Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems

Aaron Y. Lee,[1,2,3] Ryan T. Yanagihara,[1] Cecilia S. Lee,[1,2] Marian Blazes,[1] Hoon C. Jung,[1,2] Yewlin E. Chee,[1] Michael D. Gencarella,[1] Harry Gee,[4] April Y. Maa,[5,6] Glenn C. Cockerham,[7,8] Mary Lynch,[5,9] and Edward J. Boyko,[10,11]

Diabetes Care 2021;44:1168–1175 | https://doi.org/10.2337/dc20-1877

OBJECTIVE

With rising global prevalence of diabetic retinopathy (DR), automated DR screening is needed for primary care settings. Two automated artificial intelligence (AI)–based DR screening algorithms have U.S. Food and Drug Administration (FDA) approval. Several others are under consideration while in clinical use in other countries, but their real-world performance has not been evaluated systematically. We compared the performance of seven automated AI-based DR screening algorithms (including one FDA-approved algorithm) against human graders when analyzing real-world retinal imaging data.

RESEARCH DESIGN AND METHODS

This was a multicenter, noninterventional device validation study evaluating a total of 311,604 retinal images from 23,724 veterans who presented for teleretinal DR screening at the Veterans Affairs (VA) Puget Sound Health Care System (HCS) or Atlanta VA HCS from 2006 to 2018. Five companies provided seven algorithms, including one with FDA approval, that independently analyzed all scans, regardless of image quality. The sensitivity/specificity of each algorithm when classifying images as referable DR or not were compared with original VA teleretinal grades and a regraded arbitrated data set. Value per encounter was estimated.

RESULTS

Although high negative predictive values (82.72–93.69%) were observed, sensitivities varied widely (50.98–85.90%). Most algorithms performed no better than humans against the arbitrated data set, but two achieved higher sensitivities, and one yielded comparable sensitivity (80.47%, $P = 0.441$) and specificity (81.28%, $P = 0.195$). Notably, one had lower sensitivity (74.42%) for proliferative DR ($P = 9.77 \times 10^{-4}$) than the VA teleretinal graders. Value per encounter varied at \$15.14–\$18.06 for ophthalmologists and \$7.74–\$9.24 for optometrists.

CONCLUSIONS

The DR screening algorithms showed significant performance differences. These results argue for rigorous testing of all such algorithms on real-world data before clinical implementation.

[1]Department of Ophthalmology, University of Washington School of Medicine, Seattle, WA
[2]Department of Ophthalmology, Puget Sound Veteran Affairs, Seattle, WA
[3]eScience Institute, University of Washington, Seattle, WA
[4]Office of Information and Technology, Clinical Imaging, Seattle, WA
[5]Department of Ophthalmology, Emory University School of Medicine, Atlanta, GA
[6]Regional Telehealth Services, Veterans Affairs Southeast Network Veterans Integrated Service Networks (VISN 7), Duluth, GA
[7]Veterans Health Administration, Specialty Care Services, Washington, DC
[8]Ophthalmology Service, Stanford University School of Medicine, Palo Alto, CA
[9]Ophthalmology Section, Atlanta Veterans Affairs Medical Center, Atlanta, GA
[10]Seattle Epidemiologic Research and Information Center, Department of Veterans Affairs Medical Center, Seattle, WA
[11]Department of Medicine, University of Washington, Seattle, WA

Corresponding author: Aaron Y. Lee, leeay@uw.edu

Received 27 July 2020 and accepted 25 October 2020

This article contains supplementary material online at https://doi.org/10.2337/figshare.13148540.

A major microvascular complication of diabetes mellitus (DM) is diabetic retinopathy (DR), which is the leading cause of preventable blindness in working-age Americans (1,2). If detected and managed at an early stage, irreversible blindness can be avoided (3). Therefore, the American Academy of Ophthalmology Preferred Practice Pattern recommends that patients with DM undergo an annual dilated retinal fundus examination, and the American Diabetes Association recommends dilated examinations every 2 years for patients with type 2 DM without retinopathy (4,5). The global prevalence of DM has tripled over the past 20 years, affecting 151 million in 2000, 463 million in 2019, and a projected 700 million by 2045 (6). At this rate, eye care providers who deliver routine screening will become overwhelmed (7). Despite the effectiveness of teleretinal screening programs, these programs are also costly and labor intensive (2,8,9). Therefore, an inexpensive, accurate, and automated method to triage DR screening fundus photographs in the primary care clinic setting would greatly benefit providers, health care systems, and patients.

Artificial intelligence (AI)–based algorithms may provide promising solutions to alleviate the DR screening burden. Tufail and colleagues (10,11) have shown that when used in DR screening programs, AI algorithms can detect referable DR with high sensitivity and are cost-effective compared with manual grading, the current gold standard. However, these studies predated the era of deep learning, a machine learning technique that has revolutionized retinal image analysis. Currently existing deep learning algorithms have demonstrated performance similar to, or even better than, human experts at various classification tasks in DR (12,13). With significant advances in powerful deep learning algorithms, multiple companies have developed automated DR screening systems that have garnered the attention of the U.S. Food and Drug Administration (FDA), and to date, two AI-based screening algorithms have already been approved for use. As the FDA considers approval of additional automated machine learning algorithms, understanding their performance in real-world, intended-use settings is becoming increasingly important (14). In fact, the Center for Devices and Radiological Health (the FDA division responsible for regulating devices) prioritized the use of big data and real-world evidence for regulatory decision making in its 2019 regulatory science report, citing the need for validated methods of predicting device performance using real-world data (15). In line with this approach, we aimed to compare the performance of existing (either FDA approved or already in clinical use outside the U.S. and/or submitted for FDA approval) fully automated AI-based algorithms when screening for referable DR using real-world clinical data from two U.S. Veterans Affairs (VA) hospitals in geographically and demographically distinct cities. These algorithms were trained on unique, potentially limited data sets, and we hypothesized that their performance might decrease when tested with a large amount of real-world patient data. To our knowledge, this is the largest deep learning validation study to date.

