Application of deep learning algorithms for diabetic retinopathy screening

Broder Poschkamp, Andreas Stahl

Department of Ophthalmology, University Medicine Greifswald, Greifswald, Germany
Correspondence to: Broder Poschkamp. University Medicine Greifswald, Ferdinand-Sauerbruch-Straße, 17475 Greifswald, Germany.
Email: poschkamp.research@gmail.com.

Comment on: Yang Y, Pan J, Yuan M, *et al.* Performance of the AIDRScreening system in detecting diabetic retinopathy in the fundus photographs of Chinese patients: a prospective, multicenter, clinical study. Ann Transl Med 2022;10:1088.

Diabetic retinopathy (DR) is one of the most common causes of blindness in adults (1-3). The global prevalence of DR is estimated to be 34.6 % in all diabetics (4). Early stages of DR can be asymptomatic for many years. The detection of DR is important to reduce the progression of DR and preserve vision in the long term (2). In recent years, a number of procedures have evolved that use fundus photographs to make a diagnostic determination of DR (5,6). In this issue of Annals of Translational Medicine, Yang and colleagues present findings of a deep-learning algorithm (artificial intelligence DR screening, AIDRScreening system) to detect DR from a prospective, multicenter clinical evaluation study (7). First, Yang's results are compared with results from existing algorithms. Also, international experiences from deep learning-based screening of DR are addressed. Trends in the field are highlighted and the unique situation in China is discussed in the article.

The first deep learning method to be approved by the U.S. Food and Drug Administration (FDA) was the IDx-DR algorithm in 2018 (IDx Technologies Inc., Coralville, Iowa, USA) (8). Also, Google has developed an algorithm for DR detection: automated retinal disease assessment (ARDA) (9).

All three algorithms (IDx-DR, ARDA and the novel AIDRScreening algorithm published in this issue of Annals of Translational Medicine) use convolutional-neural-networks as the main method for image classification. In all three approaches, the neural networks are used as an ensemble, meaning that several networks reach decisions together, or one network is applied for only a partial decision (7-10).

The AIDRScreening algorithm was trained with 73,849 images. For comparison the training dataset from the ARDA algorithm had a volume of 128,175 images (7).

The output of the AIDRScreening system is "detected referable DR" or "nondetected referable DR". The referable DR represents a moderate or worse form of DR (7). As a screening tool, this output is sufficient because the results should be followed by an ophthalmologist to verify the diagnosis and (if necessary) treat the disease.

Similar to the AIDRScreening system but with slightly more granularity, outputs of the IDx-DR algorithm differentiate between no, mild, moderate, and severe DR (8,11).

In terms of programming, neural networks generally are used to derive a probability for a diagnosis or score from an image representing a specific degree of disease severity. This continuous probability (or score) is then transformed into an output (e.g., "detected referable DR" or "nondetected referable DR"). Transforming the score into a binary output, however, involves a loss of information. The probability for a class or a severity score represents useful information for the physician. This can have an influence on clinical decision making. Besides, a continuous score is important for clinical evaluation and improvement of the algorithm itself. An ordinal scale is derived from an interval

and rational scale (12). The statistical analysis of an ordinal scale (diagnoses) is less informative and fewer statistical tests can be applied to it. Remedy would be the introduction of two analytical modes. A user-friendly one (with classes) and an extended analysis for experts (with continuous probabilities).

In this study, no image quality analysis was performed with the AIDRScreening system (V1.0). In the meantime, however, an extension of the system with image quality analysis has been introduced (AIDRScreening system V2.0) (7). This is also possible with the IDx-DR system (8,11).

Yang et. all examined the results between three different study centers and the influence of three different camera systems (Zeiss Visucam FF450, Topcon TRC-50DX and Topcon NW-400). The algorithm was found to be robust to the different influences. This is necessary for the algorithm to be widely used. In the screening protocol, pupil dilating eye drops were applied to the patients for image assessment in two centers and in one center if the pupil diameter was below 4 mm (7).

In a German evaluation study, the IDx-DR algorithm was tested on 503 patients to investigate the screening suitability in non-mydriatic patients, i.e., without using pupil dilating eye drops. In eyes without pharmacological pupil dilation, it was significantly more difficult to obtain images that is subsequently analyzable. Only 59.1% of non-mydriatic images were of sufficient quality to yield a result from the IDx-DR algorithm (13). This example illustrates practical difficulties in applying AI algorithms to DR screening that are independent of the AI used—image acquisition is key to any valid diagnosis.

In comparison to other algorithms the AIDRScreening system could reach a comparable sensitivity of 86.72% (IDx-DR: 87.2%, ARDA algorithm: 87.0%) for referable DR and a higher specificity of 96.09% (IDx-DR: 90.7%, ARDA algorithm: 90.3%) (7-9). The specifications of the ARDA algorithm were selected for high specificity (9). Based on these results the AIDRScreening system can compete with Google's ARDA and the IDx-DR algorithm. It would be necessary to test the different algorithms against each other in an independent study.

From a doctor's view these algorithms remain a black-box and the explanation of how the algorithm found its diagnoses would be helpful in terms of trust and interpretability. The field of "explainable AI" is dedicated to the better understanding of artificial intelligence algorithms. One approach is the marking and naming of recognized pathologies in fundus images. In a comparative comparison of five algorithms for the detection of DR, which can mark pathologies in fundus photographs, however, different areas were classified as significant illustrating the current limitations of such explainability approaches (5).

