PERSPECTIVE

Artificial intelligence in retinopathy of prematurity: transfer learning and federated learning

Carolyn Yu Tung Wong[1], *MBChB; Wilson Wai Kuen Yip*[2], *MBChB, MMed, FRCSEd (Ophth), FCOphthHK, FHKAM (Ophthalmology); Henry Hing Wai Lau*[2], *MBChB, MMed, FRCSEd (Ophth), FCOphthHK, FHKAM (Ophthalmology); Carol Yim Lui Cheung*[2], *BSc (HK), MPhil (HK), PhD (UK)*

[1]*Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China*

[2]*Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China*

Correspondence and reprint requests:
Dr Carolyn Yu Tung Wong, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China.
Email: carolcarol.carolyn@gmail.com

Abstract

Artificial intelligence can help resolve the lack of specialists in retinopathy of prematurity (ROP) and its associated diagnostic subjectivity. Obstacles to the use of deep learning (DL) for ROP tasks include the condition's low prevalence, data paucity, and difficulty in multi-institutional data sharing to optimize training. Transfer learning (TL) and federated learning (FL) are advanced strategies to address DL issues. This review highlights the advantages of TL and FL applications in various ROP-related tasks. TL and FL achieve outstanding results for ROP screening, triaging, and monitoring, with certain algorithms exceeding the baseline DL models. TL assists the construction of generalizable ROP models despite little data. FL lays the groundwork for safe data exchange between institutions in TL. However, both TL and FL entail shortcomings associated with inadequate generalizability and data privacy attacks. Further research is needed to address the unsolved interpretability and liability issues in TL and FL models. Although both TL and FL have great potential to overcome DL constraints and improve the diagnosis of ROP, more work is needed to address application concerns such as model interpretability and liability.

Key words: *Artificial intelligence; Deep learning; Machine learning; Retinopathy of prematurity*

Introduction

Retinopathy of prematurity (ROP) is a vasoproliferative condition involving the aberrant formation of retinal blood vessels in premature, lightweight newborns.[1] ROP can be classified by its anteroposterior position (area), severity (stage), and vascular features,[1] as well as the extent of vascularization (ie, zone 1, 2, or 3).[2] ROP classification guides treatment in clinical practice.[2,3] ROP is a leading cause of childhood blindness; an estimated 20 000 newborns with ROP become blind annually worldwide.[4,5] ROP-related blindness is mostly avoidable with early screening, proper diagnosis, and swift intervention.[6] Unfortunately, the global burden of ROP remains high because of a lack of experts to screen for the condition, particularly in middle-to-low-income countries.[7,8] Additionally, inter-clinician differences in the qualitative evaluation of ROP characteristics in images for diagnosis and severity triaging result in fluctuating standards and questionable reliability of ROP diagnoses and classifications.[9-11] Therefore, the use of artificial intelligence (AI) for data analysis and automated diagnosis is advocated.[12]

Deep learning (DL) is widely used to diagnose and classify various retinal diseases (including ROP).[13,14] Nonetheless, ROP presents greater challenges than other retinal disorders in DL-based AI. Transfer learning (TL) and federated learning (FL) can address the challenges associated with constructing AI models for ROP and better incorporate AI into ROP diagnostic pathways.

Relationships between artificial intelligence, machine learning, deep learning, transfer learning, and federated learning

The Figure shows the relationships between AI, machine learning (ML), DL, TL, and FL. Telescreening of fundus photographs for ROP facilitates the integration of computer-based image analysis.[15,16] Early systems for ROP diagnosis use manual and semi-automated ML methods to analyze data and make decisions based on learned patterns, whereas DL learns autonomously without explicit human guidance.[15,16] ML requires defining disease-specific features and training the machine for interpretation, whereas DL uses artificial neural networks to analyze images, recognize patterns, and continuously improve.[15-17] These networks, particularly convolutional neural networks (CNNs), are adept at image processing and classifying data beyond human perception.[17]

CNN is a DL architecture that learns from data, excelling in image pattern recognition for object and category identification.[18] It comprises an input layer, an output layer, and several hidden layers that transform data to uncover features.[18] Images are made of pixels arranged in a matrix, with values from 0 to 255 representing brightness and color.[19] CNNs mimic human brain function by first identifying simple patterns before progressing to complex ones, ultimately classifying images and diagnosing conditions.[19] Various architectures such as AlexNet (for large datasets) and ResNet (for deep networks) are designed for specific tasks, with differing layer configurations and parameters.[20,21] Training an ML model involves a dataset and validation processes to assess performance.[22] Metrics such as sensitivity and area under the receiver operating characteristic curve (AUROC) evaluate the algorithm.[22] The training feeds labeled data through the CNN layers, refining the model to reduce errors and requiring substantial image input for accuracy.[22]

TL utilizes knowledge from a previous task to improve performance on related tasks.[22,23] Common CNNs for TL include ResNet, ImageNet, U-Net, and VGG-16.[22,23] Fine-tuning a pretrained network requires less data than training from scratch, enabling the use of pretrained features for new tasks.[18] For example, a network trained on millions of images can be adapted for new classification tasks with just a few hundred images.[18]

In FL, multiple clients train a model collectively while keeping their data decentralized.[24] FL trains algorithms on local datasets without sharing data samples.[25] Local models are trained on local data, and parameters such as weights and biases are exchanged periodically among nodes to create a shared global model.[25]

Transfer learning for diagnosing retinopathy of prematurity

ROP is uncommon, and collecting adequate data to generate potent DL diagnostic classifiers is challenging.[26] Many ROP data for DL training are affected by class imbalances, low label quality owing to disparities in annotators' clinical experience and interpretations, and intrinsic biases towards