RESEARCH DESIGN AND METHODS

Study Design

This was a multicenter, noninterventional device validation study that used images acquired from the VA Puget Sound Health Care System (HCS) and the Atlanta VA HCS. The institutional review board at the VA Puget Sound HCS approved the trial protocol. A waiver of informed consent was obtained for patient data used in the study. All participants had a diagnosis of DM and were not undergoing active eye care for any eye diseases and so were referred to the VA teleretinal DR screening program from 2006 to 2018. The overall study design is summarized in Supplementary Fig. 1.

Image Acquisition and Grading Process at the VA

At the VA, clinical photographs are stored in the Veterans Health Information System Technology Architecture (VISTA) Imaging system, and corresponding clinical data are deposited in the corporate data warehouse. At each encounter, at least four nonmydriatic or mydriatic color fundus photographs (at least two 45° images, one fovea centered, and at least two peripheral images) as well as an external color photograph for each eye were obtained using a Topcon TRC-NW8 fundus camera (Topcon Medical Systems, Tokyo, Japan). On average, nine photographs were available per encounter, including an average of 3.5 retinal images. All images were stored in JPG format and encapsulated using Digital Imaging and Communications in Medicine per standard VA VISTA processing for routine clinical care, with no perturbations or additional compression. The resolution of the images ranged from 4,000 × 3,000 to 4,288 × 2,848 pixels. The images were manually graded by VA-employed optometrists and ophthalmologists using the International Clinical Diabetic Retinopathy Severity Scale (ICDR) as follows: 0 = no DR, 1 = mild nonproliferative DR (NPDR), 2 = moderate NPDR, 3 = severe NPDR, 4 = proliferative DR (PDR), and 5 = ungradable image quality (16). At the VA, referable DR is defined as the presence of any DR (ICDR 1–4). The only difference in the imaging protocols between the two sites was the regular use of pharmacological pupillary dilation in all patients in Atlanta, which was not routinely performed at the Seattle site.

In this study, all images in the full data set were retrospectively acquired from each VA hospital's respective VISTA Imaging system (with the same image quality and format as available to teleretinal graders) and linked to clinical metadata from national and local VA databases, which include the original VA teleretinal grades for each image (17). None of the images had been used previously to train, validate, or test any automated diagnosis system that participated in this study. Other than removing all patient identifiers, no pre- or post-processing was applied to any image before analysis by the AI algorithms. There were no changes to the teleretinal DR screening clinical pathway. All images were available to the algorithms regardless of quality, including those that were identified as ungradable by the VA teleretinal graders. The presence of any DR was used as the threshold for referable DR per VA standards (18).

Arbitration Data Set Sampling and Grading

A subset of images was regraded using double-masked arbitration by clinical experts. Two random subsets of the full data set were created for regrading. First, a consecutive sampling of images was used, and second, a balanced set that included 50 images from each retinopathy level and ungradable class by the original VA teleretinal grade (obtained from the Seattle and Atlanta data sets evenly) was sampled. These two data sets were combined (to provide enough data to power the sensitivity analysis for the different disease thresholds described below) and presented to a board-certified comprehensive ophthalmologist and two fellowship-trained retina specialists who were masked to the original grades as well as to one another's classifications and who then graded the encounters independently. Differences in opinion were arbitrated by a retina specialist who had the two differing grades but did not know the identities of the graders to avoid confirmation bias. At no point during the arbitration process did any grader have access to the original VA teleretinal grades. The graders read the images on 22-in. 1080p monitors using the same viewing system as the VA teleretinal graders (certified Picture Archive and Communication System for the storage, viewing, and grading of medical images). The graders were allowed to manipulate the images, including changing the brightness, contrast, and zoom and generating a red-free version. This final arbitrated set of encounters was then used as the reference standard when comparing the performance of the algorithms to the VA teleretinal graders in screening for referable DR.

AI Algorithms

We invited 23 companies with automated AI-based DR screening systems to participate in this study: OphtAI, AEye, AirDoc, Cognizant, D-EYE, Diagnos, DreamUp Vision, Eyenuk, Google, IDX, Intelligent Retinal Imaging Systems, Medios Technologies, Microsoft Corporation, Remidio, Retina-AI Health, RetinAI Medical, RetinaLyze System, Retmarker, Singapore Eye Research Institute, SigTuple Technologies, Spect, VisionQuest Biomedical, and Xtend.AI. The details of the study were provided in a letter sent to each company, including the threshold for referable disease. Of the companies approached, five completed the study: OphtAI, AirDoc, Eyenuk, Retina-AI Health, and Retmarker. A total of seven algorithms were submitted for evaluation in this study. Each company sent its locked software preloaded on a workstation. Each system was masked to the original VA teleretinal grades and independently screened each image for referable DR defined as any degree of DR (ICDR grades 1–4) or unreadable encounters, without Internet connection. At the end of the study, each workstation was securely erased. As agreed upon before study initiation, the identity of each company was masked along with its submitted algorithms (labeled algorithms A–G). The study methods were provided to the participating companies upon request to give them the opportunity to adjust their software for the VA image acquisition protocol. These details included but were not limited to the camera system, image format, image resolution, aspect ratio, and number of photos per encounter. Each algorithm provided a binary classification output of each encounter as follows: 0 = does not need to be referred or 1 = should be referred for an in-person eye examination because of ungradable image quality or presence of any DR. In addition, all algorithms in the study either already had regulatory approval and/or were in active use in clinical settings around the world.

