AMERICAN ACADEMY OF OPHTHALMOLOGY®

Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy

Jonathan Krause, PhD,[1] *Varun Gulshan, PhD,*[1] *Ehsan Rahimy, MD,*[2] *Peter Karth, MD,*[3] *Kasumi Widner, MS,*[1] *Greg S. Corrado, PhD,*[1] *Lily Peng, MD, PhD,*[1,*] *Dale R. Webster, PhD*[1,*]

Purpose: Use adjudication to quantify errors in diabetic retinopathy (DR) grading based on individual graders and majority decision, and to train an improved automated algorithm for DR grading.

Design: Retrospective analysis.

Participants: Retinal fundus images from DR screening programs.

Methods: Images were each graded by the algorithm, U.S. board-certified ophthalmologists, and retinal specialists. The adjudicated consensus of the retinal specialists served as the reference standard.

Main Outcome Measures: For agreement between different graders as well as between the graders and the algorithm, we measured the (quadratic-weighted) kappa score. To compare the performance of different forms of manual grading and the algorithm for various DR severity cutoffs (e.g., mild or worse DR, moderate or worse DR), we measured area under the curve (AUC), sensitivity, and specificity.

Results: Of the 193 discrepancies between adjudication by retinal specialists and majority decision of ophthalmologists, the most common were missing microaneurysm (MAs) (36%), artifacts (20%), and misclassified hemorrhages (16%). Relative to the reference standard, the kappa for individual retinal specialists, ophthalmologists, and algorithm ranged from 0.82 to 0.91, 0.80 to 0.84, and 0.84, respectively. For moderate or worse DR, the majority decision of ophthalmologists had a sensitivity of 0.838 and specificity of 0.981. The algorithm had a sensitivity of 0.971, specificity of 0.923, and AUC of 0.986. For mild or worse DR, the algorithm had a sensitivity of 0.970, specificity of 0.917, and AUC of 0.986. By using a small number of adjudicated consensus grades as a tuning dataset and higher-resolution images as input, the algorithm improved in AUC from 0.934 to 0.986 for moderate or worse DR.

Conclusions: Adjudication reduces the errors in DR grading. A small set of adjudicated DR grades allows substantial improvements in algorithm performance. The resulting algorithm's performance was on par with that of individual U.S. Board-Certified ophthalmologists and retinal specialists. *Ophthalmology 2018;■:1–9 © 2018 by the American Academy of Ophthalmology*

Supplemental material available at www.aaojournal.org.

Diabetic retinopathy (DR) and diabetic macular edema (DME) are among the leading causes of vision loss worldwide. Several methods for assessing the severity of diabetic eye disease have been established, including the Early Treatment Diabetic Retinopathy Study Grading System,[1] the Scottish Diabetic Retinopathy Grading System,[2] and the International Clinical Diabetic Retinopathy (ICDR) disease severity scale.[3] The ICDR scale is one of the more commonly used clinical scales and consists of a 5-point grade for DR: no, mild, moderate, severe, and proliferative.

The grading of DR is a fairly complex process that requires the identification and quantification of fine features such as microaneurysms (MAs), intraretinal hemorrhages, intraretinal microvascular abnormalities, and neovascularization. As a result, there can be a fair amount of grader variability, with intergrader kappa scores ranging from 0.40 to 0.65.[1,4–7] This is not surprising because grader variability is a well-known issue with human interpretation of imaging in other medical fields such as radiology[8] or pathology.[9]

Several methods have been proposed for resolving disagreements between graders and obtaining a reference standard. One approach consists of taking the majority decision from a group of 3 or more independent graders. Another consists of having a group of 2 or more graders work independently and then having a third generally more senior grader arbitrate disagreements, with that individual's decision serving as the reference standard. Last, a group of 3 or more graders may first grade independently and then collectively discuss disagreements until there is full consensus on the final grade. The difference between various methods of resolving disagreements has not been examined extensively.

https://doi.org/10.1016/j.ophtha.2018.01.034
ISSN 0161-6420/18

Deep learning[10] is a family of machine learning techniques that allows computers to learn the most predictive features directly from images, given a large dataset of labeled examples, without specifying rules or features explicitly. It has also been applied recently in medical imaging, producing highly accurate algorithms for a variety of classification tasks, including melanoma,[11] breast cancer lymph node metastasis,[12,13] and DR.[14–16] Because the network is trained to predict labels that have been paired with the images, it is imperative that the labels accurately represent the state of disease found in the image, especially for the evaluation sets.

In this study, we examine the variability in different methods of grading, definitions of reference standards, and their effects on building deep learning models for the detection of diabetic eye disease.

