• Intelligent Ophthalmology •

Guest Editor

Prof. Wei-Hua Yang

Executive Deputy Director of Shenzhen Eye Institute; Director of the Office of Big Data and Artificial Intelligence, Shenzhen Eye Hospital; Guest Researcher at the Chinese Academy of Sciences; Vice Chairman and Secretary General of the Intelligent Medicine Special Committee of the Chinese Medicine Education Association; Vice Chairman and Secretary General of the Intelligent Ophthalmology Branch of the Chinese Medicine Education Association; Vice Chairman and Secretary General of the Ophthalmic Imaging and Intelligent Medicine Branch of the Chinese Medicine Education Association; Vice President of the Spectacles Quality Inspection and Optometry Special Committee of the China Association for Quality Inspection; Editor-in-Chief of monographs including *Intelligent Medical Engineering Series*, *Artificial Intelligence Research of Retinal Diseases*, *Introduction to Intelligent Ophthalmology*, *Artificial Intelligence in Ophthalmology*, *Medical Virtual Reality and Augmented Reality*; Author of more than 160 papers, including more than 60 papers indexed in SCI; Guest Editor and Reviewer for more than 10 journals indexed in SCI; Co-author of more than 10 Clinical Guidelines (Expert Consensus) on *Intelligent Ophthalmology*.

Guest Editor

Prof. Yi Shao

The First Affiliated Hospital of Nanchang University, Chief Physician, Postdoctoral Supervisor; Ganjiang Scholar; Vice President of the International Association of Translational Medicine and Chairman of the Ophthalmology Committee; Chairman of the Ophthalmic Imaging and Intelligent Medical Branch of the Chinese Medical Education Association; Chairman of the Corneal Disease and Eye Surface Disease Special Committee of the China Association for the Promotion of Population and Culture; National Top 100 Academic Influential Ophthalmologists and Scholars; Member of the Ophthalmology Branch of the Chinese Medical Doctor Association; Member of the American Society of Ophthalmology, the American Association for Visual and Ophthalmic Research, and the European Association for Visual and Ophthalmic Research; Guest editor of 16 SCI journals; published more than 300 papers indexed in SCI; edited/co-edited 38 monographs in ophthalmology.

Guidelines on clinical research evaluation of artificial intelligence in ophthalmology (2023)

Wei-Hua Yang[1], Yi Shao[2], Yan-Wu Xu[3,4], Expert Workgroup of Guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology (2023), Ophthalmic Imaging and Intelligent Medicine Branch of Chinese Medicine Education Association, Intelligent Medicine Committee of Chinese Medicine Education Association

[1]Shenzhen Eye Institute, Shenzhen Eye Hospital, Shenzhen 518040, Guangdong Province, China
[2]The First Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi Province, China
[3]School of Future Technology, South China University of Technology, Guangzhou 510641, Guangdong Province, China
[4]Pazhou Lab, Guangzhou 510320, Guangdong Province, China

Correspondence to: Yi Shao. Department of Ophthalmology, the First Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi Province, China. freebee99@163.com; Yan-Wu Xu. School of Future Technology, South China University of Technology, Guangzhou 510641, Guangdong Province, China. ywxu@ieee.org
Received: 2023-07-06 Accepted: 2023-07-20

Abstract

● With the upsurge of artificial intelligence (AI) technology in the medical field, its application in ophthalmology has become a cutting-edge research field. Notably, machine learning techniques have shown remarkable achievements in diagnosing, intervening, and predicting ophthalmic diseases. To meet the requirements of clinical research and fit the actual progress of clinical diagnosis and treatment of ophthalmic AI, the Ophthalmic Imaging and Intelligent Medicine Branch and the Intelligent Medicine Committee of Chinese Medicine Education Association organized experts to integrate recent evaluation reports of clinical AI research at home and abroad and formed a guideline on clinical research evaluation of AI in ophthalmology after several rounds of discussion and modification. The main content includes the background and method of developing this guideline, an introduction to international guidelines on the clinical research evaluation of AI, and the evaluation methods of clinical ophthalmic AI models. This guideline introduces general evaluation methods of clinical ophthalmic AI research, evaluation methods of clinical ophthalmic AI models, and commonly-used indices and formulae for clinical ophthalmic AI model evaluation in detail, and amply elaborates the evaluation methods of clinical ophthalmic AI trials. This guideline aims to provide guidance and norms for clinical researchers of ophthalmic AI, promote the development of regularization and standardization, and further improve the overall level of clinical ophthalmic AI research evaluations.

