tvst

https://doi.org/10.1167/tvst.8.6.40

Article

Remote Tool-Based Adjudication for Grading Diabetic Retinopathy

Mike Schaekermann[1,2], Naama Hammel[1], Michael Terry[1], Tayyeba K. Ali[3], Yun Liu[1], Brian Basham[1], Bilson Campana[1], William Chen[1], Xiang Ji[1], Jonathan Krause[1], Greg S. Corrado[1], Lily Peng[1], Dale R. Webster[1], Edith Law[2], and Rory Sayres[1]

[1] Google AI Healthcare, Google LLC, Mountain View, CA, USA
[2] University of Waterloo, Waterloo, ON, Canada
[3] Work done at Google Health via Advanced Clinical (corporate headquarters: Deerfield, IL, USA)

Correspondence: Rory Sayres, Google AI Healthcare, 3400 Hillview Ave, Palo Alto, CA 94304, USA. e-mail: sayres@google.com

Received: 18 April 2019
Accepted: 12 October 2019
Published: 18 December 2019

Keywords: retinal imaging; adjudication; diabetic retinopathy; teleophthalmology

Citation: Schaekermann M, Hammel N, Terry M, Ali TK, Liu Y, Basham B, Campana B, Chen W, Ji X, Krause J, Corrado GS, Peng L, Webster DR, Law E, Sayres R. Remote tool-based adjudication for grading diabetic retinopathy. Trans Vis Sci Tech. 2019; 8(6):40, https://doi.org/10.1167/tvst.8.6.40

Purpose: To present and evaluate a remote, tool-based system and structured grading rubric for adjudicating image-based diabetic retinopathy (DR) grades.

Methods: We compared three different procedures for adjudicating DR severity assessments among retina specialist panels, including (1) in-person adjudication based on a previously described procedure (Baseline), (2) remote, tool-based adjudication for assessing DR severity alone (TA), and (3) remote, tool-based adjudication using a feature-based rubric (TA-F). We developed a system allowing graders to review images remotely and asynchronously. For both TA and TA-F approaches, images with disagreement were reviewed by all graders in a round-robin fashion until disagreements were resolved. Five panels of three retina specialists each adjudicated a set of 499 retinal fundus images (1 panel using Baseline, 2 using TA, and 2 using TA-F adjudication). Reliability was measured as grade agreement among the panels using Cohen's quadratically weighted kappa. Efficiency was measured as the number of rounds needed to reach a consensus for tool-based adjudication.

Results: The grades from remote, tool-based adjudication showed high agreement with the Baseline procedure, with Cohen's kappa scores of 0.948 and 0.943 for the two TA panels, and 0.921 and 0.963 for the two TA-F panels. Cases adjudicated using TA-F were resolved in fewer rounds compared with TA ($P < 0.001$; standard permutation test).

Conclusions: Remote, tool-based adjudication presents a flexible and reliable alternative to in-person adjudication for DR diagnosis. Feature-based rubrics can help accelerate consensus for tool-based adjudication of DR without compromising label quality.

Translational Relevance: This approach can generate reference standards to validate automated methods, and resolve ambiguous diagnoses by integrating into existing telemedical workflows.

Introduction

Diabetic retinopathy (DR) is one of the leading causes of vision loss worldwide.[1] The process of grading DR severity involves the examination of the retina and the assessment of several features, such as microaneurysms (MAs), intraretinal hemorrhages, and neovascularization.[2] In a teleophthalmology setting for remote screening, certified graders examine retinal fundus images to determine the presence and severity of disease as it appears in a two-dimensional (2D) photograph.[3] Prior work has shown that this process of human interpretation is subject to individual grader bias, as demonstrated by high intergrader variability, with kappa scores ranging from 0.40 to 0.65.[4–9]

This moderate-to-poor agreement between graders has led to difficulties in reliable evaluation of both

individual graders as well as assistive technologies. Yet, due to limited access to skilled healthcare providers, there continues to be a surge in interest in the development of assistive technologies, such as deep-learning systems, resulting in a sharp increase in the demand for high-quality reference standards of labeled-image data.[10–19] Prior work has examined different methods for resolving disagreements among experienced graders when creating a reference standard,[11] including majority vote, arbitration of disagreements by a more senior grader, and in-person adjudication among expert panels.

In ophthalmology, a recognized method to obtain a reliable reference standard is expert adjudication of images.[11,20,21] Multiple experienced doctors independently grade images and discuss disagreements until resolved. Such "in-person" adjudication has been shown to produce higher-quality labels[11] but can be challenging to schedule: it requires coordination of multiple, highly experienced specialists for in-person sessions, and even small image sets on the order of a few thousand cases can take months to adjudicate due to clinical scheduling conflicts.

In this study, we presented and evaluated a tool-based system for adjudicating images that was suitable for remote grading and removes the need for in-person sessions. Our system allowed doctors to discuss and resolve disagreements on diagnoses remotely, without convening at a set time and place. The practices described in this paper aimed to increase the efficiency and flexibility of adjudication, while maintaining the quality of the labels produced. We evaluated our system in the context of DR severity grading based on retinal fundus images.

In addition, we proposed an adjudication system with the ability to impose an explicit structure on the adjudication process by organizing the process of image interpretation around a set of discrete, detailed evaluation criteria. We investigated the effects of such a structure on the efficiency and reliability of adjudication for DR grading. Specifically, we presented a feature-based rubric for adjudication of DR severity grades, in which graders assess individual features (MAs, hemorrhages, neovascularization, etc.) in addition to overall DR severity.

Taken together, these improvements allow high-quality reference standards to be obtained by the community, and have the further benefit of offering flexibility for individual graders to schedule their reviewing activity around their clinical duties.

Table 1. Baseline Characteristics

Characteristic	Value
Number of images	499
Number of images for which an anonymized patient code was available[a]	330
Number of unique individuals out of the images for which a patient code was available	307
DR gradeability distribution according to Baseline adjudication	
Images gradable for DR, *n*/total (%)	472/499 (94.6)
DR severity distribution according to Baseline adjudication, *n* (%)	
No apparent DR	217 (45.9)
Mild NPDR	17 (3.6)
Moderate NPDR	108 (22.9)
Severe NPDR	72 (15.3)
PDR	58 (12.3)

PDR, proliferative diabetic retinopathy.

