

Image Quality Assessment of Retinal Fundus Photographs for Diabetic Retinopathy in the Machine Learning Era: a Review

Mariana Batista Gonçalves^{1,2,3}

Luis Filipe Nakayama^{1,4}

Daniel Ferraz^{1,2,3}

Hanna Faber^{5,6}

Edward Korot^{7,8}

Fernando Korn Malerbi¹

Caio Vinicius Regatieri¹

Mauricio Maia¹

Leo Anthony Celi^{4,9,10}

Pearse A. Keane³

Rubens Belfort Jr.^{1,2}

1 Department of Ophthalmology, Sao Paulo Federal University - São Paulo (SP), Brazil

2 Instituto Paulista de Estudos e Pesquisas em Oftalmologia, IPEPO, Vision Institute, São Paulo (SP), Brazil

3 NIHR Biomedical Research Centre for Ophthalmology, Moorfield Eye Hospital, NHS Foundation Trust, and UCL Institute of Ophthalmology, London, UK

4 Massachusetts Institute of Technology, Institute for Medical Engineering and Science – Laboratory for Computational Physiology, Cambridge, MA, United States of America

5 Department of Ophthalmology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

6 Department of Ophthalmology, University of Tuebingen, Tuebingen, Germany

7 Michigan Grand Rapids, Michigan, USA

8 Stanford University Byers Eye Institute Palo Alto, California, USA

9 Harvard TH Chan School of Public Health, Department of Biostatistics, Boston, MA, United States of America

10 Beth Israel Deaconess Medical Center, Department of Medicine, Boston, MA, United States of America

Corresponding author: Luis Filipe Nakayama

luisnaka@mit.edu

<https://orcid.org/0000-0002-6847-6748>

Massachusetts Institute of Technology, Institute for Medical Engineering and Science, Cambridge, MA, United States of America

77 Massachusetts Ave., Cambridge, MA 02139

+1 (617) 253-7818

Abstract

This study aimed to evaluate the image quality assessment (IQA) and quality criteria employed in publicly available datasets for diabetic retinopathy (DR). A literature search strategy was used to identify relevant datasets, and 20 datasets were included in the analysis. Out of these, 12 datasets mentioned performing IQA, but only eight specified the quality criteria used. The reported quality criteria varied widely across datasets, and accessing the information was often challenging. The findings highlight the importance of IQA for AI model development while emphasizing the need for clear and accessible reporting of IQA

information. The study suggests that automated quality assessments can be a valid alternative to manual labeling and emphasizes the importance of establishing quality standards based on population characteristics, clinical use, and research purposes. In conclusion, image quality assessment is important for AI model development; however, strict data quality standards must not limit data sharing. Given the importance of IQA for developing, validating, and implementing deep learning (DL) algorithms, it's recommended that this information be reported in a clear, specific, and accessible way whenever possible. Automated quality assessments are a valid alternative to the traditional manual labeling process, and quality standards should be determined according to population characteristics, clinical use, and research purpose.

Keywords: dataset, ophthalmology, retinal fundus photos, artificial intelligence

Introduction

Diabetic retinopathy (DR) is a microvascular complication of diabetes and a leading cause of severe visual loss worldwide ^{1,2}. Aiming at the early diagnosis and treatment of this condition, screening programs have been established to routinely analyze retinal photographs and select the cases that need specialized evaluation ³⁻⁵. Recently, artificial intelligence (AI) and deep learning (DL) algorithms have been reported to achieve robust performance in detecting DR from retinal photographs, representing a promising tool to manage the large amount of image data generated by screening programs ^{2,6-9}. Despite recent advances in this technology, many factors should be taken into account when developing and deploying an AI system, such as the number of included images ^{9,10}, patient's characteristics ¹⁰, labels ¹¹⁻¹³, and image quality ^{9,14,15}.

Several failures in the clinical deployment of AI are a consequence of inadequate model development, the inclusion of only homogeneous populations, and non-diverse data ^{7,16}. In ophthalmology models, using exclusively high-quality images can produce higher pre-deployment accuracy results; however, these algorithms fail when they are deployed in real-world settings ^{17,18}. Datasets with more diverse demographics, diseases, and image quality are crucial for real-world simulation ¹⁶.

During the exam-capturing process, different measures can be taken to deal with low-quality images, including image reacquisition and pharmacological mydriasis ^{19,20}. A manual or automatic quality assessment in databases can be performed by addressing image parameters such as focus, image field, artifacts, and illumination.

A quality assessment can lead to more generalizable algorithms regardless of the approach chosen to handle poor-quality images. Nevertheless, there are some obstacles involved in the image quality assessment (IQA): 1) the rating of image quality is subjective, and intergrader variability can be high, especially for images in the middle of the quality scale ²¹⁻²³ 2) the criteria for the evaluation of the image quality vary according to the type of diagnosis being made since different diseases have different image features and regions of interest ^{21,23}.

Since quality control of retinal images influences model performance, fairness, and generalizability of DL algorithms, this paper reviewed publicly available DR datasets' image quality criteria and assessment.

