

Original Paper

Impact of Gold-Standard Label Errors on Evaluating Performance of Deep Learning Models in Diabetic Retinopathy Screening: Nationwide Real-World Validation Study

Yueye Wang^{1,2*}, Dr med; Xiaotong Han^{1*}, PhD, Prof Dr Med; Cong Li¹, Dr med; Lixia Luo¹, PhD, Prof Dr Med; Qiuxia Yin¹, MPH; Jian Zhang¹, MD; Guankai Peng³, BSc; Danli Shi^{2,4}, PhD, Dr med; Mingguang He^{2,4,5}, PhD, Prof Dr Med

¹State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China

²School of Optometry, The Hong Kong Polytechnic University, Kowloon, China (Hong Kong)

³Guangzhou Vision Tech Medical Technology Co, Ltd, Guangzhou, China

⁴Research Centre for SHARP Vision, The Hong Kong Polytechnic University, Kowloon, China (Hong Kong)

⁵Centre for Eye and Vision Research, Hong Kong, China (Hong Kong)

*these authors contributed equally

Corresponding Author:

Mingguang He, PhD, Prof Dr Med

School of Optometry

The Hong Kong Polytechnic University

Hung Hom

Kowloon

China (Hong Kong)

Phone: 852 34002795

Email: mingguang.he@polyu.edu.hk

Abstract

Background: For medical artificial intelligence (AI) training and validation, human expert labels are considered the gold standard that represents the correct answers or desired outputs for a given data set. These labels serve as a reference or benchmark against which the model's predictions are compared.

Objective: This study aimed to assess the accuracy of a custom deep learning (DL) algorithm on classifying diabetic retinopathy (DR) and further demonstrate how label errors may contribute to this assessment in a nationwide DR-screening program.

Methods: Fundus photographs from the Lifeline Express, a nationwide DR-screening program, were analyzed to identify the presence of referable DR using both (1) manual grading by National Health Service England-certificated graders and (2) a DL-based DR-screening algorithm with validated good lab performance. To assess the accuracy of labels, a random sample of images with disagreement between the DL algorithm and the labels was adjudicated by ophthalmologists who were masked to the previous grading results. The error rates of labels in this sample were then used to correct the number of negative and positive cases in the entire data set, serving as postcorrection labels. The DL algorithm's performance was evaluated against both pre- and postcorrection labels.

Results: The analysis included 736,083 images from 237,824 participants. The DL algorithm exhibited a gap between the real-world performance and the lab-reported performance in this nationwide data set, with a sensitivity increase of 12.5% (from 79.6% to 92.5%, $P<.001$) and a specificity increase of 6.9% (from 91.6% to 98.5%, $P<.001$). In the random sample, 63.6% (560/880) of negative images and 5.2% (140/2710) of positive images were misclassified in the pre-correction human labels. High myopia was the primary reason for misclassifying non-DR images as referable DR images, while laser spots were predominantly responsible for misclassified referable cases. The estimated label error rate for the entire data set was 1.2%. The label correction was estimated to bring about a 12.5% enhancement in the estimated sensitivity of the DL algorithm ($P<.001$).

Conclusions: Label errors based on human image grading, although in a small percentage, can significantly affect the performance evaluation of DL algorithms in real-world DR screening.

KEYWORDS

artificial intelligence; diabetic retinopathy; diabetes; real world; deep learning

Introduction

In recent years, the application of artificial intelligence (AI) in identifying ocular lesions and diseases has gained increasing popularity [1-4]. One prominent example is the application of deep learning (DL) in fundus image-based diabetic retinopathy (DR) screening [5-8]. Despite the widely reported high accuracy in existing studies, as well as the advantages in efficiency and cost-effectiveness, the real-world implementation of DL-based DR-screening models still faces challenges [9]. A major concern is the gap observed between the high laboratory performance and the unsatisfying real-world performance. For example, although most DL algorithms have achieved a sensitivity of over 90% in previous research, the reported sensitivity ranges from 55% to 85% in real-world tasks [4,7,10]. A better understanding of the reasons for the performance gap so as to enhance the real-world performance of DL-based models could help promote the implementation of AI on the ground, but related evidence is scarce.

In the development of medical AI, human expert labels usually serve as the gold standard to compare with and evaluate model results [11]. The accuracy of AI model-based tasks critically depends on high-quality labels. Assembling large data sets with human expert labels can be a huge task, and ensuring the quality of the gold standard remains an open problem [12]. In real-world image data sets, the gold standard is usually generated by trained graders who manually assign grades to the data [13-15]. However, human error is inevitable in this process, leading to problems, such as bias, subjective judgement, inadequate repeatability, and consistency. Studies have reported significant discrepancies in DR grading among human graders, with sensitivities ranging from approximately 60% to 90% [16]. As a result, using manual grading outputs as an objective reference standard can be problematic when the accuracy of manual grading is in question [11,17,18].

We hypothesized that errors in human labels may exist in the real world and exert a nonnegligible impact on real-world DL model performance. Thus, in this study, we included over 0.7 million fundus images from 0.2 million individuals from a nationwide DR-screening program, based on which we aimed to assess the gap between lab-reported and real-world performance of a previously validated DL-based DR-screening model, identify the existence of label errors, and assess the impact of those label errors on the observed performance gaps.

Methods

Ethical Considerations

The study adhered to the tenets of the Declaration of Helsinki and received approval from the Institutional Review Board of

the Zhongshan Ophthalmic Centre (2023KYPJ108) and the Chinese Foundation for Lifeline Express. All images were fully anonymized. Written informed consent was obtained from all Lifeline Express participants. Given the retrospective nature of this study using anonymized images, the ethics review board exempted this study from additional informed consent.