Statistical Analysis

To evaluate the algorithms, the screening performance of each was calculated using the original VA teleretinal grades from Seattle and Atlanta (combined and independently) as reference values. The screening performance measures included sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV). Then, a separate analysis was performed using the arbitrated set of encounters as the reference standard. The sensitivity and specificity of the original VA teleretinal grades and those of each algorithm were compared with the arbitrated data set to evaluate their relative performance using a paired exact binomial test (19). In addition, to measure the sensitivities of the algorithms for different levels of disease severity, their performance was compared with the VA teleretinal graders at different disease thresholds identified in the arbitrated data set, including moderate NPDR or worse, severe NPDR or worse, and PDR. Since each of the algorithms provided the binary output of no DR versus any DR, we calculated the sensitivity of the image subset for each disease severity without including the ungradable images.

Value-per-Encounter Analysis

Using the arbitrated data set, algorithms that performed no worse than the VA teleretinal graders in detecting referable DR in images marked as moderate NPDR or worse were selected to undergo a value-per-encounter analysis (20). The value per encounter for each individual algorithm was defined as the estimated pricing of each algorithm to make a normal profit (i.e., revenue and costs = 0) if deployed at the VA. This calculation was based on a two-stage scenario in which an AI algorithm would be used initially and then the images that screened negative would not need additional review by an optometrist or ophthalmologist. An average of 10 min per encounter was estimated as the amount of time needed to open, review, and write a report of the findings. The National Plan and Provider Enumeration System National Provider Identifier database and the FedsDataCenter database for fiscal year 2015 were used to determine the mean salary of providers (ophthalmologists or optometrists) per minute. The value per encounter was calculated as follows: value per encounter = [(10 min per encounter) × (mean salary of provider per min) × (encounter not referred)] / (total encounters). The primary outcome of our study was to evaluate the sensitivity and specificity of each algorithm compared with a human grader when determining whether the patient should be referred for an in-person ophthalmic examination on the basis of the DR screening images taken. Secondary end points included measuring the sensitivity of each algorithm against two random subsets of independently regraded images to permit estimation of the value per encounter that was due to lower reliance on expert human graders.

RESULTS

Patient Demographics and Image Characteristics

Patient demographic information and classification grades are summarized

Table 1—Demographic factors and baseline clinical characteristics of the study population

	Seattle	Atlanta	Total
Patients, *n*	13,439	10,285	23,724
Age (years)			
Mean (SD)	62.20 (10.91)	63.46 (10.14)	62.75 (10.60)
Range	21–97	24–98	21–98
Male sex	12,724 (94.68)	9,795 (95.24)	22,519 (94.92)
Race			
White	9,482 (70.56)	4,678 (45.48)	14,160 (59.69)
African American	1,642 (12.22)	5,085 (49.44)	6,727 (28.35)
Asian	383 (2.85)	34 (0.33)	417 (1.76)
Other	605 (4.50)	90 (0.88)	695 (2.93)
Unknown	1,327 (9.87)	398 (3.87)	1,725 (7.27)
Encounters, *n*	21,797	13,104	34,901
Retinopathy grade			
No DR	15,270 (70.05)	11,166 (85.21)	26,436 (75.75)
Mild NPDR	2,364 (10.85)	957 (7.31)	3,321 (9.51)
Moderate NPDR	494 (2.27)	311 (2.37)	805 (2.31)
Severe NPDR	110 (0.50)	153 (1.17)	263 (0.75)
PDR	22 (0.10)	193 (1.47)	215 (0.62)
Ungradable	3,537 (16.23)	324 (2.47)	3,861 (11.06)
Images, *n*	199,142	112,462	311,604

Data are *n* (%) unless otherwise indicated.

in Table 1. A total of 311,604 retinal images were acquired from 23,724 racially diverse patients. In the Seattle group, there was a higher prevalence of mild NPDR compared with severe NPDR. In contrast, a relatively higher number of patients presented to Atlanta with advanced stages of DR, including severe NPDR (1.17%) and PDR (1.47%), compared with Seattle (0.50% and 0.10%, respectively). More images from Seattle (16.23%) were of ungradable image quality compared with Atlanta (2.47%).

Automated DR Screening Performance

The sensitivity, specificity, NPV, and PPV for each AI screening system were calculated (using the original VA teleretinal grades as the reference standard) and are summarized in Fig. 1. In the full data set (Fig. 1*A*), sensitivity ranged from 50.98 to 85.90%, specificity from 60.42 to 83.69%, NPV from 82.72 to 93.69%, and PPV from 36.46 to 50.80%. Overall, the algorithms achieved higher NPVs using the Atlanta data set (range 90.71–98.05%) (Fig. 1*B*) compared with the Seattle data set (77.57–90.66%) (Fig. 1*C*). In contrast, PPV ranged from 24.80 to 39.07% in the Atlanta data set, which was lower than the Seattle data set (42.04–62.92%).