Methods

In this study, we evaluate the IQA and the quality criteria employed across publicly available datasets. This review compared the datasets employed in PubMed/MEDLINE Artificial Intelligence and Ophthalmology articles, using the search strategy: (("dataset" [tiab] OR "database" [tiab]) AND ("publicly available" [tiab] OR "free of charge" [tiab] OR "freely accessible" [tiab] OR "publicly accessible" [tiab]) AND ("eye"[tiab] OR "ophthalmology"[tiab] OR "ophthalmology"[MeSH] OR "retina"[tiab])) AND (("2000"[Date - Publication]: "2021"[Date - Publication])).

From the final cohort of articles, two authors (MBG and LFN) manually reviewed the applied dataset's websites and any dataset-related documentation/publication available at

these websites. These sources were considered "direct sources" of information. For incorporated datasets originating from DR screening programs, we also reviewed the publications describing the programs if the references of these publications were easily accessible (i.e., described on the websites or dataset documentation/publication). These sources were considered "indirect sources" of information. To enhance the article assessment quality, the reviewers selected were retina specialists and experts in ophthalmology and AI research. In order to minimize subjective bias, each reviewer independently evaluated every article. In discordance cases, the reviewers reached a consensus after discussion.

We included in the review publicly available color retinal fundus photos datasets with diabetic retinopathy patients. Datasets containing exclusively other ocular diseases (e.g., glaucoma or age-related macular degeneration) or imaging modalities (e.g., optical coherence tomography and fluorescein angiogram) and non-available images were excluded since image quality criteria may vary across different diseases or imaging modalities.

We reviewed direct and indirect sources by focusing on the following questions: Has any IQA been described, and if so, have the image quality criteria been specified?

Results

The search strategy yielded 130 articles. The first screening included only human and English, Spanish, and Portuguese articles, and the second consisted of title and abstract assessment. After the screening process, 39 articles were eligible for full-text assessment. We retrieved the study's included databases from the selected articles, excluding non-available datasets. (Figure 1)

From the reviewed articles, we included in this quality assessment analysis the EyePACS, DIARETDB0 and DIARETDB1, E-Ophtha, DRIVE, MESSIDOR-2, IDRiD, ROC, DR1 and DR2, REVIEW, HEI-MED, APTOS, DERIVA, DRIMDB, ROTTERDAM, and ONHSD. From the analysis of the references, we included the SUSTech-SYSU, DDR, and DRiDB datasets.

Among the datasets included in this study, image quality assessment was described in 12 (60%) datasets; however, the specific image quality criteria were specified in only 8 (40%) databases (Figure 2, Table 1).

Apply quality classification criteria

EyePACS (Eye Picture Archive Communication System)

EyePACS is a publicly available fundus image dataset* that contains 88,702 CFP. The images were provided by the Eye Picture Archive Communication System (EyePACS), which is a free platform for retinopathy screening and made available by the California Healthcare Foundation ²⁴.

In addition to DR classification, the EyePACS grading protocol ²⁵ also includes criteria for IQA. Image quality factors considered are focus, illumination, image field definition, and artefacts. After the grading of these items, the image quality is classified in one of the following categories: excellent, good, adequate, insufficient for full interpretation, insufficient for any interpretation, or other. According to the grading protocol, the grader should specify the nature of the image quality problems. Details of the quality classification are described in Tables 2 and 3.

E-Ophtha

E-ophtha is a database designed for scientific research in DR and contains 463 CFP annotated for exudates and microaneurysms. It has been generated from the OPHDIAT Tele-medical network for DR screening, in the framework of the ANR-TECSAN-TELEOPHTA project funded by the French Research Agency (ANR) ²⁶. In the OPHDIAT screening program, a photograph was considered to be of acceptable quality if (1) the centre of the fovea and the retinal vessels were clearly visible and (2) more than two-thirds of the image could be assessed ²⁷.

ROC (Retina Online Challenge)

This dataset consists of 100 CFP, which were used in the Retinopathy Online Challenge (ROC), a multi-year competition for various aspects of DR, with a focus on microaneurysm detection ²⁸. ROC images were acquired as part of the EyeCheck project, a web-based screening program for Diabetic Retinopathy that started in 2000 in the Netherlands. In this screening program, gradability criteria included focus of the large vessels, visibility of at least one arteriole or capillary in both the upper and the lower half of the image, evenness of illumination, and absence of artefacts ²⁹. Only gradable images were included in the ROC dataset ²⁸.

DR1 and DR2

DR1 and DR2 are two publicly available datasets containing, respectively, 1077 and 520 CFP. The images were provided by the Department of Ophthalmology, Federal University of Sao Paulo (UNIFESP), and graded according to the presence of DR lesions and the need for referral^{30,31}. These datasets were manually annotated for image quality according to criteria such as the image field and the presence of blurring³².

HEI-MED (The Hamilton Eye Institute Macular Edema Dataset)

This dataset consists of 169 CFP, collected as part of a telemedicine network for the diagnosis of DR developed by the Hamilton Eye Institute, the Image Science and Machine Vision Group at ORNL with the collaboration of the Université de Bourgogne. The dataset was created to train, and test algorithms for the detection of exudates and diabetic macular oedema and the images were manually annotated for the presence of exudation areas.

