• Intelligent Ophthalmology •

Artificial intelligence-guided and human grading in community eye screening: a six-year real-world experience in Jakarta, Indonesia

Valenchia[1], *Damara Andalia*[1], *Nyssa Alexandra Tedjonegoro*[1], *Viona*[1,2]

[1]JEC Eye Hospitals and Clinics, Jakarta 11520, Indonesia
[2]Department of Ophthalmology, Hasanuddin University, Makassar 90245, Indonesia
Correspondence to: Valenchia. JEC Eye Hospitals and Clinics, Jl. Terusan Arjuna Utara No. 1, Kedoya, Jakarta Barat 11520, Indonesia. valenchia@jec.co.id; dr.valenchia@gmail.com
Received: 2025-07-11 Accepted: 2026-03-31

Abstract

● **AIM:** To evaluate the diagnostic performance of human and artificial intelligence (AI)-guided graders in a community eye screening program and assess follow-up adherence over six years in Jakarta, Indonesia.

● **METHODS:** This retrospective study analyzed patients screened through the EyeCheck™ program (2018–2024). Human graders (Eye Scan™, 2018–2019) evaluated multiple anterior and posterior segment images; AI systems (DR. NOON™ and Vuno™, 2020–2024) assessed single posterior segment images only—constituting a non-concurrent comparison. Diagnostic accuracy was calculated for patients (6.2% of those with abnormal findings) who attended follow-up, using ophthalmologists' diagnoses as the reference standard.

● **RESULTS:** Of 18 628 patients screened, 9262 (49.7%) had abnormal findings, but only 574 (6.2%) sought follow-up care. Among the 574 patients who completed follow-up examination, 286 (49.8%) were male and 288 (50.2%) were female. The median age was 25y (range: 7–81y), with the majority (72.3%) aged 17–50y. Among these, human graders demonstrated higher sensitivity (93.8% *vs* 72.4%) but lower specificity (17.2% *vs* 70.1%) and accuracy (56.3% *vs* 70.8%) compared to AI graders. AI systems provided real-time results, whereas human grading required 3–7 business days. The most frequent diagnoses were refractive error (51.4%), cataract (13.1%), and dry eye syndrome (9.1%).

● **CONCLUSION:** In this non-concurrent comparison, AI-guided graders show higher specificity and accuracy while human graders achieved higher sensitivity. These findings should be interpreted cautiously given selection bias from low follow-up rates, temporal confounders, and differences in imaging protocols. The critically low follow-up rate underscores that effective screening requires robust linkage-to-care systems. Prospective studies with concurrent comparison are needed.

● **KEYWORDS:** teleophthalmology; artificial intelligence; community screening; diagnostic accuracy; eye care; human grader

DOI:10.18240/ijo.2026.08.01

Citation: Valenchia, Andalia D, Tedjonegoro NA, Viona. Artificial intelligence-guided and human grading in community eye screening: a six-year real-world experience in Jakarta, Indonesia. *Int J Ophthalmol* 2026;19(8):1431-1439

INTRODUCTION

Community-based eye screening aims to detect vision-threatening conditions in populations with limited access to eye care, often due to socioeconomic or geographic barriers[1]. Such programs have demonstrated value in identifying prevalent causes of visual impairment, including glaucoma, diabetic retinopathy, cataract, and age-related macular degeneration[2].

A persistent challenge in scaling these programs is the availability of trained personnel to interpret screening images. Human graders, while capable of high diagnostic sensitivity, require significant time to review images and may create bottlenecks when screening large populations[3-4]. This has prompted interest in artificial intelligence (AI)-guided systems, which can analyze fundus photographs rapidly and potentially reduce dependence on scarce human resources[5].

Several studies have evaluated AI systems for specific conditions, particularly diabetic retinopathy, demonstrating sensitivity and specificity comparable to or exceeding human graders in controlled settings[6-8]. However, most validation studies have used curated datasets with standardized imaging protocols and defined disease categories. Less is known about how AI-guided grading performs in real-world community screening programs, where image quality varies, disease

spectrum is broad, and the comparison may involve different imaging inputs and operational contexts.

In 2018, JEC Eye Hospitals and Clinics launched the EyeCheck™ program, a community-based screening initiative in Jakarta, Indonesia, utilizing non-mydriatic fundus photography and visual acuity assessment. From 2018 to 2019, images were interpreted by accredited human graders (Eye Scan™). Beginning in 2020, the program transitioned to AI-guided systems (DR. NOON™, followed by Vuno™), which offered real-time interpretation using fewer images. This sequential implementation provides an opportunity to examine grading performance across two eras, though the non-concurrent design and differences in imaging protocols preclude direct head-to-head comparison.

The objectives of this study were: 1) to describe the diagnostic performance of human and AI-guided graders among patients who completed follow-up examination, using ophthalmologist diagnosis as the reference standard; 2) to compare operational characteristics including turnaround time and imaging requirements; 3) to evaluate follow-up adherence among patients with abnormal screening findings over six years of program implementation.