● **KEYWORDS:** artificial intelligence; ophthalmology; evaluation; clinical research; machine learning; deep learning

DOI:10.18240/ijo.2023.09.02

Citation: Yang WH, Shao Y, Xu YW, Expert Workgroup of Guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology (2023), Ophthalmic Imaging and Intelligent Medicine Branch of Chinese Medicine Education Association, Intelligent Medicine Committee of Chinese Medicine Education Association. Guidelines on clinical research evaluation of artificial intelligence in ophthalmology (2023). *Int J Ophthalmol* 2023;16(9):1361-1372

This article is based on a study first reported in the *Guoji Yanke Zazhi (Int Eye Sci)* 2023;23(7):1064-1071.

INTRODUCTION

Artificial intelligence (AI) is a branch of computer science aimed at developing intelligent machines that can learn, reason, judge, and make decisions like humans. AI includes many sub-fields and technologies, such as natural language processing, computer vision[1], machine learning[2], and deep learning[3], which are widely applied in healthcare, finance, transportation, and manufacturing[4]. With the continuous improvement of computer technology and data processing capabilities, the development and application of AI are becoming increasingly widespread and in-depth. Ophthalmic diseases, including cataract, glaucoma, diabetic retinopathy, age-related macular degeneration, and pathological myopia, are important diseases affecting the health of the global population. Clinical research is of great significance for understanding the pathophysiological mechanisms of diseases, developing prevention and treatment strategies, improving patient quality of life, and reducing medical costs. The application of AI in the field of ophthalmic clinical research mainly includes the prediction and diagnosis of ophthalmic diseases[5-6], treatment and intervention, and prevention and management[7-8]. Among them, early screening systems for ophthalmic diseases based on ophthalmic imaging and AI technology, such as fundus screening software for diabetic retinopathy[9], fundus screening software for multiple fundus diseases (applicable to chronic glaucomatous optic neuropathy and diabetic retinopathy)[10-11], and fundus screening software for chronic glaucomatous optic neuropathy have passed the registration approval of the Class III medical device registration certificate of the National Medical Products Administration of China.

Clinical research on ophthalmic imaging and AI technology is in full swing. With the continuous increase of clinical ophthalmic AI research, it is particularly necessary to provide evaluation guidelines to ensure their quality and reliability. A high-quality guideline can not only ensure the accuracy and effectiveness of research data but also improve the repeatability and comparability of the research. In addition, the validation and authentication of AI algorithms and models are also crucial to ensure their effectiveness and reliability in real-world clinical environments[12-13]. Therefore, the Ophthalmic Imaging and Intelligent Medicine Branch and the Intelligent Medicine Professional Committee of the China Medical Education Association established the expert workgroup of Guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology (2023) for the guideline preparation. This guideline mainly focuses on clinical ophthalmic AI research based on ophthalmic imaging and AI technology[14-15], aiming to comprehensively summarize the evaluation methods, ensure the quality and reliability, and promote transparency and standardization of clinical ophthalmic AI research. Meanwhile, this guideline contributes to the stable development of clinical ophthalmic AI research and related applications with protected privacy and data security of research participants.

Development Methodology for Guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology (2023) Based on the current evaluation issues of clinical ophthalmic AI research, the Ophthalmic Imaging and Intelligent Medicine Branch and the Intelligent Medicine Professional Committee of the China Medical Education Association organized ophthalmic AI experts, ophthalmic clinical research experts, ophthalmic medical ethics experts, and ophthalmic AI product development scientists to establish a clinical ophthalmic AI research evaluation guideline expert group in July 2022. On July 25, 2022, interviews with relevant experts in clinical ophthalmic AI research were initiated to gather and organize evaluation issues pertaining to clinical ophthalmic AI research in related fields, as well as the challenges encountered in clinical research on AI technology. Due to the lack of a unified and compliant guideline for the evaluation of clinical ophthalmic AI research, the expert group of this guideline has carefully studied domestic and foreign clinical ophthalmic AI research literature and research literature, combined with practical experience in clinical ophthalmic AI research, held offline and online meetings to fully discuss and demonstrate the evaluation issues of collected clinical ophthalmic AI research. The first draft of the guideline was written by members of the writing expert group. After the first draft was formed, experts independently read it and submitted their revision suggestions to the core members of the guideline writing group through email and WeChat. The revision suggestions were discussed and summarized through WeChat, email, and online meetings. During the revision period, the guidelines fully accepted the suggestions and guidance of participating experts and ultimately reached the final version, aiming to guide the evaluation of clinical ophthalmic AI research. The development process of this guideline took approximately one year.