Instead of descriptive quality criteria, an automated quality assessment technique was applied to this dataset. During the capturing process, quality metrics for the acquired images, which varied between 0 and 1, were displayed on the operator's camera, allowing them a chance to take a new image if necessary. According to the authors, all 169 images were of sufficient quality^{33,34}.

DERIVA (Digital Extraction from Retinal Images of Veins and Arteries)

This dataset was created to evaluate an automated method for structural mapping of retinal vessels and consists of 50 CFP, acquired as part of the EyeCheck project. The gradability criteria employed by this screening program was previously described for the ROC dataset^{29,35}.

DRIMDB (Diabetic Retinopathy Image Database)

This dataset consists of 216 CFP and was created to evaluate the performance of an IQA approach. All images were classified into three classes by the same expert. Non-retinal images that could have been obtained for several reasons (e.g., wrong focus or patient absence) were graded as outliers. The remaining images were divided into two groups: good and bad-quality images, but the quality criteria for each class are not provided³⁶.

SUSTech-SYSU

This dataset contains 1219 CFPs collected from the Department of Ophthalmology, Gaoyao People's Hospital and Zhongshan Ophthalmic Center, Sun Yat-sen University. The database was created to help in the development and validation of exudate detection algorithms and DR classification algorithms. In addition to exudate annotations, DR classification and bounding boxes for the optic disc and fovea location are also provided. The dataset underwent a quality control stage, and images that were "too blurry" or "of extremely large-area lesions" were excluded³⁷.

DDR

DDR is a publicly available dataset containing 13,673 CFP, collected from 147 hospitals in China. The images have image-level annotations of DR severity, pixel-level annotations of DR lesions, and bounding-box annotations automatically generated from pixel-level annotations.

In this dataset, the image acquisition was performed taking into account parameters such as optic disc and fovea location, focus, and exposure. The detailed principles are described in Table 4. All images were graded according to DR severity and image quality and were divided into six classes: no DR, mild nonproliferative DR, moderate nonproliferative DR, severe nonproliferative DR, proliferative DR, and ungradable. Images "with a blurring degree of more than 70% and without clearly visible lesions" were considered ungradable³⁸.

DIARETDB0 and DIARETDB1 (Standard Diabetic Retinopathy Database Calibration Level 0 and 1)

DIARETDB0 and DIARETDB1 are two publicly available datasets created to evaluate methods for automatic detection of DR. These databases contain, respectively, 130 and 89 CFP taken in the Kuopio University Hospital, which were annotated for DR retinal findings, such as exudates, haemorrhages, and microaneurysms. An automated method for evaluating the partial image quality was applied to the DIARETDB0 dataset. A binary mask for every fundus image is provided, representing whether a certain location is applicable for the analysis or not. Information on the IQA is not provided for the DIARETDB1 dataset³⁹⁻⁴¹.

IDRiD (Indian Diabetic Retinopathy Image Dataset)

IDRiD is a dataset containing 516 CFP, which is available as a part of "Diabetic Retinopathy: Segmentation and Grading Challenge," organized in conjunction with IEEE International Symposium on Biomedical Imaging (ISBI-2018), Washington D.C. The dataset provides 1) pixel-level annotations for DR lesions and the optic disc, 2) image-level severity grading of DR and macular oedema, and 3) optic disc and fovea centre location ⁴².

According to the authors, "experts verified that all images are of adequate quality" ⁴², and the dataset "did not include non-gradable images" ⁴³, but further details about IQA criteria are not provided.

APTOS (Asia Pacific Tele-Ophthalmology Society)

This dataset, released by the Asia Pacific Tele-Ophthalmology Society, contains 5590 publicly available CFPs. The authors mention that the images in this dataset may have artefacts, be out of focus, underexposed, or overexposed, but information on the IQA is not provided. **

Don't apply quality classification criteria

DRIVE (Digital Retinal Images for Vessel Extraction)

This dataset was established to enable comparative studies on segmentation of blood vessels in retinal images. It consists of 40 CFPs, taken from a DR screening program in the Netherlands, that contain manual segmentation of the retinal vasculature. Information on the IQA is not provided on direct sources. Further details about the screening program are not provided, and thus indirect sources couldn't be assessed ⁴⁴.

MESSIDOR-2 (Methods to Evaluate Segmentation and Indexing Techniques in the Field of Retinal Ophthalmology)

The messidor-2 dataset consists of 1748 CFP and was established after the merge of the "Messidor-Original" database, created to facilitate studies on the computer-assisted diagnosis of diabetic retinopathy, with the "Messidor-Extension" database. Information on the IQA is not provided for this dataset ⁴⁵.

REVIEW (Retinal Vessel Image Set for Estimation of Width)

This dataset consists of 16 CFP, and it was made available online by the Department of Computing and Informatics at the University of Lincoln, Lincoln, UK. The dataset was created for the evaluation of retinal vessel diameter measurement algorithms and contains images with manually annotated vessel edges. Information on the IQA is not provided for this dataset ⁴⁶.