Methods

Development Dataset

In this work, we built upon the datasets used by Gulshan et al[14] for algorithm development and clinical validation. A summary of the various data sets and grading protocols used for this study is shown in Figure S1 (available at www.aaojournal.org). The development dataset used consists of images obtained from patients who presented for DR screening at EyePACS-affiliated clinics, 3 eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya), and the publicly available Messidor-2 dataset.[17,18] Images from the EyePACS clinics consisted of 45° retinal fundus images that were primary (or posterior pole-centered), nasal (disc-centered), and temporal field of view. The rest of the data sources only had primary field of view images. The images were captured using the Centervue DRS, Optovue iCam, Canon CR1/DGi/CR2, and Topcon NW using 45° fields of view. All images were de-identified according to Health Insurance Portability and Accountability Act Safe Harbor before transfer to study investigators. Ethics review and institutional review board exemption was obtained using Quorum Review Institutional Review Board.

There were 3 types of grades used in the development dataset: grading by EyePACS graders, grading by ophthalmologists, and adjudicated consensus grading by retinal specialists. Images were provided at full resolution for manual grading. The images provided by EyePACS clinics were graded according to the EyePACS protocol.[19] In the EyePACS protocol, all 3 images of an eye (primary, nasal, and temporal) were graded together and assigned a single grade. We assigned the grade provided for the entire eye as a grade for each image. Although the eye-level grade may not correspond exactly with the image-level grade, this method allowed for training on different fields of view, which helps the algorithm tolerate some deviation from the primary field of view. The second source of grades was the development dataset used by Gulshan et al,[14] where each image was independently graded by ophthalmologists or trainees in their last year of residency. The grading interface used is shown in Fig S2 and Fig S3 (available at www.aaojournal.org). The third type of grades was obtained using our subspecialist grading and adjudication protocol on a small subset of images. Our subspecialist adjudication protocol consisted of 3 fellowship-trained retina specialists undertaking independent grading of the dataset. Next, each of the 3 retina specialists reviewed the images with any level of disagreement in a combination of asynchronous and live adjudication sessions.

For algorithm development, the development dataset was divided into groups: a "train" set and a "tune" set. The "train" set consisted of all the images that were not adjudicated and was used to train the model parameters. The "tune" set consisted of images with adjudicated consensus grades and was used to tune the algorithm hyperparameters (e.g., input image resolution, learning rate) and make other modeling choices (e.g., network architectures). All of the images in the "tune" set were of the primary field.

Clinical Validation Dataset

The clinical validation dataset consisted of primary field of view images that were acquired from EyePACS clinics (imaged between May and October 2015). These did not overlap with any images or patients used in the development dataset.

The clinical validation set was graded using the same adjudication protocol as the tuning set (face-to-face adjudication by 3 retina specialists). In addition, the clinical validation set was graded by 3 ophthalmologists. These ophthalmologists were distinct from the retina specialists who adjudicated the set. After grading, all disagreements between the adjudicated consensus of retinal specialists and majority decision of the ophthalmologists were reviewed manually by a retinal specialist, who also assigned a likely reason for the discrepancy.

Algorithm Training

Our deep learning algorithm for predicting DR and DME was built upon the architecture used by Gulshan et al.[14] We used a convolutional neural network[20] that predicted a 5-point DR grade, referable DME, and gradability of image. The input to the neural network was an image of a fundus, and through the use of many stages of computation, parameterized by millions of numbers, the network output a real-valued number between 0 and 1 for each prediction, indicating its confidence.

The parameters of a neural network were determined by training it on a dataset of fundus images. Repeatedly, the model was given an image with a known severity rating for DR, and the model predicted its confidence in each severity level of DR, slowly adjusting its parameters over the course of the training process to become more accurate. The model was trained via distributed stochastic gradient descent and evaluated on a tuning dataset throughout the training process. The tuning dataset was used to determine model hyperparameters (parameters of the model architecture that cannot be trained with gradient descent). Finally, we created an "ensemble" of models, training 10 individual models and combining their predictions, which improves performance and robustness.

Algorithmic Improvements

We made a number of improvements to the core neural network, compared with the study by Gulshan et al.[14] First, we trained our model on a larger set of images obtained from EyePACS.[21] Because it would be prohibitively expensive to relabel all of these images with U.S.-licensed ophthalmologists (as done by Gulshan et al[14]), we also use labels determined by the EyePACS grading centers to train our DR classifier. To use both our ophthalmologist grading and the EyePACS grading effectively, we treated the labels from these 2 independent sources as separate prediction targets, training our model to predict both an EyePACS grade for DR and a grade determined by our own labeling procedure, if present.

By using this larger training set, we performed a more extensive search for well-performing hyperparameters using a Gaussian process bandit algorithm.[22] One significant change in model hyperparameters that resulted was an increase in the resolution of images the model used as input: The model used in this work