25% relative difference) between the subsets with and without CDR ground truth.

Nevertheless, AI can serve as a valuable tool to enhance the detection of retinal lesions, homogeneize CDR estimates, and guide the selection of additional examinations, referral to subspecialty and follow-up. Moreover, a synergistic approach between AI and residents showed enhanced sensitivity for both glaucoma suspects and retinal disease detection, with little effect on specificity. Therefore, its use should be considered in future clinical practice.

5. Conclusion

This study demonstrates that an AI-powered ophthalmic screening tool, outperforms first-year ophthalmology residents in several key areas of ophthalmic assessment in a Mexican tertiary care setting. Specifically:

1. CDR Estimation: AI exhibited superior accuracy in CDR estimation compared to residents, with lower mean absolute error (0.056 vs 0.105, $p < 0.001$) and higher correlation with expert measurements ($r = 0.728$ vs $r = 0.538$). This suggests that AI implementation could help standardize this critical measurement.

2. Glaucoma Suspect Detection: AI showed higher accuracy (88.6% vs 82.9%, $p = 0.016$), sensitivity (63.0% vs 50.0%, $p = 0.116$), and specificity (94.5% vs 90.5%, $p = 0.062$) compared to residents. The differences in accuracy, positive predictive value ($p = 0.02$), and F1-score ($p = 0.026$) were statistically significant, indicating a robust advantage for AI in glaucoma detection.

3. Retinal Lesion Detection: AI demonstrated superior sensitivity across all risk categories, particularly for medium and high-risk lesions (90.1% vs 63.0%), while maintaining high specificity (95.8% vs 90.4%). The statistical significance of these differences ($p < 0.001$ for accuracy, sensitivity, PPV, and F1-score; $p = 0.005$ for specificity) strongly supports the superiority of AI in retinal disease detection.

These findings highlight the potential of AI as a valuable assistive tool in ophthalmic practice, particularly for early-career clinicians. The synergistic approach of combining AI and clinical assessment shows promise for optimizing diagnostic accuracy and patient care.

However, it is crucial to emphasize that AI should complement, not replace, comprehensive clinical evaluation by ophthalmologists. Challenges such as image quality limitations and the need for clinical judgment in referral decisions underscore the continued importance of human expertise in ophthalmic care.

Future research should focus on integrating AI tools into clinical workflows, evaluating their impact on patient outcomes, and exploring their potential in ophthalmology resident training programs. As AI continues to evolve, its role in enhancing ophthalmic screening and supporting clinical decision-making is likely to expand, potentially improving access to high-quality eye care in diverse healthcare settings.

6. Acknowledgments

The authors acknowledge Diana González and Nayelli Cruz for their contributions to project definition and coordination of the optometry team. Iris Pantoja, Brenda Camarillo, and Rodolfo Pineda are recognized for their work in conducting the AI screening. The authors also acknowledge Vanessa Tirado for her contribution in annotating optic disk images.

7. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Claude 3.5 Sonnet, a large language model developed by Anthropic, PBC, from July 2nd 2024 to September 17 2024. In order to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

8. Declaration of interests

Two of the authors, Dalia Camacho-García-Formentí and Alejandro Noriega, are employees of Prosperia Salud S.A. de C.V., the enterprise that developed retinIA. Additionally, Damaris Hodelin-Fuentes and Hugo Valdez-Flores have acted as consultants of Prosperia Salud S.A. de C.V. These relationships may be perceived as potential conflicts of interest. However, the study was conducted with the objective of maintaining scientific integrity and neutrality. All other authors declare no conflicts of interest.

References

1. Basto-Abreu A, López-Olmedo N, Rojas-Martínez R, et al. Prevalencia de prediabetes y diabetes en México: Ensanut 2022. *Salud pública de México.* 2023;65:s163–s168.

2. Instituto Nacional de Estadística y Geografía. Población. Instituto Nacional de Estadística y Geografía (INEGI). (accessed June 17 2024). https://www.inegi.org.mx/temas/estructura/.