ROTTERDAM (Rotterdam Ophthalmic Data Repository)

This dataset includes 1120 CFP captured as part of the DR screening program of the Rotterdam Eye Hospital in the Netherlands and was created to evaluate the accuracy of fundus image registration methods. Information on the IQA is not provided for this dataset ⁴⁷.

ONHSD (Optic Nerve Head Segmentation Dataset)

This database contains 99 CFPs taken from patients attending the diabetic screening program at City Hospital - Birmingham. The dataset was created to evaluate an algorithm for the localization and segmentation of the optic nerve head. The authors mention that there is considerable quality variation in the images, but information on the IQA is not provided ⁴⁸.

DRiDB (Diabetic Retinopathy Image Database)

This dataset consists of 50 CFPs taken in a university hospital in Zagreb. Experts were asked to annotate areas containing DR findings and mark the blood vessels, optic disc, and macula. The authors mention that the images in this dataset contain a varying amount of noise, but information on the IQA is not provided ⁴⁹.

Discussion

Image quality assessment is important for AI model development; however, strict data quality standards must not limit dataset development. The focus on quality may have the unintended consequence of discouraging less-funded institutions from contributing data because of concerns that their images are not of sufficient quality. In this work, we review 20

publicly available DR datasets and the described image quality assessment of retinal fundus photographs. Our findings show that 12 (60%) datasets mention having performed some quality control on the images, and 8 (40%) datasets specified the quality criteria employed in the IQA. Moreover, there is a lack of consensus on the criteria of image quality, and this information is often not easily accessible (i.e., available on indirect sources). Surprisingly, even on the DRIMDB, a dataset developed for quality assessment purposes, the criteria employed to classify the images as "good" or "bad" quality are not provided ³⁶.

Public datasets and collaborative research

Sharing more diverse and representative data is needed in order to promote better AI development. Publicly available datasets represent a valuable alternative to deal with the high costs and the legal barriers often involved with data access ^{10,45}. Data sharing is essential to promote collaborative research, bias assessment, reproducibility, and generalizability ⁵⁰.

Our results show that, among publicly available DR datasets, the information on the image quality assessment is often missing, and some datasets assess image quality, but the quality criteria are not published. This is the case for IDRiD, which included only gradable images, and DRIMDB, which included good and bad-quality images, and provides the respective quality labels. For the databases that mention the image quality criteria, this information was often vague (SUSTech-SYSU, DR1, and DR2 datasets) or difficult to access (E-Ophtha, ROC, and DERIVA datasets). EyePACS provides a detailed grading protocol and specific quality criteria, but it's not clear if the images with insufficient quality were excluded from the dataset.

Importance of image quality assessment

The image quality should not be the main consideration for dataset creation and sharing; nevertheless, IQA plays a role in the development of machine learning models ⁵¹.

When discussing IQA, it's important to differentiate "image quality" from "image gradability," terms often used interchangeably in the literature. While "quality" is overall image-based, gradability is regarding a specific disease grading. According to IRIS Reading Center, a teleophthalmology group, gradability considers clinical and anatomical features that

need to be present so that the diagnosis of a specific disease can be made. Therefore, high-quality images are not always gradable. For example, a retinal image can have good focus and adequate illumination (high quality) but may be considerably ungradable for diabetic retinopathy if the centre of the retina (i.e., macula) is not included in the image ⁵². In the EyePACS' IQA, the criteria "illumination" is image-based, while the criteria "image field definition" takes into account anatomical features like the optic disc and the macula.

Automated techniques can be applied to determine the quality of the retinal images, which was the case for the DIARETDB0 and the HEI-MED datasets. This approach, also known as objective IQA, can be time and cost-saving, especially when dealing with large amounts of image data ⁵³. Despite the potential benefits of the objective IQA, there are some challenges that limit advances in this research field. Currently, IQA algorithms are developed and tested on specific fundus image databases, and the lack of a benchmark dataset for quality assessment purposes makes it difficult to compare the performance of different models ⁵³⁻⁵⁵. Therefore, datasets containing information about the quality assessment, including the image quality criteria, are crucial for the development and evaluation of automated diagnostic models and improvements in automated IQA techniques.

Image quality criteria standardization

When reported, the image quality criteria vary widely across databases. While in the EyePACS dataset, the images were graded according to four criteria and classified into six quality classes, other datasets employed fewer criteria and quality classes. The definitions for each criteria also vary across datasets. In the EyePACS dataset, the focus of an image was considered good enough if small lesions such as microaneurysms and intraretinal microvascular abnormalities could be properly assessed. On the other hand, in the E-Ophta dataset, the image was considered of acceptable quality if the centre of the fovea and the retinal vessels were clearly visible.

The standardization of image quality criteria is challenging across different datasets and for different facilities. Further research with manual or automated quality standards and model performance analysis according to various specifications is needed.