PARTICIPANTS AND METHODS

Ethical Approval This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada–Dr. Sardjito General Hospital (reference number KE/FK/1499/EC/2024). The requirement for individual informed consent was waived due to the retrospective nature of the study. All data were anonymized prior to analysis.

Study Design and Setting This retrospective observational study analyzed data from the EyeCheck™ community-based screening program conducted by JEC Eye Hospitals and Clinics in Jakarta, Indonesia, from January 2018 to June 2024. The program was suspended in 2021 due to the COVID-19 pandemic. This study is reported in accordance with the Standards for Reporting Diagnostic Accuracy Studies (STARD) guidelines for diagnostic accuracy studies.

Screening Protocol The EyeCheck™ program targeted community members with limited access to eye care services. Screening consisted of visual acuity measurement, autorefractor examination, and non-mydriatic fundus photography using a digital fundus camera (3nethra classic™). Tonometry was performed when clinically indicated. No anterior segment examination or functional testing was performed at screening sites. All examinations were conducted by designated operators who received standardized training on device operation and image acquisition prior to program deployment. Basic image quality assessment was performed at the point of grading; images deemed ungradable due to poor focus, inadequate centration, or significant media opacity were excluded from interpretation.

Grading Systems Two grading approaches were implemented sequentially (non-concurrent design): Human graders (2018–2019): Accredited graders from Eye Scan™ evaluated multiple images per eye, including anterior segment, posterior segment, and supplementary black-and-white photographs. Grading required 3–7 business days depending on reviewer availability. Results were classified dichotomously as abnormal (any detected ocular pathology) or normal.

AI-guided graders (2020–2024): Beginning in 2020, AI systems replaced human graders. DR. NOON™ was used in 2020, and Vuno™ was adopted from 2022 to 2024. Both systems analyzed single posterior segment images per eye and provided real-time results. The AI systems were designed to detect posterior segment abnormalities including diabetic retinopathy, age-related macular degeneration, and glaucoma-related findings. Results were classified dichotomously as abnormal or normal.

Importantly, the grading systems differed in both the images evaluated (anterior and posterior segments for human graders; posterior segment only for AI) and the conditions potentially detectable. This limits direct comparison between grading eras.

Study Population All patients screened through the EyeCheck™ program during the study period were included in the descriptive analysis of screening outcomes. For the diagnostic accuracy analysis, inclusion criteria were: 1) abnormal EyeCheck™ screening result, 2) follow-up ophthalmologic examination at JEC Eye Hospitals and Clinics within six months of screening, 3) complete medical records. There was no age restriction; the screened population included individuals of all ages participating in community screening events.

Reference Standard Ophthalmologists' clinical diagnoses served as the reference standard. Ophthalmologists were masked to EyeCheck™ screening results at the time of examination. Diagnoses were established through comprehensive examination including slit-lamp biomicroscopy, dilated fundus examination, and ancillary testing as clinically indicated (*e.g.*, optical coherence tomography, visual field testing, tonometry). Diagnoses were extracted from electronic medical records and were based on clinical judgment rather than standardized diagnostic criteria applied specifically for this study.

For the purposes of this analysis, an eye was classified as having disease if the ophthalmologist diagnosed any ocular pathology. An eye was classified as normal if no pathology was identified. We acknowledge that this broad definition encompasses conditions not detectable by fundus photography

alone (*e.g.*, refractive error, dry eye syndrome, asthenopia), which may affect measured specificity and the interpretation of false-positive results.

Data Collection Data were extracted from EyeCheck™ interpretation records and electronic medical records at JEC Eye Hospitals and Clinics. Variables collected included: patient demographics (age, sex), screening date, grader type, EyeCheck™ result (abnormal/normal), date of follow-up visit, and ophthalmologist diagnosis. Diagnostic accuracy was assessed per eye. For patients with multiple diagnoses affecting both eyes, each eye was analyzed separately.

Statistical Analysis Descriptive statistics were used to summarize screening program outcomes and patient characteristics. Continuous variables were reported as median and range; categorical variables were reported as frequencies and percentages. The Shapiro-Wilk test was used to assess normality of continuous data.

Diagnostic accuracy metrics were calculated at the per-eye level using 2×2 contingency tables: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, positive likelihood ratio (LR+), and negative likelihood ratio (LR−). Ninety-five percent confidence intervals were calculated using the Clopper-Pearson exact method. Results are presented separately for human graders and AI-guided graders.

Given the non-concurrent study design and differences in grading protocols, formal statistical comparison between grading systems was not performed. Instead, findings are presented descriptively to characterize performance within each grading era.

The cost comparison is presented as a descriptive summary of pricing structures and operational requirements rather than a formal health economic analysis. All analyses were performed using SPSS version 27.0 (IBM Corp., Armonk, New York, United States).

RESULTS

Screening Program Outcomes From January 2018 to June 2024, a total of 18 628 patients were screened through the EyeCheck™ program (Table 1). The program was suspended in 2021 due to the COVID-19 pandemic. Annual screening volume ranged from 939 patients (2020) to 6064 patients (2024). Of all patients screened, 9262 (49.7%) were identified as having abnormal findings based on fundus photography and visual acuity assessment.

