# Assessing Potential Bias from Metadata while Labeling Retinal Images Fundus Photographs for Diabetic Retinopathy: Preliminary Experience in the Multimodal Database of Retinal Images in Africa (MoDRIA)

Simon Arunga, Katharine Elise Morley, Teddy Kwaga, Michael Gerard Morley, Luis Filipe Nakayama, Rogers Mwavu, Fred Kaggwa, Julius Ssempiira, Leo Anthony Celi, Jessica Haberer, Celestino Obua

Submitted to: JMIR Formative Research
on: April 29, 2024

# *Table of Contents*

# Assessing Potential Bias from Metadata while Labeling Retinal Images Fundus Photographs for Diabetic Retinopathy: Preliminary Experience in the Multimodal Database of Retinal Images in Africa (MoDRIA)

Simon Arunga[1*] MD, PhD; Katharine Elise Morley[2*] MD, MPH; Teddy Kwaga[1] MD, MMed; Michael Gerard Morley[3] MD, ScM; Luis Filipe Nakayama[4, 5] MD, PhD; Rogers Mwavu[6] BIT, MSIS; Fred Kaggwa[6] PhD; Julius Ssempiira[7] PhD; Leo Anthony Celi[5, 8, 9] MD, MPH, MSc; Jessica Haberer[2] MD, MS; Celestino Obua[10] MD, MSc, PhD

[1]Department of Ophthalmology Mbarara University of Science and Technology Mbarara UG
[2]Massachusetts General Hospital Center for Global Health Department of Medicine Harvard Medical School Boston, MA US
[3]Harvard Ophthalmology AI Lab Massachusetts Eye and Ear Infirmary Harvard Medical School Boston US
[4]Ophthalmology Department Sao Paulo Federal University Sao Paulo BR
[5]Laboratory for Computational Physiology Massachusetts Institute of Technology Cambridge US
[6]Faculty of Computing and Informatics Department of Computer Science Mbarara University of Science and Technology Mbarara UG
[7]School of Public Health Makerere University Kampala UG
[8]Department of Biostatistics Harvard T.H. Chan School of Public Health Boston US
[9]Division of Pulmonary, Critical Care and Sleep Medicine Beth Israel Deaconess Medical Center Harvard Medical School Boston US
[10]Mbarara University of Science and Technology Mbarara UG
[*]these authors contributed equally

**Corresponding Author:**
Katharine Elise Morley MD, MPH
Massachusetts General Hospital Center for Global Health
Department of Medicine
Harvard Medical School
125 Nashua St. Suite 722
Boston, MA
US



## *Abstract*

**Background:** Labeling color fundus photos (CFP) is an important step in the development of artificial intelligence (AI) screening algorithms for the detection of diabetic retinopathy. The International Classification of Diabetic Retinopathy (ICDR) is used to assign labels to CFP, plus the presence or absence of macular edema. Images can be grouped as referrable or non-referrable for treatment. There is little guidance in the literature about how to collect and use clinical metadata as a part of the CFP labeling process.

**Objective:** To improve the quality of the Multimodal Database of Retinal Images in Africa (MoDRIA) by determining whether the availability of clinical metadata during the image labeling process influences the accuracy, sensitivity, and specificity of image labels.

**Methods:** This is a crossover assessment with 2 groups and 2 phases. Each group had 10 randomly assigned labelers who provided an ICDR score and presence or absence of macular edema for each of 50 CRF in a test image with and without metadata.
Specificity and sensitivity of referable retinopathy was based on ICDR scores, and macular edema calculated using 2-sided T-test. Comparison with and without metadata for each participant was calculated using the signed rank test. Statistical significance was set at $P<0.05$.