A subset of 7,379 images from 735 encounters was regraded for the arbitrated data set. Using the arbitrated grades as the new reference standard, the VA teleretinal graders achieved an overall sensitivity and specificity of 82.22% (95% CI 80.80%, 83.63%) and 84.36% (83.02%, 85.70%), respectively (Fig. 2*A*). When the algorithms were evaluated in this subset, algorithm G was the only one that did not perform significantly worse in terms of both sensitivity (80.47% [79.00%, 81.93%], $P = 0.441$) and specificity (81.28% [79.84%, 82.72%], $P = 0.195$) compared with the VA teleretinal graders. Algorithms E and F achieved higher sensitivities than the VA teleretinal graders (92.71% [91.75%, 93.67%], $P = 1.25 \times 10^{-6}$, and 92.71% [91.75%, 93.67%], $P = 7.29 \times 10^{-7}$, respectively) but were less specific. Algorithm A was the only one that achieved higher specificity (90.00% [88.89%, 91.11%], $P = 2.14 \times 10^{-2}$) than the VA teleretinal graders. In moderate and severe NPDR and PDR, the VA teleretinal graders achieved a sensitivity of 100% in gradable images. In moderate NPDR or worse (Fig. 2*B*), algorithms E, F, and G performed similarly to the VA teleretinal grader ($P = 0.500$, 0.500, and 1.000, respectively), whereas algorithms A–C had significantly lower sensitivities ($P < 0.03$). In severe NPDR or worse (Fig. 2*C*), only algorithms A and B performed worse than the VA teleretinal graders. With PDR, only algorithm A differed significantly from the VA teleretinal grader ($P = 9.77 \times 10^{-4}$) (Fig. 2*D*).

Value per Encounter

Only algorithms E, F, and G achieved comparable sensitivity to humans in detecting referable disease in encounters with moderate NPDR or worse; we report the value-per-encounter analysis of

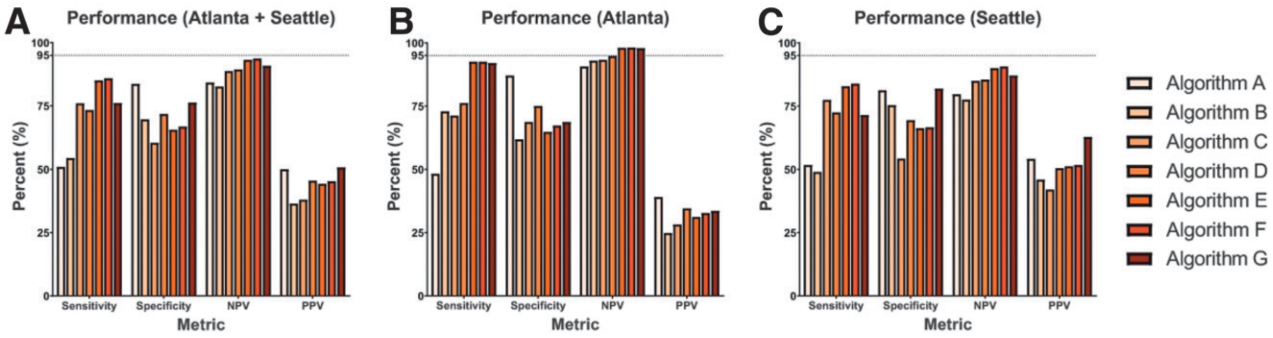

Figure 1—The relative screening performance of AI algorithms. Using the full-image data set (*A*), the sensitivity, specificity, NPV, and PPV of each algorithm are shown using the original teleretinal grader as the reference standard. These analyses were repeated separately using color fundus photographs obtained from Atlanta (*B*) and Seattle (*C*).