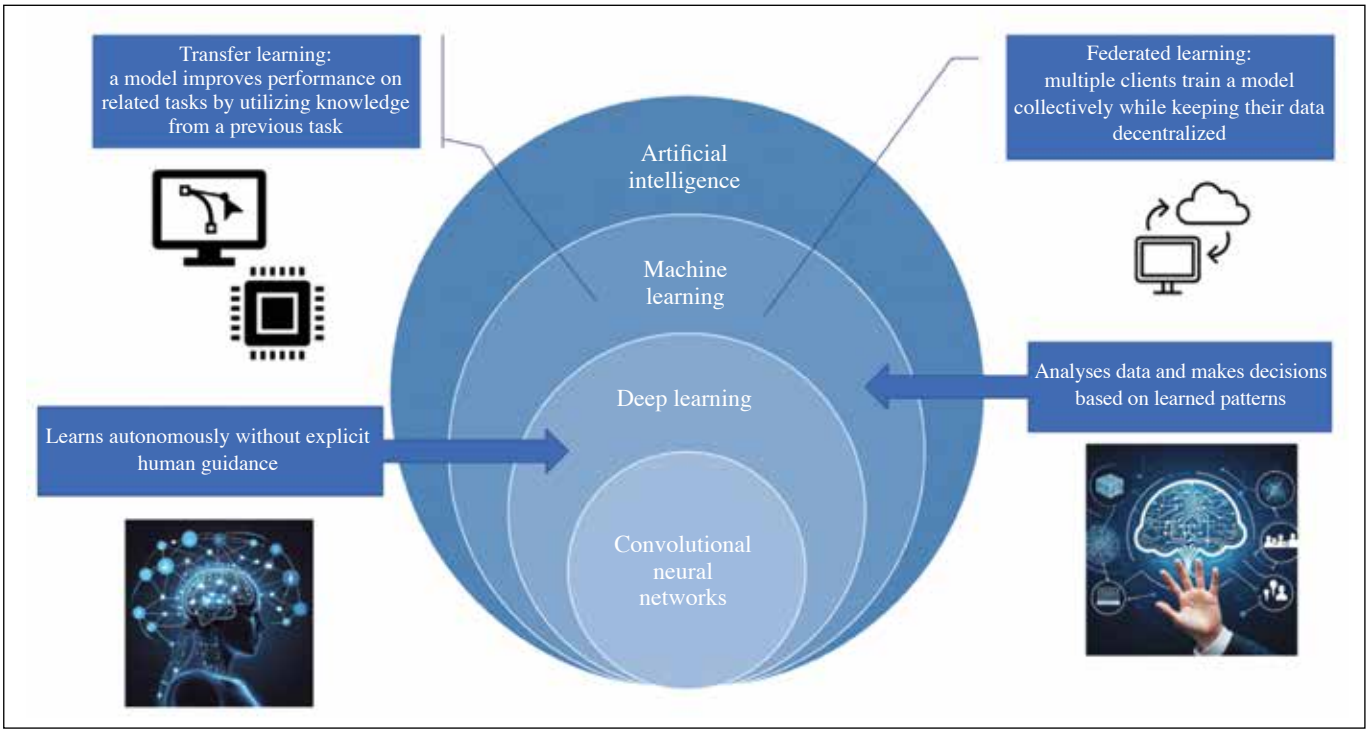

Figure. Relationships between artificial intelligence, machine learning, deep learning, transfer learning, and federated learning

the traits reflected in most single-ethnic datasets.[26,27] ROP models trained on a single population may not generalize well to different populations, whereas ROP models trained on populations in various institutions are limited because of privacy concerns.[26] Even if more ROP data can be accumulated over time to increase the database size, a new DL system will have to be built from scratch with different parameter settings owing to differences in image distribution in the new ROP training data.[28] Traditional DL requires model rebuilding, which increases expenditures on system maintenance and modifications to adapt to the population's ever-changing ROP scenario.[28] Furthermore, the process of searching for appropriate network parameters via an extensive trial-and-error procedure is time-consuming and costly while adjusting the dynamic ROP scene.[28] This decreases the incentive to create and incorporate DL systems for ROP classification.

To overcome these issues, TL enables a new model to reuse knowledge already acquired in another domain, task, or distribution in another model supplied with a more robust and diversified training dataset.[28] For example, the ImageNet dataset consists of millions of diverse images for CNNs; AlexNet trains on them to produce robustly learned model parameters and weights for training a new model using this knowledge via TL.[29] Fine-tuning the last fully connected layers using TL enables the retrained network to better adapt to the new task domain.[29] It is faster to set up a ROP system with TL, which transfers part of the established model settings, than to train a CNN from scratch, which requires developers to decide on each model setting.[29]

Table 1 summarizes studies of TL models for ROP. Rao et al[30] used TL to develop an ROP screening model using the limited number of ROP images available. The ImageNet-pretrained model was leveraged to construct the ROP diagnostic DL system using TL. The TL method yielded an AUROC of 0.970, with 91.46% sensitivity and 91.22% specificity in ROP.[30] When used correctly with appropriate networks, TL enables quicker convergence and highly accurate achievements.[30] The TL algorithm can decentralize care and increase ROP screening coverage, freeing up human experts for treatment planning.[30]

Wang et al[31] built an ROP screening and severity triaging system by using another ImageNet-pretrained model with TL. The TL-based screening and triaging systems achieved 96.62% sensitivity and 99.32% specificity for ROP screening, and 88.46% sensitivity and 92.31% specificity for ROP grading.[31]

Chen et al[32] used TL to devise AI models to classify ROP stages using the robust ImageNet-pretrained model. The TL-based models retrained on a North American dataset and a Nepali dataset demonstrated excellent staging performance when tested on data from the same population, with AUROCs of 0.99 and 0.97 and sensitivities of 94% and 73%, respectively.

Subramaniam et al[33] developed another ROP staging system (no-plus and plus categories) using the ImageNet-pretrained model via TL. The resulting TL-based ROP diagnostic algorithm achieved an AUROC of 0.9754, 96% accuracy, 100% specificity, and 71.43% sensitivity. The algorithm can be used in telemedicine for ROP using smartphone fundus photographs.

Brown et al[34] used TL to establish an ROP staging classifier based on the ImageNet-pretrained model. The TL-based algorithm attained AUROCs of 0.94 and 0.98 for normal and plus class recognition, respectively. In an independent test set, the sensitivity was 100% and specificity was 94% for pre-plus or worse diagnosis. TL helped create a continuous ROP scoring system, providing greater granularity in determining disease progression and improvement.

