AMERICAN ACADEMY OF OPHTHALMOLOGY®

AI-Human Hybrid Workflow Enhances Teleophthalmology for the Detection of Diabetic Retinopathy

Eliot R. Dow, MD, PhD,[1,2] *Nergis C. Khan, BS,*[1] *Karen M. Chen, BS,*[1] *Kapil Mishra, MD,*[1] *Chandrashan Perera, MD,*[1] *Ramsudha Narala, MD,*[1] *Marina Basina, MD,*[3] *Jimmy Dang, BSN,*[3] *Michael Kim, MD,*[3] *Marcie Levine, MD,*[3] *Anuradha Phadke, MD,*[3] *Marilyn Tan, MD,*[3] *Kirsti Weng, MD,*[3] *Diana V. Do, MD,*[1] *Darius M. Moshfeghi, MD,*[1] *Vinit B. Mahajan, MD, PhD,*[1,2] *Prithvi Mruthyunjaya, MD, MHS,*[1] *Theodore Leng, MD, MS,*[1] *David Myung, MD, PhD*[1]

Objective: Detection of diabetic retinopathy (DR) outside of specialized eye care settings is an important means of access to vision-preserving health maintenance. Remote interpretation of fundus photographs acquired in a primary care or other nonophthalmic setting in a store-and-forward manner is a predominant paradigm of teleophthalmology screening programs. Artificial intelligence (AI)-based image interpretation offers an alternative means of DR detection. IDx-DR (Digital Diagnostics Inc) is a Food and Drug Administration-authorized autonomous testing device for DR. We evaluated the diagnostic performance of IDx-DR compared with human-based teleophthalmology over 2 and a half years. Additionally, we evaluated an AI-human hybrid workflow that combines AI-system evaluation with human expert-based assessment for referable cases.

Design: Prospective cohort study and retrospective analysis.

Participants: Diabetic patients $\geq$ 18 years old without a prior DR diagnosis or DR examination in the past year presenting for routine DR screening in a primary care clinic.

Methods: Macula-centered and optic nerve-centered fundus photographs were evaluated by an AI algorithm followed by consensus-based overreading by retina specialists at the Stanford Ophthalmic Reading Center. Detection of more-than-mild diabetic retinopathy (MTMDR) was compared with in-person examination by a retina specialist.

Main Outcome Measures: Sensitivity, specificity, accuracy, positive predictive value, and gradability achieved by the AI algorithm and retina specialists.

Results: The AI algorithm had higher sensitivity (95.5% sensitivity; 95% confidence interval [CI], 86.7%−100%) but lower specificity (60.3% specificity; 95% CI, 47.7%−72.9%) for detection of MTMDR compared with remote image interpretation by retina specialists (69.5% sensitivity; 95% CI, 50.7%−88.3%; 96.9% specificity; 95% CI, 93.5%−100%). Gradability of encounters was also lower for the AI algorithm (62.5%) compared with retina specialists (93.1%). A 2-step AI-human hybrid workflow in which the AI algorithm initially rendered an assessment followed by overread by a retina specialist of MTMDR-positive encounters resulted in a sensitivity of 95.5% (95% CI, 86.7%−100%) and a specificity of 98.2% (95% CI, 94.6%−100%). Similarly, a 2-step overread by retina specialists of AI-ungradable encounters improved gradability from 63.5% to 95.6% of encounters.

Conclusions: Implementation of an AI-human hybrid teleophthalmology workflow may both decrease reliance on human specialist effort and improve diagnostic accuracy.

Financial Disclosure(s): Proprietary or commercial disclosure may be found after the references. *Ophthalmology Science 2023;3:100330*

Supplemental material available at www.ophthalmologyscience.org/

The Center for Disease Control estimates that 415 million people worldwide have diabetes mellitus.[1] Current guidelines recommend that an eye care provider perform a dilated eye examination every 1 to 2 years to screen for vision-threatening complications of diabetic retinopathy (DR) or diabetic macular edema (DME).[2] With approximately 200 000 ophthalmologists practicing worldwide, the eye care workforce is challenged to meet the rising demand of DR detection and treatment.[3] Teleophthalmology performed by remote, asynchronous interpretation of fundus photographs by experienced human readers in a store-and-forward manner has been one means of addressing this

https://doi.org/10.1016/j.xops.2023.100330
ISSN 2666-9145/23

widening gap in public health.[4] More recently, artificial intelligence (AI)-based medical devices for the detection of DR by fundus photographs have also become available for clinical use.[5-8]

IDx-DR (Digital Diagnostics Inc) is a United States Food and Drug Administration (FDA)-authorized, deep-learning-based, autonomous device for DR detection in the primary care setting. To operate IDx-DR, a trained medical assistant, typically in a nonophthalmic setting, uses a 45° nonmydriatic fundus camera to acquire 1 macula-centered image and 1 optic disc-centered image from each eye of a patient with diabetes. The images are then submitted to a cloud-based algorithm that returns a result of whether or not the patient has more-than-mild-DR (MTMDR), defined as meeting or exceeding the ETDRS level 35 or having clinically significant DME.[9] Patients receiving an MTMDR result are recommended by the autonomous system to be referred to an ophthalmologist for an in-person examination, whereas those without MTMDR are instructed to repeat testing in 12 months.[5]

The pivotal trial for IDx-DR, which led to FDA authorization of the device, compared the results of 900 participants tested by the AI system against the interpretation by reading center personnel of widefield stereoscopic photographs and macula-centered spectral-domain (SD) OCT. The trial used a reflexive dilation protocol in which patients were pharmacologically dilated if the device was unable to assess fundus images acquired without mydriasis. The device achieved a rate of gradability of 96.1% along with a sensitivity of 87.2% and specificity of 90.7% for diagnosing MTMDR.

Since this trial, several other studies have reported on the performance of IDx-DR outside of the setting of a formal clinical study. Among 1616 reflexively dilated patients at a hospital in the Netherlands, IDx-DR had an estimated sensitivity of 79.4% and specificity of 93.8% compared with retina specialists.[10] A second study in the Netherlands using image interpretation based on the International Clinical Diabetic Retinopathy Severity Scale (ICDRSS) as a standard of comparison found that the device had a sensitivity of 68% and specificity of 86%.[11,12] Results from a study conducted in Germany found that IDx-DR had a sensitivity of 65.2% and specificity of 66.7%.[13] When used in a hospital in Spain, IDx-DR achieved 100% sensitivity and 81.8% specificity for MTMDR.[14] Finally, when used with 310 children $<$ 21 years old, the sensitivity and specificity were 85.7% and 79.3%, respectively, compared with retina specialists.[15] Outside of the device's pivotal trial, the performance of IDx-DR has not yet been reported in an adult population in the United States.