[a] Patient codes were available for images from two hospitals (Sankara Nethralaya and Narayana Nethralaya) of three.

Methods

Experimental Design

The experiment conducted for this study compared three different adjudication procedures for assessing DR severity based on retinal fundus images as follows: in-person adjudication (Baseline); remote, tool-based adjudication (TA) for assessing DR severity alone; and remote, tool-based adjudication using a feature-based rubric to assess DR severity (TA-F). The experiment implemented a between-subjects design in which independent panels of three retina specialists each graded and adjudicated the same set of images following one of three adjudication procedures (Baseline, TA, TA-F). We describe the image set, each of the three adjudication procedures, and details about the retina specialist graders below. For each design, graders were primarily assessing DR severity, but not diabetic macular edema (DME).

Image Set

We used a subset of 499 images (Table 1) from the development dataset used by Krause et al.[11] The

image set consisted of central field-of-view images obtained from patients who presented for DR screening at three eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya). The image set was sampled to include approximately 50% of cases that had some level of DR.[11] Anonymized patient codes were provided from two of three hospitals, allowing us to verify no patient duplication. For the third hospital, patient codes were not provided; this allows for the possibility that 169 images from this hospital may contain multiple images from the same patient; given that these were sampled from a much larger set of images, duplication is unlikely. Image sizes ranged from 640×480 to 2588×3388 pixels, and were presented to adjudicators at the original resolutions. All images were de-identified according to the Health Insurance Portability and Accountability Act Safe Harbor before transfer to study investigators. Ethics review and institutional review board exemption were obtained through the Quorum Review institutional review board (Seattle, WA).

Adjudication Procedures

Baseline Adjudication

Following the practices described in Krause et al.,[11] our Baseline adjudication procedure consisted of the following three stages: (1) an initial independent evaluation; (2) remote review of disagreements; and (3) in-person discussion and final resolution of remaining cases.

For the first stage, three fellowship-trained retina specialists undertook independent grading of the image set. Images in which the independent grades agreed were considered resolved. Next, each of the three retina specialists independently reviewed one-third of the remaining images with any level of disagreement. This independent review procedure was facilitated through the use of online spreadsheets. Cases that remained unresolved after the independent review round were discussed by all three retina specialists in person. During the in-person sessions, all three retina specialists were present at a set time and place, and conflicting grades were reviewed and adjudicated within the panel until all specialists came to an agreement. The time from start of independent grading to full adjudication for the image set was around 3 months. While the total time each grader spent on grading and adjudication activities was not tracked precisely, a substantial portion of the 3-month period was due to difficulties in scheduling the retina specialists to physically convene for in-person discussions.

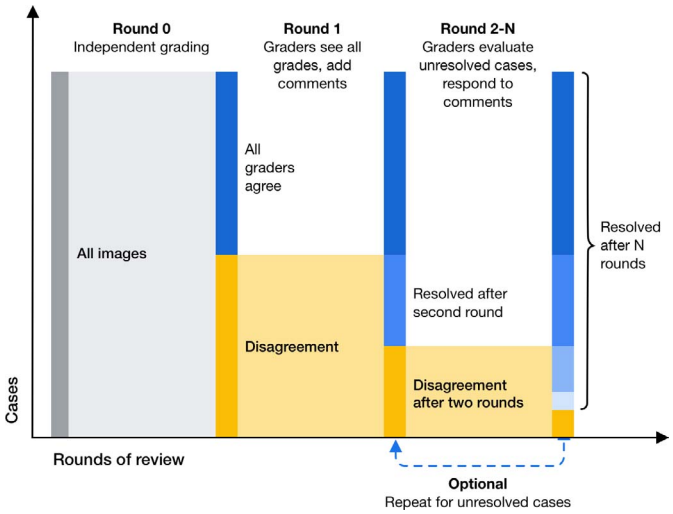

Figure 1. Process diagram illustrating remote TA; images are first graded independently by each panel member (round 0); cases with any level of disagreement after independent grading are reviewed by all graders in a round-robin fashion (rounds 1–N); the procedure ends after N review rounds.

Tool-Based Adjudication (TA)

To ensure the continuity of the adjudication process and to reduce the logistic overhead associated with in-person adjudication, we designed and implemented a tool-based system for remote adjudication that removes the need for in-person sessions (Fig. 1). Similar to the Baseline procedure, the TA procedure commences with independent grading: each panel member first assesses each image for DR severity. Next, those images with any level of disagreement are reviewed by one panel member at a time in a round-robin fashion until agreement is reached for the given case (Fig. 2). For each review round, the active grader reviews all grades and comments provided in previous rounds, re-grades the given image for DR severity, and provides more detailed comments, or replies to other graders' comments. To handle cases with persistent disagreement, the TA procedure imposes a limit on the number of review rounds for each case. In our studies, each case was limited to a maximum of 15 review rounds (i.e., 5 reviews per grader for a panel of 3 graders). See Table 2 for a comparison of the Baseline and TA adjudication procedures.