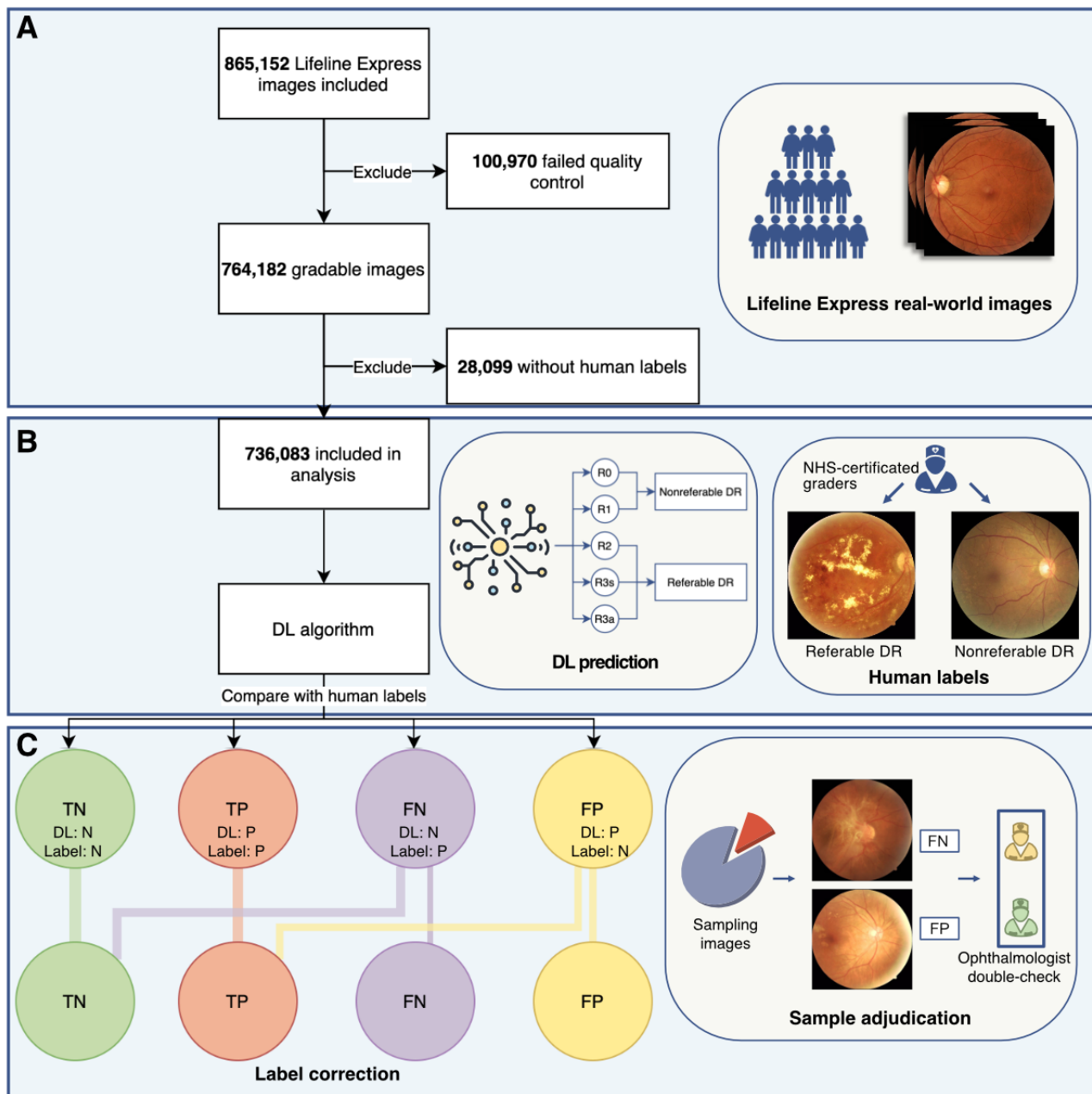
Real-World Data Set

The pipeline of this study is depicted in [Figure 1](#). The overall design of this study comprised 3 sections. First, more than 0.8 million images were collected from a real-world data set, and all images were included except those that failed quality control or lacked human labels. Second, fundus images were reviewed by National Health Service (NHS)-certificated human graders and manually classified as nonreferable DR (R0 and R1) or referable DR (R2, R3s, and R3a) according to NHS DR-screening guidelines. Next, a previously validated DL model was deployed for identifying referable DR, and the performance of the DL model was evaluated against human labels. Third, a random sample of images, comprising false-negative (FN) and false-positive (FP) subsets, was extracted for adjudication by 2 ophthalmologists. The details of the process are given next.

When adjudication yielded consistent results with DL instead of human labels, the errors of the labels were documented. The extracted sample's error rates were used to recalculate the number of positive and negative cases in the entire data set, as postcorrection labels. The diagnostic performance of the DL algorithm was reassessed against the postcorrection labels.

We included 865,152 color fundus images of participants who participated in a nationwide real-world DR screening program, the Lifeline Express DR Screening Program, between 2014 and 2019. For each participant, nonmydriatic, 45° field color fundus images were captured using locally available imaging devices (including AFC-230, NIDEK; Canon CR-DGi; FundusVue, Cystalvue) with at least 1 image centered on the macula or optic disc. The size of the images was 512 × 512 pixels, at a resolution of 72 pixels per inch. Images were subsequently sent to 5 different central grading centers (Peking Union Medical College Hospital, Beijing Tongren Hospital, Peking University Third Hospital, Joint Shantou International Eye Center, and Zhongshan Ophthalmic Center) and graded for DR by human graders who possessed English NHS-certified qualifications. The following images were excluded: (1) those that failed the quality control procedure and (2) those without human labels. Details regarding the quality control and DR-grading criteria can be found in [Table S1](#) in [Multimedia Appendix 1](#).

Figure 1. Schematic pipeline of the study design. DL: deep learning; DR: diabetic retinopathy; FN: false negative; FP: false positive; NHS: National Health Service; TN: true negative; TP true positive.



Grading Criteria

Based on NHS DR-screening guidelines, each fundus image was classified into 1 of 5 grades: R0 (no DR), R1 (background DR), R2 (preproliferative DR), R3s (static proliferative DR), or R3a (active proliferative DR). Referable DR was defined as preproliferative DR or worse (ie, R2, R3s, and R3a) and recorded as positive, whereas nonreferable DR (ie, R0 and R1) was recorded as negative. A participant was diagnosed as having referable DR if at least 1 of their gradable images was classified as positive.

The manual grading process involved the following stages. Initially, all images were assessed by a primary grader. Images classified as positive and a random sample of 15% of images classified as negative were then subjected to a second round of grading by another primary grader. Images with any discrepancy

between the assessments of the 2 primary graders were referred to a secondary grader, whose conclusion served as a final determination. The resulting manual grading outputs formed human labels.