Apart from fully autonomous DR screening, the cooperation between human experts and AI has not been well explored. Although until recently most tasks in health care could be better performed by an experienced human than an algorithm, increasingly, AI models are able to exceed human experts.[16,17] In other tasks, humans and AI models have complementary strengths and weaknesses across different cases and their combined judgments exceed the individual performance of either.[18] However, in some situations, human interaction with an AI can result in an overall worse performance than either one alone.[19,20] There are a variety of ways that humans and AI models may interact to optimize performance and resources. Further exploration of this interaction is important because a simulated economic analysis suggested that a semiautomated approach to DR screening involving human experts and AI may be less expensive than a solely human-based or fully automated screening system, although these figures may vary depending on the cost of human labor and screening adherence rates.[21-23]

The present study evaluates the performance of an AI system as a substitute for or complement to a human-based teleophthalmology screening program for DR. The investigation occurs in the context of the Stanford Teleophthalmology Autonomous Testing and Universal Screening (STATUS) program, a screening program for DR at 7 primary care sites in the San Francisco Bay Area. The first 18 months of the STATUS program involved store-and-forward teleophthalmology performed by retina specialists in an academic reading center. This phase of the program was followed by a 12-month study period in which an AI system (IDx-DR) was implemented at the same primary care sites with the same imaging hardware operated by largely the same personnel. Performance of the AI system was evaluated against a standard of care for teleophthalmology in DR screening, adjudicated consensus reading by retina specialists, as well as against an in-person examination by a retina specialist supported by multimodal imaging (Figs S1, S2, available at www.ophthalmologyscience.org/). Finally, an AI-human hybrid workflow was evaluated both for DR diagnosis and for AI-ungradable encounters.

Methods

Clinical Sites

The STATUS program is a DR screening network involving the Byers Eye Institute at Stanford University, 5 Stanford-affiliated regional primary care sites (Santa Clara, Los Gatos, Hayward, Castro Valley, and Pleasanton), and 2 Stanford-affiliated endocrinology clinics (Palo Alto and Emeryville) in the San Francisco Bay Area. For 18 months before the implementation of IDx-DR, these primary care sites participated in a human-based store-and-foward teleophthalmology program that obtained fundus images with the same nonmydriatic fundus camera that is paired with IDx-DR (TopCon NW400, TopCon) without the use of autonomous AI in the evaluation of images. Immediately after the human-based teleophthalmology phase of the program, the IDx-DR system was introduced, and the study period spanned 12 months.

Consent was waived because the study was retrospective or adhered to standard of care treatment for the patient without introduction of any additional risk. All patients were verbally informed of the DR screening workflow and elected to participate in the procedure. The study was approved by the Institutional Review Board of Stanford University. The described research adheres to the tenets of the Declaration of Helsinki.

Image Acquisition and Patient Evaluation

Patients $\geq$ 18 years old with type 1 or type 2 diabetes mellitus without a prior DR diagnosis or a DR examination in the past 12

months who presented to the primary care clinic for diabetes management were offered the opportunity to undergo DR teleophthalmology screening. These criteria were adopted to be in compliance with FDA labeling and intended use for IDx-DR and reflect current DR screening recommendations. Screening was performed with human-based teleophthalmology evaluated by retina specialists at the Stanford Ophthalmic Reading Center during the first phase of the program and with evaluation by IDx-DR during the second phase of the program. During both phases, fundus imaging was performed by a trained medical assistant using a TopCon NW400 nonmydriatic fundus camera to acquire 1 45° macula-centered image and 1 optic disc-centered image for each eye. Because patients were not pharmacologically dilated in the program, for 2 minutes before imaging, patients sat in a windowless room with the lights turned off, computers and other ambient light sources turned off or blocked, and eyelids closed. After acquisition of 4 fundus images per session, the images were submitted for diagnosis.

During the teleophthalmology phase, single-reader evaluation occurred by a rotating pool of fellowship-trained retina specialists (T.L. and V.M.) at the Stanford Ophthalmic Reading Center who judged whether the images had sufficient quality for evaluation, and if so, evaluated the images for DR (no DR, mild nonproliferative DR [NPDR], moderate NPDR, severe NPDR, and proliferative DR) and DME (DME present or absent) based on the categories of the ICDRSS (N = 790 patient encounters). Because it was not possible to produce a full count of the number of intraretinal hemorrhages per retinal quadrant in the two 45° image fields of view, a key differentiation between moderate and severe NPRD in the ICDRSS, the 2 categories were combined into a single category, moderate-severe NPDR. Image interpretation was performed using Picture Archiving and Communication Software that allowed readers to manipulate image brightness and contrast (Zeiss Forum, Carl Zeiss Meditec). Patients who received an MTMDR or ungradable result were referred to an ophthalmologist for in-person examination. During the teleophthalmology phase, assessment of image quality and the decision of whether to repeat image acquisition was left to the discretion of the medical assistant.

During the AI phase, the device's secure cloud-based AI systems assessed image quality before clinical evaluation. If ≥ 1 of the images was judged by IDx-DR as having insufficient quality for diagnosis of MTMDR, all 4 images were rejected without clinical evaluation per the device's standard operation. Medical assistants were trained to repeat acquisition up to 4 times before accepting a patient encounter as ungradable. If IDx-DR judged the submitted images to be of sufficient quality for clinical evaluation, it rendered a clinical result of the presence or absence of MTMDR defined as either ETDRS level 35 or higher or DME in ≥ 1 eye. Patients who received an MTMDR or ungradable result were referred to an ophthalmologist for in-person examination (N = 1222 patient encounters; 776 AI-gradable, 446 AI-ungradable).

Human Consensus Overread and In-Person Examination

Overread of a subset of images from AI-gradable patients (n = 199 encounters) was performed by 2 fellowship-trained retina specialists at the Stanford Ophthalmic Reading Center who determined whether the images were gradable and, if so, assigned a diagnosis of DR or DME based on the ICDRSS ("human consensus"). The subset of 199 patients included all AI-gradable patients who were seen for an in-person examination (n = 122) as well as a random subset of AI-gradable patients not seen for an in-person examination (n =77) selected by random number generation. In cases where the 2 readers differed in their assessment, a third reader (E.D.) evaluated the images. The overreads were binned into a patient-level grade of the presence or absence of MTMDR in ≥ 1 eye so that they could be directly compared with the output from IDx-DR. In 3% (6 of 199) of cases, the 3 readers were unable to render a patient-level consensus (1 ungradable, 1 MTMDR-positive, and 1 MTMDR-negative evaluation), and an independent fourth assessment by an ophthalmologist resolved the disagreement. During the AI phase of the program, nearly all AI-ungradable images were prospectively read by a single retina specialist (n = 438), and a subset of these were further overread by an adjudicated consensus of 3 retina specialists (n = 223) in the manner described above.