**Results:** The sensitivity for identifying referrable diabetic retinopathy with metadata was 92.8% (95% CI: 87.6-98.0) compared with 93.3% (95% CI: 87.6-98.9) without metadata, and the specificity was 84.9% (95% CI: 75.1-94.6) with metadata compared with 88.2% (95% CI: 79.5-96.8) without metadata. The sensitivity for identifying the presence of macular edema was 64.3% (95% CI: 57.6-71.0) with metadata, compared with 63.1% (95% CI: 53.4-73.0) without metadata, and the specificity was 86.5% (95% CI: 81.4-91.5) with metadata compared with 87.7% (95% CI: 83.9-91.5) without metadata. Sensitivity and specificity of

ICDR score and presence or absence of ME were also calculated for the 20 individual labelers with and without metadata. No findings were statistically significant.

**Conclusions:** In this quality improvement project, clinical metadata availability did not influence labeling quality. Additional studies are needed to understand the potential implications of the process and components of labeling with and without metadata more thoroughly with regards to accuracy and bias. These issues have far reaching implications given the rapidly expanding use of AI with clinical images, including on the African continent.

(JMIR Preprints 29/04/2024:59914)
DOI: https://doi.org/10.2196/preprints.59914

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

# Original Manuscript

## Introduction

Imaging exams in ophthalmology serve as a tool for diagnosing and following up ocular pathologies and play a critical role in the management of diabetic retinopathy (DR). Retinal color fundus photos specifically capture the ocular posterior segment, comprising the retina, optic disc, macula, and vessels, offering crucial information about ocular and systemic health during ophthalmological examinations.[1] Diabetes is a global epidemic, affecting more than 500 million people in 2021 and a projected 783 million by 2045, with diabetic retinopathy as the most common complication of systemic diabetes.[2] Retinal color fundus photographs (CFP) have been used for screening of referable cases, optimizing the referral process worldwide, and more recently they have been employed in the development of artificial intelligence (AI) algorithms for automatic diabetic retinopathy screening. [3]

### Classification of diabetic retinopathy

In DR screening algorithms developed using supervised machine learning,[4] an important step in the process is labeling the CFPs; these labels indicate the presence and severity of DR and macular edema (ME) for training the AI model. Most studies use a two-image capturing protocol using the International Classification of Diabetic Retinopathy (ICDR), [5] which has 5 levels of severity (Table 1): 0- no retinopathy, 1-microaneurysms only, 2- hemorrhages 3- proliferative 4- proliferative retinopathy. It has been proven effective in comparison with the gold standard Early Treatment Diabetic Retinopathy Study (ETDRS) field protocol.[6] Individuals with pre-proliferative (3) and proliferative (4) retinopathy are candidates for treatment intervention with laser, anti-VEGF drugs or surgery. The presence of ME is another important criterion for treatment intervention. A key goal for AI screening algorithms is to identify patients with DR who need referral for potential treatment.

**Table 1: International Classification of Diabetic Retinopathy (ICDR)**[5]

| | | |
|---|---|---|
| ICDR severity level | 0 | No retinopathy - No abnormalities |
| | 1 | Mild non-proliferative retinopathy - Microaneurysm(s) only |
| | 2 | Moderate non-proliferative retinopathy - More than just microaneurysm(s) but less than severe non-proliferative diabetic retinopathy |
| | 3 | Severe non-proliferative or pre-proliferative retinopathy: Any of the following: >20 intra-retinal hemorrhages in each of 4 quadrants, venous beading in ≥ 2 quadrants, intraretinal microvascular abnormalities in ≥1 quadrant, and no signs of proliferative retinopathy |
| | 4 | Proliferative retinopathy - One or more of the following: neovascularization and/or vitreous or pre-retinal hemorrhages |
| Macula edema | | Exudates or apparent thickening within one disc diameter from the fovea |

### Background on fundus image labeling and use of metadata for AI algorithm development

Labeling large numbers of CFPs has many challenges. Strategies used include recruiting highly trained retinal specialists, comprehensive ophthalmologists, [7] professional labelers, and crowdsourcing using labelers with different backgrounds and experience, [8] and more recently, unsupervised learning with deep learning algorithms.[9] Another variable is the availability and use of metadata during the labeling process. Metadata for medical imaging can include information generated from the imaging device and process itself such as order codes and image files, along

with other biomarkers, demographics, and clinical information related to the image [10]. When an electronic medical record is available, the medical history, diagnostic results, and the clinical assessment and plan may be linked to the image. The actual image interpretation may also be present as in the case of radiology or pathology reports. In the absence of an integrated electronic record, as is typically the case in low resource settings, any additional clinical information must be collected separately and linked to the image.