Among the 9262 patients with abnormal screening findings, only 574 (6.2%) attended follow-up examination at JEC Eye Hospitals and Clinics within 6mo. Annual follow-up rates ranged from 1.6% (2018) to 20.9% (2023). This low follow-up rate represents a significant limitation for the diagnostic accuracy analysis, as only this self-selected subgroup could be evaluated against the reference standard.

Characteristics of Patients Completing Follow-up Among the 574 patients who completed follow-up examination, 286 (49.8%) were male and 288 (50.2%) were female (Table 2). The median age was 25y (range: 7–81y), with the majority (72.3%) aged 17–50y. The young median age reflects the community screening setting, which included workplace and school-based events targeting working-age populations; this differs from hospital-based populations where older patients with age-related diseases predominate.

Human graders (Eye Scan™) evaluated 95 patients (190 eyes) during 2018–2019. AI-guided graders assessed 479 patients (958 eyes) during 2020–2024, comprising 59 patients evaluated by DR. NOON™ (2020) and 420 patients evaluated by Vuno™ (2022–2024).

Diagnoses at Follow-up Examination Among patients completing follow-up, ophthalmologists identified a total of 658 diagnoses in 574 patients; some patients had multiple diagnoses (Table 3). Each eye was evaluated independently. The most frequent diagnoses were refractive error (338 patients, 51.4%), cataract (86 patients, 13.1%), dry eye syndrome (60 patients, 9.1%), and glaucoma (42 patients, 6.4%). Seventy-eight patients (11.9%) were found to have normal eyes on comprehensive examination.

Notably, refractive error—the most prevalent diagnosis—can reduce visual acuity without causing fundus abnormality. Patients flagged as abnormal based on visual acuity but with normal fundus findings may represent appropriate referrals for refraction rather than grading errors. This overlap complicates the interpretation of specificity for both grading systems.

Diagnostic Performance of Human Graders Human graders evaluated 190 eyes from 95 patients during 2018–2019, with each eye evaluated independently (Tables 4 and 5). Sensitivity was 93.8% [95% confidence interval (CI): 87.0%–97.7%], indicating that human graders correctly identified most eyes with pathology. However, specificity was low at 17.2% (95%CI: 10.2%–26.4%), reflecting a high false-positive rate. Of 190 eyes evaluated, human graders classified 168 (88.4%) as abnormal, suggesting a low threshold for flagging abnormalities. Overall accuracy was 56.3% (95%CI: 48.9%–63.5%).

Diagnostic Performance of AI-Guided Graders AI-guided graders evaluated 958 eyes from 479 patients during 2020–2024, with each eye evaluated independently (Tables 6 and 7). Sensitivity was 72.4% (95%CI: 66.8%–77.5%), lower than human graders. Specificity was 70.1% (95%CI: 66.5%–73.5%), substantially higher than human graders. Of 958 eyes evaluated, AI graders classified 408 (42.6%) as abnormal, indicating a more balanced classification threshold. Overall accuracy was 70.8% (95%CI: 67.8%–73.6%).

Table 1 Annual screening volume and follow-up attendance in the EyeCheck™ program, 2018–2024

Year	Patients screened	Abnormal findings, *n* (%)	Attended follow-up, *n* (percentage of abnormal)
2018	3274	2645 (80.8)	42 (1.6)
2019	2650	1749 (66.0)	53 (3.0)
2020	939	818 (87.1)	59 (7.2)
2021	—	—	—
2022	2641	1055 (40.0)	68 (6.4)
2023	3060	1187 (38.8)	248 (20.9)
2024[a]	6064	1808 (29.8)	104 (5.8)
Total	18628	9262 (49.7)	574 (6.2)

[a]Through June 2024. Program suspended in 2021 due to COVID-19 pandemic.

Table 2 Demographic characteristics of patients completing follow-up examination (*n*=574)

Characteristic	*n* (%)
Sex	
Male	286 (49.8)
Female	288 (50.2)
Age, y	
Median (range)	25 (7–81)
<17	15 (2.6)
17–50	415 (72.3)
>50	144 (25.1)
Grading system	
Human graders (Eye Scan™), 2018–2019	95 (16.6)
AI graders (DR. NOON™), 2020	59 (10.3)
AI graders (Vuno™), 2022–2024	420 (73.2)

AI: Artificial intelligence.

Table 3 Diagnosis among patients completing follow-up examination (*n*=574 patients)

Diagnosis	*n* (%)
Refractive error	338 (51.4)
Cataract	86 (13.1)
Dry eye syndrome	60 (9.1)
Glaucoma	42 (6.4)
Normal	78 (11.9)
Age-related macular degeneration	9 (1.4)
Vitreous degeneration	6 (0.9)
Conjunctivitis	6 (0.9)
Diabetic retinopathy	4 (0.6)
Other diagnoses[b]	29 (4.4)

[b]Other diagnoses (each <0.5%): hordeolum, posterior capsule opacification, asthenopia, hypertensive retinopathy, retinal detachment, ocular hypertension, anisometropia, arcus senilis, bullous keratopathy, branch retinal vein occlusion, chorioretinal scars, congenital optic disc malformation, corneal scar, macular hole, myasthenia gravis, pinguecula, peripapillary atrophy, pterygium, visual field defect, vitreous detachment. Multiple diagnoses per patient were possible; total diagnoses exceed patient count.