Mao et al[35] used the ImageNet-pretrained network to train an ROP staging system using TL. The TL-based system could diagnose plus-stage disease with 95.1% sensitivity and 97.8% specificity, and pre-plus-stage disease or worse with 92.4% and 97.4% sensitivity, respectively. The quadratic weighted kappa of the system, a metric that measures the agreement between the predicted and actual outcome (varying from 0 [random agreement] to 1 [complete agreement]), was 0.9244.[36] The TL method may provide useful second opinions after ophthalmologists' diagnoses.

Yildiz et al[37] used TL to generate an ROP staging system based on the robust ImageNet-pretrained network. The resulting TL system achieved an AUROC of 99% and 94% for plus and not-plus stages on the internal and external datasets, respectively. The internally and externally verified AUROC for predicting pre-plus or worse stages versus normal stages were 99% and 88%, respectively.

Tong et al[38] used TL to create an ROP triage system that classifies patients based on disease severity (normal, mild, semi-urgent, or urgent) and stage (1 to 4). The TL method obtained an accuracy of 0.903 for severity and 0.957 for ROP staging. TL effectively improved the competency, efficiency, and consistency of ROP screening.

TL enables accurate ROP identification and triaging for personalized treatment and follow-up plans. It reduces the need to gather vast quantities of ROP data and conduct high-quality image capturing and expert-led data labeling as in traditional DL model development.[28] TL models can screen and classify patients with ROP across various subgroups, despite the limited volume of training data.[29]

However, TL has limitations to ROP AI tasks. Despite ImageNet's ability to shape robust model settings for the initial phase of ROP model training, the image properties differ between ROP medical images and natural images. This could result in unrealistic and inaccurate knowledge transferring from the pretrained network.[39] ImageNet may fail to depict complex pathological structures in two or three

Table 1. Transfer learning (TF) retinopathy of prematurity (ROP) models

Study	Task	Algorithm	Dataset	Operations
Rao et al,[30] 2023	Differentiate between ROP (stage 1-3) and non-ROP eyes	EfficientNet-B0	227 326 wide-field retinal images	Involved pretraining on ImageNet; the EfficientNet-BO initialized was retrained as a binary.
Wang et al,[31] 2018	Classifying ROP and non-ROP, followed by grading on identified ROP-positives into minor or severe ROP	Inception BN as the ROP identification network (ROP vs non-ROP); Gr-Net as the ROP severity grading network (minor vs severe ROP)	Id-Net trained on an identification dataset and Gr-Net trained on a grading dataset	Training processes of Id-Net and Gr-Net were identical except for the training datasets. An Inception–BN network was first pretrained on Image-Net. The learned parameters were used as the initial parameters of Id-Net and Gr-Net, which were later fine-tuned.
Chen et al,[32] 2021	Classifying non-staged and staged ROP (1 to 3) images	Three CNNs trained on three different image datasets	5943 images from 711 patients in the North American dataset and 5009 images from 541 patients from the Nepali dataset	Three types of CNN models were trained using three datasets: North American alone, Nepali alone, and both. A ResNet-152 architecture pretrained on the ImageNet dataset was incorporated into the training of the three CNNs through TL.
Subramaniam et al,[33] 2023	Classifying plus and no plus	Algorithm developed from pretrained GoogleLeNet	5000 images from EyePACS	The ImageNet- pretrained GoogLeNet was leveraged. In the training, the first 20 layers were frozen, and the last learnable layer and the final classification layer were replaced with layers relevant to our dataset.
Brown et al,[34] 2018	Classifying normal, pre-plus, and plus disease in ROP	U-Net architecture as the vessel segmentation network; Inception version 1 architecture as the network to classify preprocessed images into normal, pre-plus, and plus	Pretraining of Inception version 1 involves ImageNet. Training of the TL algorithm involves 5511 retinal photographs, with 4535 (82.3%) being normal, 805 (14.6%) being pre-plus disease, and 172 (3.1%) being plus disease.	The second CNN (Inception version 1) architecture was previously pretrained on the ImageNet database of 1.2 million images from 1000 classes. In the training of the algorithm, the two networks were presented with corresponding reference standard diagnoses, which were used to adjust the network's numerous internal parameters to output the correct diagnoses.
Mao et al,[35] 2020	Classifying normal, pre-plus, and plus disease	Three deep CNNs, including a U-Net for vessel segmentation, another U-Net for optic disc segmentation and a DenseNet for the three-class classification	5711 images for training, 450 images for testing, and 63 images for treatment	The input of the DenseNet was the output of the modified U-net that segmented the blood vessels. For the DenseNet, the weights were initialized using TL based on the ImageNet.
Yildiz et al,[37] 2020	Classifying plus vs not-plus; classifying pre-plus or worse vs normal	Three classifiers: logistic regression, support vector machine, and neural networks	5512 images (163 plus, 802 pre-plus, and 4547 normal) for training and validation; 100 images (15 plus, 34 pre-plus, and 51 normal) for external validation	The pipeline of the system begins with segmenting the vessels in a color retina image. The system incorporated a pretrained U-Net CNN architecture for segmentation. Using the segmented images and optic disc centers, vessels were traced and vessel tree information was extracted. Using the outputs from previous steps, features of the retina were extracted, and these features were used for classification in the system.
Tong et al,[38] 2020	Classifying ROP severity (normal, mild, semi-urgent and urgent); classifying ROP into stages (stages 1 to 5) and detecting plus disease	Two CNNs: the 101-layer ResNet as classification network) and the Faster R-CNN as an identification network	36 231 fundus images	A ResNet-101 CNN architecture was pretrained on the ImageNet. It was then retrained on the study's dataset using TL, while the fine-tuning technique was applied to transfer the connection weights from the pretrained model to the study's model, and the model was retrained to the present task.