Development of the DR-Grading DL Algorithm

The DR-grading algorithm is based on a convolutional neural network. The algorithm was trained using 71,043 retinal images from 36 ophthalmic departments and validated using an additional 106,244 retinal images from 3 population-based studies. Image preprocessing involved scaling image pixel values within a range of 0-1 and subtracting the local-space average color. Images were resized to 299 × 299 pixels. We performed data augmentation techniques on training data. This involved randomly shifting each image horizontally by 0-3 pixels and rotating it by 90°, 180°, or 270°.

The original version of the algorithm has been reported previously [19]. The original DL algorithm achieved an area under the receiver operating characteristic curve (AUROC) of 0.955 (sensitivity: 92.5%, specificity: 98.5%) in validation for the detection of referable DR. The algorithm used in our study was retrained using the Xception architecture to improve model performance, while using the same training data set. Max pooling of each Inception module and global average pooling were used in the Xception architecture. For training, a minibatch gradient descent size of 32 was used. The learning rate adopted an inverse time decay schedule with an initial value of 0.001.

Adjudication and Human Label Correction

We adjudicated images that showed discrepancies between DL and human labels in referable DR classification, including FN (DL: negative, label: positive) and FP (DL: positive, label: negative). Adjudication was performed by 2 experienced ophthalmologists who were masked to the DL and human grading results; they used a back-to-back method in accordance with NHS DR-screening guidelines. Images with any discrepancy between the 2 ophthalmologists were sent to a senior ophthalmologist for reassessment, whose grading results were deemed as final.

Given the large number of images in this study, reviewing all images with discrepancies would have been impractical. Therefore, random samples were drawn for adjudication. To ascertain the representativeness of these random samples across the entire data set, a series of gradual samplings were conducted, varying in size from 100 to 500 FN images. Notably, a similar distribution of the estimated error rate was observed across all sample sizes (Figure S1 in [Multimedia Appendix 1](#)), suggesting that even a subset as small as 100 images can effectively represent the label distribution within the entire data set. Considering the total volume of FN and FP cases, we examined a 10% sample from FN cases and a 5% sample from FP cases, at both image and participant levels, aiming for a representative and manageable size of the sample. When adjudication yielded consistent results with DL instead of human labels, the label errors were documented in the FN and FP samples, respectively. Using the label error rates in the FN and FP samples, we recalculated the number of positive and negative cases in the entire data set, serving as postcorrection labels.

Statistical Analysis

We evaluated the performance of the DL algorithm using pre- and postcorrection labels as the gold standard. Primary evaluation metrics for diagnostic performance included accuracy, precision, sensitivity (recall), specificity, and the F_1 -score. AUROC was calculated at the image level. These evaluations were conducted at both image and individual levels.

Label error rates in adjudicated samples were calculated as follows:

$$\text{fnr} = \frac{\text{Negative cases in adjudication}}{10\% \times \text{FN cases}},$$

$$\text{fpr} = \frac{\text{Positive cases in adjudication}}{5\% \times \text{FP cases}},$$

where fnr and fpr represent label error rates in FN and FP samples, respectively. Label changes during label correction included:

$$\text{pn} = \text{fnr} \times \text{FN cases},$$

$$\text{np} = \text{fpr} \times \text{FP cases},$$

where pn is the number of labels changing from positive to negative, while np is the number of labels changing from negative to positive. The total positive and negative cases after label correction were calculated as follows:

$$\text{Total positive cases} = \text{Positive labels} + \text{np} - \text{pn}$$

$$\text{Total negative cases} = \text{Negative labels} + \text{pn} - \text{np}$$

Comparisons of diagnostic performance were conducted using chi-square tests. Adjudication was conducted on the open source LabelMe platform [20]. All statistical analyses were performed using R version 4.1.2 (Foundation for Statistical Computing). A 2-sided P value of $<.05$ was considered statistically significant.

Results

Precorrection Model Performance

Of the 865,152 images acquired from 251,535 participants, 736,083 (85.1%) images from 237,824 (94.5%) participants were included in this analysis. [Table 1](#) shows the results of the DL algorithm and precorrection labels for detecting referable DR. At the image level, DL-based grading yielded 9339 (1.3%) FN, 57,886 (7.9%) FP, 632,457 (85.9%) true negative (TN), and 36,401 (5.0%) true positive (TP) images against precorrection labels. The DL algorithm yielded an AUROC of 0.927, a sensitivity (recall) of 79.6%, and a specificity of 91.6%, lower than the lab-reported performance values (AUROC: 0.955, sensitivity: 92.5%, specificity: 98.5%). The corresponding gaps in sensitivity and specificity were 12.9% and 6.9%, respectively. The precision and F_1 -score at the image level were 38.6% and 52.0%, respectively. Evaluation at the individual level yielded 2062 (0.9%) FN, 33,878 (14.2%) FP, 186,648 (78.5%) TN, and 15,236 (6.4%) TP images. The sensitivity, specificity, and accuracy for identifying individuals with referable DR were 88.1%, 84.6%, and 84.9%, respectively. The precision and F_1 -score at the individual level were 31.0% and 45.9%, respectively. Subgroup analysis for each DR grade indicated label discrepancies between the DL algorithm and human labels, mainly accumulated in R1, R2, and R3a ([Figure S2 in Multimedia Appendix 1](#)).

Table 1. Grading results of the DL^a algorithm and the ground truth in identifying the presence of referable DR^{b,c}.

Performance	Images (N=736,083)		Participants (N=237,824)	
	DL model	Human labels ^d	DL model	Human labels
Positive, n (%)	98,127 (13.3)	45,740 (6.2)	49,114 (20.7)	17,298 (7.3)
Negative, n (%)	666,055 (90.5)	690,343 (93.8)	188,710 (79.4)	220,526 (92.7)
TP ^e , n (%)	36,401 (5.0)	— ^f	15,236 (6.4)	—
FP ^g , n (%)	57,886 (7.9)	—	33,878 (14.2)	—
TN ^h , n (%)	632,457 (85.9)	—	186,648 (78.5)	—
FN ⁱ , n (%)	9,339 (1.3)	—	2,062 (0.9)	—
Accuracy, %	90.9	—	84.9	—
Precision, %	38.6	—	31.0	—
Sensitivity (recall), %	79.6	—	88.1	—
Specificity, %	91.6	—	84.6	—
F ₁ -score, %	52.0	—	45.9	—

^aDL: deep learning.