Introduction to International Guidelines for Artificial Intelligence Clinical Research Evaluation At present, there are no evaluation guidelines for clinical research on ophthalmic AI internationally. However, there are some general guidelines for regulating AI clinical research or trials that can be referenced. For example, "Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence (SPIRIT-AI)"[16] and "Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI)"[17] were released in 2020, while "Standards for Reporting of Diagnostic accuracy studies-Artificial Intelligence (STARD-AI)"[18] and "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-Artificial Intelligence (TRIPOD-AI)"[19] were released in 2021. Among them, SPIRIT-AI is a normative guideline for clinical trials of intervention measures involving AI. To enhance transparency in the design and methods of AI clinical trials, it is crucial to utilize specific information that should be reported in conjunction with SPIRIT2013 and other SPIRIT extended guidelines. This aims to promote transparency in the design of AI clinical experiments and methods, ultimately optimizing understanding, interpretation, and peer review[16]. Similarly, CONSORT-AI is used to standardize clinical trial reports on interventions involving AI. It is recommended to provide a clear description of AI interventions, including guidance and skills required for use, environment for AI intervention integration, input and output processing of AI interventions, interaction between AI and humans, and error case analysis, to promote transparency and completeness of AI intervention clinical trial reports[17]. STARD-AI is a guideline for standardizing diagnostic testing accuracy research reports centered on AI, proposing the need to report on data preprocessing methods, AI testing development methods (such as dataset partitioning, model calibration, training stop criteria, use of external validation sets), fairness metrics, non-standard performance indicators, interpretability, and interaction between humans and AI testing. The aim is to improve the transparency and fairness of research on the accuracy of AI diagnostic testing[18]. TRIPOD-AI is a guideline for research reports on multivariate AI prediction models, aimed at helping researchers report research content transparently and helping reviewers understand research methods and results, thereby reducing research waste[19].

Evaluation Methods for Clinical Ophthalmic Artificial Intelligence Research The clinical research of ophthalmic AI includes four key links: data collection and management of ophthalmic examinations, model development, clinical trials, and clinical applications. This guideline will introduce evaluation methods for these key links. It is worth noting that clinical ophthalmic AI research models can be divided into three types according to the clinical application tasks: intervention models, diagnostic models, and predictive models[20-21]. Specifically, ophthalmic AI intervention models can be used as independent interventions or in combination with conventional interventions for the treatment, prevention, or management of specific diseases or symptoms. The ophthalmic AI diagnostic model is used to determine the presence, classification, and grading of a certain disease or lesion. The ophthalmic AI prediction model is used to predict the risk of future diseases or the effectiveness of treatment based on the characteristics of research participants. Therefore, the introduction of model evaluation methods will be carried out separately based on these three clinical ophthalmic AI research models. In addition, as clinical trials are a necessary condition for the domestic

and foreign marketing of medical devices[22-23], this guideline will separately introduce the evaluation methods of clinical ophthalmic AI trials in section "Evaluation Methods for Clinical Ophthalmic Artificial Intelligence Trials".

General Evaluation Methods of Clinical Ophthalmic Artificial Intelligence Research

Evaluation of data collection and management The evaluation of data collection and management in clinical ophthalmic AI research aims to ensure the quantity, quality, completeness, safety, and reliability of research data[24]. Specific evaluation methods are recommended to cover the following aspects: 1) Data quantity evaluation: Evaluate the quantity of collected data to ensure that it meets the requirements of model development and performance validation in clinical research. 2) Data quality evaluation: Evaluate the quality of data[25-26], including the completeness, accuracy, logicality, consistency, and usability of the data, to ensure that the quality of the data meets the requirements[27]. 3) Data cleansing evaluation: Evaluate whether the data cleansing process is desensitized, logical, effective, *etc*. 4) Data label evaluation: Evaluate the construction process and label quality of data labels, that is, reference standards[28], to ensure the reliability of data labels. For labels generated by relying on manual labeling, it is necessary to evaluate the standardization of the labeling process, labeling personnel and equipment, labeling process, and labeling quality[29]. 5) Data storage evaluation: Evaluate the quality of data storage to ensure that data storage is secure and meets requirements. Common methods include checking the storage location, storage medium, and storage method of data. 6) Data management evaluation: Evaluate the quality of data management to ensure that data management is safe and meets requirements. Common methods include checking the data management process and the abilities of data management personnel[30]. 7) Data usage evaluation: Evaluate the quality of data usage to ensure the safety and compliance of data usage and sharing processes. Commonly used methods include checking the purpose, scope, ethics[31], and legality of data usage, as well as policies, methods, and purposes for data sharing.