Results	Remarks	Added value to the ROP clinical pathway	Limitation
91.46% sensitivity, 91.22% specificity, 0.970 AUROC, 81.72% positive predictive value, and 96.14% negative predictive value	TL enabled faster convergence, potential higher diagnostic accuracy acquisitions, and lower operational demand for training networks.	The highly accurate TL model can be used as an objective ROP screening tool, facilitating broader screening coverage and care decentralization (ie, moving triaging away from an ophthalmologist-led system).	Limited generalizability owing to single ethnicity (South Asian) population; uncertain generalizability for stage 4 & 5 ROP disease detection owing to remaining class imbalance (ie, inadequate representation of the very advanced stages)
In test dataset, the Id-Net achieved 96.62% sensitivity and 99.32% specificity for ROP identification, whereas Gr-Net achieved 88.46% sensitivity and 92.31% specificity for ROP grading. In 552 cases, the deep neural networks outperformed some human experts and achieved 84.91% sensitivity and 96.90% specificity for ROP identification, whereas the corresponding values for ROP grading were 93.33% and 73.63%, respectively.	TL demonstrated its capability to derive high-performing deep neural networks. Although TL is less sensitive to the size of training dataset, the quality and quantity of pretraining and retraining datasets are still relevant. The dataset plays a crucial role in avoiding overfitting the algorithm.	TL helped the ROP screening system to reach ROP-expert comparable performance, facilitating ROP diagnostic consistency. The system could be used to detect and triage ROP into different grades for individualized management plans.	Limited generalizability owing to insufficient severe ROP cases (ie, remaining class imbalance) and persistently limited data availability
Both the North American- and Nepali-trained models demonstrated high performance on the test set from the same population, with 0.99 AUROC, 0.98 area under precision-recall curve, and 94% and 73% sensitivity, respectively. Compared to the models trained on individual datasets, the model trained on a combined dataset had better performance on each test set, with 98% sensitivity in the North American test set and 82% sensitivity in the Nepali test set.	Internal and external performance of the algorithm was most improved by increasing the heterogeneity of the training dataset features (eg, by combining images from different populations and cameras).	CNNs could be trained to detect stages 1 to 3 ROP with high accuracy, facilitating triaging patients with ROP for differential interventions and management.	Limited generalizability due to the filtered images not being adequately representative of the real-world population; absence of expandability provided for the CNN imaging features associated with stages 1 to 3
In the held-out test set of 50 images, 0.9754 AUC, 96% accuracy, 100% specificity, and 71.43% sensitivity were obtained. Precision was 1, recall was 0.7143, and the F1 score was 0.833.	Despite TL, the limited number of cases of plus disease necessitated further collaboration with other countries in the SP-ROP project and improved access to more smartphone images.	The successful performance of the TL-trained solution suggested the potential for more accurate diagnosis using smartphone fundus images and the future utilization of these images for telemedicine-based diagnoses.	Limited generalizability remained caused by the limited access to diverse and vast amounts of images.
The mean AUROC was 0.94 for the normal diagnosis and 0.98 for the plus-stage diagnosis. In an independent test set of 100 retinal images, the algorithm achieved 93% sensitivity and 94% specificity for plus-stage diagnosis. For the detection of pre-plus disease or worse, sensitivity was 100% and specificity was 94%. On the same test set, the algorithm reached a quadratic weighted κ coefficient of 0.92, outperforming 6 of 8 ROP experts.	By enabling the network to learn highly generalizable image features from an unrelated yet large and highly diverse dataset of images, TL was conducive to the acceleration of classification performance in the algorithm.	The TL-based system allowed continuous scoring, providing more granularity in determining relative disease progression or regression and facilitating the formulation of management plans for patients with progression monitoring. Incorporating the TL model into fundus camera systems and telemedicine platforms for ROP and other image-based diseases may improve the objectivity, accuracy, and efficiency of healthcare delivery.	The robustness of the TL algorithm remained constrained by persistent data limitations (eg, image quality, resolution, camera systems, and field of view). Biased training might have remained owing to the limited data variations. It is a poorly interpretable 'black box' model.
The trained network achieved 95.1% sensitivity with 97.8% specificity for the diagnosis of plus disease. For the detection of pre-plus or worse, the sensitivity was 92.4% and specificity was 97.4%. The quadratic weighted κ was 0.9244.	The TL training in DenseNet contributed to better generalizability and robustness in the classifier. The selection of the appropriate network was equally important for allowing easier training and alleviating overfitting issues.	The proposed system provides an accurate diagnosis of plus disease and can act as an assistive diagnostic tool to validate ophthalmologists' judgements.	Limited generalizability remained owing to the persistent lack of labelled images.
The AUCs on the first and second datasets were 99% and 94%, respectively, for predicting plus vs not-plus categories, and 99% and 88%, respectively, for predicting pre-plus or worse vs normal categories.	The CNN can identify more discriminative features given the exposure to a large dataset for training in pretraining during the TL training process.	This fully automated system, which combines retinal vessel segmentation, tracing, feature extraction and classification stages, diagnosed plus disease in ROP with performance on par with recent publications reporting on the use of CNNs. It is likely to benefit holistic ROP diagnostic practices.	Uninterpretable black-box nature; quality of input images distorted by image resizing
The system achieved 90.3% accuracy for classification of ROP severity. Specifically, the accuracies in discriminating normal, mild, semi-urgent, and urgent were 88.3%, 90.0%, 95.7%, and 87.0%, respectively, whereas the corresponding accuracies of the two experts were 90.2% and 89.8%. The model also achieved 95.7% accuracy for detecting the ROP stage and 89.6% accuracy for detecting plus disease. The accuracies in discriminating stages I to stage V were 0.876, 0.942, 0.968, 0.998, and 0.999, respectively.	After TL, the proposed ROP classifier system performed well without requiring a novel database of millions of images. It allowed the two CNN models to be applied as the study's classification and identification algorithms, which perform screening functions with proficiency comparable to or better than that of ROP experts.	The most prominent advantage is that the TL attempt promoted the identification of ROP stage and plus disease presence, along with disease severity. This functionality enables clinical review and verification of the automated diagnosis, rather than simply identifying the presence of ROP. TL benefits also include consistent prediction and instantaneous reporting of results.	Limited generalizability owing to the persistently limited number of ROP stage V fundus images (ie, bias) and an inadequately large patient cohort.