A subset of patients who underwent IDx-DR evaluation were seen at the Byers Eye Institute at Stanford for an in-person eye examination by a fellowship-trained retina specialist ("in-person examination," n = 180). The examination occurred at a median of 53 days after the IDx-DR encounter. The patient encounter involved a dilated fundus examination, SD-OCT imaging (Cirrus, Carl Zeiss Meditec), and other retinal imaging. The diagnosis was based on a documented examination and assessment. Because the criteria for moderate NPDR per the ICDRSS and ETDRS 35 or greater differ based on the inclusion of cotton wool spots in ETDRS, a separate grade was assigned for the 2 systems (i.e., the presence of absorbent cotton wool spots and microaneurysms indicated ETDRS 35 but only mild NPDR per ICDRSS), although this only resulted in a different patient-level MTMDR diagnosis in 1 instance. In 52.2% of encounters, 200° ultrawide-field color fundus imaging (Optos Inc) was performed. The images were evaluated using ICDRSS and ETDRS criteria. In cases where the objective evaluation of the Optos image differed from the documented diagnosis from in-person examination, a second reader (E.D.) adjudicated the evaluation. In 2 cases (2% of encounters), the adjudicated Optos evaluation changed the patient-level MTMDR grade from that documented in the in-person examination.

Statistics

Data were managed and analyzed using Python (version 3.9.0) with Pandas (version 1.3.0) and Microsoft Excel (version 16.16.27). Figures were created in Google Sheets and Adobe Illustrator. Student *t* tests were conducted as 2-tailed tests.

Data Availability

Data without patient identifiers used in the analyses for the manuscript will be made available to investigators upon request.

Results

Performance of the AI System and Human-based Teleophthalmology

During the teleophthalmology phase of the DR screening program, 5.6% of gradable encounters (5.1% of all encounters) resulted in a diagnosis of MTMDR. In comparison, during the AI phase, the AI system identified MTMDR in 19.0% of gradable encounters (11.9% of all encounters). The increased proportion of patients diagnosed with MTMDR by the AI system began the first month that the system was introduced and persisted throughout the study period (Figs 3, S4, available at www.ophthalmologyscience.org/). This suggested that the more than threefold increase in the rate of MTMDR resulted from either

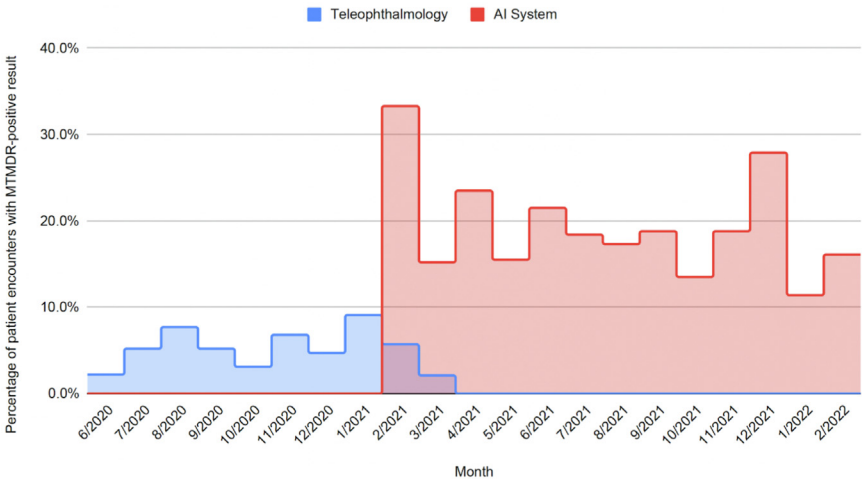

Figure 3. Percentage of monthly encounters with a more-than-mild diabetic retinopathy (MTMDR)-positive diagnosis identified by human-based teleophthalmology (blue) or artificial intelligence (AI) system (red) among gradable encounters.

higher specificity of the human teleophthalmology readers, higher sensitivity of the AI system for MTMDR, or both.

An adjudicated consensus overread by retina specialists of the fundus images ("human consensus") from the AI phase of the program showed agreement between the human consensus and AI system on the presence or absence of MTMDR in 78.4% of cases. Both the human consensus and the AI system were compared to in-person examination by a retina specialist ("in-person examination") that was supported by a combination of SD-OCT and widefield fundus photography. The human consensus compared with the in-person examination had an accuracy of 91.8%, including 69.5% sensitivity (95% confidence interval [CI], 50.7%−88.3%) and 96.9% specificity (95% CI, 93.5%−100%) for the presence or absence of MTMDR (Figs 5 and 6 and S7 and S8, available at www.ophthalmologyscience.org/). When the DR assessments were expanded beyond being positive or negative for MTMDR to a modified 4-stage classification based on ICDRSS, the human consensus and in-person examination had the same diagnosis in 83.3% of encounters and were within 1 stage of each other in 96.5% of encounters (Figs S9, S10, available at www.ophthalmologyscience.org/). The comparison of the AI system to in-person examination had a 70.0% accuracy including 95.5% sensitivity (95% CI, 86.7%−100%) and 60.3% specificity (95% CI, 47.7%−72.9%) (Figs 5, 6, S7, S8).

Evaluation of Discrepancies in Disease Assessment

Encounters in which there was a discrepancy in the DR diagnosis between the AI system and the in-person examination result did not show any statistically significant predilection for age, sex, or self-reported race or ethnicity (Student's 2-tailed t test for age, $P = 0.74$; chi-square statistic for sex, 3.51, $P = 0.061$; chi-square statistic for White vs. non-White race, 0.446, $P = 0.50$; N = 1222). There was also no association with ocular comorbidities including cataract or ocular surface disease when analyzed for the subset of patients who had a documented in-person examination (chi-square statistic for cataract, 0.0419, $P = 0.84$; chi-square statistic for ocular surface disease, 0.0546, $P = 0.815216$; n = 80).

Close examination of the fundus images from discrepant cases between the AI system and in-person examination showed a single case in which there was an intraretinal hemorrhage in the midperipheral retina outside of the fields of view of the AI system's fundus images; all human consensus reads for this case were also MTMDR-negative (Fig S11, available at www.ophthalmologyscience.org/). Inspection of the cases in which the AI system gave an MTMDR-positive assessment compared with an MTMDR-negative diagnosis on in-person examination showed 1 instance of a branched retinal vein occlusion without DR found on in-person examination; human readers also made a diagnosis of MTMDR-positive while noting that the pattern of hemorrhages on 1 side of the fundus midline made a retinal vein occlusion more likely than DR (Figs S11, 12, available at www.ophthalmologyscience.org/). Close inspection of 2 other encounters in which the AI system was MTMDR-positive showed an intraretinal hemorrhage that was not described on the in-person examination; there was no widefield fundus image taken during those in-person examinations to confirm the presence or absence of the hemorrhages at that encounter (Fig S13, available at www.ophthalmologyscience.org/).