Evaluation of artificial intelligence model development in ophthalmology The evaluation of the model development process in clinical ophthalmic AI research aims to ensure that the models developed in the study have high quality, reliability, and stability. Specific evaluation methods are suggested to cover the following aspects: 1) Evaluation of development data set: Evaluate whether the quality, quantity, and balance of the data set used to develop the AI model are sufficient, how representative the data set is, and whether the division of training set, verification set, and test set is reasonable; is there a sufficient clinical basis for the definition method of evaluation labels. 2) Feature selection and extraction evaluation: If it is necessary to select features manually, evaluate whether the selected features can have an important impact on the model performance and whether the feature extraction method is appropriate. 3) Ophthalmic AI model performance evaluation: Use common indicators to evaluate the performance of the model, ensuring that the model can accurately predict target variables, as detailed in Section "Evaluation Methods of Clinical Ophthalmic Artificial Intelligence Models". 4) Cross-validation: Use cross-validation methods (such as k-fold cross-validation) to evaluate the model's generalization ability, ensuring that the model can make accurate predictions on new data. 5) Model interpretive evaluation: Evaluate the interpretability of the model to ensure that the predicted results of the model can be clinically explained and understood. 6) Model stability evaluation: Evaluate the stability of the model against data noise and randomness, ensuring that the model can produce consistent results when facing different datasets. 7) Model adaptability evaluation: Evaluate the model's adaptability to different groups and environments, ensuring that the model can produce accurate results in practical applications.

Evaluation of clinical application of artificial intelligence models in ophthalmology The evaluation of the clinical application of ophthalmic AI models is to ensure the safety, effectiveness, and repeatability of clinical applications. Specific evaluation methods are recommended to cover the following aspects: 1) Security evaluation: Evaluating whether there are issues with data privacy and security in the clinical application process to protect the privacy and personal information of research participants. 2) Internal effectiveness evaluation: Evaluate the accuracy, credibility, and applicability of research results. The level of internal effectiveness depends on factors such as the rationality of the research design, the selection and allocation of research and control groups, blind design, control and management during the research process, and the reliability of data analysis. 3) External effectiveness evaluation: Evaluate the promotion ability and universality of research results. The level of external effectiveness depends on factors such as the representativeness of the research sample, the authenticity of the experimental environment, the universality of the research method, and the applicability of the research results. 4) Repeatability evaluation: Evaluate whether the research results can be repeatedly verified, that is, evaluate whether the AI model's performance is stable on different datasets, whether the performance fluctuation range is acceptable, whether the performance is consistent on different devices, and whether the prediction results are consistent under multiple inputs of the same data. The level of

repeatability depends on factors such as the representativeness of data during the model development stage, transparency of the research process, clarity of research methods, openness of data, and repeatability of analysis. 5) Application effectiveness evaluation: Evaluate the effectiveness of clinical applications, including the degree of guidance and improvement for patient diagnosis and treatment. 6) Analysis and evaluation of health economics: Evaluate the value of health economics in clinical application, including cost-effectiveness analysis, cost-utility analysis, cost-benefit analysis, *etc*. Cost includes human, material, and economic costs. Output indicators include clinical effects, quality-adjusted life years, and saved medical expenses generated in the actual application process.

Evaluation Methods of Clinical Ophthalmic Artificial Intelligence Models

Evaluation of ophthalmic artificial intelligence intervention models The ophthalmic AI intervention model can be used as an independent intervention measure or combined with conventional intervention measures for the treatment, prevention, or management of specific diseases or symptoms. To prove the effectiveness of the ophthalmic AI intervention models for the target disease treatment, the clinical research of the ophthalmic AI intervention models is usually evaluated through two aspects: the intervention process and the intervention effect. The intervention process can be directly evaluated by comparing the ophthalmic AI intervention models with the conventional intervention measures. According to the types of indicators derived from different aspects, such as the duration, safety, and effectiveness of the intervention process and health economics, appropriate statistical methods can be selected for comparison[32-34]. The evaluation of intervention effect is usually measured by clinical outcome indicators, such as mortality, disease recurrence rate, and survival time, which can be evaluated by the results of symptom relief, disease progression, or survival rate after the intervention. See Section "Commonly-used indices and formulae for ophthalmic artificial intelligence intervention model evaluation" for details.