dimensions.[39] Thus, leveraging the power of large amounts of non-realistic big data may shift the model's focus to non-generalizable image attributes that are unrelated to the disease or clinical anatomy in a given collection of medical images.[39] Consequently, biased training may undermine the algorithm's capacity to recognize disease patterns in new data.[39] Therefore, gathering massive volumes of raw ROP medical images is paramount in generalizable training. Furthermore, the selection of the TL technique, the backbone architecture, and the fine-tunable layers are critical in guiding the success of a smooth and accurate conveyance of generalizable characteristics.[40,41] However, most key decisions in TL were based primarily on ophthalmologists' intuition or personal experience.[42] There is presently no agreement or defined procedure for making such training decisions in TL. This puts TL-trained models at risk of incurring uncalculated decision making, which may limit the performance boost that TL provides.[42]

Federated learning for diagnosing retinopathy of prematurity

Data sharing among institutions fosters generalizable training. Multicenter research is increasingly important in developing robust ROP DL algorithms that can be used in real-world applications.[43,44] The most common approach is centralized learning, in which data from multiple institutions are pooled in a centralized model. However, exchanging data among institutions increases the risk of data breaches and privacy violations.[45]

To protect data privacy and reduce the risk of raw ROP data leakage, the distributed learning method allows ROP data to be disseminated among institutions rather than combined into a single pool.[46] FL enables multiple medical institutions to cooperatively train AI systems without sharing data.[47-49] Thus, FL promotes ROP AI research and development, necessitating multicenter collaboration and access to large-scale data for performance excellence.

Table 2 summarizes studies of FL models for ROP. Lu et al[50] used FL to create a no plus, pre-plus, and plus ROP classification model. After training with 1145 color fundus photographs, both the FL and centralized learning models performed well, with AUROCs ranging from 0.93 to 0.96. In four of seven sites, FL outperformed locally trained models that used a single-institution dataset. FL can generate higher-quality models by leveraging larger and more diverse training datasets from multiple sources while maintaining a high standard of data privacy.

Hanif et al[51] applied FL to develop a vascular severity scoring system for three ROP subclasses (no plus, pre-plus, and plus). They compared the average DL-generated vascular severity score for each ROP class across institutions to determine inter-institutional variability. There were no significant inter-institutional disparities between the pre-plus and plus groups; however, there were significant disparities in the mean vascular severity score for the no-plus group. The vascular severity scoring system for ROP can be an objective measure for inter-clinician diagnostic variability. FL has the potential to improve ROP screening uniformity, objectivity, and consistency, as well as intervention triage.

However, FL for ROP has limitations. Despite the FL models' superior overall performance, locally trained DL models are not inferior to the FL models. 43% of the participating sites likely gained no advantage from the FL approach, especially when FL models are more vulnerable to data breaches (inference attacks and model inversion attacks). Furthermore, despite the inclusion of real-life clinical data, FL models remain stimulated models that fail to address practical FL challenges such as different privacy restrictions among participating institutions, varying accessibility, extensive communication and overhead expenses, and unpredictable reliability among participating sites in terms of supplying high-quality input data.

Real-world use of models

The i-ROP DL is the first DL model to receive breakthrough designation from the United States Food and Drug Administration for real-world ROP screening. The i-ROP DL system was trained on over 5000 deidentified posterior retinal images captured using a RetCam camera as part of the multicenter Imaging and Informatics in Retinopathy of Prematurity (i-ROP) cohort study.[34] All eight study centers used a standard imaging protocol, capturing images in five standard fields of view (posterior, nasal, temporal, superior, and inferior).[34] However, the analysis focused solely on wide-angle images of the posterior retina, excluding images without an optic nerve present.[34] Subsequent studies developing ROP DL systems also adopted this approach of analyzing only posterior retinal images.[52,53]

To ensure quality control in the images used for model training, it is recommended that a reference standard diagnosis is assigned to each image.[54] This involves independent image-based diagnoses by three trained graders (two ophthalmologists and one study coordinator) and a clinical diagnosis by an expert ophthalmologist.[54] The images are classified as normal, pre-plus disease, or plus disease according to the international classification of ROP, which includes zone, stage, and plus disease.[34,52] The reference standard diagnosis serves as the foundation for training the DL system.[34] Images are excluded if at least two of the three graders deem them unacceptable for diagnosis or if they exhibit stage 4 or 5 ROP.[52] In these advanced stages, the relevance of diagnosing plus disease for ROP screening diminishes, as visualizing retinal blood vessels becomes challenging.[34] The effectiveness of incorporating a reference standard diagnosis consensus procedure for training highly accurate ROP DL systems has been demonstrated.[52,53]

Brown et al[34] reported that the i-ROP-DL system demonstrated expert-level automated performance, with 91% accuracy in diagnosing plus disease in premature infants from a North American population.[34] The i-ROP DL

Table 2. Federated learning (FL) retinopathy of prematurity (ROP) models

Study	Task	Algorithm	Dataset	Operations	Results	Remarks	Added value to the ROP clinical pathway	Limitation
Lu et al,[50] 2022	Classifying three level plus in ROP	ResNet-18	5255 wide-angle retinal images in the neonatal intensive care units of 7 institutions	Model averaging was used. At the start of each round of training, a copy of the global model was shared with each of the 7 institutions to initialize that institution's local model. Each local model was then trained for 10 epochs, and the model checkpoint with the best validation performance, as measured by AUROC, was chosen to be aggregated into the global model at the end of the round. Upon the completion of all the federation rounds, a copy of the global model was finally transferred to each individual site for the prediction of new, unseen data.	Four (57%) of the seven models trained on local institutional data performed inferiorly to the FL models.	FL has been promising for allowing multiple institutions to leverage their individual data resources to collaboratively develop DL models without directly sharing data while reducing time and costs and protecting patient privacy. Each contributing entity eventually benefits from an aggregated model trained on a larger and more heterogeneous distribution of cases, which often gives far more generalizable models with greater performance than that of standalone models trained using only a single institution's data.	FL trained on point-of-care labels performed comparably to models trained on centralized datasets with consensus-expert labels, supporting the feasibility of the approach in clinical settings. FL approach is more secure and convenient than the more commonly used collaborative approach and is a beneficial alternative.	Datasets contributed by each institution were not entirely based on the population, causing insufficient generalizability; absence of assessment in the practical obstacles to implementing this method, such as limited communication bandwidth; sole focus on the performance aspect of FL without taking into account practical attacks on FL in a real-world setting that may require additional privacy-preserving measures
Hanif et al,[51] 2022	Classifying plus, pre-plus and no plus	-U-Net	Wide-angle retinal images in 5245 patients from the neonatal intensive care units of 7 institutions	Model averaging was used. At the beginning of each round of training, a copy of the global model was shared with each institution to initialize that institution's local model. The local model is then trained for a set number of epochs, and the model checkpoint from the epoch with the best validation performance, (as measured by AUROC), is chosen to be aggregated into the global model at the end of the round.	The proportion of patients diagnosed with pre-plus disease varied significantly between institutions (p<0.00l). A significant inverse relationship between the institutional vascular severity score and the mean gestational age was found (p=0.049, adjusted R2=0.49).	This is likely to represent a generalizable approach to assess clinician diagnostic paradigms as well as disease severity for epidemiological evaluation across institutions without the need for sharing sensitive patient data.	In comparison to a centrally hosted trained model, FL model training eliminates the need for a central process or consensus of ROP experts, making it more feasible to implement.	Each institution's datasets were not completely population based. There were also variations in the participating sites' enrolment protocols, potentially introducing population bias into the cohorts. These factors led to persistent inadequate generalizability.