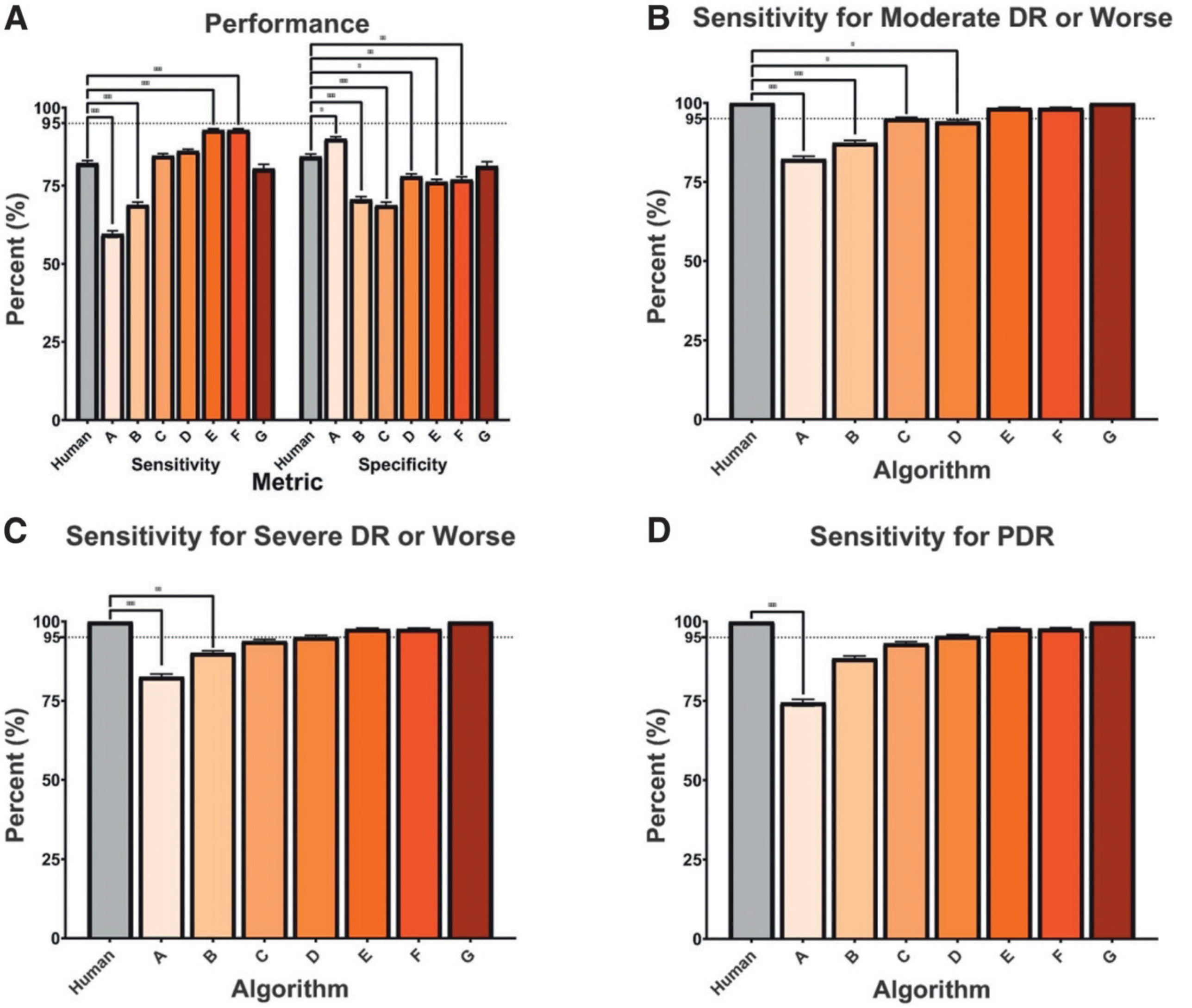

Figure 2—Relative performance of human grader compared with AI algorithms. The relative performance of the VA teleretinal grader (Human) and algorithms A–G in screening for referable DR using the arbitrated data set at different thresholds of DR. *A*: Sensitivity and specificity of each algorithm compared with a human grader with 95% CI bars against a subset of double-masked arbitrated grades in screening for referable DR in images with mild NPDR or worse and ungradable image quality. *B*–*D*: Only gradable images were used. The VA teleretinal grader is compared with the AI sensitivities, with 95% CIs, at different thresholds of disease, including moderate NPDR or worse (*B*), severe NDPR or worse (*C*), and PDR (*D*). **P* ≤ 0.05, ***P* ≤ 0.001, ****P* ≤ 0.0001.

these algorithms only (Fig. 3). The value per encounter of each algorithm was similar regardless of location for both ophthalmologists and optometrists. In the combined Atlanta and Seattle data set, the estimated value per encounter for ophthalmologists was $15.14 (95% CI $12.33, $17.95), $15.35 ($12.50, $18.20), and $18.06 ($14.71, $21.41) for algorithms E, F, and G, respectively. For optometrists, the approximate value of each respective algorithm on the combined data set was $7.74 ($6.43, $9.05), $7.85 ($6.52, $9.18), and $9.24 ($7.67, $10.80).

CONCLUSIONS

In this independent, external, head-to-head automated DR screening algorithm validation study, we found that the screening performance of state-of-the-art algorithms varied considerably, with substantial differences in overall performance, even though all the tested algorithms are currently being used clinically around the world and one has FDA approval. Using the arbitrated data set as the ground truth, the performance of the VA teleretinal graders was excellent, and no case of referable DR in images of moderate NPDR or worse was missed. In contrast, most of the algorithms performed worse, with only three of seven (42.86%) and one of seven (14.29%) of them having comparable sensitivity and specificity to the VA teleretinal graders, respectively. Only one algorithm (G) had similar performance to that of VA graders.

Overall, the algorithms had low PPVs compared with the human teleretinal grades, especially in the Atlanta data set. Both NPV and PPV should be considered when evaluating the performance of algorithms. For the purpose of screening, high NPV has foremost