Tool-Based Adjudication With Feature Rubric (TA-F)

Disagreements over DR severity can arise for various reasons (e.g., due to divergent assessments of

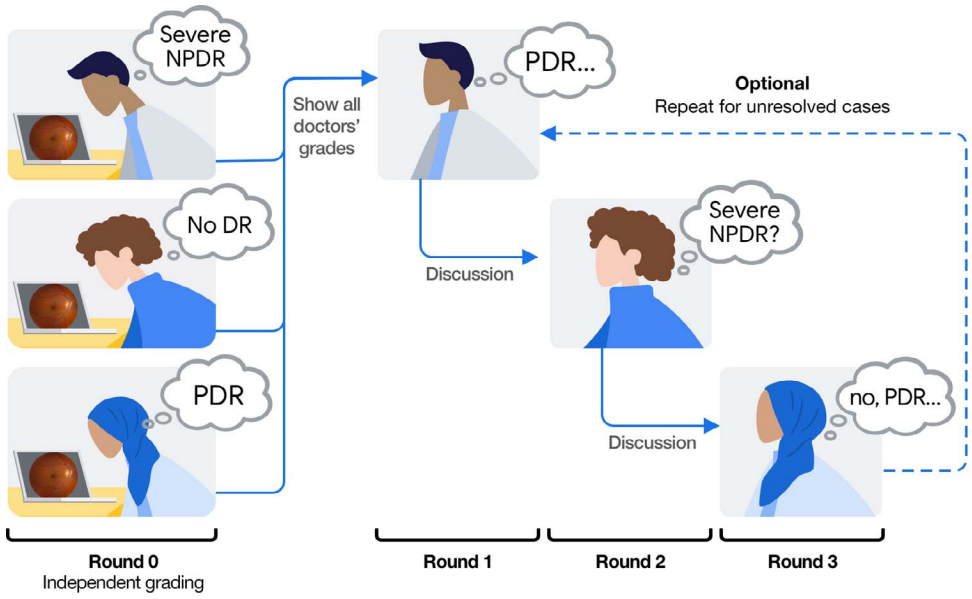

Figure 2. Illustration of the round-robin approach for remote TA in the context of DR severity grading.

the presence and extent of individual features or due to divergent interpretations of whether a retinal pathology is diabetic in nature or not). One benefit of the tool-based adjudication procedure proposed in this work is the ability to impose an explicit structure to the adjudication process by introducing prompts for individual, detailed evaluation criteria. This ability can be leveraged to remind graders of the specific criteria they should apply to assess an image (e.g., from standardized grading guidelines) so that discussions over potential disagreements are grounded in predefined factors relevant to the overall diagnostic decision.

In our experiment, we developed a feature-based rubric in which graders were first prompted to assess each image for a set of DR-related features before assessing the image for overall DR severity. Following the International Clinical Diabetic Retinopathy (ICDR) disease severity scale,[2] we included the following 10 features in the TA-F procedure: MAs, cotton-wool spots, hard exudates, retinal hemorrhage (heme), venous beading (VB), intraretinal microvascular abnormalities (IRMA), neovascularization or fibrous proliferation, preretinal or vitreous hemorrhage, laser scars from panretinal photocoagulation, and laser scars from focal photocoagulation. Graders assessed whether each feature was present, not present, or ungradable. For heme, graders also assessed whether any retinal hemorrhage was extensive in four quadrants, based on standard photo 2A from the Early Treatment Diabetic Retinopathy Study (ETDRS).[22] For VB, graders also assessed whether definite venous beading, if present, was observed in two or more quadrants, based on ETDRS standard photo 6A.[22] Similarly, for IRMA, graders assessed whether any IRMA was prominent, based on ETDRS standard photo 8A.[22] Intergrader disagreement may not only arise over the presence or severity of disease, but also over the specific classification and etiology of an observed pathology. In particular, the appearance of DR may resemble other forms of retinal disease, such as hypertensive retinopathy (HTN), retinal vein occlusion (RVO), and retinal artery occlusion (RAO).[23,24] Graders were therefore prompted to assess for the presence of HTN, RVO, and RAO in addition to providing a DR severity assessment. In the adjudication interface (Fig. 3), disagreements were visualized for both feature- and diagnosis-level decisions to inform adjudicators about assessments from other panel members. Full agreement within a panel was only required regarding the overall gradeability of an image as well as for the diagnosis decisions (DR, HTN, RVO, RAO) in order to resolve a case; cases could be resolved despite disagreements on individual features. See Supplementary Table S1 for a detailed list of the questions and answer options used in the TA and TA-F procedures.

Graders

We recruited 14 American Board of Ophthalmology–certified fellowship-trained retina specialists to

Table 2. Comparison of Adjudication Procedures

Property	Adjudication Procedure	
	Baseline	Tool-Based (TA and TA-F)
Image viewer	Web-based image viewer with built-in tools to adjust zoom level and contrast settings; graders submitted their independent assessments using prompts embedded into the image viewer	
Aggregation of grades and identification of disagreements	Exporting results into spreadsheet to manually identify disagreements	Automated process to identify images with disagreement in the grades database
First review round	Remotely in spreadsheet	Remotely, using the web-based image viewer; one grader at a time in a round-robin fashion
Subsequent review rounds	In-person session; all panel members convene at a set time	
Channel for discussion	In-person verbal discussion	Discussion thread integrated into the image viewer; up to one written comment per grader per review round
Scheduling of review rounds	Manual process	No manual scheduling required; grading and review tasks automatically queue up for individual graders in the online platform
Anonymization of graders	Possible only in the first review round, but not during live discussion	Possible throughout the entire procedure
Organization of the disagreement discussion around a set of explicit diagnostic criteria (e.g., lesions)	Challenging to implement during live discussion	Possible using prompt structure integrated into the image viewer

form five adjudication panels, including one panel for the Baseline procedure, two panels for the TA procedure (Panels A and B), and two panels for the TA-F procedure (Panels C and D). Due to the limited availability of retina specialists, one of 14 graders participated in two panels (Baseline and TA Panel B); otherwise, each grader participated in one panel only. Participating retina specialist graders completed their fellowship training between the years 2009 and 2017 and the number of years in practice (post fellowship) at the time of participation in the study ranged from 0.5 to 8.5 years.