^bDR: diabetic retinopathy.

^cReferable DR was defined as preproliferative and proliferative DR.

^dHuman labels came from the final DR-grading results of human graders, based on the NHS DR-screening guidelines.

^eTP: true positive.

^fNot applicable.

^gFP: false positive.

^hTN: true negative.

ⁱFN: false negative.

Adjudication and Label Correction

Table 2 presents the adjudication results of extracted samples with inconsistent grading results between DL and human labels. After adjudication, 560 (63.6%) of 880 images labeled positive were deemed negative, leading to a mean estimated FN error rate of 63.6% (SD 1.6%, 95% CI 60.4%-66.8%). However, 140

(5.2%) of 2710 images labeled negative were deemed positive, with a mean estimated FP error rate of 5.2% (SD 0.4%, 95% CI 4.4%-6.1%). The adjudication at the individual level yielded a similar finding (FN error rate: mean 65.3%, SD 3.4%, 95% CI 58.2%-72.0%; FP error rate: mean 5.1%, SD 0.5%, 95% CI 4.1%-6.3%). Detailed classifications are listed in Table S2 in [Multimedia Appendix 1](#).

Table 2. Adjudicated results of sample images.

Sample adjudication	Image level		Participant level	
	Images labeled positive (n=880)	Images labeled negative (n=2710)	Images labeled positive (n=195)	Images labeled negative (n=1657)
Erroneous labels	560 (63.6%)	140 (5.2%)	128 (65.3%)	85 (5.1%)
Accurate labels	320 (36.4%)	2570 (94.8%)	68 (34.7%)	1572 (94.9%)

As shown in Table 3, the top reasons for misclassifying R0 images as referable DR images in precorrection labels included high myopia (n=106, 28.5%), a normal-appearing fundus (n=82, 22%), artifacts (n=40, 10.8%), and age-related macular degeneration (AMD; n=37, 10%). The predominant reasons for misclassified referable DR cases in precorrection labels included undetected laser spots (n=56, 41.8%), potential venous anomalies (n=25, 18.7%), and intraretinal microvascular

anomalies (IRMAs; n=22, 16.4%). Typical images with erroneous labels are presented in Figure 2: Figures 2A and 2B show referable DR images misclassified as negative in human labels, with laser spots and retinal neovascularization, while Figures 2C and 2D show nonreferable DR images misclassified as positive in labels, probably due to a high-myopia fundus and artifacts.

Table 3. Potential reasons for errors^a in precorrection labels.

Probable reasons	Count, n (%)
Images misclassified as positive	
High myopia	106 (28.5)
Normal-appearing fundus	82 (22.0)
Artifacts	40 (10.8)
AMD ^b	37 (10.0)
Underexposure	33 (8.9)
Other retinopathies	20 (5.4)
Poor quality	11 (3.0)
Hypertensive retinopathy	8 (2.2)
Media opacity	6 (1.6)
Images misclassified as negative	
Laser spots	56 (41.8)
Venous anomalies	25 (18.7)
IRMAs ^c	22 (16.4)
Neovascularization	14 (10.5)
Multiple blot hemorrhages	11 (8.2)
Fibrous membrane	3 (2.2)
Vitreous hemorrhage	1 (0.8)

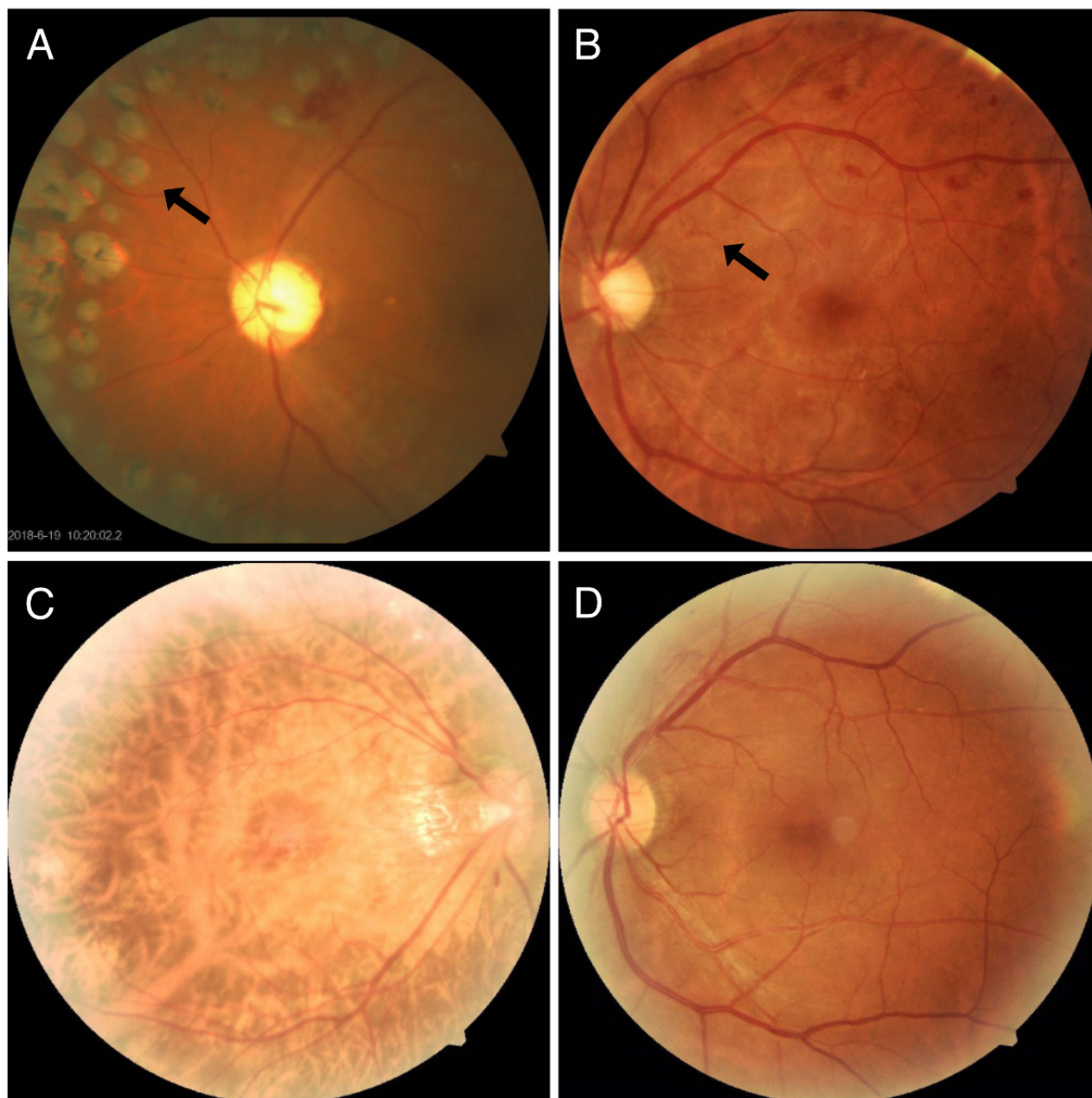
^aErrors in precorrection labels were estimated against the results of the adjudication. Images misclassified as positive represented those graded as referable diabetic retinopathy (DR) in the labels but turned out to be adjudicated as R0; images misclassified as negative represented those graded as nonreferable DR in the labels but turned out to be referable DR after adjudication.