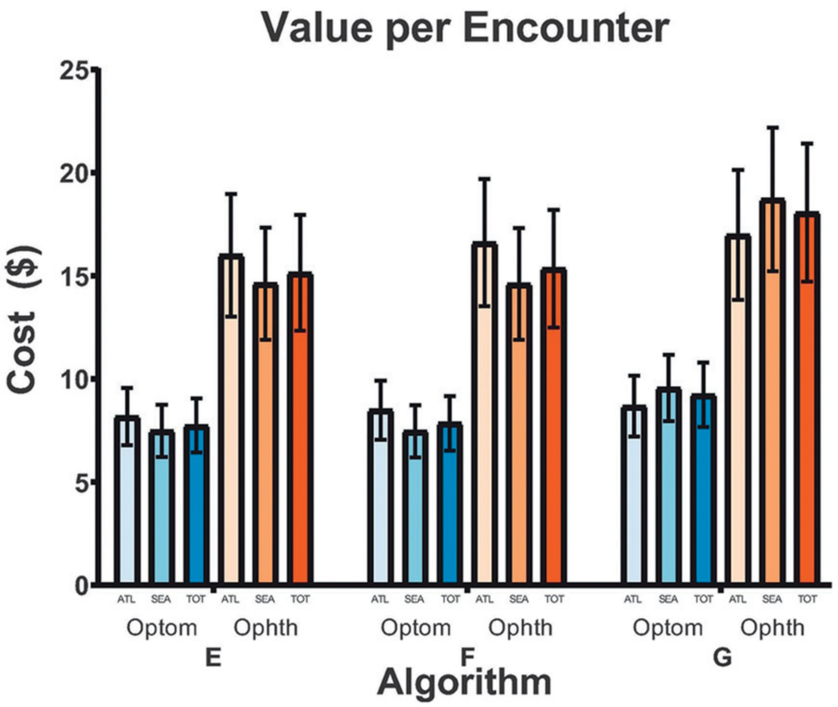

Figure 3—Value per encounter of AI algorithms meeting the sensitivity threshold. The value per encounter with 95% CI bars of algorithms E, F, and G. Only algorithms that achieved equivalent sensitivity to the VA teleretinal graders in screening for referable DR in images regraded as moderate NPDR or worse in the arbitrated data set were carried forward. The value per encounter of each algorithm if optometrists (Optom) or ophthalmologists (Ophth) were to implement this system into their clinical practice to make a normal profit on the basis of geographical location or the combined data set is shown. ATL, Atlanta; SEA, Seattle; TOT, total (Atlanta and Seattle).

importance to ensure that negative cases indeed do not have DR, while there should be a low threshold for in-person evaluation for unclear cases, possibly leading to low PPV. The different predictive value results in the two populations may also reflect differences in disease prevalence: 14.79% of the Atlanta population had DR compared with 29.95% of the Seattle population. The lower DR prevalence in Atlanta likely influenced the lower PPV on the basis of Bayes theorem, even though Atlanta had a higher rate of more severe DR.

The algorithms performed better overall on the Atlanta data set compared with Seattle, with fewer ungradable images, which is likely associated with the use of pharmacologic pupillary dilation (21). All patients were routinely dilated before screening in Atlanta (2.47% ungradable) but not in Seattle (16.23% ungradable). The majority of the participating algorithms are designed for nonmydriatic retinal imaging, but dilation requirements may vary between screening centers, and algorithms must be able to generalize. The difference in the number of ungradable images (fewer in the Atlanta data set) may also be due to the Atlanta VA's imaging protocol, which involves extensive technician training and retaking of poor-quality images. In addition, while the Seattle VA population was predominantly White (70.56%), nearly 50% of the Atlanta patients were African American. Different ethnic backgrounds may have affected the quality of fundus photos because the background retinal and choroidal pigmentation can vary substantially (22). The performance difference between Atlanta and Seattle is significant and highlights the potential lack of generalizability of some algorithms.

Several reasons may explain the discrepancy between our study results and previously reported findings (13,23–28). If studies use training data that are limited to a certain geographic and/or ethnic group, performance can decrease when the algorithm is tested in a different population (26). In addition, many studies process or remove lower-quality images from their analysis (13,23–25,28). Studies that exclude ungradable images and/or patients with comorbid eye disease (glaucoma, etc.) do not reflect the real-world data set where all images from all patients are analyzed, which can lower the performance of an algorithm (13,27). We made all images available to the algorithms, although some may analyze more images per encounter than others. Details of how most of these algorithms are trained and developed are not publicly available except for the two FDA-approved algorithms, which require two fundus images per eye.

The limited performance of most of the algorithms in our study emphasizes the need for external validation of screening algorithms before their clinical application. One of the seven algorithms in our study has FDA approval, four are in clinical use outside the U.S. and have been submitted to the FDA for approval, and several have a CE marking. Nevertheless, most algorithms performed similarly or even worse than the VA teleretinal graders. The two algorithms (E and F) that achieved superior sensitivities than the VA teleretinal graders had worse specificity for mild DR or worse and ungradable image quality. Additionally, none were better than the human graders in identifying referable disease when analyzed by DR severity. In fact, the performance of algorithm A was significantly worse than that of the VA teleretinal graders at all levels of DR severity. In this group of patients, algorithm A would miss 25.58% of advanced retinopathy cases, an error that can potentially result in severe vision loss. Because most of these algorithms are already in clinical use, these results are concerning. Implementation of such algorithms in a real-world clinical setting would pose a serious patient safety hazard (28).