The use of local data is crucial for AI development and validation, yet automated systems face a critical risk of biased decisions based on this information.[11] In practice, the clinician makes a diagnosis using *all* the available information about the patient including history, exam findings, diagnostic tests, and imaging. But labeling is frequently done with only the image (i.e., no additional clinical metadata).[12,13] In their paper on image labeling quality control, Freeman, et al, reported that the gap between the clinical and labeling contexts is a challenge in optimizing the accuracy of labels [14]. The label tends to be given as an overall impression of the findings. They stressed the importance of having labeling criteria and guidelines explicitly focused on the labeling task to improve consistency and inferred that it does not include other clinical information. Alternatively, Kondylakis et al, [10] state that metadata are essential for the correct use and interpretation of medical images and stressed the importance of data harmonization to use this information in the development of AI models. The importance of incorporating clinical information as a multimodal data stream has been increasingly recognized in the development of radiology algorithms. [15,16] The availability of correct clinical information has been shown to improve the interpretations of diagnostic tests[17] accuracy of computerized tomography interpretation by radiologists, [18] and interpretation of radiological imaging [19] in addition to the impact of including age and gender in DR screening algorithms.[20]

AI algorithms have been touted as a means of improving health care access in low resource settings. [21] Many existing algorithms have been developed from images obtained from only the United States, Europe, and China. There is a near lack of such data from the African continent raising concerns about generalizability, accuracy, and bias.[22] However, collecting even basic clinical information in low resource settings is difficult, as existing medical records typically have less detailed information than those in high resource settings and may be paper based; the available results and findings are often incomplete and less accurate. Prospective clinical metadata collection at the time of image capture is also limited by patient health literacy and knowledge about their health conditions.

**Project objective**

Despite the importance of high-quality labels for optimizing algorithm performance, [23] there is little guidance in the literature about how to collect and use clinical metadata for algorithm development in low resource settings. The Multimodal Database of Retinal Images in Africa (MoDRIA) is an one of the inaugural research projects of the Mbarara University Data Science Research Hub (MUDSReH), [24] part of the Data Science for Health Discovery and Innovation in Africa (DS-I Africa)[25] initiative to “advance Data Science and related innovations in Africa to create an ecosystem that can begin to provide local solutions to countries’ most immediate public health problems through advances in research”. Here, we present an analysis to determine whether the availability of clinical metadata during the image labeling process influences the accuracy, sensitivity and specificity of image labels provided by newly trained labelers when using a known set of properly labeled images.

## Methods

### Ethics Approval

This work was part of the ongoing MoDRIA study (Mbarara University of Science and Technology IRB approval number: MUST-2021-239 and Uganda National Council of Science and Technology number: HS2094ES) as a quality improvement project to improve the training of CFP readers and optimize the labeling protocol of the MoDRIA fundus image database Uganda.

### Setting

This project was conducted at the Mbarara University of Science and Technology (MUST) in Mbarara, Uganda in November 2023. MUST is the site of the Mbarara University Data Science Research Hub (MUDSReH) and the MoDRIA research project. MUST is also the parent institution for the Mbarara Regional Referral Hospital in southwestern Uganda and located 268 kilometers southwest from the capital of Kampala.

### Project participants

The project participants were 20 Ugandan pre-interns (i.e., medical school graduates awaiting commencement of their internship). These “MoDRIA labelers” completed a labeling training course consisting of 40 hours of teaching, training, supervised labeling, and testing by Ugandan ophthalmologists and ophthalmology residents, and 2 international visiting retinal specialists. The training course content included 1) Review of the BRSET image reading training manual, [26] 2). Videos and didactic lectures on retinal anatomy, macular edema, DR abnormalities in each ICDR category, and macular edema and a 4-day hands-on workshop in which MoDRIA labelers practiced labeling a minimum of 200 CFPs followed by tests to confirm labeler competency and accuracy by test set labeling.