Comparison of Grading Performance Table 8 summarizes the diagnostic performance of human and AI-guided graders.

Table 4 Diagnostic performance of human graders (Eye Scan™), 2018–2019 (*n*=190 eyes)

Items	Ophthalmologist: disease	Ophthalmologist: no disease	Total
EyeCheck: abnormal	91	77	168
EyeCheck: normal	6	16	22
Total	97	93	190

Table 5 Statistical performance of human graders (Eye Scan™), 2018–2019

Metric	Value	95%CI
Sensitivity	93.8%	87.0%–97.7%
Specificity	17.2%	10.2%–26.4%
PPV	54.2%	46.3%–61.9%
NPV	72.7%	49.8%–89.3%
Accuracy	56.3%	48.9%–63.5%
LR+	1.13	—
LR−	0.36	—

PPV: Positive predictive value; NPV: Negative predictive value; LR+: Positive likelihood ratio; LR−: Negative likelihood ratio; CI: Confidence interval.

Table 6 Diagnostic performance of AI-guided graders (DR. NOON™ and Vuno™), 2020–2024 (*n*=958 eyes)

Items	Ophthalmologist: disease	Ophthalmologist: no disease	Total
EyeCheck: abnormal	207	201	408
EyeCheck: normal	79	471	550
Total	286	672	958

AI: Artificial intelligence.

Table 7 Statistical performance of AI-guided graders (DR. NOON™ and Vuno™), 2020–2024

Metric	Value	95%CI
Sensitivity	72.4%	66.8%–77.5%
Specificity	70.1%	66.5%–73.5%
PPV	50.7%	45.8%–55.7%
NPV	85.6%	82.4%–88.5%
Accuracy	70.8%	67.8%–73.6%
LR+	2.42	—
LR−	0.39	—

PPV: Positive predictive value; NPV: Negative predictive value; LR+: Positive likelihood ratio; LR−: Negative likelihood ratio; CI: Confidence interval; AI: Artificial intelligence.

Table 8 Summary comparison of diagnostic performance between human and AI-guided graders

Metric	Human graders (*n*=190 eyes)	AI-guided graders (*n*=958 eyes)
Sensitivity	93.8% (87.0%–97.7%)	72.4% (66.8%–77.5%)
Specificity	17.2% (10.2%–26.4%)	70.1% (66.5%–73.5%)
PPV	54.2% (46.3%–61.9%)	50.7% (45.8%–55.7%)
NPV	72.7% (49.8%–89.3%)	85.6% (82.4%–88.5%)
Accuracy	56.3% (48.9%–63.5%)	70.8% (67.8%–73.6%)
LR+	1.13	2.42
LR−	0.36	0.39

Values are presented as percentage (95%CI). PPV: Positive predictive value; NPV: Negative predictive value; LR+: Positive likelihood ratio; LR−: Negative likelihood ratio; CI: Confidence interval; AI: Artificial intelligence.

Table 9 Operational characteristics of grading systems

Characteristic	Human graders (Eye Scan™)	AI-guided graders (DR. NOON™)	AI-guided graders (Vuno™)
Turnaround time	3–7 business days	Real-time	Real-time
Images required	Multiple per eye (anterior+posterior)	1 per eye (posterior)	1 per eye (posterior)
Segments evaluated	Anterior and posterior	Posterior only	Posterior only
Cost	USD 1/patient	IDR 10 million/6mo (subscription)	USD 2/patient

AI: Artificial intelligence; USD: United States Dollar; IDR: Indonesian Rupiah.

The 95%CIs for sensitivity do not overlap (human: 87.0%–97.7%; AI-guided: 66.8%–77.5%), suggesting a meaningful difference favoring human graders for disease detection. Similarly, CIs for specificity do not overlap (human: 10.2%–26.4%; AI-guided: 66.5%–73.5%), indicating AI-guided graders had meaningfully higher specificity. However, these comparisons should be interpreted with caution given the non-concurrent design, different imaging protocols, and potential temporal confounders.

Operational Characteristics Table 9 summarizes the operational differences between grading systems. Human graders required 3–7 business days for interpretation and evaluated multiple images per eye including anterior and posterior segment photographs. AI-guided systems provided real-time results using a single posterior segment image per eye. Human grading (Eye Scan™) cost approximately United States Dollar (USD) 1 per patient, while Vuno™ cost USD 2 per patient. DR. NOON™ used a subscription model [Indonesian Rupiah (IDR) 10 million per 6-month package] (Figure 1).

DISCUSSION

This retrospective study examined the diagnostic performance and operational characteristics of human and AI-guided graders implemented sequentially in a community-based eye screening program in Jakarta over 6y. Among patients who completed follow-up examination, human graders demonstrated higher sensitivity (93.8% *vs* 72.4%) while AI-guided graders showed higher specificity (70.1% *vs* 17.2%) and overall accuracy (70.8% *vs* 56.3%). AI systems offered substantial operational advantages including real-time results and reduced imaging requirements. However, only 6.2% of patients with abnormal screening findings attended follow-up care, highlighting a critical gap between screening and treatment.