An important question regarding the clinical implementation of these algorithms is estimating their economic value (29). As an initial screening tool, the appropriately selected algorithm could reduce the burden on human graders by eliminating images without retinopathy; fewer images/encounters requiring evaluation reduces costs. We only performed an economic analysis of the algorithms that did not perform worse than the human VA teleretinal grader (algorithms E, F, and G, which had higher or equivalent sensitivities compared with the teleretinal graders) in screening for referable DR in images regraded as moderate NPDR or worse because the performance of these three algorithms was closest to the current standard of care. In addition, although these models must achieve a level of sensitivity that is safe for clinical use, a model with high specificity translates to additional labor savings that could be interpreted as higher value per encounter. On the basis of the performance of the three best algorithms and the mean salary of eye care providers, we approximated the value of each DR screening encounter to range from $15.14 to $18.06 for a system with ophthalmologists as human graders and from $7.74 to $9.24 when optometrists are the graders. Thus, if there are 100,000 annual cases to be screened for DR, using an acceptable automated algorithm as the first step and relying on ophthalmologists to review only the ungradable and abnormal cases detected by the algorithm, the resulting annual labor savings would be $1,500,000–$1,800,000. Interestingly, we found that the value per encounter did not differ significantly between Seattle and Atlanta, despite the difference in PDR prevalence.

Several limitations exist in our study. First, although the patients were from geographically different sites and had varying ethnic backgrounds, they were predominantly older male (94.68%) veterans, and almost all of them had type 2 DM. These factors may have affected the performance of the algorithms, and additional validation in different populations will be important. Second, it is possible that in clinical practice, images may be graded by both an algorithm and a human grader in a semiautomated fashion. While this setup may improve sensitivity and specificity, the semiautomated system relies on the algorithm to first identify patients who require further screening by the physician; hence, it is important to evaluate its performance independently. Furthermore, each tested algorithm was designed to be fully automated, so we used an all-AI scenario to evaluate the performance and value of each algorithm. Third, the threshold for referable DR in the VA system does not distinguish between mild versus higher levels of DR, while the referral threshold in many health care systems is moderate DR or worse. The results of the sensitivity analysis for different thresholds of disease, in which several algorithms were equivalent to the human teleretinal graders, indicate that these algorithms would be applicable for health care settings that do not refer for mild DR. In addition, the presence of macular edema is a positive indicator of diabetic disease, but unlike human readers, the tested algorithms do not provide an output for the presence/absence of macular edema. Fourth, the results suggest that the human graders may have had lower sensitivity for mild NPDR compared with the algorithms given that a single microaneurysm or dot-and-blot hemorrhage would cross the threshold into mild NPDR. The use of double-masked, arbitrated expert human grades as the benchmark when comparing the algorithms' performance to the teleretinal graders may be considered as a limitation. However, current accepted reference standards by regulatory bodies, such as the FDA and AI literature, use expert human grading. We used double-masked, arbitrated regrading by experts as our reference standard, and our experts had no access to previous teleretinal grades, the same as with the AI algorithms.

Another limitation of our study was the relatively small number of companies that participated. We agreed to mask the identity of the algorithms to encourage participation, but of the 23 companies we approached, only 5 (21.74%) agreed to participate (providing seven algorithms for evaluation). Studies like ours that validate algorithms using real-world data sets will ultimately accelerate their subsequent performance but will need buy-in from all companies. New reporting guidelines recommend increasing transparency about how AI devices are trained and evaluated, including plans for anticipating and mitigating risks upon implementation (14,30,31). With greater openness and participation, progress in AI efficacy, safety, and science will advance the field, inspire innovation, and benefit the global population.

The value-per-encounter analysis in our study did not factor in fringe benefits and indirect costs for optometrists and ophthalmologists, and costs of graphics processing unit servers were not included. Additionally, the estimated value per encounter is specific to the VA. Although these automated systems are intended for use with the teleretinal screening system, we did not estimate the cost of adding automated grading to the existing VA teleretinal system, which would include the costs associated with integration and any additional computing hardware needs. Many of the AI companies offer a cloud-based solution so that the latter is less of a concern.

To our knowledge, this is the largest AI-based DR screening algorithm validation study to date, modeling real-world conditions by analyzing 311,604 color fundus photographs from two geographically diverse populations regardless of quality and without any preprocessing or filtering. Our study was powered to evaluate the presence of referable disease in images with undiagnosed PDR. Unlike other studies in which too few severe cases can lead to oversampling of mild disease, our large database did not require balancing since it covers >10 years of clinical data (16). Thus, unlike many studies reported previously (12,13,32), we were able to assess both PPV and NPV and acquire insights into the relative performance of the algorithms in regions with different prevalence rates.

Although some algorithms in our study performed well from a screening perspective, others would pose safety concerns if implemented within the VA. These results demonstrate that automated devices should undergo prospective, interventional trials to evaluate their efficacy as they are integrated into clinical practice, even after FDA approval. Ideally, validation data sets should include real-world data sets representative of where the algorithms will be deployed so that they function well regardless of variables such as race, image quality, dilation practices, and coexisting disease. Automated screening systems are not limited to DR and may be applicable for other conditions, such as age-related macular degeneration and glaucoma, where earlier detection would likely improve clinical outcome. Rigorous pre- and postapproval testing of all such algorithms is needed to sufficiently identify and understand the algorithms' characteristics to determine suitability for clinical implementation.

Funding. A.Y.L. is supported by the FDA. This material is the result of work supported with resources and the use of facilities at VA Puget Sound and the Veterans Health Administration Innovation Ecosystem. This study was supported by National Institutes of Health grants K23-EY-029246 and R01-AG-060942 and an unrestricted grant from Research to Prevent Blindness.

The sponsors/funding organizations had no role in the design or conduct of this research. The contents do not represent the views of the U.S. Department of Veterans Affairs, the FDA, or the United States Government.