can also provide a quantitative severity score (1 to 9) based on a single photograph, correlating with full zone, stage, and plus disease classification.[52,55-57] It is highly sensitive in detecting treatment-requiring ROP in a North American population.[52,58]

Campbell et al[59] assessed the diagnostic performance of the i-ROP-DL system-based ROP severity score using data from an Indian ROP telescreening program.[59] The system demonstrated sustained high diagnostic accuracy at the individual eye examination level, with an AUROC of 0.98, 100% sensitivity, and 78% specificity for identifying treatment-requiring ROP in diverse patient groups. At the population level, the system identified higher ROP severity in neonatal care units lacking resources for oxygen monitoring and titration.[59] These findings suggest that the system can enhance ROP screening efficiency and serve as an epidemiological tool for monitoring ROP severity across different regions and time periods.[59] As an autonomous screening device, the system can effectively expand ROP screening coverage across large geographic areas, provide automated real-time referral decisions, and reduce the screening workload by 60% to 80%.[59] Additionally, the i-ROP-DL system could aid in resource allocation to neonatal care units caring for high-risk infants susceptible to treatment-requiring ROP and aggressive posterior ROP.[59]

Table 3 summarizes the advantages and disadvantages of TL and FL, as well as the overall effect of AI, ML, and DL on ROP screening.

Future directions

Poor transparency is seen in advanced AI techniques' decision-making. Most ROP models are black boxes, making it difficult for ophthalmologists to understand how they arrive at their decisions.[60] To completely trust AI's clinical rationality, future ROP AI research should

Table 3. Pros and cons of artificial intelligence (AI), machine learning (ML), deep learning (DL), transfer learning (TL), and federated learning (FL) for retinopathy of prematurity (ROP) screening

	Pros	Cons
AI	Improved decision support: AI can aid healthcare providers by offering diagnostic assistance, thereby enhancing the precision of ROP screening.	Interpretability concerns: AI systems can be intricate and may lack clarity, making it challenging for clinicians to comprehend the reasoning behind decisions.
	Process automation: AI can streamline routine tasks including image analysis, saving time and lessening the burden on clinicians.	Reliance on data quality: The success of AI systems depends significantly on the quality and quantity of the training data.
	Multimodal data integration: AI can evaluate different types of data (such as clinical, imaging, and demographic information) to deliver a thorough assessment of ROP risk.	Challenges in implementation: Incorporating AI into current clinical workflows can be difficult and may necessitate substantial adjustments to existing practices.
ML	Predictive analytics: ML algorithms can detect patterns in data that may elude human observers, potentially facilitating earlier identification of ROP.	Data constraints: ML models need substantial, well-annotated datasets for training, which can pose challenges in specialized areas like ROP.
	Flexibility: ML models can be updated with new data, enabling them to enhance their performance over time as additional information becomes available.	Overfitting concerns: Without proper management, ML models may become overly tailored to the training data, resulting in poor performance on new, unseen data.
	Resource efficiency: ML can swiftly analyze extensive datasets, allowing for the efficient screening of a large number of patients.	Data bias: If the training dataset contains biases, the ML model may generate biased results, negatively impacting patient care.
DL	High accuracy in image analysis: DL, especially convolutional neural networks, is highly effective in evaluating medical images and often achieves impressive accuracy in identifying ROP.	High computational demand: DL requires considerable computational resources for both training and inference, which may not be accessible in every clinical environment.
	Automated feature learning: DL models can automatically extract relevant features from raw data, minimizing the need for manual feature selection.	Data intensive: DL models generally need substantial labelled datasets to perform effectively, which can be a drawback in ROP screening contexts.
	Scalability: DL models are capable of processing large datasets and recognizing complex patterns, making them ideal for extensive ROP screening initiatives.	Limited interpretability: DL models often function as 'black boxes', making it challenging to understand their decision-making processes, a critical aspect in clinical settings.
TL	Shorter training time: Utilizing pretrained weights from existing models significantly reduces the time needed to train a new model on ROP data.	Domain shift challenges: If the source domain (where the model was pretrained) differs considerably from the target domain (ROP images), the model's performance might suffer.
	Enhanced performance: Leveraging knowledge from related tasks can improve the model's effectiveness, particularly when the dataset for ROP is limited.	Risk of overfitting: Overfitting is a possibility if the model is not adequately fine-tuned for the specific ROP dataset.
	Reduced data dependency: The TL model can yield satisfactory results with smaller datasets, which is advantageous in medical imaging where annotated data may be hard to obtain.	Limited explainability: Employing pretrained models can complicate the understanding of how decisions are made in relation to ROP.
FL	Data privacy: FL facilitates the training of ROP models on decentralized data while keeping sensitive patient information secure, which is essential.	Complex setup: Implementing federated learning systems can be technically demanding and necessitates strong infrastructure.
	Varied data sources: It allows the local ROP model to learn from diverse datasets across multiple institutions, enhancing the model's generalization and robustness.	Increased communication needs: Regular interactions between the central server and local devices may result in higher latency and greater resource usage.
	Ongoing learning: The ROP model can be continuously updated as new data from various sources becomes available, improving its performance over time.	Challenges of heterogeneous data: Differences in the quality and distribution of ROP data across institutions can complicate the training process and impact model performance.

follow the explainable AI strategy, in which the AI system is divided into numerous components (eg, pre-diagnostic module, picture segmentation module, and final diagnosis module) for ophthalmologists to visualize.