Automatic Quality Assessment

To address the lack of standards and laborious processes in quality assessments, AI algorithms have been developed and are promising to automatically qualify an image's gradability. The deep-learning AutoMorph is a publicly available pipeline that uses retinal vasculature morphology on fundus photographs through image preprocessing, quality grading, anatomical segmentation, and morphological feature measurement ⁵⁶. The open-source Nderitu et al. algorithm detects laterality, retinal field, retinal presence, and gradability ⁵⁷.

Recommendations

We recommend diversity in datasets, including image quality, for fairness and generalizability in AI development. Data quality is important during model development; however, it should not be a deterrent to dataset creation and sharing. Retinal fundus photos quality standardization is limited, and automated processes for quality assessment are a viable alternative.

Given the importance of quality assessment when applied in fundus photos and machine learning in the DR context, we propose that, whenever possible, information on the IQA and specific grading used for labeling be reported together with other metadata.

In algorithms, we recommend the report of imaging quality criteria definitions and the parameter to include or exclude images for the model development, to avoid and address biases. The image quality criteria should be clear and as specific as possible, which contributes to greater reproducibility of the methodology.

Implications and Limitations

This is the first study to review publicly available DR datasets with a focus on the quality assessment and the quality criteria for retinal images. With this review, we hope to reinforce the importance of IQA and the need to address image quality for better algorithms. Moreover, its significance extends beyond the evaluation of DR grading alone. It sheds light on the inherent risks of biases within AI and the intricate tradeoff between model performance metrics and the effective deployment of AI systems in real-world scenarios. We also hope this

study can encourage similar research involving other diseases, imaging modalities, and medical specialties.

The Flexner Report is an example of the unintended consequence of applying strict quality standards for medical schools that led to the closure of almost all Black medical schools in America at the turn of the twentieth century⁵⁸. Quality assessment is important in imaging exams, but the construct of quality is manifold. The collection technique and the equipment characteristics, among others, contribute to the image quality. More studies are needed to address quality assessment standards, which should vary according to the research objectives, clinical use, and deployment setting.

This study has three main limitations. Firstly, there were only 20 datasets included in the review, and the study was also limited to a specific ocular disease and imaging modality. Secondly, the search was restricted to sources (e.g., websites, documentation, and publications) related to the datasets and/or screening programs. Third, we focused on the dataset's quality assessment report, with further studies evaluating models needed.

Conclusion

Image quality information is crucial not only for the development and validation of diagnostic DL algorithms but also for the improvement of automated IQA techniques. Publicly available DR datasets are valuable research tools. Feedback during the data curation, modeling, and deployment contributes to improvement in data quality assessment. Assessing image quality during model development in a clear and transparent way can improve the value of machine learning models, decrease biases, and contribute to advances in this research field.

Figure and tables

Figure 1. Flowchart of articles assessment.

Figure 2. Image quality assessment across 20 publicly available DR datasets

Table 1. Image quality criteria across publicly available DR datasets

Table 2. EyePACS quality factors for IQA

Table 3. Description of retinal image quality classification according to EyePACS

Table 4. Principles applied during the acquisition of images of the DDR dataset

References

1. Davila JR, Sengupta SS, Niziol LM, Sindal MD, Besirli CG, Upadhyaya S, et al. Predictors of Photographic Quality with a Handheld Nonmydriatic Fundus Camera Used for Screening of Vision-Threatening Diabetic Retinopathy. *Ophthalmologica*. 2017; 238(1–2): 89–99. Available at: <http://dx.doi.org/10.1159/000475773>.
2. Heydon P, Egan C, Bolter L, Chambers R, Anderson J, Aldington S, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. *Br. J. Ophthalmol*. 2020. Available at: <http://dx.doi.org/10.1136/bjophthalmol-2020-316594>.
3. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003–2016. *Acta Diabetol*. 2017; 54(6): 515–525. Available at: <https://doi.org/10.1007/s00592-017-0974-1>.
4. Nguyen HV, Tan GSW, Tapp RJ, Mital S, Ting DSW, Wong HT, et al. Cost-effectiveness of a National Telemedicine Diabetic Retinopathy Screening Program in Singapore. *Ophthalmology*. 2016; 123(12): 2571–2580. Available at: <http://dx.doi.org/10.1016/j.ophtha.2016.08.021>.
5. Huemer J, Wagner SK, Sim DA. The Evolution of Diabetic Retinopathy Screening Programmes: A Chronology of Retinal Photography from 35 mm Slides to Artificial Intelligence. *Clin. Ophthalmol*. 2020; 14: 2021–2035. Available at: <http://dx.doi.org/10.2147/OPHTH.S261629>.
6. Grzybowski A, Brona P, Lim G, Ruamviboonsuk P, Tan GSW, Abramoff M, et al. Artificial intelligence for diabetic retinopathy screening: a review. *Eye*. 2020; 34(3): 451–460. Available at: <http://dx.doi.org/10.1038/s41433-019-0566-0>.
7. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases

Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*. 2017; 318(22): 2211–2223. Available at: <http://dx.doi.org/10.1001/jama.2017.18152>.

8. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*. 2016; 316(22): 2402–2410. Available at: <http://dx.doi.org/10.1001/jama.2016.17216>.