Images were examined from 5 encounters in which the human consensus was MTMDR-negative, whereas both the AI system and the in-person examinations were MTMDR-positive. In 4 of the cases, there were microaneurysms or small intraretinal hemorrhages, while the fifth had exudates confirmed on SD-OCT imaging (Fig S14, available at www.ophthalmologyscience.org/). There were also 2 encounters that the AI system assessed as MTMDR-positive, whereas both the human consensus and in-person examinations were MTMDR-negative; however, close

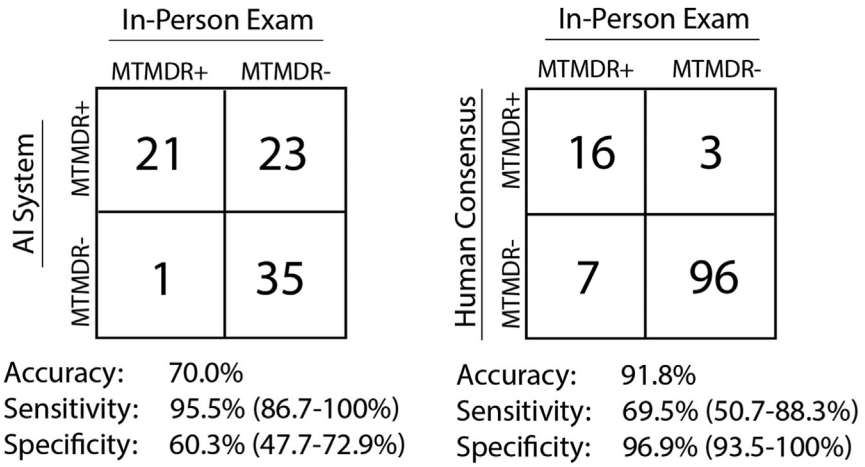

Figure 5. Confusion matrices comparing more-than-mild diabetic retinopathy (MTMDR) positive or negative results of screening encounters by the artifical intelligence (AI) system versus the in-person examination, and the human consensus overread by retina specialists versus the in-person examination. All numbers represent a patient-level assessment within a screening encounter (n = 80 patient encounters evaluated by the AI system and in-person examination; n = 122 encounters evaluated by both the human consensus and the in-person examination; note that these groups of patients are not mutually exclusive).

inspection of the fundus photographs and ultrawide-field fundus images showed 1 clear intraretinal hemorrhage in each encounter (Fig S13).

Patients who were screened by the AI system could have up to 6 human assessments for DR (MTMDR-positive, MTMDR-negative, or ungradable), including up to 3 human overreads of the fundus images acquired during the AI screening encounter, up to 2 evaluations of ultrawide-field fundus images, and 1 in-person dilated fundus examination. The level of agreement between these independent human assessments was calculated with a Fleiss statistic, a measure of interrater reliability across > 2 raters and multiple categories. For cases in which the AI-system agreed with the in-person examination, the Fleiss statistic was 0.71, indicating a high level of interrater agreement among independent human evaluations. In contrast, cases in which the AI system and in-person examination had discordant diagnoses had a Fleiss statistic of 0.06, indicating a low level of agreement among the independent human evaluations.

AI-Human Hybrid

When judged against the in-person examination diagnosis as an estimate of the true state of disease, the AI system had a positive predictive value of 47.7%, indicating that fewer

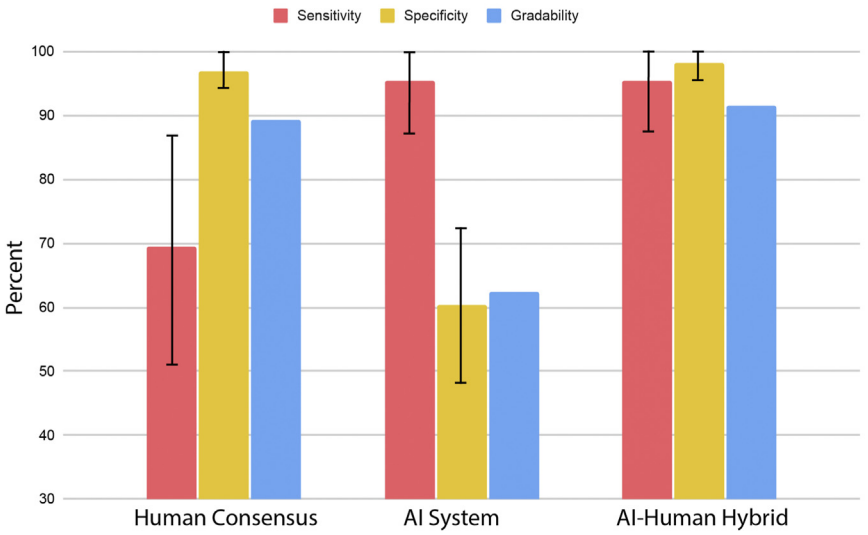

Figure 6. Sensitivity, specificity, and gradability of encounters for human consensus adjudicated among retina specialists, artificial intelligence (AI) system, and a 2-step AI-human hybrid workflow. 95% confidence intervals are displayed.