In an ideal world, a deep learning system would provide physicians and patients not only with output variables ("treatment required" or "no treatment required") but would also report the level of certainty (or uncertainty) of its decisions. Reporting uncertainty when communicating diagnoses is common practice between human physicians and patients and it would help to integrate such communication also in future AI tools. The AI system should be allowed a certain error variance if it also reports this uncertainty. The reported uncertainties can have several dimensions. Thus, the circumstances of the examination (image quality) or the uncertainty in the diagnosis (severity of the disease, differential diagnoses) could be described. In a clinical context, a physician would seek a second opinion from a colleague if uncertain. In the future, this colleague could be an AI system.

Most of the DR screening algorithms today can only distinguish between DR and non-DR (6). Thus, all other diseases are always classified as DR or non-DR—a true differential diagnosis is currently not possible. It is important to keep this in mind when examining AI screened patients as a physician.

As a further development, not only the *diagnosis* of DR will play a role, but also the *prognosis* of the disease. Thus, it may be possible to assess whether the disease will progress and a shorter time interval for follow-up should be scheduled. One paper from Arcadu *et al.* describes the application of multiple deep learning models, on the output of which a random forest was built to subsequently predict disease progression at 6, 12, and 24 month time intervals (14).

A smartphone-based screening would be a widely available and cost-effective DR screening method. The smartphone-based fundus cameras, however, have the problem that the field-of-view of the retina is smaller. In a study on synthetic data, the diagnostic power was reduced dramatically with a smaller retinal field-of-view (15).

It has been observed that the performance of deep learning algorithms for DR detection varies between different ethnicities and populations (16). Yang and colleagues discussed the need of a fully automated DR screening system specifically designed for Chinese patients (7). The prevalence of DR in China differs much depending on the region observed. In rural regions, the prevalence of patients

with type 2 diabetes is higher (29.1–43.1%) compared to urban regions (18.1%) (17,18). Vision-threatening DR in rural China is thus estimated to affect a very high number of 1.3 million patients (17). In addition, there is an increasing inequality in the accessibility of doctors between rural and urban areas in China (19). This emphasizes the need for DR screening and disease management programs especially in rural areas of China (17).

The IDx-DR algorithm was investigated as a telemedical screening tool in primary health care in Germany. It demonstrated a 72% increase in referrals to ophthalmologists. At the same time, only 32% of those screened were found to be willing to pay for the service themselves (20). In addition, screening is only useful if there are enough ophthalmologists available to examine and treat patients referred to them based on (AI-based) screening programs. When used properly, however, an AI-based screening tool can help reduce unnecessary medical examinations and focus availability of medical care to those patients who need it most. In their work, Yang and colleagues were able to present a functional algorithm that can detect DR in a population in order to initiate preventive actions. Proper implementation of Deep Learning-based DR screening in China will be a technical and logistical challenge but has the potential to preserve the vision of many people.

Acknowledgments

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Annals of Translational Medicine*. The article did not undergo external peer review.

Conflicts of Interest: Both authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-2022-73/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Open Access Statement: This is an Open Access article

References

1. 1. Lightman S, Towler HM. Diabetic retinopathy. Clin Cornerstone 2003;5:12-21.
2. Wong TY, Cheung CM, Larsen M, et al. Diabetic retinopathy. Nat Rev Dis Primers 2016;2:16012.
3. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. Eye Vis (Lond) 2015;2:17.
4. Yau JW, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 2012;35:556-64.
5. Bellemo V, Lim G, Rim TH, et al. Artificial Intelligence Screening for Diabetic Retinopathy: the Real-World Emerging Application. Curr Diab Rep 2019;19:72.
6. Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. Eye (Lond) 2020;34:451-60.
7. Yang Y, Pan J, Yuan M, et al. Performance of the AIDRScreening system in detecting diabetic retinopathy in the fundus photographs of Chinese patients: a prospective, multicenter, clinical study. Ann Transl Med 2022;10:1088.
8. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39.
9. 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;316:2402-10.
10. Niemeijer M, Abramoff MD, van Ginneken B. Information fusion for diabetic retinopathy CAD in digital color fundus photographs. IEEE Trans Med Imaging 2009;28:775-85.
11. Niemeijer M, Abràmoff MD, van Ginneken B. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. Med Image Anal 2006;10:888-98.
12. Stevens SS. On the theory of scales of measurement. Science 1946;103:677-80.

13. Paul S, Tayar A, Morawiec-Kisiel E, et al. Use of artificial intelligence in screening for diabetic retinopathy at a tertiary diabetes center. Ophthalmologie 2022;119:705-13.
14. Arcadu F, Benmansour F, Maunz A, et al. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. NPJ Digit Med 2019;2:92.
15. Hacisoftaoglu RE, Karakaya M, Sallam AB. Deep Learning Frameworks for Diabetic Retinopathy Detection with Smartphone-based Retinal Imaging Systems. Pattern Recognit Lett 2020;135:409-17.
16. Ting DSW, Cheung CY, Lim G, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. JAMA 2017;318:2211-23.
17. Wang FH, Liang YB, Zhang F, et al. Prevalence of diabetic retinopathy in rural China: the Handan Eye Study. Ophthalmology 2009;116:461-7.
18. Liu L, Wu X, Liu L, et al. Prevalence of diabetic retinopathy in mainland China: a meta-analysis. PLoS One 2012;7:e45264.
19. Cao X, Bai G, Cao C, et al. Comparing Regional Distribution Equity among Doctors in China before and after the 2009 Medical Reform Policy: A Data Analysis from 2002 to 2017. Int J Environ Res Public Health 2020;17:1520.
20. Wintergerst MWM, Bejan V, Hartmann V, et al. Telemedical Diabetic Retinopathy Screening in a Primary Care Setting: Quality of Retinal Photographs and Accuracy of Automated Image Analysis. Ophthalmic Epidemiol 2022;29:286-95.