^bAMD: age-related macular degeneration.

^cIRMA: intraretinal microvascular anomaly.

Based on the FN and FP error rates in the sample, we assumed that approximately 8933 images (from 3084 individuals) were misclassified in the precorrection labels in the entire data set, resulting in a mean estimated image error rate of 1.2% (SD 0.01%, 95% CI 1.1%-1.3%; individual level: mean 1.3%, SD

0.02%, 95% CI 1.1%-1.5%). The postcorrection labels of the entire data set were estimated to include 42,787 truly positive images and 693,296 truly negative images (from n=17,689, 7.4%, referable DR participants and n=220,135, 92.6%, nonreferable DR participants, respectively).

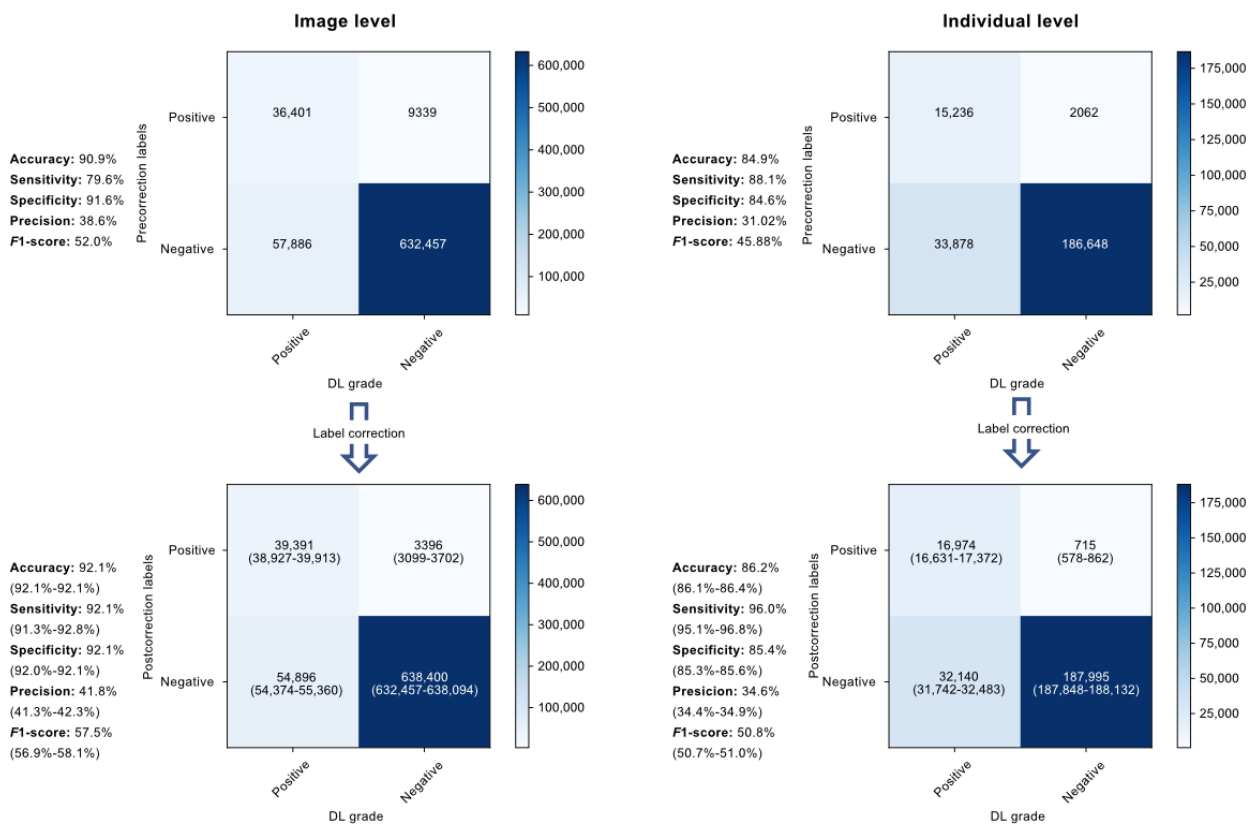
Figure 2. Typical images with human label errors before correction.

Postcorrection Model Performance

The confusion matrixes in Figure 3 show the comparison of DL model performance before and after label correction. After correction at the image level, sensitivity significantly improved by 12.5% (from 79.6% to 92.1%, $P < .001$), and the corresponding improvement in specificity and accuracy was 0.5% (from 91.6% to 92.1%, $P < .001$) and 1.2% (from 90.9% to 92.1%, $P < .001$), respectively. Similarly, postcorrection results

at the individual level showed a 7.9% increase in sensitivity (from 88.1% to 96%, $P < .001$), a 0.8% increase in specificity (from 84.6% to 85.4%, $P < .001$), and a 1.3% increase in accuracy (from 84.9% to 86.2%, $P < .001$). The lab/real-world performance gap was significantly smaller after label correction, with a significantly reduced sensitivity gap from 12.9% to 0.4% ($P < .001$), as well as a reduced specificity gap from 6.9% to 6.4% ($P < .001$).

Figure 3. Performance of the DL model before and after label correction. DL: deep learning.