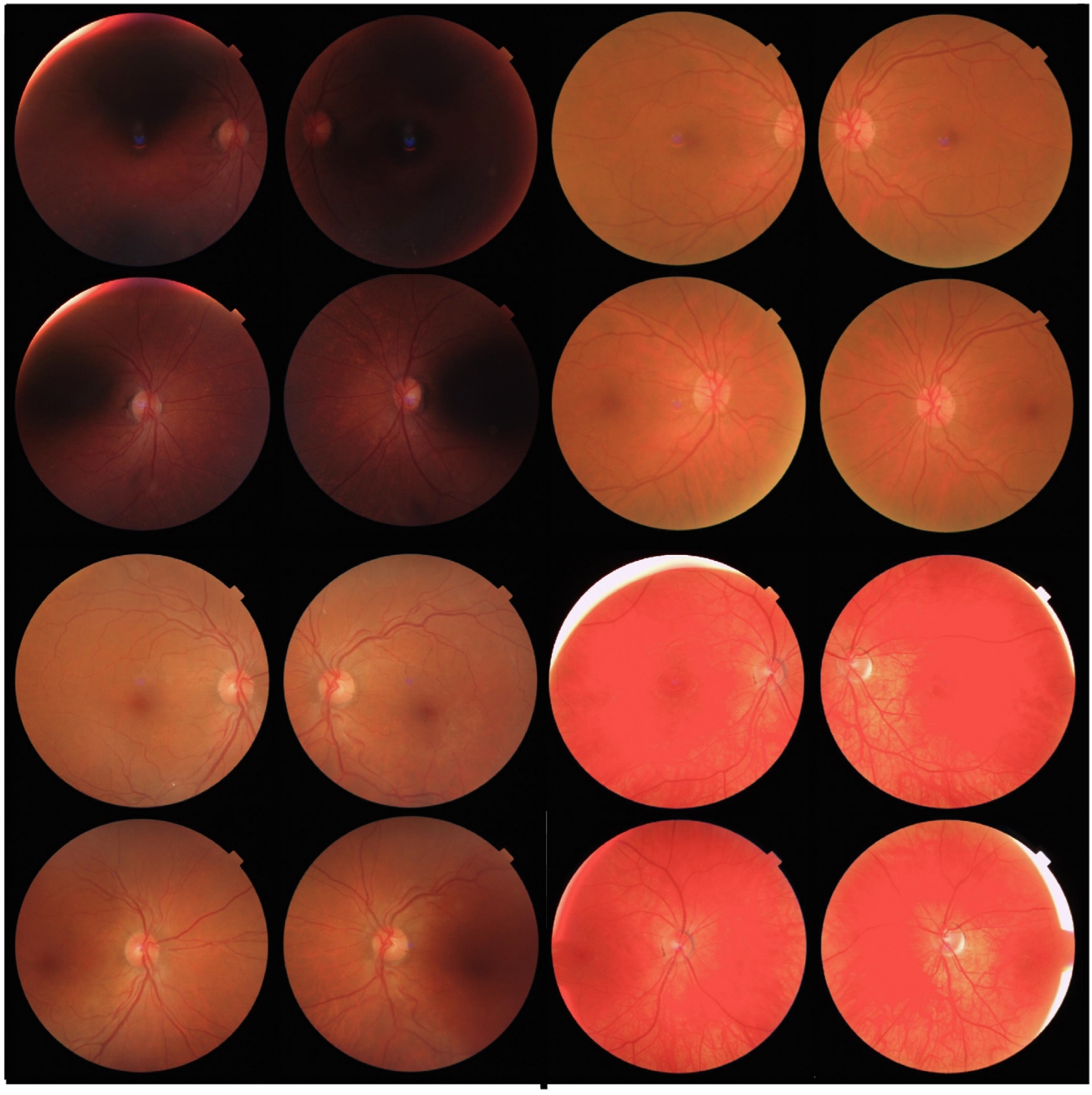

Figure 12. Sets of fundus images from 4 representative patient encounters. In each case, the artificial intelligence system gave a more-than-mild diabetic retinopathy (MTMDR)-positive diagnosis and both the human consensus and in-person examination were MTMDR-negative.

than half of patients referred to an ophthalmologist would be MTMDR-positive. We considered whether an AI-human hybrid workflow for DR screening could improve the positive predictive value and accuracy of the program. We analyzed the data as a 2-step workflow in which the DR screening encounter is first performed by the AI system followed by the most experienced retina specialist overreading all MTMDR-positive results before referral to an ophthalmologist. Under this 2-step workflow, the positive predictive value increased to 95.5% (95% CI, 86.7%−100%), along with accuracy of 97.4%, sensitivity of 95.5% (95% CI, 86.7%−100%), and specificity of 98.2% (95% CI, 94.6%−100%) (Figs 6, 15). The adjudicated consensus overread only changed the diagnosis in 2 cases out of 80 compared with the single experienced reader.

Additionally, in this nonpharmacologically dilated patient population, the AI system rendered an ungradable result in 36.5% of encounters. A single experienced retina specialist was able to interpret 93.1% of all patient encounters as well as 88.1% of AI-ungradable encounters, and

an adjudicated consensus agreed that the image was interpretable in 77.7% of cases. Thus, we evaluated a second AI-human hybrid workflow in which fundus images from AI-ungradable encounters were prospectively assessed by a retina specialist. Under this 2-step AI-human hybrid workflow, a diagnosis was rendered in 91.6% of encounters (Fig 6). Finally, patients who were ungradable by both the AI system and the retina specialist were referred for an in-person examination, and only 1 of 36 eyes (2.8%) in this group could not be evaluated for DR or DME owing to a dense cataract. Among the remaining 35 eyes examined in person, 3 were found to have MTMDR (8.5%).

Discussion

As the number of individuals with diabetes mellitus increases both within the United States and worldwide, there is a looming public health crisis of delivering necessary eye care to these patients. Even at current prevalence of diabetes, only an estimated 15.3% of patients with diabetes receive recommended eye examinations.[24] Teleophthalmology and AI-based screening offer potential alternatives to in-person examination by an ophthalmologist to increase access to eye care in a resource-conscious manner. This study describes the results of a DR screening program at primary care clinics in the San Francisco Bay Area that combines both AI- and human-based teleophthalmology screening methods into an AI-human hybrid workflow. The program initially employed teleophthalmology by having retina specialists perform remote, asynchronous interpretation of fundus images in a store-and-forward manner; subsequently, screening was carried out using an FDA-approved autonomous AI system. This continuous transition from teleophthalmology to AI at the same clinical sites using the same fundus imaging protocol and hardware allowed us to compare the performance of human-based teleophthalmology and AI for remote DR screening. It additionally allowed us to implement a hybrid workflow that combined AI system and expert-human evaluation.

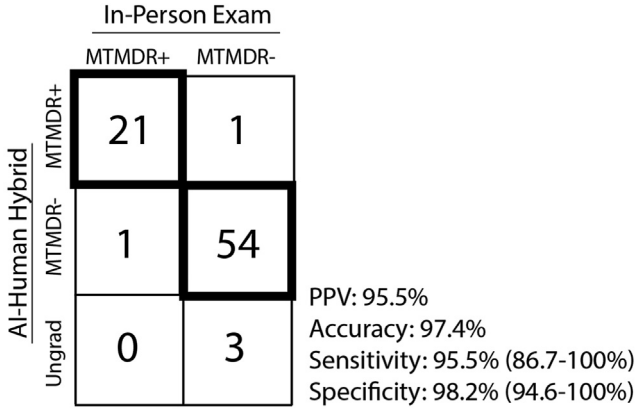

Figure 15. Confusion matrix of a 2-step, artificial intelligence (AI)-human hybrid workflow involving an experienced retinal specialist overreading all patients judged as more-than-mild diabetic retinopathy (MTMDR)-positive by the AI system. PPV = positive predictive value.

We found that during the AI phase of the program, > 3 times the number of MTMDR patients were identified than during the human teleophthalmology phase. We investigated this difference by comparing the results of the AI system and an adjudicated consensus overread of the encounters by retina specialists to in-person examination by retina specialists supported by multimodal retinal imaging. This analysis showed that although both approaches demonstrate high accuracy, the sensitivity of the AI system in our patient population exceeded that of human readers, while the specificity of human readers exceeded the AI system.

There are several explanations for the diagnostic discrepancies between IDx-DR, expert human assessment, and in-person examination. First, diagnostic errors by the AI device may have been due in part to poor image quality and image artifacts. Unlike the IDx-DR pivotal trial, screening in our system was performed exclusively with patients who were not pharmacologically dilated, likely leading to a higher frequency of darker images and other pupil-related image anomalies. Artificial intelligence performance did not differ across age, sex, or self-reported race or ethnicity.