Interpretation of Diagnostic Performance The marked difference in specificity between grading systems warrants careful interpretation. Human graders classified 88.4% of eyes as abnormal compared to 42.6% for AI graders, suggesting fundamentally different classification thresholds. Human graders appeared to adopt a highly sensitive approach, flagging nearly all eyes for further evaluation. This strategy minimizes missed pathology but generates substantial false-positive burden, potentially overwhelming follow-up capacity and causing unnecessary patient anxiety. AI graders applied a more balanced threshold, reducing false positives at the cost of missing some pathology.

The difference in imaging inputs may partially explain these patterns. Human graders evaluated both anterior and posterior segment photographs, potentially identifying a broader range of abnormalities. AI systems analyzed posterior segment images only, limiting detection to retinal and optic nerve pathology. Presence of abnormal screening results, such as reduced visual acuity or poor image quality even in the absence of clear posterior pathology, often led human graders to refer these conditions. This also accounts for the high prevalence of refractive error and cataract among confirmed diagnoses at follow-up: reduced visual acuity on automated refraction or Snellen testing triggered an abnormal screening result regardless of fundus findings, such that patients with uncorrected refractive error or visually significant lens opacity entered the referral pathway even when posterior segment photographs were normal or inconclusive. The EyeCheck™ program thus functioned as a broad vision health assessment

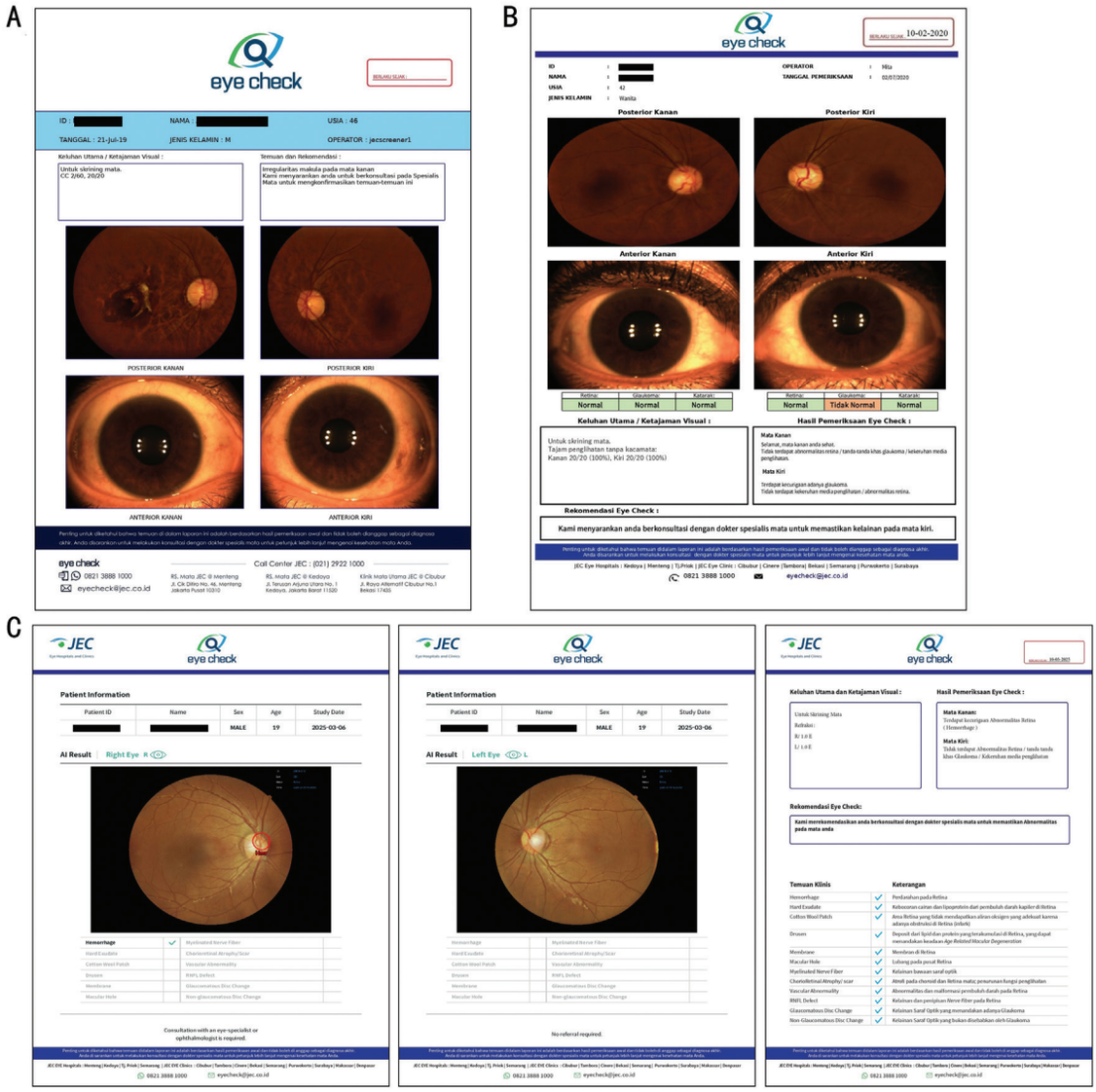

Figure 1 Comparison of reports for EyeCheck™ human grader and AI-guided grader A: Eye Scan™ report; B: DR. NOON™ report; C: Vuno™ report. AI: Artificial intelligence.

rather than a purely posterior segment-focused tool, which is important context when interpreting the diagnostic accuracy metrics reported here. Conditions affecting the anterior segment would be missed by AI but potentially flagged by human graders. This fundamental difference in scope complicates direct comparison of diagnostic performance.