Evaluating Tool-Based Adjudication

We evaluated the tool-based adjudication procedures (TA, TA-F) for reliability and efficiency. Reliability was assessed in terms of agreement with the Baseline adjudication procedure, using Cohen's quadratically weighted kappa score.[25] A nonparametric bootstrap procedure[26] with 2000 samples was used to compute confidence intervals (CIs) for the kappa scores. The weighting function for the calculation of kappa scores was the square of the stepwise distance between DR grades on a five-point ordinal scale (e.g., a disagreement between no DR and severe nonproliferative diabetic retinopathy [NPDR], which are three steps apart on the ICDR scale, would receive a weight of $3^2 = 9$ when calculating kappa; larger disagreements would more strongly reduce this metric). Images unanimously deemed ungradable and those with persistent disagreement after 15 review rounds in any of the panels were excluded from kappa score

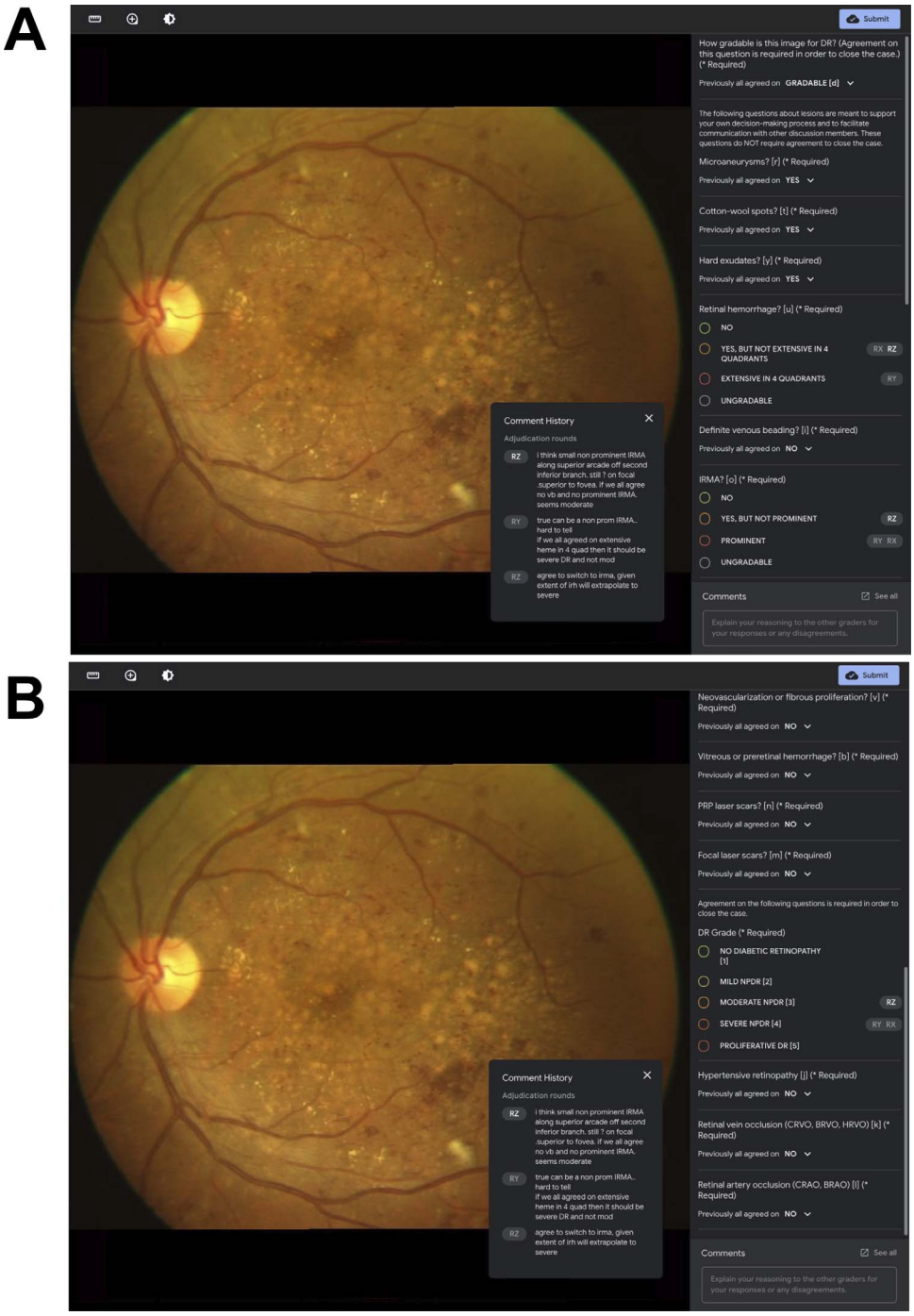

Figure 3. Grading interface for remote TA-F for DR severity assessment. Grader pseudonyms (RX, RY, RZ) are used to associate grading decisions and discussion comments from previous rounds with specific (anonymized) grader identities. The current grader's pseudonym is highlighted with *bold white font* (see RZ). The panel on the *right-hand side* lists all prompts included in the TA-F procedure and allows for vertical scrolling between the *top half* (A) and the *bottom half* (B).

calculations. Exact agreement rates and strikeout rates (i.e., the fraction of images for which grades differed by >2 steps) were calculated as additional measures of agreement for each panel pair.

The efficiency of TA versus TA-F was evaluated using the number of review rounds required to resolve each case in a given panel, and using the cumulative percentage of cases resolved in each round, including independent grading (round 0) and the subsequent review rounds (rounds 1–15). We used the standard

Table 3. Interpanel Agreement Between all Adjudication Panels

	TA		TA-F	
Parameter	Panel A	Panel B	Panel C	Panel D
Baseline	0.948 (0.931–0.964)	0.943 (0.919–0.962)	0.921 (0.886–0.948)	0.963 (0.949–0.975)
TA				
Panel A	/	0.932 (0.911–0.950)	0.917 (0.885–0.944)	0.939 (0.916–0.960)
Panel B	/	/	0.911 (0.873–0.942)	0.936 (0.914–0.953)
TA-F				
Panel C	/	/	/	0.919 (0.882–0.949)

Values are quadratically weighted Cohen's Kappa (95%CI).

permutation test to assess the statistical significance of these differences.[26] Due to software-related irregularities, in which the full-adjudication discussions were not recorded, 11 images (2%) were excluded from the analysis. Results are based on the remaining 488 cases. For TA-F specifically, the relative efficiency of resolving disagreements on each of the rubric criteria was assessed as the number of review rounds required to reach agreement on a given criterion, or as the round number in which a case was closed despite disagreement on the criterion.