### Data collection

*Metadata*

This project used clinical metadata only and included blood pressure; visual acuity; blood glucose; presence of diabetes, hypertension and/or HIV; and class of medications taken. To ensure all metadata elements were available for all test images, metadata values were synthesized to align with the ICDR scores of the test image. The metadata for each image was presented in a spreadsheet with the image number and fields to enter the ICDR score and ME assessment. The images appeared on the same screen.

*Image sets*

The MoDRIA database contains 14,000 CFP from 3,500 individuals. Each study participant has 4 CFPs (disc center and macular center view from right and left eyes). The MoDRIA database will be used to develop AI algorithms to screen patients for posterior segment retinal diseases such as diabetic retinopathy. The MoDRIA CFP labeling protocol was based on the Brazilian Diabetic Retinopathy fundus image dataset (BRSET) labeling protocol.[27] It is a publicly available collection of 16,000 retinal fundus images collected and labeled in Brazil.

MoDRIA CFPs were collected on 3-Nethra Classic (Forus Royal, India) fundus cameras by ophthalmic technicians trained in fundus photography. BRSET images were collected on Nikon NF505 (Nikon, Japan) and a Canon CR-2 (Canon Inc, USA) in JPEG format, and no preprocessing techniques were

applied. There were 50 CFPs in the test set for this study, 20 from MoDRIA and 30 from the BRSET. The ICDR and DME scores of the BRSET and MoDRIA test set images were reviewed and confirmed by the international retina specialists participating in the study (LN and MM). The distribution of ICDR scores and presence/absence of DME in the test set is presented in Table 2, with approximately half the images being normal.

Table 2: Description of Color Fundus Photos Used in Labeling Test Set (N=50, each image scored for ICDR and ME)

| | Non-referable N=28 | | | Referable N=22 | | | |
|---|---|---|---|---|---|---|---|
| | ICDR ≤ 1 | | Macular edema | ICDR > 2 | | | Macular edema |
| Score | 0 | 1 | Absent | 2 | 3 | 4 | Present |
| # of images | 26 | 4 | 40 | 6 | 6 | 8 | 10* |

*Labeling protocol*

Each CFP was individually labeled for diabetic retinopathy with an ICDR score of 0-4, ranging from 0 (no retinopathy) to 4 (proliferative DR). (Table 1) These scores were grouped into 2 categories: non-referable (ICDR ≤ 1 and no ME) and referrable (ICDR > 2 and/or with ME). The same CFPs were also labeled with the presence or absence of macular edema.

*Project design*

This is a crossover assessment with 2 groups and 2 phases. Each group had 10 randomly assigned labelers who labeled the same image test set of 50 CFP twice (Phase 1 and Phase 2) with the ICDR score and presence or absence of ME. Group 1 ("with/without") labeled the CFPs with metadata in Phase 1 and without metadata in Phase 2. Group 2 ("without/with") labeled the CFP in Phase 1 without metadata and with metadata in Phase 2. In Phase 2, the order of presentation for the same CFPs was scrambled for both groups (Figure 1).