Sensitivity-Specificity Trade-off in Screening In population screening, sensitivity is generally prioritized to minimize missed pathology[7-10]. The lower sensitivity of AI-guided graders (72.4%) compared to human graders (93.8%) is therefore clinically significant. AI graders missed 79 of 286 eyes (27.6%) with ophthalmologist-confirmed disease, compared to 6 of 97 eyes (6.2%) missed by human graders. In a community screening context where the goal is early detection, this difference raises concern.

However, the clinical significance depends on what pathology was missed. If AI primarily missed mild or slowly progressive conditions while detecting sight-threatening disease, the lower overall sensitivity might be acceptable. Conversely, if vision-threatening conditions were missed, the implications are more serious. Our study did not analyze sensitivity by disease severity or category, which limits interpretation. Future studies should examine whether AI systems maintain adequate sensitivity for high-priority conditions such as diabetic retinopathy, glaucoma, and macular degeneration specifically.

It should also be noted that the 93.8% sensitivity of human graders came at substantial cost to specificity (17.2%), meaning the vast majority of patients flagged as abnormal did not have confirmed disease. A screening strategy that refers almost everyone is sensitive but not useful. The optimal balance between sensitivity and specificity depends on factors including disease prevalence, consequences of missed diagnosis, follow-up capacity, and cost, considerations that extend beyond diagnostic accuracy metrics alone.

Operational Considerations Beyond diagnostic performance, AI-guided systems offered meaningful operational advantages. Real-time interpretation eliminates the 3–7d delay required for human grading, enabling immediate patient feedback and potentially improving engagement with follow-up recommendations. Single-image requirements simplify the screening workflow and reduce technical demands at screening sites. These efficiencies may be particularly valuable in resource-limited settings where trained graders are scarce.

The cost comparison favors human graders on a per-patient basis (USD 1 *vs* USD 2 for Vuno™)[11-12]. However, this comparison does not capture the full economic picture. Human grading costs do not account for the infrastructure required to manage delayed results, track patients awaiting reports, and handle the administrative burden of a multi-day workflow. A comprehensive health economic analysis incorporating indirect costs, scalability, and throughput capacity would be needed to determine true cost-effectiveness[13-14].

The Follow-up Gap Perhaps the most striking finding is that only 6.2% of patients with abnormal screening results attended follow-up examination. This attrition represents a fundamental challenge to the effectiveness of community screening programs. A screening test, regardless of its diagnostic accuracy, provides no benefit if patients do not receive subsequent care.

This low follow-up rate may reflect limited awareness of eye health and the importance of early intervention among the screened population. Many individuals may not perceive eye conditions as urgent, particularly if symptoms are absent or mild. In a community screening context—where participants are identified opportunistically rather than presenting with complaints—motivation to pursue further evaluation may be inherently low. This finding suggests that community screening programs must address not only disease detection but also health literacy and awareness if they are to achieve meaningful impact.

Additional factors likely contributing to low follow-up include: financial barriers, geographic distance to eye care facilities, competing time demands, and inadequate linkage-to-care systems within the screening program. The variation in follow-up rates across years (ranging from 1.6% to 20.9%) suggests that program-level factors may also play a role, though we could not identify specific determinants from available data.

Interestingly, the years with AI-guided grading (2020–2024) generally showed higher follow-up rates than the human-graded years (2018–2019), though this observation is confounded by temporal trends, the COVID-19 pandemic, and potential changes in program implementation. The immediate availability of AI results may have facilitated same-day counselling and referral, though this hypothesis requires prospective evaluation.

Improving follow-up adherence should be a priority for community screening programs. Strategies may include: point-of-care health education emphasizing the importance of early treatment, assisted referral systems, reminder mechanisms, reduction of financial barriers, and integration with primary care networks. Point-of-care appointment scheduling—where patients are booked for follow-up before leaving the screening site—has been shown to substantially improve referral adherence compared to passive referral models. Real-time AI result delivery may further support this by enabling immediate counselling; evidence from analogous programs suggests that referral adherence can be nearly three-fold higher when patients receive same-day results compared to delayed teleophthalmology-based grading. Digital reminder systems *via* short message service (SMS) or mobile health applications, community health worker–assisted navigation, and subsidized or co-located ophthalmology services represent additional evidence-informed approaches in low- and middle-income country settings. Without effective linkage to care and improved community awareness of eye health, investments in screening technology—whether human or AI-guided—will yield limited population health benefit.