Results

Reliability

Remote TA grades showed high agreement with the Baseline adjudication procedure (Table 3, Supplementary Tables S2–S5), with Cohen's kappa scores of 0.943 (95%CI, 0.919–0.962) and 0.948 (95%CI, 0.931–0.964) for the two panels assessing DR severity alone without the use of a feature rubric (TA), and 0.921 (95%CI, 0.886–0.948) and 0.963 (95%CI, 0.949–0.975) for the two panels using the feature-based rubric (TA-F). Both TA and TA-F)showed high rates of reproducibility, as measured by the Cohen's kappa score between the two independent panels for each procedure. The kappa score for agreement was at 0.932 (95%CI, 0.911–0.950; Supplementary Table S6) between the two panels in the TA procedure and at 0.919 (95%CI, 0.882–0.949; Supplementary Table S7) for TA-F. Exact agreement rates (Table 4) and strikeout rates (Table 5) are reported as additional measures of agreement for each pair of panels.

Efficiency

Cases adjudicated using TA-F were resolved in significantly fewer rounds compared with assessing DR severity without the rubric (TA; $P < 0.001$; permutation test, Fig. 4). During independent grading (round 0), graders were in agreement for 72% of all cases using TA-F, compared with 67% TA, and to 58% in Baseline in-person adjudication. Using TA-F, only 3% of the cases required more than one full "round-robin" of reviews from the panel (round 3), compared with 9% of the cases in the absence of the feature-based rubric (Fig. 5). Both differences were statistically significant under a permutation test of two panels for TA versus two panels for TA-F ($P = 0.004$ for round 0, $P < 0.001$ for round 3). The only two cases with persistent disagreement after 15 rounds

Table 4. Interpanel Agreement (Exact Agreement Rate) Between All Adjudication Panels

	TA		TA-F	
Parameter	Panel A	Panel B	Panel C	Panel D
Baseline	0.820	0.828	0.789	0.857
TA				
Panel A	/	0.811	0.811	0.852
Panel B	/	/	0.822	0.816
TA-F				
Panel C	/	/	/	0.820

Table 5. Interpanel Agreement (Strikeout Rate) Between All Adjudication Panels

	TA		TA-F	
Parameter	Panel A	Panel B	Panel C	Panel D
Baseline	0.026	0.026	0.027	0.017
TA				
Panel A	/	0.039	0.041	0.038
Panel B	/	/	0.042	0.034
TA-F				
Panel C	/	/	/	0.033

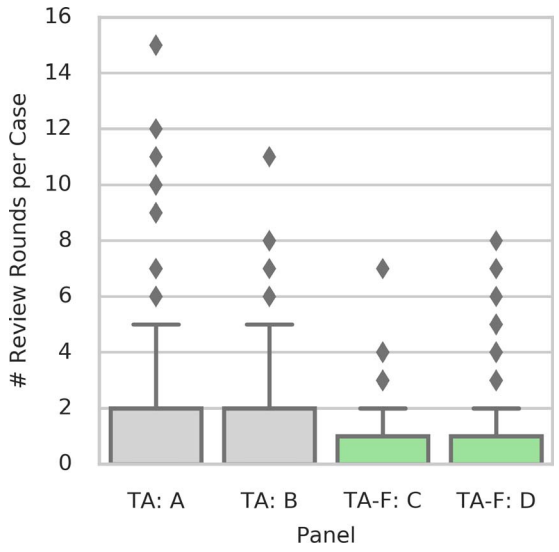

Figure 4. Number of review rounds required per case (i.e., number of rounds until agreement or 15 in case of persistent disagreement) for each of the four adjudication panels.

of review were observed in the TA procedure (Panel A). See Supplementary Table S8 for details about graders' DR assessments and discussion comments in each review round for both cases. For comparison, we also provide discussions of two example cases in Panel A that were resolved after three rounds of review (Supplementary Table S9). Overall, cases assessed as mild NPDR or severe NPDR in the Baseline adjudication procedure showed the lowest rates of independent agreement (i.e., before adjudication), with agreement rates of 31.7% and 44.6%, respectively (Supplementary Table S10). Mild NPDR and severe NPDR were also the only two categories with any persistent disagreement (1 case each), and with the highest proportion of cases requiring more than two rounds of review for at least one of three graders (3.3% and 1.8%, respectively) in order to reach a consensus.

Among the 10 feature criteria included in the TA-F rubric, assessments of the presence and extent of heme, VB, and IRMA required the greatest number of review rounds on average (Fig. 6). As for the differential diagnosis section of the TA-F rubric, assessment of HTN required more review rounds on average than assessments of RVO and RAO.

Finally, each of the four panels conducting tool-based adjudication completed all 499 images within 58 days from initial grading to full adjudication, with the fastest panel completing in 19 days. Note that these durations also include intervals of idle time in which the system waited for graders to complete their review passes, and that graders performed other labeling tasks during their own idle intervals. The total amount of time spent on grading and reviewing activities is therefore substantially lower than the corresponding end-to-end durations per panel (Supplementary Table 11).

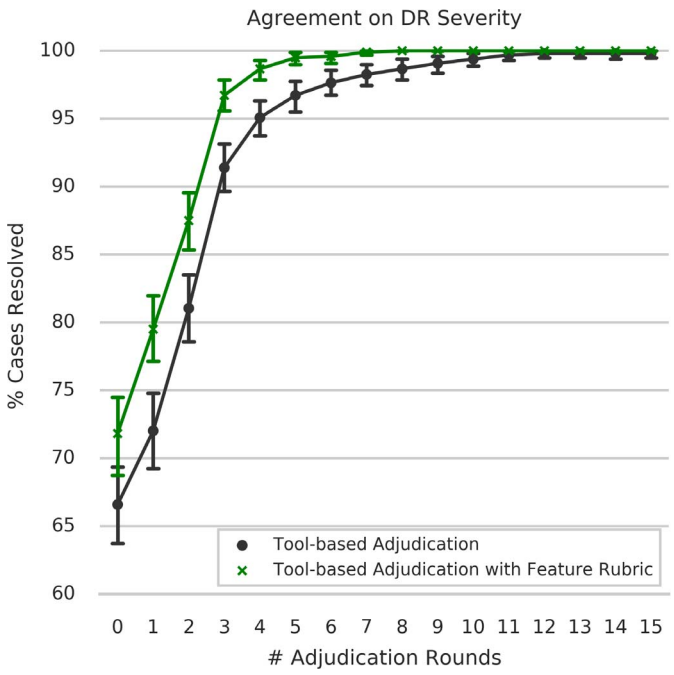

Figure 5. Cumulative percentage of cases resolved per adjudication round for TA procedures.