9. Yip MYT, Lim G, Lim ZW, Nguyen QD, Chong CCY, Yu M, et al. Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. *NPJ Digit Med*. 2020; 3: 40. Available at: <http://dx.doi.org/10.1038/s41746-020-0247-1>.

10. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*. 2020. Available at: <http://www.sciencedirect.com/science/article/pii/S2589750020302405>.

11. Schaekermann M, Hammel N, Terry M, Ali TK, Liu Y, Basham B, et al. Remote Tool-Based Adjudication for Grading Diabetic Retinopathy. *Transl. Vis. Sci. Technol*. 2019; 8(6): 40. Available at: <http://dx.doi.org/10.1167/tvst.8.6.40>.

12. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology*. 2018; 125(8): 1264–1272. Available at: <http://dx.doi.org/10.1016/j.ophtha.2018.01.034>.

13. Hsu J, Phene S, Mitani A, Luo J, Hammel N, Krause J, et al. Improving Medical Annotation Quality to Decrease Labeling Burden Using Stratified Noisy Cross-Validation. *arXiv [cs.LG]*. 2020. Available at: <http://arxiv.org/abs/2009.10858>.

14. Fu H, Wang B, Shen J, Cui S, Xu Y, Liu J, et al. Evaluation of Retinal Image Quality Assessment Networks in Different Color-spaces. *arXiv [cs.CV]*. 2019. Available at: <http://arxiv.org/abs/1907.05345>.

15. Li Z, Guo C, Nie D, Lin D, Zhu Y, Chen C, et al. Deep learning from “passive feeding” to “selective eating” of real-world data. *NPJ Digit Med*. 2020; 3: 143. Available at: <http://dx.doi.org/10.1038/s41746-020-00350-y>.

16. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. *Br. J. Ophthalmol*. 2019; 103(2): 167–175. Available at: <http://dx.doi.org/10.1136/bjophthalmol-2018-313173>.

17. Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, et al. Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems. *Diabetes Care*. 2021; 44(5): 1168–1175. Available at: <http://dx.doi.org/10.2337/dc20-1877>.

18. Heaven WD. Google’s medical AI was super accurate in a lab. Real life was a different story. *Technol. Rev*. 2020. Available at: <https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab->

real-life-clinic-covid-diabetes-retina-disease/ [Accessed February 2, 2023].

19. Lamirel C, Bruce BB, Wright DW, Delaney KP, Newman NJ, Biousse V. Quality of nonmydriatic digital fundus photography obtained by nurse practitioners in the emergency department: the FOTO-ED study. *Ophthalmology*. 2012; 119(3): 617–624. Available at: <http://dx.doi.org/10.1016/j.ophtha.2011.09.013>.
20. Malerbi FK, Morales PH, Farah ME, Drummond KRG, Mattos TCL, Pinheiro AA, et al. Comparison between binocular indirect ophthalmoscopy and digital retinography for diabetic retinopathy screening: the multicenter Brazilian Type 1 Diabetes Study. *Diabetol. Metab. Syndr*. 2015; 7(1): 116. Available at: <http://dx.doi.org/10.1186/s13098-015-0110-8>.
21. Paulus J, Meier J, Bock R, Hornegger J, Michelson G. Automated quality assessment of retinal fundus photos. *Int. J. Comput. Assist. Radiol. Surg*. 2010; 5(6): 557–564. Available at: <http://dx.doi.org/10.1007/s11548-010-0479-7>.
22. Karlsson RA, Jonsson BA, Hardarson SH, Olafsdottir OB, Halldorsson GH, Stefansson E. Automatic fundus image quality assessment on a continuous scale. *Comput. Biol. Med*. 2020: 104114. Available at: <http://www.sciencedirect.com/science/article/pii/S0010482520304455>.
23. Anon. New Developments in Biomedical Engineering.
24. Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *J. Diabetes Sci. Technol*. 2009; 3(3): 509–516. Available at: <http://dx.doi.org/10.1177/193229680900300315>.
25. EyePACS. EyePACS DIGITAL RETINAL IMAGE GRADING PROTOCOL NARRATIVE. Available at: <https://www.eyepacs.org/consultant/Clinical/grading/EyePACS-DIGITAL-RETINAL-IMAGE-GRADING.pdf>.
26. Decencière E, Cazuguel G, Zhang X, Thibault G, Klein J-C, Meyer F, et al. TeleOphta: Machine learning and image processing methods for teleophthalmology. *IRBM*. 2013; 34(2): 196–203. Available at: <http://www.sciencedirect.com/science/article/pii/S1959031813000237>.
27. Erginay A, Chabouis A, Viens-Bitker C, Robert N, Lecleire-Collet A, Massin P. OPHDIAT: quality-assurance programme plan and performance of the network. *Diabetes Metab*. 2008; 34(3): 235–242. Available at: <http://dx.doi.org/10.1016/j.diabet.2008.01.004>.
28. Niemeijer M, van Ginneken B, Cree MJ, Mizutani A, Quellec G, Sanchez CI, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE Trans. Med. Imaging*. 2010; 29(1): 185–195. Available at: <http://dx.doi.org/10.1109/TMI.2009.2033909>.
29. Abramoff MD, Suttrop-Schulten MSA. Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. *Telemed. J. E. Health*. 2005; 11(6): 668–674. Available at: <http://dx.doi.org/10.1089/tmj.2005.11.668>.
30. Pires R, Jelinek HF, Wainer J, Goldenstein S, Valle E, Rocha A. Assessing the Need for

Referral in Automatic Diabetic Retinopathy Detection. *IEEE Trans. Biomed. Eng.* 2013; 60(12): 3391–3398. Available at: <http://dx.doi.org/10.1109/TBME.2013.2278845>.