3. Campos-Nonato I, Hernández-Barrera L, Oviedo-Solís CI, Ramírez-Villalobos D, Hernández B, Barquera S. Epidemiología de la hipertensión arterial en adultos mexicanos: diagnóstico, control y tendencias. Ensanut 2020. *Salud pública de México.* 2021;63(6):692–704.

4. McMonnies CW. Glaucoma history and risk factors. *J Optom.* 2017;10(2):71–78.

5. Yang Z, Tan TE, Shao Y, Wong TY, Li X. Classification of diabetic retinopathy: Past, present and future. *Front Endocrinol (Lausanne).* 2022;13:1079217.

6. Fraser-Bell S, Symes R, Vaze A. Hypertensive eye disease: a review. *Clin Exp Ophthalmol..* 2017;45(1):45–53.

7. Thompson J, Lakhani N. Cataracts. *Prim Care.* 2015;42(3):409–423.

8. Instituto Nacional de Estadística y Geografía. Estadísticas a propósito del día internacional de las personas adultas mayores. *Comunicado de prensa. Instituto Nacional de Estadística y Geografía (INEGI).* 2022;22(568):1–6.

9. Flaxel CJ, Adelman RA, Bailey ST, et al. Age-Related Macular Degeneration Preferred Practice Pattern®. *Ophthalmology.* 2020;127(1):P1-P65.

10. López-Medina A, Arias-Mendoza A, Vega-Morales AQ, Andrade-Estrada I, Oviedo-Díaz F, Hernández EF. Impact of Primary Open-Angle Glaucoma: Early Diagnosis and Timely Treatment. *International Journal of Medical Science and Clinical Research Studies.* 2022;2(12):1414–1420.

11. Instituto Nacional de Estadística y Geografía. Censo de Población y Vivienda 2020. Instituto Nacional de Estadística y Geografía (INEGI). accessed August 2 2024). https://www.inegi.org.mx/programas/ccpv/2020/#datos_abiertos.

12. Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado. Población Mexicana es propensa a PADECER glaucoma: ISSSTE. Gobierno de México. (accessed August 1 2024). https://www.gob.mx/issste/prensa/poblacion-mexicana-es-propensa-a-padecer-glaucoma-issste?idiom=es-MX.

13. Consejo Mexicano de Oftalmología. Listado de Médicos. Consejo Mexicano de Oftalmología. (accessed July 22 2024). https://cmoftalmologia.org/listado-de-medicos/.

14. González-Gonzalo C, Sánchez-Gutiérrez V, Hernández-Martínez P, et al. Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration. *Acta Ophthalmol.* 2020;98(4):368–377.

15. Gonzalez-Briceno G, Sanchez A, Ortega-Cisneros S, Contreras MSG, Diaz GAP, Moya-Sanchez EU. Artificial intelligence-based referral system for patients with diabetic retinopathy. *Computer.* 2020;53(10):77–87.

16. Rogers TW, Gonzalez-Bueno J, Garcia Franco R, et al. Evaluation of an AI system for the detection of diabetic retinopathy from images captured with a handheld portable fundus camera: the MAILOR AI study. *Eye (Lond).* 2021;35(2):632–638.

17. Cen LP, Ji J, Lin JW, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat Commun.* 2021;12(1):4828.

18. Lin D, Xiong J, Liu C, et al. Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. *Lancet Digit Health.* 2021;3(8):e486–e495.

19. Li B, Chen H, Zhang B, et al. Development and evaluation of a deep learning model for the detection of multiple fundus diseases based on colour fundus photography. *Br J Ophthalmol.* 2021;106(8):1079-1086.

20. Arenas-Cavalli JT, Abarca I, Rojas-Contreras M, Bernuy F, Donoso R. Clinical validation of an artificial intelligence-based diabetic retinopathy screening tool for a national health system. *Eye (Lond).* 2022;36(1):78–85.

21. Dong L, He W, Zhang R, et al. Artificial Intelligence for Screening of Multiple Retinal and Optic Nerve Diseases. *JAMA Netw Open.* 2022;5(5):e229960.