Discussion

As machine-learning methods become more common in ophthalmology, the need to accurately assess diagnostic performance grows. Algorithms that may be used to automate or augment aspects of eye care should be subjected to rigorous evaluation of their performance, against trusted reference standards. This in turn motivates the development of high-quality reference standards, a process that has received relatively little attention in the literature.

Previous studies suggest that adjudication can not only reliably be used to evaluate DR severity, but should be the reference standard used in deep-learning algorithms.[11] While several methods may be used for this process, such as in-person adjudication among expert panels and arbitration of disagreements by a senior grader (Domalpally A, et al. *IOVS* 2018;59:ARVO E-Abstract 4676) these methods rely on the time and expertise of certain physicians. In the present study, we present a tool-based system for remote expert adjudication of image-based interpre-

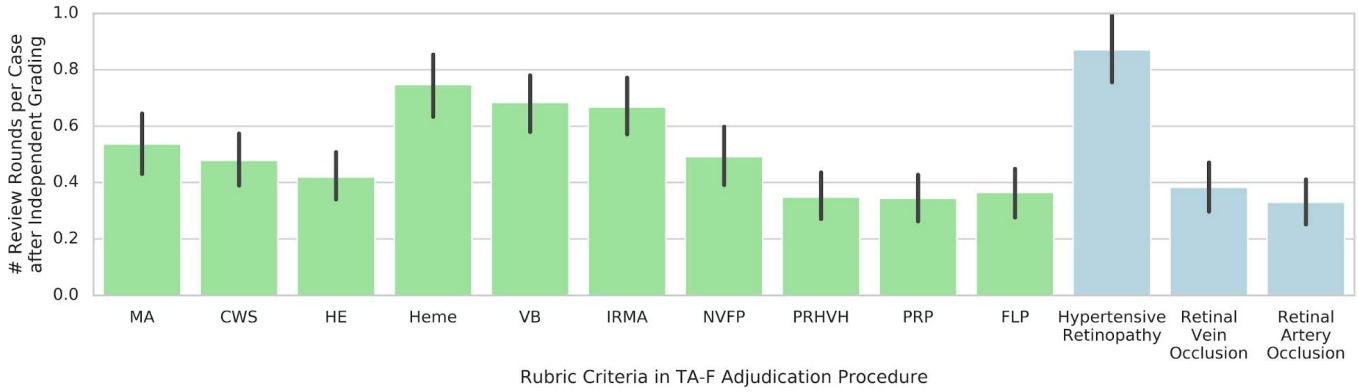

Figure 6. Mean number of review rounds required per rubric criterion in remote TA-F. The Y axis indicates the number of rounds after independent grading until either agreement was reached for the given criterion; or the case was closed due to overall agreement on the diagnosis level. Note that the mean number of review rounds may be below 1 because cases not requiring adjudication due to independent agreement were considered to have 0 review rounds. *Green bars* correspond to feature criteria, *blue bars* correspond to differential diagnosis criteria. *Error bars* indicate the 95% confidence intervals. CWS, cotton-wool spot; HE, hard exudate; NVFP, neovascularization or fibrous proliferation; PRHVH, Preretinal or vitreous hemorrhage; PRP, pan-retinal photocoagulation scars; FLP, focal laser photocoagulation scars.

tations and evaluate the system in the context of DR severity assessment. In this tool-based method, anonymity allows for an unbiased review of the image, with further clarity added by the rubric feature. Furthermore, the flexibility inherent in the design increases its appeal and ease of use.

Performance of Tool-Based Adjudication

Our study suggests that remote, tool-based adjudication procedures can produce DR grades that are in high agreement with the reference standard of in-person adjudication while offering a range of benefits: an increase in flexibility to accommodate graders' schedules, the possibility to anonymize graders throughout the adjudication process to avoid potential biases grounded in grader identity or seniority, and the option to explicitly structure the adjudication process around detailed evaluation criteria.

Research into efficient and reliable procedures to produce high-quality grading decisions can be applied to manual screening in teleophthalmology settings and to the validation of automated methods, such as deep-learning systems. In both cases, reliable classification decisions are required to avoid potentially devastating consequences, such as missing cases of advanced disease. Validating the classification performance against a reliable reference standard may be of particular importance for automated methods as, once deployed into a clinical screening setting, these methods can affect large patient populations in a short amount of time.

Beyond the evaluation of our remote TA procedure for adjudicating fundus images for DR severity assessment, we demonstrate how our proposed tool-based procedure can provide structure to the adjudication process itself using explicit prompts for detailed evaluation criteria. The resulting TA-F procedure leads to a significant reduction in the number of rounds needed to resolve disagreements. One possible explanation for the observed efficiency improvement may be that the feature rubric helped graders communicate their rationale and the specific source of disagreement more efficiently than was otherwise achieved through free-form comments (e.g., by focusing communication on the specific guideline criteria that graders are instructed to factor into a diagnosis). Besides efficiency in communication, the guideline-centric rubrics may serve as a lightweight checklist, leading graders to be more consistent in their individual practices. This may reduce variance or allow graders to externalize the diagnostic criteria in a way that reduces their task-related mental workload. As supporting evidence (Fig. 5), the first-round agreement rates were significantly higher with the use of the rubric, even before further adjudication. Finally, the rubrics lead to the production of structured information (i.e., the specific evaluation criteria applied in each case), facilitating detailed quantitative analyses to examine how and why disagreements arise both across a set of images and for individual cases.