Duality of Interest. A.Y.L. reports grants from Santen, Carl Zeiss Meditec, and Novartis and personal fees from Genentech, Topcon, and Verana Health outside the submitted work. A.Y.M. reports consulting fees from Warby Parker and Click Therapeutics outside of the submitted work. E.J.B. reports personal fees from Bayer AG. No other potential conflicts of interest relevant to this article were reported.

Author Contributions. A.Y.L., C.S.L., and E.J.B. conceived and designed the study. A.Y.L. developed and validated the algorithms with C.S.L., Y.E.C., and E.J.B. providing clinical input. A.Y.L., H.C.J., M.D.G., A.Y.M., and M.L. collected and prepared the data. A.Y.L., R.T.Y., C.S.L., and M.B. wrote the initial draft. A.Y.L., R.T.Y., C.S.L., M.B., H.C.J., Y.E.C., M.D.G., H.G., A.Y.M., G.C.C., M.L., and E.J.B. subsequently critically edited the report. All authors read and approved the final report. A.Y.L. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Prior Presentation. Parts of this study were presented in abstract form at the 38th Annual Scientific Meeting of the American Society of Retina Specialists, Virtual, 24–26 July 2020.

References

1. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. Eye Vis (Lond) 2015;2:17

2. Liew G, Michaelides M, Bunce C. A comparison of the causes of blindness certifications in England and Wales in working age adults (16-64 years), 1999-2000 with 2009-2010. BMJ Open 2014;4:e004015

3. Jampol LM, Glassman AR, Sun J. Evaluation and care of patients with diabetic retinopathy. N Engl J Med 2020;382:1629–1637

4. Flaxel CJ, Adelman RA, Bailey ST, et al. Diabetic retinopathy preferred practice pattern®. Ophthalmology 2020;127:66–P145

5. American Diabetes Association. 11. Microvascular complications and foot care: *Standards of Medical Care in Diabetes—2020*. Diabetes Care 2020;43(Suppl. 1):S135–S151

6. International Diabetes Federation. Diabetes Facts & Figures. Accessed 26 April 2020. Available from https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

7. Resnikoff S, Felch W, Gauthier T-M, Spivey B. The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200,000 practitioners. Br J Ophthalmol 2012;96:783–787

8. Kirkizlar E, Serban N, Sisson JA, Swann JL, Barnes CS, Williams MD. Evaluation of telemedicine for screening of diabetic retinopathy in the Veterans Health Administration. Ophthalmology 2013;120:2604–2610

9. Joseph S, Kim R, Ravindran RD, Fletcher AE, Ravilla TD. Effectiveness of teleretinal imaging-based hospital referral compared with universal referral in identifying diabetic retinopathy: a cluster randomized clinical trial. JAMA Ophthalmol 2019;137:786–792

10. Tufail A, Kapetanakis VV, Salas-Vega S, et al. An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness. Health Technol Assess 2016;20:1–72

11. Tufail A, Rudisill C, Egan C, et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. Ophthalmology 2017;124:343–351

12. Ting DSW, Cheung CY-L, 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;318:2211–2223

13. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39

14. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 2018;1:40

15. Center for Devices and Radiological Health. CDRH Regulatory Science Priorities. U.S. Food and Drug Administration, 2019. Accessed 23 July 2020. Available from https://www.fda.gov/medical-devices/science-and-research-medical-devices/cdrh-regulatory-science-priorities

16. Ogunyemi O, Kermah D. Machine learning approaches for detecting diabetic retinopathy from clinical and public health records. AMIA Annu Symp Proc 2015;2015:983–990

17. Kuzmak P, Demosthenes C, Maa A. Exporting diabetic retinopathy images from VA VISTA Imaging for research. J Digit Imaging 2019;32:832–840

18. Conlin PR, Fisch BM, Orcutt JC, Hetrick BJ, Darkins AW. Framework for a national teleretinal imaging program to screen for diabetic retinopathy in Veterans Health Administration patients. J Rehabil Res Dev 2006;43:741–748

19. Stock C, Hielscher T. DTComPair: Comparison of Binary Diagnostic Tests in a Paired Study Design, 2014. Accessed 20 April 2020. Available from https://rdrr.io/cran/DTComPair/man/dtcompair-package.html

20. 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:e240–e249

21. Wintergerst MWM, Brinkmann CK, Holz FG, Finger RP. Undilated versus dilated monoscopic smartphone-based fundus photography for optic nerve head evaluation. Sci Rep 2018;8:10228

22. Silvar SD, Pollack RH. Racial differences in pigmentation of the fundus oculi. Psychon Sci 1967;7:159–159

23. Abràmoff MD, Lou Y, 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;57:5200–5206

24. 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

25. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 2017;124:962–969

26. Romero-Aroca P, Verges-Puig R, de la Torre J, et al. Validation of a deep learning algorithm for diabetic retinopathy. Telemed J E Health 2020;26:1001–1009

27. Bhaskaranand M, Ramachandra C, Bhat S, et al. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. Diabetes Technol Ther 2019;21:635–643

28. Natarajan S, Jain A, Krishnan R, Rogye A, Sivaprasad S. Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone. JAMA Ophthalmol 2019;137:1182–1188

29. Xie Y, Gunasekeran DV, Balaskas K, et al. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. Transl Vis Sci Technol 2020;9:22

30. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351–1363

31. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020;370:m3164

32. Ranganathan P, Aggarwal R. Common pitfalls in statistical analysis: understanding the properties of diagnostic tests - part 1. Perspect Clin Res 2018;9:40–43