31. Pires R, Jelinek HF, Wainer J, Valle E, Rocha A. Advancing bag-of-visual-words representations for lesion classification in retinal images. *PLoS One.* 2014; 9(6): e96814. Available at: <http://dx.doi.org/10.1371/journal.pone.0096814>.

32. Pires R, Jelinek HF, Wainer J, Rocha A. Retinal Image Quality Analysis for Automatic Diabetic Retinopathy Detection. In: 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images.; 2012. pp. 229–236. Available at: <http://dx.doi.org/10.1109/SIBGRAPI.2012.39>.

33. Giancardo L, Meriaudeau F, Karnowski TP, Li Y, Garg S, Tobin KW Jr, et al. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. *Med. Image Anal.* 2012; 16(1): 216–226. Available at: <http://dx.doi.org/10.1016/j.media.2011.07.004>.

34. Giancardo L, Meriaudeau F, Thomas P, Chaum E, Tobi K. Quality Assessment of Retinal Fundus Images using Elliptical Local Vessel Density. *New Developments in Biomedical Engineering.* 2010. Available at: <http://dx.doi.org/10.5772/7618>.

35. Joshi VS, Reinhardt JM, Garvin MK, Abramoff MD. Automated method for identification and artery-venous classification of vessel trees in retinal vessel networks. *PLoS One.* 2014; 9(2): e88061. Available at: <http://dx.doi.org/10.1371/journal.pone.0088061>.

36. Şevik U, Köse C, Berber T, Erdöl H. Identification of suitable fundus images using automated quality assessment methods. *J. Biomed. Opt.* 2014; 19(4): 046006. Available at: <http://dx.doi.org/10.1117/1.JBO.19.4.046006>.

37. Lin L, Li M, Huang Y, Cheng P, Xia H, Wang K, et al. The SUSTech-SYSU dataset for automated exudate detection and diabetic retinopathy grading. *Sci Data.* 2020; 7(1): 409. Available at: <http://dx.doi.org/10.1038/s41597-020-00755-0>.

38. Li T, Gao Y, Wang K, Guo S, Liu H, Kang H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf. Sci.* . 2019; 501: 511–522. Available at: <http://www.sciencedirect.com/science/article/pii/S0020025519305377>.

39. Kauppi T, Kamarainen VKJ-K, Lensu L, Uusitalo ISH, Kalviainen H, Pietila J. DIARETDB0: Evaluation database and methodology for diabetic retinopathy algorithms. Available at: https://www.it.lut.fi/project/imageret/diaretdb0/doc/diaretdb0_techreport_v_1_1.pdf [Accessed December 29, 2020].

40. Kauppi T, Kalesnykiene V, Kamarainen J-K, Lensu L, Sorri I, Raninen A, et al. the DIARETDB1 diabetic retinopathy database and evaluation protocol. In: Proceedings of the British Machine Vision Conference 2007. British Machine Vision Association; 2007. Available at: http://www2.it.lut.fi/project/imageret/diaretdb1/doc/diaretdb1_techreport_v_1_1.pdf [Accessed December 29, 2020].

41. Kauppi T, Kamarainen J-K, Lensu L, Kalesnykiene V, Sorri I, Uusitalo H, et al. A framework for constructing benchmark databases and protocols for retinopathy in medical image analysis. In: Intelligent Science and Intelligent Data Engineering. Lecture notes in computer

science. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. pp. 832–843. Available at: <http://vision.cs.tut.fi/data/publications/iscide2012.pdf> [Accessed January 1, 2021].

42. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. *Brown Univ. Dig. Addict. Theory Appl.* 2018; 3(3): 25. Available at: <https://www.mdpi.com/2306-5729/3/3/25> [Accessed December 29, 2020].

43. Porwal P, Pachade S, Kokare M, Deshmukh G, Son J, Bae W, et al. IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge. *Med. Image Anal.* 2020; 59: 101561. Available at: <http://dx.doi.org/10.1016/j.media.2019.101561>.

44. Staal J, Abramoff MD, Niemeijer M, Viergever MA, van Ginneken B. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging.* 2004; 23(4): 501–509. Available at: <http://dx.doi.org/10.1109/TMI.2004.825627>.

45. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, et al. Feedback on a publicly distributed image database: The Messidor database. *Image Anal. Stereol.* 2014; 33(3): 231. Available at: <https://www.ias-iss.org/ojs/IAS/article/view/1155>.