It is unclear if healthcare providers, developers, suppliers, or regulators should be held responsible for AI system errors in real-world clinical practice despite being properly clinically verified. The distribution of liability clarifies not just whether and from whom patients should seek restitution but also if AI systems may find their way into clinical practice.[60] Before AI may be used in clinical settings, ethical frameworks must be developed that outline the legal responsibility of various parties in ensuring that AI functions in a specified manner and implementing proper compensating actions if and when harm happens.

Conclusion

TL and FL are useful in improving generalizability,

repeatability, and data safety when creating ROP models. However, the benefits of TL and FL for ROP AI training may not always result in improved ROP diagnosis and triaging. Future attempts to narrow the interpretability and liability gaps are needed.

Contributors

CYTW and HHWL designed the study, acquired the data, analyzed the data, and drafted the manuscript. All authors critically revised the manuscript for important intellectual content. All authors had full access to the data, contributed to the study, approved the final version for publication, and take responsibility for its accuracy and integrity.

Conflicts of interest

All authors have disclosed no conflict of interest.

Funding/support

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

All data analyzed in this study are available from the corresponding author upon reasonable request.

References

1. *Dogra MR, Katoch D, Dogra M. An update on retinopathy of prematurity (ROP). Indian J Pediatr 2017;84:930-6.*
2. *Kumar V, Patel H, Paul K, Surve A, Azad S, Chawla R. Deep learning assisted retinopathy of prematurity screening technique. HEALTHINF 2021;234-43.*
3. *Multicenter trial of cryotherapy for retinopathy of prematurity. One-year outcome--structure and function. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Arch Ophthalmol 1990;108:1408-16.*
4. *Good WV, Early Treatment for Retinopathy of Prematurity Cooperative Group. Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial. Trans Am Ophthalmol Soc 2004;102:233-50.*
5. *Blencowe H, Lawn JE, Vazquez T, Fielder A, Gilbert C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. Pediatr Res 2013;74(Suppl 1):35-49.*
6. *Campbell JP, Chiang MF, Chen JS, et al. Artificial intelligence for retinopathy of prematurity: validation of a vascular severity scale against international expert diagnosis. Ophthalmology 2022;129:e69-76.*
7. *Campbell JP, Ryan MC, Lore E, et al. Diagnostic discrepancies in retinopathy of prematurity classification. Ophthalmology 2016;123:1795-801.*
8. *Prakalapakorn SG, Greenberg L, Edwards EM, Ehret DEY. Trends in retinopathy of prematurity screening and treatment: 2008-2018. Pediatrics 2021;147:e2020039966.*
9. *Fleck BW, Williams C, Juszczak E, et al. An international comparison of retinopathy of prematurity grading performance within the Benefits of Oxygen Saturation Targeting II trials. Eye 2018;32:74-80.*
10. *Gschließer A, Stifter E, Neumayer T, et al. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. Am J Ophthalmol 2015;160:553-60.e3.*
11. *Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol 2007;125:875-80.*
12. *Peng Y, Chen Z, Zhu W, et al. ADS-Net: attention-awareness and deep supervision based network for automatic detection of retinopathy of prematurity. Biomed Opt Express 2022;13:4087-101.*
13. *Tan Z, Simkin S, Lai C, Dai S. Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease. Transl Vis Sci Technol 2019;8:23.*
14. *Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019;103:167-75.*
15. *Scruggs BA, Chan RVP, Kalpathy-Cramer J, Chiang MF, Campbell JP. Artificial intelligence in retinopathy of prematurity diagnosis. Transl Vis Sci Technol 2020;9:5.*
16. *Gensure RH, Chiang MF, Campbell JP. Artificial intelligence for retinopathy of prematurity. Curr Opin Ophthalmol 2020;31:312-7.*
17. *Ramanathan A, Athikarisamy SE, Lam GC. Artificial intelligence for the diagnosis of retinopathy of prematurity: a systematic review of current algorithms. Eye 2023;37:2518-26.*
18. *Indolia S, Goswami AK, Mishra SP, Asopa P. Conceptual understanding of convolutional neural network: a deep learning approach. Procedia Comput Sci 2018;132:679-88.*
19. *Taye MM. Theoretical understanding of convolutional neural network: concepts, architectures, applications, future directions. Computation 2023;11:52.*
20. *Anwar A. Difference between AlexNet, VGGNet, ResNet, and inception. Towards Data Science. Accessed 27 July 2024. Available from: https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96.*
21. *Nabil M. Unveiling the diversity: A comprehensive guide to types of CNN architectures Medium. Accessed 27 July 2024. Available from: https://medium.com/@navarai/unveiling-the-diversity-a-comprehensive-guide-to-types-of-cnn-architectures-9d70da0b4521.*
22. *Maitra P, Shah PK, Campbell PJ, Rishi P. The scope of artificial intelligence in retinopathy of prematurity (ROP) management. Indian J Ophthalmol 2024;72:931-4.*
23. *Bai A, Carty C, Dai S. Performance of deep-learning artificial intelligence algorithms in detecting retinopathy of prematurity: a systematic review. Saudi J Ophthalmol 2022;36:296-307.*
24. *Kairouz P, McMahan HB, Avent B, et al. Advances and open problems in federated learning. Foundations Trends Machine Learning 2021;14:1-210.*
25. *Federated learning. Wikipedia. Accessed 27 July 2024. Available from: https://en.wikipedia.org/w/index.php?title=Federated_learning&oldid=1236419947.*
26. *Coyner AS, Chen JS, Chang K, et al. synthetic medical images for robust, privacy-preserving training of artificial intelligence: application to retinopathy of prematurity*