Evaluation of ophthalmic artificial intelligence diagnostic models A diagnostic model is used to determine whether a certain disease or lesion exists. The main objective of evaluating diagnostic models is to examine their diagnostic accuracy. The evaluation indicators that can be used include sensitivity, specificity, accuracy, and Kappa consistency coefficient, as detailed in Section "Commonly-used indices and formulae for ophthalmic artificial intelligence diagnostic model evaluation".

Evaluation of ophthalmic artificial intelligence prediction models Prediction models are used to predict the risk of diseases, changes in physiological structures, or treatment outcomes based on the characteristics of research participants. The evaluation prediction model can include classification results for evaluating the future occurrence of diseases and regression results for evaluating future physiological structure measurement parameters. If there is a clear prediction label (reference standard), the available evaluation indicators include root-mean-square error, mean absolute error, sensitivity, and specificity. In the absence of a clear prediction label (reference standard), the evaluation indicators include positive compliance rate, negative compliance rate, and total compliance rate compared with other state-of-the-art methods, as detailed in Section "Commonly-used indices and formulae for ophthalmic artificial intelligence prediction model evaluation".

Commonly-used Indices and Formulae for Clinical Ophthalmic Artificial Intelligence Models Evaluation This guideline provides commonly used evaluation indices and formulae for ophthalmic AI models[28,35]. Clinical studies of different models should select different indicators for evaluation based on tasks.

Commonly-used indices and formulae for ophthalmic artificial intelligence intervention model evaluation 1) Intervention model mortality rate: the rate of death among research participants after intervention:

$$\text{Mortality rate}=\frac{\text{Number of deaths after intervention}}{\text{Number of interventions}}\times 100\% \quad (1)$$

2) Intervention model disease recurrence rate: The rate of disease recurrence among research participants after intervention:

$$\text{Disease recurrence rate}=\frac{\text{Number of disease relapses after intervention}}{\text{Number of interventions}}\times 100\% \quad (2)$$

3) The intervention model survival period: the number of days between the start of intervention and death or loss of follow-up for research participants.

Commonly-used indices and formulae for ophthalmic artificial intelligence diagnostic model evaluation 1) Confusion matrix, a special visual matrix with two dimensions, can be used to compare classification results and actual measured values in supervised learning evaluation. Each row of the confusion matrix represents the instances in an actual class while each column represents the instances in a predicted class (Table 1).

2) Sensitivity (Sen), also known as Recall (R), is the proportion of true positive samples to all positive samples:

$$Sen=R=\frac{TP}{TP+FN} \quad (3)$$

3) Specificity (Spe), the proportion of true negative cases to all negative cases:

$$Spe=\frac{TN}{TN+FP} \quad (4)$$

Table 1 Confusion matrix

Predictive values	Actual values		Total
	Positive	Negative	
Positive	True positive (TP)	False positive (FP)	R_1
Negative	False negative (FN)	True negative (TN)	R_2
Total	C_1	C_2	N

TP: The number of samples that are actually positive and correctly predicted to be positive; FP: The number of samples that are actually negative and incorrectly predicted as positive; FN: The number of samples that are actually positive but are incorrectly predicted to be negative; TN: The number of samples that are actually negative and correctly predicted to be negative; R_1: The number of true positive and false positive cases. R_2: The sum of false negative and true negative cases; C_1: The sum of true positive and false negative cases; C_2: The sum of false positive and true negative cases; N: The total number of samples.

4) Likelihood ratio, a composite indicator that reflects both sensitivity and specificity, is the ratio of the probability of obtaining a certain screening test result among the diseased individuals to the probability of obtaining the same result among the non-diseased individuals.

Positive likelihood ratio is the ratio of the true positive rate to the false positive rate of the screening results. As the ratio increases, the probability of a positive screening result being a true positive also increases:

$$Positive\ likelihood\ ratio=\frac{Sen}{1\text{-}Spe} \tag{5}$$

Negative likelihood ratio is the ratio of the false negative rate to the true negative rate of the screening test result, with a smaller ratio indicating a higher likelihood of true negativity when the research result is negative:

$$Negative\ likelihood\ ratio=\frac{1\text{-}Sen}{Spe} \tag{6}$$

5) Accuracy is the proportion of samples with correct algorithm diagnosis to all samples:

$$Accuracy=\frac{TP+TN}{N} \tag{7}$$

6) Precision, also known as positive prediction value, is the proportion of true positive samples to the positive samples determined by the algorithm:

$$Precision=Positive\ prediction\ value=\frac{TP}{TP+FP} \tag{8}$$

7) Negative prediction value is the proportion of true negative samples to the negative samples determined by the algorithm:

$$Negative\ prediction\ value=\frac{TN}{TN+FN} \tag{9}$$

8) Miss rate, also known as missed report rate, missed diagnosis rate, missed alarm rate, false negative rate, refers to the proportion of undetected positive samples in the test to all positive samples:

$$Miss\ rate=1-\frac{TP}{TP+FN}=1\text{-}Sen \tag{10}$$

9) False alarm, also known as false alarm rate, misdiagnosis rate, false alarm rate, false positive rate, refers to the proportion of all negative samples that are incorrectly predicted as positive:

$$False\ alarm=1-\frac{TN}{TN+FP}=1\text{-}Sen \tag{11}$$

10) F_1 score is the harmonic mean of recall and precision:

$$F_1=\frac{2\times P\times R}{P+R}=\frac{2\times TP}{2\times TP+FP+FN} \tag{12}$$

In the formula, P represents the precision and R represents the recall.

11) The Youden index, also known as the correct index, assuming that false negative (rate of missed diagnosis) and false positive (rate of misdiagnosis) are equally harmful, is the sum of sensitivity and specificity minus 1. The larger the index, the better the screening effect.

$$Youden\ index=Sen+Spe\text{-}1 \tag{13}$$

12) Kappa value is an indicator used to evaluate the consistency between screening systems and reference labeled diagnostic results:

$$Kappa=\frac{p_o-p_e}{1-p_e} \tag{14}$$

In the formula, $p_0=(TP+TN)/N$, $p_e=(R_1C_1+R_2C_2)/N^2$. Therefore:

$$Kappa=\frac{N(TP+TN)-(R_1C_1+R_2C_2)}{N^2-(R_1C_1+R_2C_2)} \tag{15}$$

13) Area under curve (AUC) of receiver operating characteristic (ROC): ROC is the sensitivity and specificity of the screening system on the test set estimated under preset thresholds to generate a set of specificity-sensitivity pairs. These operating points are connected in turn to form a curve. As shown in Figure 1, AUC is the area enclosed by the curve and the X-axis. It can be used to measure the performance of classification models, with values generally ranging from 0.5 to 1. A larger AUC value suggests better classification performance.

14) Precision recall (PR) curve: Similar to ROC, the PR curve estimates the accuracy and recall of a screening system with preset thresholds on a testing set, resulting in a set of recall-precision pairs that are sequentially connected to form a curve.

Commonly-used indices and formulae for ophthalmic artificial intelligence prediction model evaluation If the prediction model outputs classification category results, the evaluation indices and formulae provided in Section

"Commonly-used indices and formulae for ophthalmic artificial intelligence diagnostic model evaluation" can be used for evaluation. If the output is a continuous numerical result, the following evaluation indices and formulae can be used:

1) Root mean square error can measure the deviation between the predicted value and the true value, and can reflect the accuracy of measurement. The closer the root-mean-square deviation is to 0, the better the model can predict the target value:

$$Root\ mean\ square\ error=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{16}$$

In the formula, N is the total number of samples, y_i is the true value of the i-th sample, and $\hat{y}_i$ is the predicted value of the i-th sample.

2) Mean absolute error is the average of the absolute value of the deviation between each measured value and the reference standard. The average absolute error can avoid the problem of mutual cancellation of errors and accurately reflect the size of actual prediction errors:

$$Mean\ absolute\ error=\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i-y_i| \tag{17}$$

3) Mean absolute percentage error (MAPE) is a relative measure. Compared with mean absolute error, MAPE calculates the percentage of the deviation between the predicted value and its reference standard relative to the reference standard:

$$MAPE=\frac{100\%}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i-y_i|}{|y_i|} \tag{18}$$

The range of MAPE is [0, ∞) that a value of 0 represents a perfect model, and a value greater than 100% represents a poor-quality model. Note that the formula is not available when the value of the reference standard is 0.

4) Symmetric mean absolute percentage error (SMAPE): Compared with MAPE, the reference standard absolute value in the denominator of the calculation formula is replaced with the average of the reference standard absolute value and the predicted value absolute value:

$$SMAPE=\frac{100\%}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i-y_i|}{(|\hat{y}_i|+|y_i|)/2} \tag{19}$$

The value range of SMAPE is [0, 200%], and the formula is not available when the reference standard and predicted value are both 0.

5) R square, also known as the coefficient of determination, is the statistical coefficient of the degree of fitness between the regression prediction value and the calibration value. The square value of R is between 0 and 1. An R square value closer to 0 suggests that the prediction results of the model are close to randomness. As R square value increases, the model fits better to the target.