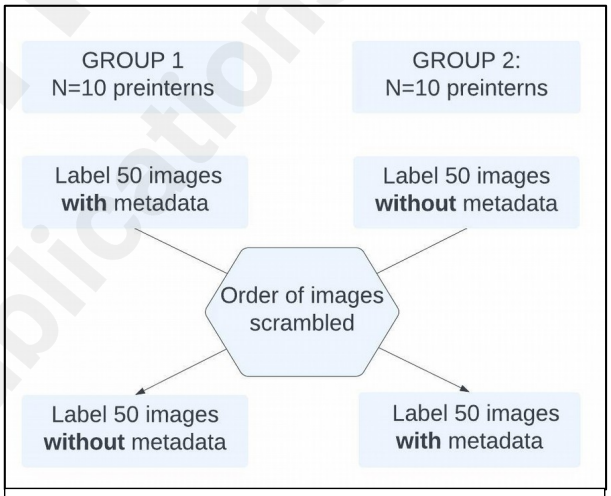


Figure 1: Project design

*Statistical analysis*

Statistical analysis was conducted using STATA 17.0. ICDR scores were grouped into referable (ICDR 2-4 +/- ME) and non-referable categories (ICDR 0-1 and no ME) for statistical analysis. Specificity and sensitivity of referable retinopathy was based on ICDR scores and ME calculated using 2-sided T-test. Comparison of sensitivity and specificity for ICDR and ME with and without metadata for each participant was calculated using the signed rank test. Statistical significance was set at $P<0.05$.

## Results

Table 3 lists the sensitivity and specificity of referable retinopathy based on both ICDR scores calculated with and without metadata. Sensitivity and specificity of ICDR score and presence or absence of ME were also calculated for the 20 individual labelers with and without metadata. (see supplemental data) There were no statistically significant differences with and without metadata for any of the labelers.

**Diabetic retinopathy**

The sensitivity for identifying referrable DR with metadata was 92.8% (95% CI: 87.6-98.0) compared

with 93.3% (95% CI: 87.6-98.9) without metadata, and the specificity was 84.9% (95% CI: 75.1-94.6) with metadata compared with 88.2% (95% CI: 79.5-96.8) without metadata. The improvements in sensitivity and specificity without metadata were not statistically significant.

**Macular edema**

The sensitivity for identifying the presence of macular edema was 64.3% (95% CI: 57.6-71.0) with metadata, compared with 63.1% (95% CI: 53.4-73.0) without metadata, and the specificity was 86.5% (95% CI: 81.4-91.5) with metadata compared with 87.7% (95% CI: 83.9-91.5) without metadata. The improvements in sensitivity and specificity without metadata were not statistically significant.

Table 3: Comparison of Sensitivity and Specificity of Labeling Color Fundus Photo as Referable or Non-referable With and Without Metadata for all Labelers (N=20)

| Diagnostic measure | ICDR: referable vs non-referable Mean (95%CI) | | | Macular edema: present vs. absent Mean (95%CI) | | |
|---|---|---|---|---|---|---|
| | With metadata | No metadata | P-value | With metadata | No metadata | P-value |
| Sensitivity | 92.8 (87.6, 98.0) | 93.3 (87.6, 98.9) | 0.90 | 64.3 (57.6, 71.0) | 63.1 (53.4, 73.0) | 0.60 |
| Specificity | 84.9 (75.1, 94.6) | 88.2 (79.5, 96.8) | 0.84 | 86.5 (81.4,91.5) | 87.7 (83.9,91.5) | 0.69 |

## Discussion

**Principal results**

The objective of our project was to understand the impact of metadata on CFP labeling for DR and ME and to use this information to improve the quality of the labeling process. This assessment serves as a baseline for future iterative improvements in the training of labelers and the labeling process. Our results can also inform a more rigorous investigation of the role of metadata in the labeling process for the MoDRIA dataset as well as other datasets developed through MUDSReH, the DS-I for Africa, and others.

As a group, the labelers detected referrable DR reasonably well (92.8%) but detected ME only 64.3% of the time. This difference may be a result of the subtle appearance of hard exudates on ME when there are only cystic changes or a blunted foveal reflex rather than the presence of more obvious lipid. In screening programs, the false negative rate (failing to identify the condition when it is present) is the most potentially dangerous error. Given the more subtle presentation on ME CFPs, optical coherence tomography, which easily identifies ME, is a valuable complementary tool to CFPs in screening for referrable DR if available.