Comparison with Prior Literature Our findings align with prior studies demonstrating trade-offs between AI and human grading performance. Ting *et al*[15] reported that a deep learning system achieved sensitivity of 90.5% and specificity of 91.6% for detecting referable diabetic retinopathy, comparable to trained graders. Abràmoff *et al*[7] demonstrated that an autonomous AI system achieved 87.2% sensitivity and 90.7% specificity for diabetic retinopathy detection in primary care settings. Ruan *et al*[16] observed a high sensitivity (88.2%) and low specificity (40.7%) of the AI system compared to the conventional method in referable diabetic retinopathy screening. Therefore, individuals without or with mild diabetic retinopathy may receive a referral despite not requiring one. The sensitivity–specificity trade-off is not unique to our findings and has been consistently demonstrated across platforms and real-world settings. A 2024 direct comparison of two commercial AI systems for referable diabetic retinopathy found that IDx-DR achieved sensitivity of 99.3% but specificity of only 68.9%, while RetCAD offered a more balanced profile of 89.4% sensitivity and 94.8% specificity, with both maintaining sensitivity above 95% for sight-threatening disease[17-18]. Real-world implementations illustrate similar patterns: the DAIRET system achieved 100% sensitivity for moderate-or-worse diabetic retinopathy but specificity of only 0.59 owing to a high false-positive burden, while the CARA system reported 87.5% sensitivity and 66.2% specificity in a tertiary care setting[19-20]. A 2025 Meta-analysis of 82 studies encompassing 887 244 examinations

across 25 regulator-approved deep learning devices reported pooled sensitivity of 0.93 and specificity of 0.90 per patient; meta-regression identified ungradable images, lower-income settings, and lower disease severity thresholds as key drivers of higher false-positive rates—all factors directly relevant to our community-based program[21]. Camera hardware further modulates performance: a 2024 study comparing three non-mydriatic devices found that sensitivity ranged from 90.5% to 95.7% and specificity from 95.9% to 97.2% for the same algorithm across different cameras, reinforcing that variable image acquisition technology is a key source of real-world performance heterogeneity[22]. Collectively, these studies confirm that the performance gap between our AI graders and benchmark figures in the literature most likely reflects methodological, technological, and contextual heterogeneity rather than a fundamental limitation of AI-guided grading in community settings.

The same heterogeneity applies when interpreting the low specificity of human graders in our study (17.2%), which is notably lower than typically reported in the literature. This likely reflects the specific training or operational instructions given to Eye Scan™ graders, who appear to have maintained a very low threshold for flagging abnormalities—a strategy that maximizes sensitivity at the cost of high false-positive burden. A 2024 study further demonstrated that label errors in human reference grading can materially affect measured AI performance, with corrected labels improving estimated deep learning sensitivity by 12.5%; this serves as an important reminder that some apparent performance discrepancies between AI systems and reference standards may reflect inter-grader variability rather than true model error[23]. Differences in disease prevalence, imaging quality, and reference standard definitions also contribute, and direct comparison across studies remains limited by methodological heterogeneity.

This study has several important limitations that affect interpretation. First, the non-concurrent design prevents direct comparison between grading systems. Human and AI graders operated in different time periods with potentially different patient populations, screening contexts, and disease prevalence. The COVID-19 pandemic, which occurred during the AI-grading era, may have altered both the screened population and care-seeking behavior. Any observed differences between grading systems may reflect temporal confounders rather than true performance differences.

Second, verification bias substantially limits the generalizability of diagnostic accuracy estimates. Only 6.2% of patients with abnormal screening findings completed follow-up and contributed to the accuracy analysis. Patients who attended follow-up likely differ systematically from those who did not—they may have had more severe symptoms, greater health motivation, or better access to care. The true sensitivity and specificity in the overall screened population cannot be determined from these data.

Third, the grading systems differed in imaging inputs. Human graders evaluated anterior and posterior segment photographs while AI systems assessed posterior segment images only. This difference in scope confounds comparison of diagnostic performance. AI systems cannot be expected to detect anterior segment pathology they were not designed or provided images to evaluate.

Fourth, the reference standard, while masked, was based on clinical judgment rather than standardized diagnostic criteria. Inter-ophthalmologist variability in diagnosis may introduce measurement error. Additionally, the broad definition of "disease" (any ocular pathology) encompasses a heterogeneous spectrum from mild refractive error to sight-threatening conditions, limiting the clinical interpretability of aggregate sensitivity and specificity estimates.

Fifth, we pooled results from two different AI systems (DR. NOON™ and Vuno™) due to limited sample size for separate analysis. These systems may have different performance characteristics that are obscured by pooling.

Finally, this study was conducted at a single institution in Jakarta, and findings may not generalize to other populations, screening contexts, or AI platforms.

Future Directions Prospective studies with concurrent grading, where both human and AI systems evaluate the same images simultaneously, are needed to establish comparative diagnostic performance while controlling for temporal and population confounders. Such studies should examine performance stratified by disease category and severity to determine whether AI systems maintain adequate sensitivity for sight-threatening conditions specifically.