Discussion

Principal Findings

Based on a nationwide DR-screening program, we identified a significant gap between the lab-reported and real-world performance of a DL-based DR-screening model, which could be largely reduced by checking and correcting the errors in human labels. These results highlight the significance of acknowledging potential inaccuracies in human labels, which is often regarded as the gold standard. In this study, we proposed an approach of sampling adjudication to address this issue for a large data set. By evaluating the error rate within sampled subsets, we extrapolated and approximated the accurate distribution of labels across the entire data set. This approach enables a more precise evaluation of AI performance in real-world practice by providing estimates that consider and mitigate the impact of label errors.

Currently, there are 2 predominant approaches to generate a manually labeled gold standard in image-based real-world investigations [6-8,21-28]. The first approach involves a 2-stage grading process. In this process, 2 primary graders review all images, and a senior grader intervenes to make a final decision in the case of a disagreement between the primary graders [21]. The second method is a 1-stage grading process, wherein a single grader reviews all images to obtain human labels. In some instances, the 1-stage grading process might need additional verification for positive cases or a subsample of images [6]. Although accuracy may be somewhat compromised, the 1-stage grading method has advantages in terms of efficiency, convenience, time-effectiveness, and cost-effectiveness.

Therefore, 1-stage grading is more commonly used in real-world practice, such as in the case of the Lifeline Express in this study. However, regardless of which grading method is used, we recommend double-checking human labels, even in a subsample of the data, before assessing the performance of the DL model in the real world.

Despite increasing endeavors aimed at enhancing the accuracy of human labels through measures such as standardized training for graders, the issue of mislabeling remains inevitable. Even if human graders make just 1 mistake in 100 cases, such as misclassifying different retinal conditions as DR or missing laser scars on the peripheral retina, it affects the gold standard. These small flaws can significantly influence the development and evaluation of AI algorithms. However, the amount of research available that evaluates the accuracy of manual grading outputs in DR screening is limited. In one study, 7379 images collected from 735 individuals were used to assess the performance of manual grading, which achieved a sensitivity of 82.2% and a specificity of 84.4% [7]. In another study involving 2384 images collected from 1208 participants in a nationwide DR-screening program in Thailand, the grading performance of human graders was deemed relatively reliable (accuracy: 93.5%, sensitivity: 84.8%, specificity: 95.5%) [28]. Moreover, Krause et al [29] compared DR-grading outcomes of a majority decision made by 3 ophthalmologists with the standard reference of adjudication by retinal specialists [29]. A discrepancy rate of around 0.8% can be drawn for identifying vision-threatening DR from their conclusion. These previous studies, in conjunction with our findings, support the need to estimate the error rate associated with manual grading in various real-world DR-screening settings. These estimates range from

approximately 1% to 7%. Variations in the DR prevalence in the target population, grading workflow, and grader certifications may serve as contributing factors.

The label error rate in our study was around 1%, primarily attributed to the misclassification of nonreferable cases as referable DR. This is not anticipated to impede prompt referral and intervention for patients truly requiring medical attention. However, even such a small error rate can largely deteriorate the performance of DL algorithms in the real world. To address this concern, 1 previously proposed solution is to establish a rigorous reference standard during the human labeling process—for example, incorporating adjudication consensus from retinal specialists, rather than relying on major decisions made by multiple graders [29]. However, implementing a rigorous grading workflow in large-scale real-world screening practices can be time-consuming and resource intensive. Additionally, this process cannot be executed after study completion. Our study introduced an optimization strategy based on sample adjudication and label correction. Importantly, this strategy does not require modifications to the grading workflow or retraining of the DL model. Despite addressing only around 1% of label errors in the entire data set, this optimization strategy has been shown to bring about 12% enhancement in the AI model's sensitivity.

The label correction strategy in our study can be considered a post hoc quality control procedure. It has been reported that label corrections for model tuning could effectively enhance the performance of AI algorithms [29]. Our findings further support the potential beneficial effect of adding this quality control procedure in existing and future DR-screening programs, especially when a large gap is observed between the real-world and lab-reported performance of the AI algorithm. However, implementation of this strategy in real-world practice may present challenges. Even if only a small subsample of images is selected for review, this would increase the workload of ophthalmologists, necessitating additional time and resources. Based on our study, we recommend that sampling approximately 5% of images with disagreement is reasonable for post hoc quality control. We recommend that future studies determine an appropriate sample size based on their expected error rate, resources, and other real-world considerations. We believe that this label correction strategy offers a standardized quality control pipeline across diverse image-based AI applications. Future investigations could explore the potential effect of this strategy in other AI-based models (eg, other diseases or imaging modalities) and other health care scenarios (eg, disease diagnosis or prognosis).

Acknowledgments

We would like to acknowledge the Chinese Foundation for Lifeline Express and the Lifeline Express Hong Kong Foundation for authorizing access to the data set. We would also like to acknowledge Ms Nellie Fong and Ms Mingyue Wang for their kind assistance during data collection.

This study was funded by the National Natural Science Foundation of China (82000901, 82101171), the Global STEM Professorship Scheme (P0046113), the Fundamental Research Funds of the State Key Laboratory of Ophthalmology (303060202400201209), and the Outstanding PI Research Funds of the State Key Laboratory of Ophthalmology (3030902113083).

Strengths

This study possesses several notable strengths. First, it was embedded in a nationwide DR-screening program, ensuring a large and robust data set comprising over 0.7 million fundus images collected from more than 0.2 million participants. This considerable scale enhanced the reliability of our findings. Second, to the best of our knowledge, this study represents a pioneering effort in evaluating the impact of label errors on the performance of a DR-grading DL algorithm. In addition, the proposal and validation of an effective optimization strategy in this study contribute significantly to the understanding and advancement of DL implementation in real-world DR screening.