22. Pinos-Vélez E, Encalada S, Gamboa E, Robles-Bykbaev V, Ipanque W, Chacón C. in *Development of a support system for the presumptive diagnosis of glaucoma through the processing of biomedical images of the human eye fundus in Ecuador*. Advances in Intelligent Systems and Computing:100–109 2018. Advances in Intelligent Systems and Computing ; Conference date: 01-01-2015.

23. 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.

24. Yu S, Xiao D, Frost S, Kanagasingam Y. Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Comput Med Imaging Graph.* 2019;74:61–71.

25. Latif J, Tu S, Xiao C, Ur Rehman S, Imran A, Latif Y. ODGNet: a deep learning model for automated optic disc localization and glaucoma classification using fundus images. *SN Appl. Sci.* 2022;4:98.

26. Neto A, Camara J, Cunha A. Evaluations of deep learning approaches for glaucoma screening using retinal images from mobile device. *Sensors (Basel).* 2022;22(4):1449.

27. Hemelings R, Elen B, Schuster AK, et al. A generalizable deep learning regression model for automated glaucoma screening from fundus images. *NPJ Digit Med.* 2023;6(1):112.

28. Zhao J, Lu Y, Qian Y, Luo Y, Yang W. Emerging Trends and Research Foci in Artificial Intelligence for Retinal Diseases: Bibliometric and Visualization Study. *J Med Internet Res.* 2022;24(6):e37532.

29. Almazroa A, Alodhayb S, Osman E, et al. Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images. *Int Ophthalmol.* 2017;37:701–717.

30. Yu S, Xiao D, Frost S, Kanagasingam Y. Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Comput Med Imaging Graph.* 2019;74:61–71.

31. Wang J, Bai Y, Xia B. Feasibility of diagnosing both severity and features of diabetic retinopathy in fundus photography. *IEEE access.* 2019;7:102589–102597.

32. Ruamviboonsuk P, Teerasuwanajak K, Tiensuwan M, Yuttitham K, for Diabetic Retinopathy Study Group TS, others. Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. *Ophthalmology.* 2006;113(5):826–832.

33. Noriega A, Meizner D, Camacho D, et al. Screening Diabetic Retinopathy Using an Automated Retinal Image Analysis System in Independent and Assistive Use Cases in Mexico: Randomized Controlled Trial. *JMIR Form Res.* 2021;5(8):e25290.

34. Han R, Yu W, Chen H, Chen Y. Using artificial intelligence reading label system in diabetic retinopathy grading training of junior ophthalmology residents and medical students. *BMC Med Educ.* 2022;22(1):258.

35. Fang Z, Xu Z, He X, Han W. Artificial intelligence-based pathologic myopia identification system in the ophthalmology residency training program. *Front Cell Dev Biol.* 2022;10:1053079.

36. PROSPERiA. retinIA. (accessed: February 12 2024 to March 14 2024) . [Online Software]. https://retinia.prosperia.health/login.

37. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. :2818–2826 2016.

38. Graham B. Kaggle diabetic retinopathy detection competition report. *University of Warwick.* 2015;22(9).

39. Ronneberger O, Fischer P, Brox T. in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. :234–241Springer 2015.

40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. in *2017 IEEE International Conference on Computer Vision (ICCV)*. :618-626 2017.

41. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825.* 2017.

42. LinkedAI. LinkedAI Platform. (accessed: March 25, 2024 to June 15, 2024) . [Online Software]. https://platform.linkedai.co.

43. Harizman N, Oliveira C, Chiang A, et al. The ISNT rule and differentiation of normal from glaucomatous eyes. *Arch Ophthalmol.* 2006;124(11):1579–1583.

44. PROSPERiA. Televal. (accessed: February 12 2024 to June 18 2024) . [Online Software]. https://televal.prosperia.health/prosperia/login.

45. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).* 2012;22(3):276–282.

46. Armaly M. Cup-disc ratio in early open-angle glaucoma. *Doc Ophthalmol.* 1969;26:526–533.