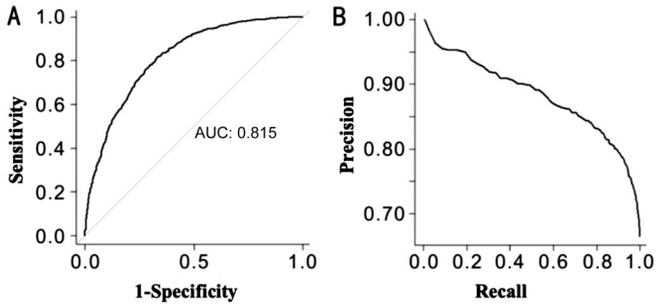

Figure 1 Receiver operating characteristic (ROC) curve, area under curve (AUC) index and precision recall (PR) curve.

Table 2 2×2 table with unknown reference standards[36]

Method to be evaluated	Comparison method		Total
	Positive	Negative	
Positive	a	b	a+b
Negative	c	d	c+d
Total	a+c	b+d	n

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{20}$$

In the formula, N represents the total number of samples, y_i is the actual value of the i-th sample, $\hat{y}_i$ is the predicted value of the i-th sample, and $\bar{y}$ is the average of the actual values of all samples.

6) When the reference standard output by the prediction model is unknown, the conformity rate can be calculated between the results of the evaluation method and those obtained by other state-of-the-art methods, such as positive conformity rate, negative conformity rate, and total conformity rate, as shown in Table 2[36] and the formula.

$$Positive\ coincidence\ rate=\frac{a}{a+c}\times100\% \tag{21}$$

$$Negative\ coincidence\ rate=\frac{d}{b+d}\times100\% \tag{22}$$

$$Total\ coincidence\ rate=\frac{a+d}{n}\times100\% \tag{23}$$

7) In addition to evaluating the accuracy of predictive models, it is also important to access their calibration. Calibration is considered one of the most important attributes of predictive models, which reflects the degree to which the predictive model correctly estimates absolute risk. Poorly calibrated predictive models may underestimate or overestimate the target outcome[37]. The evaluation of calibration usually uses the Hosmer-Limeshow goodness of fit test and calibration curve.

Hosmer-Lemeshow goodness of fit test[37] is used to determine the difference between predicted values and true values. If the P-value is less than or equal to 0.05, it indicates that the difference between the predicted value and the true value is statistically significant, indicating poor fitting of the model; If the P-value is greater than 0.05, it indicates passing the Hosmer-Lemeshow goodness of fit test[38].

The calibration curve[37] is used to assist in observing whether the predicted probability of the model is close to the true probability. It is a scatter plot of the actual occurrence rate to the predicted occurrence rate. Essentially, the calibration curve is a visualization of the goodness of fit test results.

Other commonly-used evaluation indices and formulae in clinical ophthalmic artificial intelligence research 1) The effective utilization rate of data refers to the proportion of data that is ultimately effectively used in the process of data collection and processing to the total data volume:

$$Effective\ data\ usage\ rate=\frac{Effective\ data\ volume}{Total\ data\ volume}\times 100\% \quad (24)$$

2) The sample size estimation formula can derive the required quantity of data for each category in the test set based on the expected effect of the ophthalmic AI model:

$$N=\frac{[Z_{1-\alpha/2}]^2\,P\,(1-P)}{\Delta^2} \quad (25)$$

In the formula, Z is the Z statistical measure of confidence level, Δ is the allowable error, P is the expected evaluation index (sensitivity, specificity, *etc*.), and N is the required sample size. The confidence level of the parameter estimation bilateral confidence interval is usually set to 95% (*i.e.,* Class I error α=0.05, bilateral), then $Z_{1-\alpha/2}$=1.96 and the expected evaluation index estimation accuracy (confidence interval half width) Δ is usually set to 5%.

3) When evaluating multi-category classification ophthalmic AI research tasks, if multiple categories are independent of each other, the evaluation of multiple categories can be transformed into the evaluation of multiple binary classification problems. The negative samples of each category are defined as all samples in the total sample except for the positive samples of that category. Computable evaluation indicators include Micro/Macro F_1 value, Micro/Macro AUC, and Kappa value.