Limitations

This study has some limitations. In our evaluation, images with identical outputs between DL and labels (TP and TN cases) were considered correctly classified, without undergoing further adjudication. In smaller-scale trials, we adjudicated a limited subset of TP (200 images) and TN (1000 images) cases. These assessments revealed a minimal error rate in human labeling, at 2.5% for TP and 0.1% for TN (further details in Table S3 in [Multimedia Appendix 1](#)). Given the substantial quantity of TP and TN images within our data set, refraining from label correction in these subsets would unlikely cause a significant alteration in the estimated performance of the DL model. Additionally, during the adjudication process, we conducted random sampling instead of reviewing all images with discrepancies. To ensure the representativity of the random sample, we performed multiple samplings with varying sample sizes and validated consistent distributions of the obtained results. Moreover, the Lifeline Express program exclusively recruited Chinese participants. It is imperative to conduct additional validation studies encompassing diverse ethnicities to corroborate our findings. Lastly, our analysis focused solely on a single DR-screening model. Nevertheless, it is reasonable to infer that our conclusions are applicable to alternative models used in the screening of other ocular diseases. Therefore, further investigations are warranted to substantiate this discovery.

Conclusion

In conclusion, we found that even a small percentage of label errors could have a substantial impact on the performance of DL algorithms in real-world DR screening. Attention should be paid to minimizing and correcting label errors in future studies and implementations of AI-based DR-screening models in the real world.

Authors' Contributions

Study concept and design were handled by YW, XTH, and MH; acquisition, analyses, interpretation, and critical revision of the manuscript for important intellectual content by all authors; drafting of the manuscript by YW, CL, XTH, and DLS; statistical analyses by YW and JZ; funding by MH and XTH; administrative, technical, and material support by GKP and DLS; and study supervision by MH.

Conflicts of Interest

The authors declare no proprietary interest in any aspect of the study. The Chinese Foundation for Lifeline Express and the Guangzhou Vision Tech Medical Technology Co hereby are not associated directly or indirectly with the subject or design of this study.

Multimedia Appendix 1

Supplementary tables and figures.

[\[DOCX File, 749 KB-Multimedia Appendix 1\]](#)

References

1. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. Sep 2018;24(9):1342-1350. [doi: [10.1038/s41591-018-0107-6](https://doi.org/10.1038/s41591-018-0107-6)] [Medline: [30104768](https://pubmed.ncbi.nlm.nih.gov/30104768/)]
2. 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*. Dec 13, 2016;316(22):2402-2410. [doi: [10.1001/jama.2016.17216](https://doi.org/10.1001/jama.2016.17216)] [Medline: [27898976](https://pubmed.ncbi.nlm.nih.gov/27898976/)]
3. Milea D, Najjar RP, Jiang Z, Ting D, Vasseneix C, Xu X, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med*. Apr 30, 2020;382(18):1687-1695. [doi: [10.1056/nejmoa1917130](https://doi.org/10.1056/nejmoa1917130)]
4. Ting DSW, Cheung CY, 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*. Dec 12, 2017;318(22):2211-2223. [FREE Full text] [doi: [10.1001/jama.2017.18152](https://doi.org/10.1001/jama.2017.18152)] [Medline: [29234807](https://pubmed.ncbi.nlm.nih.gov/29234807/)]
5. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet*. Jul 10, 2010;376(9735):124-136. [doi: [10.1016/S0140-6736\(09\)62124-3](https://doi.org/10.1016/S0140-6736(09)62124-3)] [Medline: [20580421](https://pubmed.ncbi.nlm.nih.gov/20580421/)]
6. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39. [FREE Full text] [doi: [10.1038/s41746-018-0040-6](https://doi.org/10.1038/s41746-018-0040-6)] [Medline: [31304320](https://pubmed.ncbi.nlm.nih.gov/31304320/)]
7. 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*. May 2021;44(5):1168-1175. [FREE Full text] [doi: [10.2337/dc20-1877](https://doi.org/10.2337/dc20-1877)] [Medline: [33402366](https://pubmed.ncbi.nlm.nih.gov/33402366/)]
8. Zhang Y, Shi J, Peng Y, Zhao Z, Zheng Q, Wang Z, et al. Artificial intelligence-enabled screening for diabetic retinopathy: a real-world, multicenter and prospective study. *BMJ Open Diabetes Res Care*. Oct 21, 2020;8(1):e001596. [FREE Full text] [doi: [10.1136/bmjdc-2020-001596](https://doi.org/10.1136/bmjdc-2020-001596)] [Medline: [33087340](https://pubmed.ncbi.nlm.nih.gov/33087340/)]
9. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res*. Apr 22, 2021;23(4):e25759. [FREE Full text] [doi: [10.2196/25759](https://doi.org/10.2196/25759)] [Medline: [33885365](https://pubmed.ncbi.nlm.nih.gov/33885365/)]
10. Bellefleur V, Lim G, Rim TH, Tan GSW, Cheung CY, Sadda S, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Curr Diab Rep*. Jul 31, 2019;19(9):72. [doi: [10.1007/s11892-019-1189-3](https://doi.org/10.1007/s11892-019-1189-3)] [Medline: [31367962](https://pubmed.ncbi.nlm.nih.gov/31367962/)]
11. Lebovitz S, Levina N, Lifshitz-Assa H. Is AI ground truth really 'true'? The dangers of training and evaluating AI tools based on experts' know-what. *MISQ*. Sep 01, 2021;45(3):1501-1526. [doi: [10.25300/misq/2021/16564](https://doi.org/10.25300/misq/2021/16564)]
12. Khattak F, Salleb-Aouissi A. Quality control of crowd labeling through expert evaluation. 2011. Presented at: NIPS 2011: 2nd Workshop on Computational Social Science and the Wisdom of Crowds; December 17, 2011; Sierra Nevada, Spain.
13. Jatton F. Assessing biases, relaxing moralism: on ground-truthing practices in machine learning design and application. *Big Data Soc*. May 05, 2021;8(1):205395172110135. [doi: [10.1177/20539517211013569](https://doi.org/10.1177/20539517211013569)]
14. Lemoigne Y, Caner A. *Molecular Imaging: Computer Reconstruction and Practice*. Berlin. Springer Science & Business Media; 2008.
15. Søgaard A, Plank B, Hovy D. Selection bias, label bias, and bias in ground truth. 2014. Presented at: COLING 2014: 25th International Conference on Computational Linguistics: Tutorial Abstracts; August 2014; Dublin, Ireland.
16. Lin D, Xiong J, Liu C, Zhao L, Li Z, Yu S, et al. Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. *Lancet Digital Health*. Aug 2021;3(8):e486-e495. [doi: [10.1016/s2589-7500\(21\)00086-8](https://doi.org/10.1016/s2589-7500(21)00086-8)]