Still, there were specific features of the disease that

required more discussion. While the explicit reasons for why heme, VB, and IRMA required the greatest number of review rounds are not clear, it is possible that the overlap between the objective (i.e., simple presence or absence) and subjective (i.e., extent and prominence) features of these particular anatomic abnormalities led to more disagreement. Despite standard reference photographs to help guide whether or not the heme is extensive, VB is definite, or the IRMA is prominent, there is an inherent subjectivity to the process. Ultimately, the physician's gestalt leads her to define disease severity. This same overall impression or pattern recognition may explain why venous and arterial occlusions resolved in fewer rounds, as these diseases have a hallmark appearance. HTN, on the other hand, can overlap with and mimic several other eye diseases, DR being the most common, and giving a definitive diagnosis based on a fundus photograph alone can be challenging. Exploring these feature-based discrepancies may provide more insight on how the model synthesizes the information within the image and also on how to continue to improve it.

Utility in Clinical Practice

We believe the technology we describe here may have several clinical applications. First, our approach for remote adjudication is well suited for integration into existing telemedical workflows, which face the same problem of high intergrader variability as is the case for on-site clinical grading.[3] Here, our proposed system can help resolve ambiguous cases through group decision-making[27] on demand to improve clinical outcomes on a patient-by-patient basis. Apart from adjudication, our tool's functionality of integrating feature-level rubrics into the image interpretation process may facilitate grading by individual graders in difficult cases, by helping list and systematize the image findings.

Second, expanding TA and TA-F use for rare conditions or difficult to diagnose cases, where a patient may otherwise be advised to travel to seek a second or third opinion, could potentially have an important impact on time to diagnosis and treatment, which are likely to impact quality of life and healthcare costs.

Third, and perhaps most importantly, our adjudication tool lends itself naturally for generating highly reliable and trusted reference standards for the validation of automated methods, such as deep-learning models. The process of building and evaluating deep-learning models typically involves at least the three following distinct datasets: a 'development' dataset used to train the model, a 'tuning' dataset used to select high-performing model candidates during the training phase, and a 'validation' dataset used to benchmark the performance of the final model. While development datasets, in many cases, consist of tens or hundreds of thousands of training examples, the datasets used for tuning and validation are typically smaller scale, on the order of several hundred up to a few thousand cases. The methods presented here can enable the creation of tuning and validation sets with a substantially reduced overhead, due to lower time and coordination requirements. In this study, we demonstrated the feasibility of remote, TA for a set of 499 images, positioning it as a useful procedure especially for generating tuning and validation datasets. The availability of a highly trusted validation dataset is of critical importance especially for so-called "black box" systems, where there is limited ability to understand how the model makes its diagnosis. As methods for remote adjudication in clinical decision-making scale, it may become feasible to produce adjudicated datasets large enough to be used for training, which would extend the current state-of-the-art in model development.

Limitations and Future Work

Our study is not without limitations. First, while we quantify the reliability of each adjudication method using consensus grades from two independent expert panels, the metrics reported in this work remain relative ones given the lack of an absolute, objective gold standard for DR severity assessment in the context of our study. To alleviate this issue, further work may benchmark adjudication decisions from digital fundus images against more rigorous diagnostic procedures (e.g., dilated fundus exam by a retina specialist)[28] or objective outcomes, such as any future development of blindness. Second, graders participating in this study were practicing retina specialists rather than research-grade reading center graders. While reading center gradings may be a more standardized gold standard, the incorporation of insights from clinical practice into the grading may render our results more applicable to real-life scenarios than may otherwise be the case with research-grade readings. Third, we only adjudicated DR severity, but did not adjudicate DME. Agreement levels may be lower overall given difficulties in diagnosing DME on 2D fundus photos. Finally, the grading decisions in this study were based on fundus images without accompanying patient information or

clinical records. In practice, DR severity assessment based on digital fundus photography should consider patient history and be complemented by more rigorous diagnostic procedures including dilated fundus examination by a trained eye care professional and optical coherence tomography (OCT) or other imaging techniques when indicated to confirm the diagnosis.[28]

Our TA-F procedure included a mechanism to assign pseudonyms to graders to avoid biases grounded in grader identity. Anonymization of graders was not possible during the in-person discussions of the Baseline adjudication procedure, and could not be done for the TA procedure because the functionality for grader anonymization was added at a later stage of our tool's development. Anonymization of members in group-based decision processes generally reduces incentives for groupthink behavior, and thus tends to slow down consensus formation rather than accelerating it.[29,30] Thus, we reason that our reported benefit of TA-F is an underestimate of the true benefit, relative to comparing TA and TA-F when neither (or both) is anonymized.

Our results show that remote, tool-based adjudication can help organize the consensus formation process especially for those cases that can be resolved in the first few review rounds, but falls short of fully alleviating the problem of small portions of disagreement cases persisting over several review rounds. Future work may explore methods to accelerate resolution for such hard cases, for example, by investigating if aggregation methods like majority vote after the first two review rounds are sufficient proxies for final adjudicated decisions, or by implementing automatic techniques to schedule video conference calls to discuss small collections of hard cases among panelists without the need to involve a human coordinator.

Other promising avenues for future research revolve around the development of feature rubrics for improved efficiency and reliability of adjudication procedures. Understanding which strategies and practices for rubric development generally result in the biggest improvements across various diagnostic tasks would be helpful for the community so that other researchers can reliably produce effective rubrics for different areas of medical image interpretation.

Conclusions

Remote, tool-based adjudication presents a reliable alternative to in-person adjudication for DR severity assessment. The system allows flexibility so that graders can schedule their reviewing around their clinical duties. Additional benefits include the option of blinding graders from the identity of other panel members and the ability to structure the discussion of controversial cases around a set of discrete evaluation criteria. We found that feature-based rubrics for DR can help accelerate consensus formation for tool-based adjudication without compromising label quality.