Overall, the sensitivity and specificity scores tended to be slightly better without metadata, but the difference was not statistically significant. The wide confidence intervals noted in the data reflect the variation in our labelers. We cannot make definitive conclusions about the impact of clinical metadata on the sensitivity and specificity of image labels in our study. However, given the

importance of metadata in clinical situations (and its enhancement of model performance in many cases), we believe that it may benefit labeling quality as well. For example, mild DR, HTN retinopathy, and HIV retinopathy can have a similar appearance on CFP and be difficult to differentiate with just a single image. Another consideration is whether knowing the metadata ahead of determining the labels may have introduced bias on the part of the labelers. For example, if the labeler sees the individual has a history of diabetes and elevated blood glucose, they may be more likely to give a higher ICDR score.

**Comparison with previous work**

Few other studies have been published on the impact of using metadata in labeling CFP. We conducted MEDLINE search using Medical Subject Headings: “fundus image” and “metadata”, “Image grading” and “metadata”, “fundus photo” and “metadata”, “image grading” and “clinical information” to search for previous studies evaluating the impact of using metadata or clinical information in the CFP labeling process. Additional free text topics heading searches with the same terms were also conducted without finding other dedicated studies using metadata in the CFP labeling process. We also examined the labeling protocols for the following large open source fundus photo datasets - Messidor[28], BRSET[27], Eye Pacs[29], and IDRiD[30] and did not find documentation indicating whether metadata was used or not used in the labeling process.

**Limitations**

We acknowledge several important limitations of our project. First, our assessment design did not include a defined step in the process where the labelers confirmed review of the metadata. It was provided on the screen at the time of labeling, and they were encouraged to use it, but there was no step confirming whether it was viewed. Second, we selected a sample size of 50 images, which may not have been large enough given that half the images were normal exams. This distribution of ICDR categories was intentionally chosen to better reflect the composition of the MoDRIA database; however, it may have introduced some bias as the distribution across categories was not even. Third, the focus of labeler training was to familiarize themselves with CFPs of normal and DR images, as well as other common retinal pathology. The use of metadata to inform labeling decisions tended to be subsumed by learning retinal image pathology. This process may have influenced if and how they used the metadata. Fourth, the images were labeled with ICDR scores 0-4, but our analysis was based on a binary classification of referable or non-referable DR. Finally, our metadata was synthesized based on the ICDR score and presence or absence of ME therefore may not the same is using available clinical metadata.

**Strengths**

Our project also has several strengths. To our knowledge, it is the first attempt to understand the role of metadata in CFP image labeling by a cadre of non-ophthalmologists in Africa. It is critically important in building local image labeling capacity to support the development and implementation of data science research and technologies in Africa and avoid the expansion on digital sweatshops in Africa. [31] It also provided experience using a quality improvement approach to improve image labeling and training for the researchers and clinicians at the MUDSReH. An advantage of a quality improvement approach is the ability to rapidly identify actionable results, such as the need for additional training on recognizing ME. Finally, this project highlighted the importance of understanding metadata and the need to conduct further rigorous investigations.

**Opportunities for improvement and future study**

As this was a quality improvement project, we sought opportunities for improvement in our labeling process. Specifically, we identified the following items 1) defined guidelines for reviewing metadata in the labeling process, including when it should be reviewed; 2) add a field confirmation review of metadata in the MoDRIA Data Collection and Management application developed by the MUDSReH hub team; and 3) enhanced training on appearance of ME on CFP. We also identified several areas for future study. First, we intend to perform a more rigorous, sufficiently powered study to determine the sensitivity and specificity of CFP labels with and without metadata using a cohort of images from patients with DM without HIV or hypertension with higher percentage of abnormal images. This approach will also allow analysis by individual ICDR scores rather than referable/non-referable category, so we have a more nuanced understanding of the impact of metadata on labels and algorithm performance. Given the challenge of metadata collection in this low resource environment, we also plan to determine which metadata variables are most informative in accurately predicting referrable DR. Finally, we will assess the optimal timing and method to present metadata to labelers, as well as determine intra-rater reliability with and without metadata.