diagnosis. Ophthalmol Sci 2022;2:100126.
27. *Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. Ophthalmology 2016;123:2345-51.*
28. *Ruamviboonsuk P, Kaothanthong N, Ruamviboonsuk V, Theeramunkong T. Transfer Learning for Artificial Intelligence in Ophthalmology. In: Yogesan K, Goldschmidt L, Cuadros J, Ricur G, editors. Digital Eye Care and Teleophthalmology: a Practical Guide to Applications. Springer International Publishing; 2023: 181-98.*
29. *Al-Timemy AH, Ghaeb NH, Mosa ZM, Escudero J. Deep transfer learning for improved detection of keratoconus using corneal topographic maps. Cognit Comput 2022;14:1627-42.*
30. *Rao DP, Savoy FM, Tan JZE, et al. Development and validation of an artificial intelligence based screening tool for detection of retinopathy of prematurity in a South Indian population. Front Pediatr 2023;11:1197237.*
31. *Wang J, Ju R, Chen Y, et al. Automated retinopathy of prematurity screening using deep neural networks. EBioMedicine 2018;35:361-8.*
32. *Chen JS, Coyner AS, Ostmo S, et al. Deep learning for the diagnosis of stage in retinopathy of prematurity: accuracy and generalizability across populations and cameras. Ophthalmol Retina 2021;5:1027-35.*
33. *Subramaniam A, Orge F, Douglass M, et al. Image harmonization and deep learning automated classification of plus disease in retinopathy of prematurity. J Med Imaging (Bellingham) 2023;10:061107.*
34. *Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol 2018;136:803-10.*
35. *Mao J, Luo Y, Liu L, et al. Automated diagnosis and quantitative analysis of plus disease in retinopathy of prematurity based on deep convolutional neural networks. Acta Ophthalmol 2020;98:e339-45.*
36. *Understanding the quadratic weighted kappa. Kaggle. Accessed 17 July 2024. Available from: https://www.kaggle.com/code/reighns/understanding-the-quadratic-weighted-kappa.*
37. *Yildiz VM, Tian P, Yildiz I, et al. Plus disease in retinopathy of prematurity: convolutional neural network performance using a combined neural network and feature extraction approach. Transl Vis Sci Technol 2020;9:10.*
38. *Tong Y, Lu W, Deng QQ, Chen C, Shen Y. Automated identification of retinopathy of prematurity by image-based deep learning. Eye Vis (Lond) 2020;7:40.*
39. *Cadrin-Chênevert A. Moving from ImageNet to RadImageNet for improved transfer learning and generalizability. Radiol Artif Intell 2022;4:e220126.*
40. *Ayana G, Dese K, Choe SW. Transfer learning in breast cancer diagnoses via ultrasound imaging. Cancers (Basel) 2021;13:738.*
41. *Iman M, Arabnia HR, Rasheed K. A review of deep transfer learning and recent advancements. Technologies 2023;11:40.*
42. *Li W, Huang R, Li J, et al. A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: theories, applications and challenges. Mech Syst Signal Process 2022;167:108487.*
43. *Campbell JP, Lee AY, Abràmoff M, et al. Reporting guidelines for artificial intelligence in medical research. Ophthalmology 2020;127:1596-9.*
44. *Ting DSW, Wong TY, Park KH, Cheung CY, Tham CC, Lam DSC. Ocular Imaging Standardization for Artificial Intelligence Applications in Ophthalmology: the joint position statement and recommendations from the Asia-Pacific Academy of Ophthalmology and the Asia-Pacific Ocular Imaging Society. Asia Pac J Ophthalmol (Phila) 2021;10:348-9.*
45. *Tom E, Keane PA, Blazes M, et al. Protecting data privacy in the age of ai-enabled ophthalmology. Transl Vis Sci Technol 2020;9:36.*
46. *Chang K, Balachandar N, Lam C, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc 2018;25:945-54.*
47. *Yang Z, Chen M, Wong KK, Poor HV, Cui S. Federated learning for 6G: applications, challenges, and opportunities. Proc Est Acad Sci Eng 2022;8:33-41.*
48. *Konečný J, Brendan McMahan H, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: strategies for improving communication efficiency. Accessed 27 July 2024. Available from: http://arxiv.org/abs/1610.05492.*
49. *Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3:119.*
50. *Lu C, Hanif A, Singh P, et al. Federated learning for multicenter collaboration in ophthalmology: improving classification performance in retinopathy of prematurity. Ophthalmol Retina 2022;6:657-63.*
51. *Hanif A, Lu C, Chang K, et al. Federated learning for multicenter collaboration in ophthalmology: implications for clinical diagnosis and disease epidemiology. Ophthalmol Retina 2022;6:650-6.*
52. *Redd TK, Campbell JP, Brown JM, et al. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. Br J Ophthalmol 2018:bjophthalmol-2018-313156.*
53. *Sharafi SM, Ebrahimiadib N, Roohipourmoallai R, Farahani AD, Fooladi MI, Khalili Pour E. Automated diagnosis of plus disease in retinopathy of prematurity using quantification of vessels characteristics. Sci Rep 2024;14:6375.*
54. *Ryan MC, Ostmo S, Jonas K, et al. Development and evaluation of reference standards for image-based telemedicine diagnosis and clinical research studies in ophthalmology. AMIA Annu Symp Proc 2014;2014:1902-10.*
55. *Bellemo V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. Lancet Digit Health 2019;1:e35-44.*
56. *Ting DSW, Wu WC, Toth C. Deep learning for retinopathy of prematurity screening. Br J Ophthalmol 2018:bjophthalmol-2018-313290.*
57. *Taylor S, Brown JM, Gupta K, et al. Monitoring disease progression with a quantitative severity scale for retinopathy of prematurity using deep learning. JAMA Ophthalmol 2019;137:1022-8.*
58. *Greenwald MF, Danford ID, Shahrawat M, et al. Evaluation of artificial intelligence-based telemedicine screening for retinopathy of prematurity. J AAPOS 2020;24:160-2.*
59. *Campbell JP, Singh P, Redd TK, et al. Applications of artificial intelligence for retinopathy of prematurity screening. Pediatrics 2021;147:e2020016618.*
60. *Li Z, Wang L, Wu X, et al. Artificial intelligence in ophthalmology: The path to the real-world clinic. Cell Rep Med 2023;4:101095.*