Acknowledgments

From Google AI Healthcare: Abi Jones, MEd, Kasumi Widner, MS, Cristhian Cruz, MS, Quang Duong, PhD, Olga Kanzheleva, MS.

From Shri Bhagwan Mahavir Vitreoretinal Services Sankara Nethralaya, Chennai, Tamil Nadu, India: Rajiv Raman, MD for permission to use the fundus photograph shown in Figure 3.

Retina specialist graders: Peter A. Karth, MD, MBA, Loh-Shan B. Leung, MD, Jesse J. Jung, MD, Ehsan Rahimy, MD, Jeffrey J. Tan, MD, Rajeev S. Ramchandran, MD, MBA, Jesse M. Smith, MD, Rahul N. Khurana, MD, Ali Akbar Zaidi, MD, Margaret Greven, MD, Steven J. Ryder, MD, Joshua N. Carlson, MD, Courtney Crawford, MD, FACS, Nathan Haines, MD.

Disclosure: **M. Schaekermann**, Google LLC (F, E); **N. Hammel**, Google LLC (F, I, E); **M. Terry**, Google LLC (F, I, E); **T.K. Ali**, Google LLC (F, C); **Y. Liu**, Google LLC (F, I, E); **B. Basham**, Google LLC (F, I, E); **B. Campana**, Google LLC (F, I, E); **W. Chen**, Google LLC (F, I, E); **X. Ji**, Google LLC (F, I, E); **J. Krause**, Google LLC (F, I, E); **G.S. Corrado**, Google LLC (F, I, E); **L. Peng**, Google LLC (F, I, E); **D.R. Webster**, Google LLC (F, I, E); **E. Law**, None; **R. Sayres**, Google LLC (F, I, E)

References

1. Ting DSW, Cheung GCM, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin Exp Ophthalmol*. 2016;44:260–277.
2. International Council of Ophthalmology. International clinical diabetic retinopathy disease severity

scale, detailed table. Available at: http://www.icoph.org/resources/45/International-Clinical-Diabetic-Retinopathy-Disease-Severity-Scale-Detailed-Table-.html. Accessed January 6, 2019.
3. Shi L, Wu H, Dong J, Jiang K, Lu X, Shi J. Telemedicine for detecting diabetic retinopathy: a systematic review and m21eta-analysis. *Br J Ophthalmol*. 2015;99:823–831.
4. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs–an extension of the modified Airlie House classification. ETDRS report number 10. *Ophthalmology*. 1991;98(5 Suppl):786–806.
5. Scott IU, Bressler NM, Bressler SB, et al. Agreement between clinician and reading center gradings of diabetic retinopathy severity level at baseline in a phase 2 study of intravitreal Bevacizumab for diabetic for diabetic macular edema. *Retina*. 2008;28:36–40.
6. Li HK, Hubbard LD, Danis RP, et al. Digital versus film Fundus photography for research grading of diabetic retinopathy severity. *Invest Ophthalmol Vis Sci*. 2010;51:5846–5852.
7. Gangaputra S, Lovato JF, Hubbard L, et al. Comparison of standardized clinical classification with fundus photograph grading for the assessment of diabetic retinopathy and diabetic macular edema severity. *Retina*. 2013;33:1393–1399.
8. Ruamviboonsuk P, Teerasuwanajak K, Tiensuwan M, Yuttitham K; for the Thai Screening for Diabetic Retinopathy Study Group. Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. *Ophthalmology*. 2006;113:826–832.
9. Lichter PR. Variability of expert observers in evaluating the optic disc. *Trans Am Ophthalmol Soc*. 1976;74:532–572.
10. 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;304:649–656.
11. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125:1264–1272.
12. Hannun AY, Rajpurkar P, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. *Nat Med*. 2019;25:65–69.
13. Guan M, Gulshan V, Dai A, Hinton G. Who said what: modeling individual labelers improves classification. *32nd AAAI Conference on Artificial Intelligence*. Palo Alto, CA; 2018;3109–3118. https://arxiv.org/pdf/1703.08774.pdf.
14. 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.
15. Sayres R, Taly A, Rahimy E, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*. 2019;126:552–564.
16. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
17. Golden JA. Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. *JAMA*. 2017;318:2184–2186.
18. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124:962–969.
19. 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.
20. Trucco E, Ruggeri A, Karnowski T, et al. Validating retinal fundus image analysis algorithms: issues and a proposal. *Invest Ophthalmol Vis Sci*. 2013;54:3546–3559.
21. Fleming AD, Goatman KA, Philip S, Prescott GJ, Sharp PF, Olson JA. Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts. *Br J Ophthalmol*. 2010;94:1606–1610.
22. Diabetic retinopathy PPP 2014: standard photographs 2A, 6A, 8A. Available at: https://www.aao.org/asset.axd?id=f29258e4-9744-463a-86eb-6822d6ff107b. Accessed January 6, 2019.
23. Barbazetto IA. Retinal physician - diabetic retinopathy: the masqueraders. retinal physician. Available at: https://www.retinalphysician.com/issues/2010/july-aug/diabetic-retinopathy-the-masqueraders. Accessed August 6, 2019.
24. Bhavsar AR. Diabetic retinopathy differential diagnoses. Available at: https://emedicine.medscape.com/article/1225122-differential. Published May 30, 2019. Accessed August 6, 2019.
25. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37–46.
26. Chihara LM, Hesterberg TC. *Mathematical Statistics with Resampling and R*. Hoboken: John Wiley & Sons; 2018.

27. Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA Netw Open*. 2019; 2:e190096.
28. Diabetic retinopathy PPP - updated 2017. American Academy of Ophthalmology. Available at: https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp-updated-2017. Accessed August 1, 2019.
29. Scott CR. The impact of physical and discursive anonymity on group members' multiple identifications during computer-supported decision making. *West J Speech Commun*. 1999;63:456–487.
30. Postmes T, Lea M. Social processes and group decision making: anonymity in group decision support systems. *Ergonomics*. 2000;43:1252–1274.