Second, IDx-DR was calibrated against images analyzed in a reading center applying strict ETDRS criteria to fundus photographs and aided by SD-OCT for the assessment of DME. Although the prognostic value of the ETDRS system for DR outcomes has been well established, in the course of routine care, retina specialists may or may not evaluate patients by those criteria instead favoring ICDRSS or another grading system. Experts, including retina specialists, may also have imperfect inter- or intrarater reliability.[25] Several fundus photographs in this study contained intraretinal hemorrhages that were not identified by most human graders nor were documented on the in-person eye exam, suggesting that in some cases the AI system may catch lesions missed by experts. However, there were other cases in which the AI system was discordant with both the human consensus and the in-person examination, and exhaustive examination of the fundus images could not identify retinal biomarkers to support the AI system's diagnosis of MTMDR, showing that the AI system is also not without error. Of note, there were no cases in which IDx-DR missed DME that was subsequently detected by SD-OCT.

Moreover, the discordant images may have sat at the borderline of mild and moderate DR or had other attributes that made the diagnosis difficult at a ground-truth level. This was reflected in the Fleiss statistic calculated across all available human diagnoses: for cases in which the AI system and the in-person examination agreed, there was very high agreement across the independent human diagnoses, whereas when the AI system disagreed with the in-person examination, humans evaluating the patient across multiple modalities also tended to disagree. For instance, there were several cases in which a fundus image with a lesion that could be construed as a microaneurysm or a small intraretinal hemorrhage was judged as either mild or moderate NPDR. Thus, although strict adherence to a prognostic standard like ETDRS may increase interrater reliability, routine clinical practice may demonstrate higher agreement

in clinically meaningful diagnoses (microaneurysms only versus neovascularization) than in clinical distinctions with less prognostic value (interpretation of a single lesions as a microaneurysm versus intraretinal hemorrhage).

A fourth possibility is that IDx-DR may be calibrated to prioritize sensitivity even if there is a trade-off in specificity. One motivation for greater sensitivity is that the clinical consequences of missing 1 case of MTMDR typically exceed the psychosocial effect on the patient and the economic impact on the health care system of a false positive result. Additionally, just as a car accident in a self-driving vehicle may be judged more harshly by the public than one caused by a human driver, the autonomous AI screening device may use a conservative operating point to avoid a similar outcry. Although in the pivotal trial IDx-DR demonstrated sensitivity and specificity without statistically significant difference, this balance may be different in pharmacologically dilated patients or other wider data distributions.[5]

The high sensitivity and comparatively lower specificity of the AI system resulted in a positive predictive value of 47.7% among gradable results, which offers value to a DR screening program because it would reduce the number of patients who need to be seen by a provider by more than half. Additionally, at this positive predictive value, approximately 1 of 2 referred patients will be positive for MTMDR, whereas without this screening step approximately 1 in 10 or fewer patients referred for a diabetic eye examination will be positive.[26] Yet, the value of AI screening for DR can be enhanced even more by layering AI-based examinations into a teleophthalmology program rather than fully replacing human readers. We simulated a 2-step, AI-human hybrid screening system in which patients were first evaluated by the AI system, and then MTMDR-positive results were overread by a single, experienced retina specialist. Under this system, the positive predictive value increased to 95.5% at a sensitivity of 95.5% and specificity of 98.2%. Previous work has also shown this hybrid model to be more cost-effective than either fully automated or human-based assessment alone.[21]

A second application of the AI-human workflow is the use of expert readers to interpret fundus images from AI-ungradable encounters. In our study, gradability was 62.5%, which is in line with other reports from use outside of the clinical-trial setting, particularly studies in which patients were not pharmacologically dilated.[10,12-14] Because ungradable patients are referred for in-person examination, a low rate of gradability can overwhelm referral pathways, confuse patients, and decrease provider reliance on the screening system. We found that for more than three-quarters of encounters ungradable for AI, retina specialists judged that they were sufficient to evaluate for DR. These results support a 2-step, AI-human hybrid workflow involving all IDx-DR encounters with a referable result, MTMDR-positive or AI-ungradable, being delegated to human readers for further evaluation before referral for in-person examination (Fig 16).

A hybrid workflow may also offer a transition of existing human-based teleophthalmology screening programs into AI-based screening programs. Current teleophthalmology programs may have significant investments in personnel and capital equipment that may disincentivize an outright change to the use of an AI device. As AI systems improve in specificity and as pharmacologic pupil dilation becomes increasingly within the scope of practice at primary care clinics, AI may gradually take on a greater share of the workload.

Areas both within and outside healthcare have benefited from productive AI-human interactions. For instance, in basic neuroscience research, a human-in-the-loop approach to the reconstruction of axons and dendrites in volumetric electron microscopy data has advanced the field of connectomics.[27] In a more everyday example, the braking speed of a self-driving vehicle that detects an impending collision on the freeway far exceeds human reaction time. However, the human driver may be alerted to control vehicle steering in a low-confidence navigational situation such as maneuvering through a field scattered with people as they depart from an outdoor concert. Although AI is becoming increasingly abundant in health care, it is not always clear when the provider should take over the wheel. Additional research on out-of-distribution awareness and anomaly detection for medical image analysis may be helpful.[28,29] Human factors and usability engineering should be an important consideration in the design of medical devices, but as yet, there are no FDA-authorized diagnostic medical devices intended for use in a hybrid pathway with human-in-the-loop analysis of medical images.[30]

Finally, to the best of our knowledge, this study represents the first results of routine clinical use of IDx-DR in an adult population in the United States. The patients served by the program reflect the demographically diverse communities of the San Francisco Bay Area. Consequently, our study included $>$ 20 times the proportion of East Asian patients compared with the IDx-DR pivotal trial.[5] The cohort also included significant numbers of patients of Hispanic/Latino, Black, White, Pacific Islander, or Native American self-reported race or ethnicity.

There are several limitations to this study. One important qualification of the study results is that all encounters during the screening program were performed on nonmydriatic patients. It is likely that pharmacologic pupil dilation not only increases the rate of gradability for IDx-DR as documented in the pivotal trial, but also increases the accuracy of the analysis among gradable images due to fewer imaging artifacts.[5] However, pharmacologic pupil dilation may not be available in screening sites due to concerns for risk of adverse events, scope-of-practice issues, and limited resources for performing and monitoring dilated patients. Of note, with additional training of personnel to improve fundus image acquisition, gradability has improved to over 70% now in the second year of the AI phase of the STATUS program.