46. Al-Diri B, Hunter A, Steel D, Habib M, Hudaib T, Berry S. REVIEW - a reference data set for retinal vessel profiles. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2008; 2008: 2262–2265. Available at: <http://dx.doi.org/10.1109/IEMBS.2008.4649647>.

47. Adal KM, van Etten PG, Martinez JP, van Vliet LJ, Vermeer KA. Accuracy assessment of intra- and intervisit fundus image registration for diabetic retinopathy screening. *Invest. Ophthalmol. Vis. Sci.* 2015; 56(3): 1805–1812. Available at: <http://dx.doi.org/10.1167/iovs.14-15949>.

48. Lowell J, Hunter A, Steel D, Basu A, Ryder R, Fletcher E, et al. Optic nerve head segmentation. *IEEE Trans. Med. Imaging.* 2004; 23(2): 256–264. Available at: <http://dx.doi.org/10.1109/TMI.2003.823261>.

49. Prentasic P, Loncaric S, Vatauvuk Z, Bencic G, Subasic M, Petkovic T, et al. Diabetic retinopathy image database(DRiDB): A new database for diabetic retinopathy screening programs research. 2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA). 2013. Available at: <http://dx.doi.org/10.1109/ispa.2013.6703830>.

50. Seastedt KP, Schwab P, O'Brien Z, Wakida E, Herrera K, Marcelo PGF, et al. Global healthcare fairness: We should be sharing more, not less, data. *PLOS Digit Health.* 2022; 1(10): e0000102. Available at: <https://journals.plos.org/digitalhealth/article/file?id=10.1371/journal.pdig.0000102&type=printable>.

51. Cheng J, Li Z, Gu Z, Fu H, Wong DWK, Liu J. Structure-preserving Guided Retinal Image Filtering and Its Application for Optic Disc Analysis. *arXiv [cs.CV]*. 2018. Available at: <http://arxiv.org/abs/1805.06625>.

52. iris_dev_. Image quality vs. Gradeability: The IRIS difference. *IRIS.* 2019. Available at: <https://retinalscreenings.com/blog/image-quality-vs-gradeability-the-iris-difference/>

[Accessed December 7, 2022].

53. Raj A, Tiwari AK, Martini MG. Fundus image quality assessment: survey, challenges, and future scope. *IET Image Proc.* 2019; 13(8): 1211–1224. Available at: <https://digital-library.theiet.org/content/journals/10.1049/iet-ipr.2018.6212> [Accessed November 22, 2020].

54. Zago GT, Andreão RV, Dorizzi B, Teatini Salles EO. Retinal image quality assessment using deep learning. *Comput. Biol. Med.* 2018; 103: 64–70. Available at: <http://dx.doi.org/10.1016/j.combiomed.2018.10.004>.

55. Lin J, Yu L, Weng Q, Zheng X. Retinal image quality assessment for diabetic retinopathy screening: A survey. *Multimed. Tools Appl.* 2020; 79(23): 16173–16199. Available at: <https://doi.org/10.1007/s11042-019-07751-6>.

56. Zhou Y, Wagner SK, Chia MA, Zhao A, Woodward-Court P, Xu M, et al. AutoMorph: Automated Retinal Vascular Morphology Quantification Via a Deep Learning Pipeline. *Transl. Vis. Sci. Technol.* 2022; 11(7): 12. Available at: <http://dx.doi.org/10.1167/tvst.11.7.12>.

57. Nderitu P, Nunez do Rio JM, Webster ML, Mann SS, Hopkins D, Cardoso MJ, et al. Automated image curation in diabetic retinopathy screening using deep learning. *Sci. Rep.* 2022; 12(1): 11196. Available at: <http://dx.doi.org/10.1038/s41598-022-15491-1>.

58. Flexner A. Medical Education in the United States and Canada: A Report to the Carnegie Foundation for the Advancement of Teaching (PDF). OCLC. 1910; 9795002.

Disclosure:

Dr. Keane has acted as a consultant for Google, DeepMind, Roche, Novartis, Apellis, and BitFount and is an equity owner in Big Picture Medical. He has received speaker fees from Heidelberg Engineering, Topcon, Allergan, and Bayer.

Edward Korot: Sanro Health: Owner, Alimera sciences: Consultant, Genentech: consultant (in past six months, not current)

Acknowledgments:

Dr. Keane is supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1).

Luis Filipe Nakayama is a researcher supported by Lemann Foundation, Instituto da Visão-IPEPO, São Paulo, Brazil

Authors roles:

MBG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing-original draft, Writing-review & editing

LFN: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing-original draft, Writing-review & editing

DF: Writing-review & editing

HF: Writing-review & editing

EK: Writing-review & editing

FKM: Writing-review & editing

CVSR: Writing-review & editing

MM: Writing-review & editing

LAC: Conceptualization, Methodology, Supervision, Writing-review & editing

PK: Supervision, Writing-review & editing

RBJ: Conceptualization, Methodology, Supervision, Writing-review & editing