## Conclusion

In this quality improvement project, clinical metadata availability did not influence labeling quality. Additional studies are needed to understand the potential implications of the process and components of labeling with and without metadata more thoroughly with regards to accuracy and bias. These issues have far reaching implications given the rapidly expanding use of AI with clinical images, including on the African continent.

## Funding

The research reported in this publication was supported by the Fogarty International Center of the National Institutes of Health under Award Number 1U54TW012043. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Acknowledgments

The authors wish to acknowledge the hard work and dedication of the following groups and individuals:

**MoDRIA and MUDReSH staff:** Gerald Ddumba (MoDRIA Research assistant), Vicent Balitema (MoDRIA coordinator), Lawrance Tebandeke (MoDRIA coordinator), Amos Baryashaba (Infrastructure Engineer)

**Mbarara University of Science and Technology ophthalmology residents***:* Dr Apap Jocef, Dr Angela Birungi, Dr Flora Patrice, Dr Jessica Kabejja.

**MoDRIA image labelers:** Josephine G. Ajolorwoth, Ebenezer Asiimwe, Lorna Atimango, Namara Boaz, William Byansi, Andrew Kasagga, Edmund Katambira, Evelyn B Kirabo, Moses Kwesiga, Racheal Nagasha, Abraham Nduhukire, Racheal Ninsiima, Saulo Nkuratiire, David Nyombi, Ronald Awanii Okii, Jordan Ssemwogerere, Francis Ssengoba, Tophias Tumwebaze, Lenus Tumwekwatse, Joy Queen Uwihirwe.

## References


1. Benet D, Pellicer-Valero OJ. Artificial intelligence: the unstoppable revolution in ophthalmology. *Surv Ophthalmol*. 2022;67(1):252-270. doi:10.1016/j.survophthal.2021.03.003

2. Sun H, Saeedi P, Karuranga S, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract*. 2022;183:109119. doi:10.1016/j.diabres.2021.109119

3. 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(22):2402-2410. doi:10.1001/jama.2016.17216

4. Saraswat P. Supervised Machine Learning Algorithm: A Review of Classification Techniques. In: *International Conference on Intelligent Emerging Methods of Artificial Intelligence & Cloud Computing*. Springer International Publishing; 2022:477-482. doi:10.1007/978-3-030-92905-3_58

5. Wilkinson CP, Ferris FL 3rd, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110(9):1677-1682. doi:10.1016/S0161-6420(03)00475-5

6. Esmaeilkhanian H, Liu H, Fasih-Ahmed S, et al. The relationship of diabetic retinopathy severity scales with frequency and surface area of diabetic retinopathy lesions. *Graefes Arch Clin Exp Ophthalmol*. 2023;261(11):3165-3176. doi:10.1007/s00417-023-06145-7

7. Laurik-Feuerstein KL, Sapahia R, Cabrera DeBuc D, Somfai GM. The assessment of fundus image quality labeling reliability among graders with different backgrounds. *PLoS One*. 2022;17(7):e0271156. doi:10.1371/journal.pone.0271156

8. Mitry D, Zutis K, Dhillon B, et al. The Accuracy and Reliability of Crowdsource Annotations of Digital Retinal Images. *Transl Vis Sci Technol*. 2016;5(5):6. doi:10.1167/tvst.5.5.6

9. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng*. 2022;6(12):1346-1352. doi:10.1038/s41551-022-00914-1

10. Kondylakis H, Ciarrocchi E, Cerda-Alberich L, et al. Position of the AI for Health Imaging (AI4HI) network on metadata models for imaging biobanks. *Eur Radiol Exp*. 2022;6(1):29. doi:10.1186/s41747-022-00281-1

11. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. *Nat Med*. 2023;29(11):2686-2687. doi:10.1038/s41591-023-02540-z

12. 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 macular edema. *Retina*. 2008;28(1):36-40. doi:10.1097/IAE.0b013e31815e9385

13. Ruamviboonsuk P, Teerasuwanajak K, Tiensuwan M, Yuttitham K, 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(5):826-832. doi:10.1016/j.ophtha.2005.11.021