An additional limitation includes the fact that patients who underwent an in-person examination were not randomly chosen from the screened population and thus may represent a biased subset. This represents an inherent

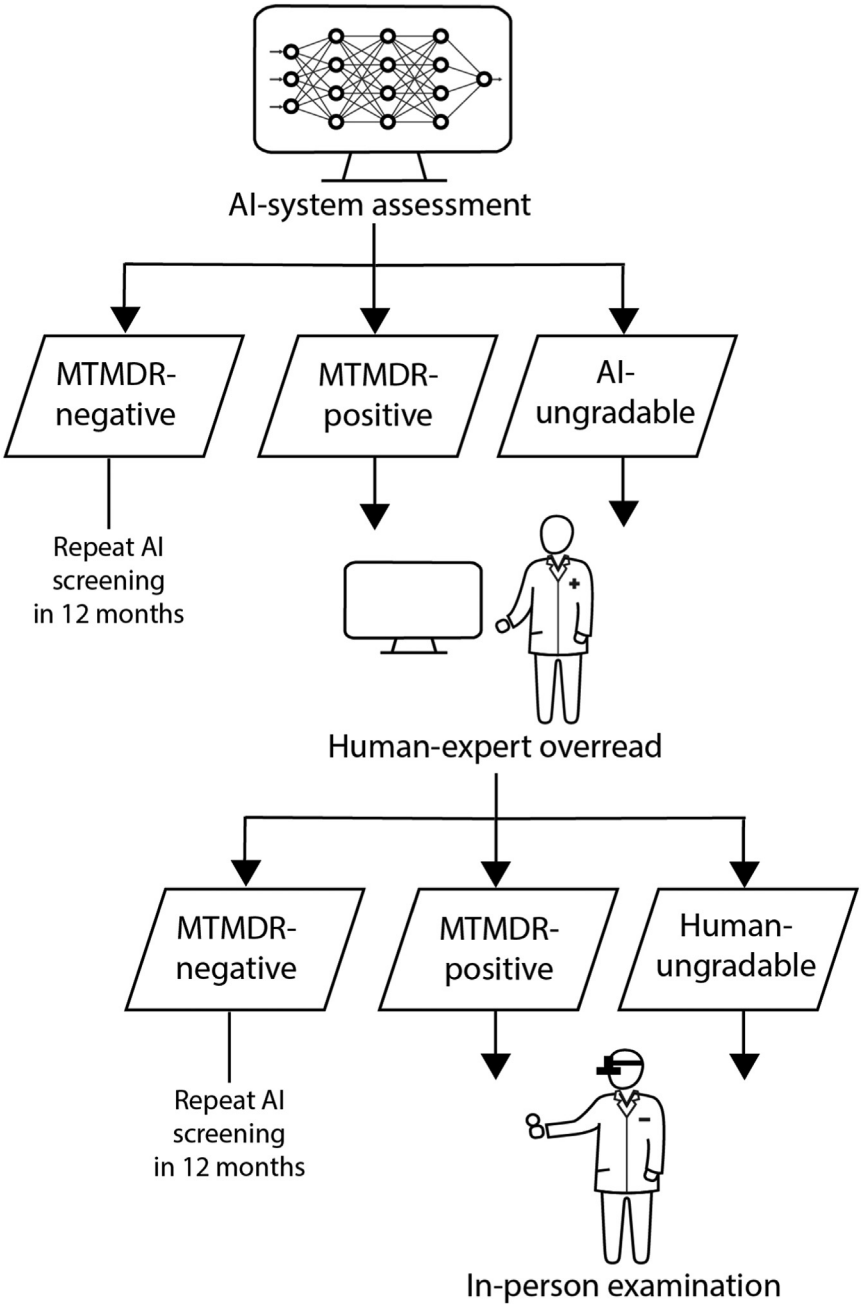

Figure 16. Diagram of a proposed artificial intelligence (AI)-human hybrid workflow in which fundus images from AI-screening encounters with an output of more-than-mild diabetic retinopathy (MTMDR)-positive or AI-ungradable are overread by a human expert in a traditional teleophthalmology store-and-forward system. Patients with a MTMDR-negative outcome by the human assessment can be rescreened by the AI system in 12 months, whereas patients with a MTMDR-positive assessment or who are ungradable by the human should be referred for an in-person eye examination.

limitation of the observational nature of the study, and the findings could be further validated through a prospective, randomized trial. Although many patients with an MTMDR-negative IDx-DR result subsequently underwent in-person examination, most patients had been referred for in-person examination because of an ungradable or MTMDR-positive IDx-DR result. The group of patients undergoing in-person examination therefore likely had a higher prevalence of MTMDR than the general population, which may have resulted in an overestimation of the positive predictive value. Also of note, the 52-day median lag between the referable screening result and in-person examination may be overly long, particularly for patients with severe NPDR or PDR. This underscores the importance of approaches that improve the positive predictive value of screening results so as not to overwhelm referral pathways, in addition to clear communication to patients regarding the importance of prompt follow-up care.

Finally, although the retina specialists involved in the study are highly qualified, their training in and use of strict ETDRS criteria was unknown, whereas the AI system was tested under a clinical trial setting against readers using this prognostically validated scale. Further prospective investigations should perform in-person DR assessments using ETDRS-grading criteria.

In conclusion, this study demonstrates the feasibility of a hybrid caseload sharing model combining both autonomous AI and store-and-forward teleophthalmology for diabetic retinopathy screening. The system starts with autonomous AI-based screening, which renders a diagnosis for most patients and catches almost all true-positive cases of MTMDR. The lower specificity compared with human experts can be compensated by an AI-human hybrid workflow involving expert overreading of positive cases before referral for in-person examination. Likewise, nearly all patients who have an AI-ungradable encounter may be evaluated remotely and asynchronously using an experienced reader. The remaining patients who are either MTMDR-positive or ungradable by both AI and the human reader may present for an in-person dilated fundus examination by a retina specialist (Figure 16). The provision of necessary eye care to the hundreds of millions of individuals worldwide with diabetes mellitus, a figure that grows each year, is a significant public health challenge that may not be addressed by ophthalmologists alone. An AI-human hybrid workflow may allow scalable software solutions to meet the rising need for essential preventative eye care while maintaining the standard level of care offered by human assessment.

Acknowledgments

The authors thank Michael Abramoff for comments on the manuscript. The authors thank Jill Terrell, Dena Weissman, and Laurisa Perlberg of Digital Diagnostics LLC for their contributions to and support of the STATUS Program. The authors thank Christopher Corbett for artwork used in the figures.

Footnotes and Disclosures

Originally received: December 31, 2022.
Final revision: May 4, 2023.
Accepted: May 8, 2023.
Available online: May 12, 2023. Manuscript no. XOPS-D-22-00279R2.

[1] Byers Eye Institute at Stanford, Stanford University School of Medicine, Palo Alto, California.

[2] Veterans Affairs Palo Alto Health Care System, Palo Alto, California.

[3] Stanford Healthcare, Stanford University, Palo Alto, California.

Disclosures:

All authors have completed and submitted the ICMJE disclosures form.

The authors made the following disclosures:

V.B.M.: Stock – Digital Diagnostics, LLC; Supported by NIH grants (R01 EY031952, R01 EY030151, R01NS98950, and R01EY031360), Stanford Center for Optic Disc Drusen. D.M.M.: Grants – Research to Prevent Blindness, Inc. (grant no. NEI P30-EY026877); Consulting fees – Ainsly, Alexion, Akebia, Regeneron, Clinical trials research group; Payment or honoraria – Slack, Vindico, University of Miami; Data Safety Monitoring Board – AffaMed; Leadership or fiduciary role - ASRS Credentialing Committee; Stock or stock options – Ainsly, DSENTZ, Plenoptika, PR3VENT, Promisight, Pykus, Visunex.