17. Kats E, Goldberger J, Greenspan H. A soft STAPLE algorithm combined with anatomical knowledge. 2019. Presented at: MICCAI 2019: 22nd International Conference Medical Image on Computing and Computer Assisted Intervention; October 13-17, 2019; Shenzhen, China. [doi: [10.1007/978-3-030-32248-9_57](https://doi.org/10.1007/978-3-030-32248-9_57)]
18. Jacob J, Ciccarelli O, Barkhof F, Alexander D, editors. Disentangling human error from the ground truth in segmentation of medical images. 2020. Presented at: NIPS 2020: 34th International Conference on Neural Information Processing Systems; December 6-12, 2020; Online.
19. Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. *Diabetes Care*. Dec 2018;41(12):2509-2516. [doi: [10.2337/dc18-0147](https://doi.org/10.2337/dc18-0147)] [Medline: [30275284](https://pubmed.ncbi.nlm.nih.gov/30275284/)]
20. Kentaro WM, Martijn B, Zhang C, Kubovčik M, Myczko A, Zhu L, et al. wkentaro/labelme: v4.6.0. Zenodo. Nov 18, 2021. URL: <https://zenodo.org/records/5711226> [accessed 2024-07-01]
21. Verbraak FD, Abramoff MD, Bausch GCF, Klaver C, Nijpels G, Schlingemann RO, et al. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. *Diabetes Care*. Apr 2019;42(4):651-656. [doi: [10.2337/dc18-0148](https://doi.org/10.2337/dc18-0148)] [Medline: [30765436](https://pubmed.ncbi.nlm.nih.gov/30765436/)]
22. 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*. May 2021;105(5):723-728. [FREE Full text] [doi: [10.1136/bjophthalmol-2020-316594](https://doi.org/10.1136/bjophthalmol-2020-316594)] [Medline: [32606081](https://pubmed.ncbi.nlm.nih.gov/32606081/)]
23. Ming S, Xie K, Lei X, Yang Y, Zhao Z, Li S, et al. Evaluation of a novel artificial intelligence-based screening system for diabetic retinopathy in community of China: a real-world study. *Int Ophthalmol*. Apr 03, 2021;41(4):1291-1299. [doi: [10.1007/s10792-020-01685-x](https://doi.org/10.1007/s10792-020-01685-x)] [Medline: [33389425](https://pubmed.ncbi.nlm.nih.gov/33389425/)]
24. Scheetz J, Koca D, McGuinness M, Holloway E, Tan Z, Zhu Z, et al. Real-world artificial intelligence-based opportunistic screening for diabetic retinopathy in endocrinology and indigenous healthcare settings in Australia. *Sci Rep*. Aug 04, 2021;11(1):15808. [FREE Full text] [doi: [10.1038/s41598-021-94178-5](https://doi.org/10.1038/s41598-021-94178-5)] [Medline: [34349130](https://pubmed.ncbi.nlm.nih.gov/34349130/)]
25. Li N, Ma M, Lai M, Gu L, Kang M, Wang Z, et al. A stratified analysis of a deep learning algorithm in the diagnosis of diabetic retinopathy in a real-world study. *J Diabetes*. Feb 2022;14(2):111-120. [FREE Full text] [doi: [10.1111/1753-0407.13241](https://doi.org/10.1111/1753-0407.13241)] [Medline: [34889059](https://pubmed.ncbi.nlm.nih.gov/34889059/)]
26. Malerbi FK, Andrade RE, Morales PH, Stuchi JA, Lencione D, de Paulo JV, et al. Diabetic retinopathy screening using artificial intelligence and handheld smartphone-based retinal camera. *J Diabetes Sci Technol*. May 2022;16(3):716-723. [FREE Full text] [doi: [10.1177/1932296820985567](https://doi.org/10.1177/1932296820985567)] [Medline: [33435711](https://pubmed.ncbi.nlm.nih.gov/33435711/)]
27. Mehra AA, Softing A, Guner MK, Hodge DO, Barkmeier AJ. Diabetic retinopathy telemedicine outcomes with artificial intelligence-based image analysis, reflex dilation, and image overread. *Am J Ophthalmol*. Dec 2022;244:125-132. [doi: [10.1016/j.ajo.2022.08.008](https://doi.org/10.1016/j.ajo.2022.08.008)] [Medline: [35970206](https://pubmed.ncbi.nlm.nih.gov/35970206/)]
28. Ruamviboonsuk P, Tiwari R, Sayres R, Nganthavee V, Hemarat K, Kongprayoon A, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. *Lancet Digit Health*. Apr 2022;4(4):e235-e244. [FREE Full text] [doi: [10.1016/S2589-7500\(22\)00017-6](https://doi.org/10.1016/S2589-7500(22)00017-6)] [Medline: [35272972](https://pubmed.ncbi.nlm.nih.gov/35272972/)]
29. 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*. Aug 2018;125(8):1264-1272. [doi: [10.1016/j.ophtha.2018.01.034](https://doi.org/10.1016/j.ophtha.2018.01.034)] [Medline: [29548646](https://pubmed.ncbi.nlm.nih.gov/29548646/)]

Abbreviations

- AI:** artificial intelligence
- AUROC:** area under the receiver operating characteristic curve
- DL:** deep learning
- DR:** diabetic retinopathy
- FN:** false negative
- FP:** false positive
- NHS:** National Health Service
- TN:** true negative
- TP:** true positive

Edited by B Puladi; submitted 11.09.23; peer-reviewed by M Liu, C Zheng, J Jaskari; comments to author 06.11.23; revised version received 30.12.23; accepted 22.03.24; published 14.08.24

Please cite as:

Wang Y, Han X, Li C, Luo L, Yin Q, Zhang J, Peng G, Shi D, He M

Impact of Gold-Standard Label Errors on Evaluating Performance of Deep Learning Models in Diabetic Retinopathy Screening: Nationwide Real-World Validation Study

J Med Internet Res 2024;26:e52506

URL: <https://www.jmir.org/2024/1/e52506>

doi: [10.2196/52506](https://doi.org/10.2196/52506)

PMID:

©Yueye Wang, Xiaotong Han, Cong Li, Lixia Luo, Qiuxia Yin, Jian Zhang, Guankai Peng, Danli Shi, Mingguang He. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 14.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.