14. Freeman B, Hammel N, Phene S, et al. Iterative Quality Control Strategies for Expert Medical Image Labeling.

15. Khader F, Müller-Franzes G, Wang T, et al. Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters: A Case for Transformers. *Radiology*. 2023;309(1):e230806. doi:10.1148/radiol.230806

16. Li Y, Han Y, Li Z, Zhong Y, Guo Z. A transfer learning-based multimodal neural network combining metadata and multiple medical images for glaucoma type diagnosis. *Sci Rep*. 2023;13(1):12076. doi:10.1038/s41598-022-27045-6

17. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. ., 292(13):1602-9. doi: 10.1001/jama.292.13.1602. PMID: 15467063. Accuracy of Diagnostic Tests Read With and Without Clinical Information. *JAMA*. Published online October 6, 2004. https://watermark-silverchair-com.ezp-prod1.hul.harvard.edu/jrv40009.pdf

18. Leslie A, Jones AJ, Goddard PR. The influence of clinical information on the reporting of CT by radiologists. *BJR Suppl*. 2000;73(874):1052-1055. doi:10.1259/bjr.73.874.11271897

19. Castillo C, Steffens T, Sim L, Caffery L. The effect of clinical information on radiology reporting: A systematic review. *J Med Radiat Sci*. 2021;68(1):60-74. doi:10.1002/jmrs.424

20. Bai L, Chen S, Gao M, Abdelrahman L, Ghamdi MA, Abdel-Mottaleb M. The Influence of Age and Gender Information on the Diagnosis of Diabetic Retinopathy: Based on Neural Networks. *Conf Proc IEEE Eng Med Biol Soc*. 2021;2021:3514-3517. doi:10.1109/EMBC46164.2021.9629607

21. Cleland CR, Rwiza J, Evans JR, et al. Artificial intelligence for diabetic retinopathy in low-income and middle-income countries: a scoping review. *BMJ Open Diabetes Res Care*. 2023;11(4). doi:10.1136/bmjdrc-2023-003424

22. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health*. 2021;3(1):e51-e66. doi:10.1016/S2589-7500(20)30240-5

23. Yip MYT, Lim G, Lim ZW, et al. Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. *NPJ Digit Med*. 2020;3:40. doi:10.1038/s41746-020-0247-1

24. Mbarara University Data Science Research Hub. Mbarara University of Science & Technology. Published June 25, 2022. Accessed August 27, 2022. https://www.must.ac.ug/collaboration/projects-and-studies-at-must/mudsreh/

25. Home. Accessed April 7, 2024. https://dsi-africa.org/

26. Nakayama LF, Gonçalves MB, Ribeiro LZ, Malerbi FK, Regatieri CVS. Diabetic Retinopathy Labeling Protocol for the Brazilian Multilabel Ophthalmological Dataset. Published online August 2023. doi:10.31219/osf.io/puznm

27. Nakayama LF, Goncalves M, Zago Ribeiro L, et al. A Brazilian Multilabel Ophthalmological Dataset (BRSET). Published online March 8, 2023. doi:10.13026/XCXW-8198

28. Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: The Messidor database. *Image Anal Stereol*. 2014;33(3):231. doi:10.5566/ias.1155

29. Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *J Diabetes Sci Technol*. 2009;3(3):509-516. doi:10.1177/193229680900300315

30. Porwal P, Pachade S, Kokare M, et al. IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge. *Med Image Anal*. 2020;59:101561. doi:10.1016/j.media.2019.101561

31. Behind the AI boom, an army of workers in 'digital sweatshops.' THE AFRICAN. Published August 30, 2023. Accessed April 24, 2024. https://theafrican.co.za/featured/behind-the-ai-boom-an-army-of-workers-in-digital-sweatshops-b25527d9-bfbb-428b-b21a-f3df1e3880f3/

# Supplementary Files

# Multimedia Appendixes

Individual reader sensitivity and specificity results.
URL: http://asset.jmir.pub/assets/2ee9437ff8bb2f65ae1e973fc83579ee.docx