The other authors have no proprietary or commercial interest in any materials discussed in this article.

This research was funded in part by Research to Prevent Blindness (T.L., D.M., R.N., D.D., P.M., V.M.), National Eye Institute/NIH grant P30-EY026877 (T.L., D.M., R.N., D.D., P.M., V.M.), and the Stanford Diabetes Research Center. The funding sources had no involvement in the study design, data collection, analysis, interpretation, or manuscript preparation.

HUMAN SUBJECTS: Human subjects were included in this study.

Consent was waived because the study was retrospective or adhered to standard of care treatment for the patient without introduction of any additional risk. All patients were verbally informed of the diabetic retinopathy screening workflow and elected to take part in the procedure. The study was approved by the Institutional Review Board at Stanford University. The described research adheres to the tenets of the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Dow, Leng, Moshfeghi, Myung.

Data collection: Dow, Khan, Chen, Mishra, Narala, Basina, Dang, Levine, Kim, Phadke, Tan, Weng, Do, Moshfeghi, Mahajan, Mruthyunjaya, Leng, Myung.

Analysis and interpretation: Dow, Khan, Chen, Mishra, Moshfeghi, Myung.

Obtained funding: Dow, Tan, Do, Mahajan, Mruthyunjaya, Leng, Myung.

Overall responsibility: Dow, Khan, Perera, Do, Mahajan, Mruthyunjaya, Leng, Moshfeghi, Myung.

Abbreviations and Acronyms:
AI = artificial intelligence; **CI** = confidence interval; **DME** = diabetic macular edema; **DR** = diabetic retinopathy; **FDA** = Food and Drug Administration; **ICDRSS** = International Clinical Diabetic Retinopathy Severity Scale; **MTMDR** = more-than-mild diabetic retinopathy; **NPDR** = nonproliferative diabetic retinopathy; **SD-OCT** = spectral-domain OCT; **STATUS** = Stanford Teleophthalmology Autonomous Testing and Universal Screening.

Keywords:
Artificial intelligence, Deep learning, Diabetic retinopathy, Human-in-the-loop, Teleophthalmology.

Correspondence:
David Myung, MD, PhD, Byers Eye Institute at Stanford, Stanford University School of Medicine, Palo Alto, CA 94303. E-mail: djmyung@stanford.edu.

References

1. Center for Disease Control and Prevention (CDC). Diabetes: Facts and Statistics. March 9, 2022. https://www.cdc.gov/diabetes/data/index.html. Accessed March 9, 2022.

2. Solomon SD, Chew E, Duh EJ, et al. Diabetic retinopathy: a position statement by the American Diabetes Association. *Diabetes Care*. 2017;40(3):412–418.

3. Resnikoff S, Felch W, Gauthier TM, et al. The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200,000 practitioners. *Br J Ophthalmol*. 2012;96(6):783−787.
4. Surendran TS, Raman R. Teleophthalmology in diabetic retinopathy. *J Diab Sci Technol*. 2014;8(2):262−266.
5. 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 Digital Med*. 2018;1:39.
6. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402−2410.
7. Ipp E, Liljenquist D, Bode B, et al. for the EyeArt Study Group. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. *JAMA Netw Open*. 2021;4(11):e2134254.
8. Ting DSW, Cheung CYL, 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):2211−2223.
9. Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy: ETDRS report number 12. *Ophthalmology*. 1991;98(5):823−833.
10. van der Heijden AA, Abramoff MD, Verbraak F, et al. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*. 2018;96:63−68.
11. Wilkinson CP, Ferris FL, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110(9):1677−1682.
12. Verbraak FD, Abramoff MD, Bausch GCF, et al. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. *Diabetes Care*. 2019;42:651−656.
13. Paul S, Tayar A, Morawiec-Kisiel E, et al. Einsatz von künstlicher Intelligenz im Screening auf diabetische Retinopathie an einer diabetologischen Schwerpunktklinik. *Der Ophthalmologe*. 2021;119:705−713.
14. Shah A, Clarida W, Amelon R, et al. Validation of automated screening for referable diabetic retinopathy with an autonomous diagnostic artificial intelligence system in a Spanish population. *J Diab Sci Technol*. 2021;15(3):655−663.
15. Wolf RM, Liu ATY, Thomas C, et al. The SEE study: safety, efficacy, and equity of implementing autonomous artificial intelligence for diagnosing diabetic retinopathy in youth. *Diab Care*. 2021;44(3):781−787.
16. Yamashita T, Asaoka R, Terasaki H, et al. Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Transl Vis Sci Technol*. 2020;9(2):4.
17. Yim J, Chopra R, Spitz T, Winkens J, Obika A, Kelly C, et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nat Med*. 2020;26(6):892−899.
18. Patel BN, Rosenberg L, Willcox G, et al. Human−machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit Med*. 2019;2:111.
19. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*. 2007;356:1399−1409.
20. Tschandl P, Rinner C, Apalla Z, et al. Human−computer collaboration for skin cancer recognition. *Nat Med*. 2020;26:1229−1234.
21. Xie Y, 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.
22. Wolf RM, Channa R, Abramoff MD, et al. Cost-effectiveness of autonomous point-of-care diabetic retinopathy screening for pediatric patients with diabetes. *JAMA Ophthalmol*. 2020;138(10):1063−1069.
23. Abràmoff MD, Roehrenbeck C, Trujillo S, et al. A reimbursement framework for artificial intelligence in healthcare. *NPJ Digit Med*. 2022;5(1):72−75.
24. Benoit SR, Bonnielin S, Geiss LS, et al. Eye care utilization among insured people with diabetes in the U.S., 2010-2014. *Diabetes Care*. 2019;42(3):427−433.
25. Lin DY, Blumenkranz MS, Brothers RJ. The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. *Am J Ophthalmol*. 2002;134(2):204−213.
26. Eye Diseases Prevalence Research Group. The prevalence of diabetic retinopathy among adults in the United States. *Arch Ophthalmol*. 2004;122(4):552−563.
27. Motta A, Boergens KM, et al. Dense connectomic reconstruction in layer 4 of the somatosensory cortex. *Science*. 2019;366(6469).
28. Burlina P, Paul W, Liu TYA, et al. Detecting anomalies in retinal diseases using generative, discriminative, and self-supervised deep learning. *JAMA Ophthalmol*. 2022;140(2):185−189.
29. Cao T, Huang CW, Hui DYT, et al. A benchmark of medical out of distribution detection. *ArXiv*. 2020. https://doi.org/10.48550/arXiv.2007.04250.
30. Food and Drug Administration. *Applying Human Factors and Usability Engineering to Medical Devices*; 2016. https://www.fda.gov/media/80481/download. Accessed March 1, 2022.