Research on the follow-up gap is equally important. Studies examining barriers to follow-up care and interventions to improve linkage, such as patient navigation, telemedicine consultation, or integrated referral systems, may have greater impact on population health outcomes than incremental improvements in screening accuracy.

Health economic analyses incorporating throughput, scalability, indirect costs, and downstream outcomes would help inform implementation decisions. The optimal role of AI in screening workflows, whether as autonomous graders, triage tools, or adjuncts to human review, remains to be determined.

ACKNOWLEDGEMENTS

Authors' Contributions: Conceptualization and study design: Valenchia and Andalia D; Data collection: Valenchia, Andalia D, and Tedjonegoro NA; Data analysis and interpretation: all authors; Manuscript writing: all authors; Critical revision of

the manuscript: all authors; Final approval of manuscript: all authors.

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

AI-Generated Content Disclosure: No generative AI tools were used in the creation of content for this manuscript.

Conflicts of Interest: Valenchia, None; **Andalia D,** None; **Tedjonegoro NA,** None, **Viona,** None.

REFERENCES

1 Shahid K, Kolomeyer AM, Nayak NV, *et al*. Ocular telehealth screenings in an urban community. *Telemed E Health* 2012;18(2):95-100.

2 Brinks M, Zaback T, Park DW, *et al*. Community-based vision health screening with on-site definitive exams: design and outcomes. *Cogent Med* 2018;5(1):1560641.

3 WHO and Vanessa BC. A vision and eye screening implementation handbook. 2024.

4 Popescu Patoni SI, Muşat AAM, Patoni C, *et al*. Artificial intelligence in ophthalmology. *Rom J Ophthalmol* 2023;67(3):207-213.

5 Kim TN, Aaberg MT, Li P, *et al*. Comparison of automated and expert human grading of diabetic retinopathy using smartphone-based retinal photography. *Eye (Lond)* 2021;35(1):334-342.

6 Harapan BN, Harapan T, Theodora L, *et al*. From Archipelago to pandemic battleground: unveiling Indonesia's COVID-19 crisis. *J Epidemiol Glob Health* 2023;13(4):591-603.

7 Abràmoff MD, Lavin PT, Birch M, *et al*. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 2018;1:39.

8 Abràmoff MD, Lou YY, Erginay A, *et al*. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci* 2016;57(13):5200.

9 Ting DSW, Cheung CY, Nguyen Q, *et al*. Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: a multi-ethnic study. *NPJ Digit Med* 2019;2:24.

10 Lim JI, Regillo CD, Sadda SR, *et al*. Artificial intelligence detection of diabetic retinopathy. *Ophthalmol Sci* 2023;3(1):100228.

11 Xie YC, Nguyen QD, Hamzah H, *et al*. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit Health* 2020;2(5):e240-e249.

12 Keel S, Lee PY, Scheetz J, *et al*. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. *Sci Rep* 2018;8:4330.

13 De Fauw J, Ledsam JR, Romera-Paredes B, *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24(9):1342-1350.

14 Kufel J, Bargieł-Łączek K, Kocot S, *et al*. What is machine learning, artificial neural networks and deep learning—examples of practical applications in medicine. *Diagnostics (Basel)* 2023;13(15):2582.

15 Ting DSW, Cheung CY, Lim G, *et al*. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017;318(22):2211.

16 Ruan S, Liu Y, Hu WT, *et al*. A new handheld fundus camera combined with visual artificial intelligence facilitates diabetic retinopathy screening. *Int J Ophthalmol* 2022;15(4):620-627

17 Cao YT, Che DY, Pan YL, *et al*. Artificial intelligence improves accuracy, efficiency, and reliability of a handheld infrared eccentric autorefractor for adult refractometry. *Int J Ophthalmol* 2022;15(4): 628-634.

18 Grzybowski A, Brona P, Krzywicki T, *et al*. Diagnostic accuracy of automated diabetic retinopathy image assessment software: IDx-DR and RetCAD. *Ophthalmol Ther* 2025;14(1):73-84.

19 Burlina S, Radin S, Poggiato M, *et al*. Screening for diabetic retinopathy with artificial intelligence: a real world evaluation. *Acta Diabetol* 2024;61(12):1603-1607.

20 Antaki F, Hammana I, Tessier MC, *et al*. Implementation of artificial intelligence–based diabetic retinopathy screening in a tertiary care hospital in Quebec: prospective validation study. *JMIR Diabetes* 2024;9:e59867.

21 Wang TW, Luo WT, Tu YK, *et al*. Systematic review and meta-analysis of regulator-approved deep learning systems for fundus diabetic retinopathy detections. *NPJ Digit Med* 2026;9:110.

22 Doğan ME, Bilgin AB, Sari R, *et al*. Head to head comparison of diagnostic performance of three non-mydriatic cameras for diabetic retinopathy screening with artificial intelligence. *Eye (Lond)* 2024;38(9): 1694-1701.

23 Hu CL, Wang YC, Wu WF, *et al*. Evaluation of AI-enhanced non-mydriatic fundus photography for diabetic retinopathy screening. *Photodiagn Photodyn Ther* 2024;49:104331.