Macro F_1 and Macro AUC are calculated separately for each predicted F_1 value and AUC value, and then averaged for each category:

$$Macro\ F_1=\frac{\sum_{i=1}^{C}F_{1i}}{C}=\frac{1}{C}\sum_{i=1}^{C}\frac{2\times P_i\times R_i}{P_i+R_i}=\frac{1}{C}\sum_{i=1}^{C}\frac{2\times TP_i}{2\times TP_i+FP_i+FN_i} \quad (26)$$

$$Macro\ AUC=\frac{\sum_{i=1}^{C}AUC_i}{C} \quad (27)$$

In the formula, C is the total number of categories for the classification task.

Micro F_1 and Micro AUC are calculated by first calculating the number of true positive, false positive, true negative, and false negative samples in the population, and then based on the definition of F_1 and AUC, namely:

$$R_m=\frac{\sum_{i=1}^{C}TP_i}{\sum_{i=1}^{C}TP_i+\sum_{i=1}^{C}FN_i} \quad (28)$$

$$P_m=\frac{\sum_{i=1}^{C}TP_i}{\sum_{i=1}^{C}TP_i+\sum_{i=1}^{C}FP_i} \quad (29)$$

$$Micro\ F_1=\frac{2\times P_m\times R_m}{P_m+R_m} \quad (30)$$

Micro AUC relies on the global confusion matrix. When drawing the global receiver operating characteristic, the horizontal and vertical coordinates represent the global 1-specificity and sensitivity respectively, that is:

$$\left(1-\frac{\sum_{i=1}^{C}TN_i}{\sum_{i=1}^{C}TN_i+\sum_{i=1}^{C}FP_i},\ \frac{\sum_{i=1}^{C}TP_i}{\sum_{i=1}^{C}TP_i+\sum_{i=1}^{C}FN_i}\right) \quad (31)$$

Micro/Macro F_1 and Micro/Macro AUC are values ranging from 0 to 1, and the closer the value is to 1, the better the performance of the multi-classification model.

When evaluating multi-classification tasks, the Kappa consistency coefficient:

$$Kappa=\frac{p_o-p_e}{1-p_e} \quad (32)$$

In the formula,

$$p_o=\sum_{i=1}^{C}TP_i/N,\ p_e=\sum_{i=1}^{C}(TP_i+FN_i)\times(TP_i+FP_i)/N^2$$

4) In clinical ophthalmic AI research, the evaluation indicators for structural (*e.g.*, physiological structure, and lesion) region segmentation results include Dice coefficient and Jaccard coefficient:

Dice coefficient is the ratio of the intersection between the segmented contour of the structural area and the reference standard contour to the average value of the segmented contour and the reference standard contour (Figure 2):

$$DICE\ (X,\ Y)=\frac{|X\cap Y|}{1/2\,(\,|X|+|Y|\,)}=\frac{2\times TP}{(TP+FN)+(TP+FP)} \quad (33)$$

In the formula, and $|X|$ and $|Y|$ respectively represent the number of elements of X and Y, and $|X\cap Y|$ is the intersection between X and Y.

Jaccard coefficient, also known as Intersection over Union (IoU), refers to the proportion of the intersection between the segmented contour of the structural area and the reference standard contour to the union of the segmented contour and the target contour (Figure 3):

$$Jaccard\ (X,Y)=\mathrm{IoU}\ (X,Y)=\frac{|X\cap Y|}{|X\cup Y|}=\frac{TP}{TP+FN+FP} \quad (34)$$

Evaluation Methods for Clinical Ophthalmic Artificial Intelligence Trials Clinical trials are an important component of clinical research, used to verify the safety and effectiveness of drugs or medical devices. The evaluation methods for clinical ophthalmic AI trials are suggested to cover the following aspects: Design of experiments, research participants, ethical issues, sample size, control and blind design, trial results, data analysis, and adverse events. 1)

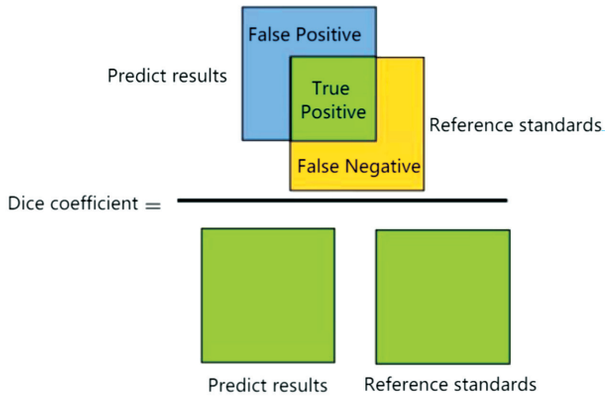

Figure 2 Dice coefficient calculation.

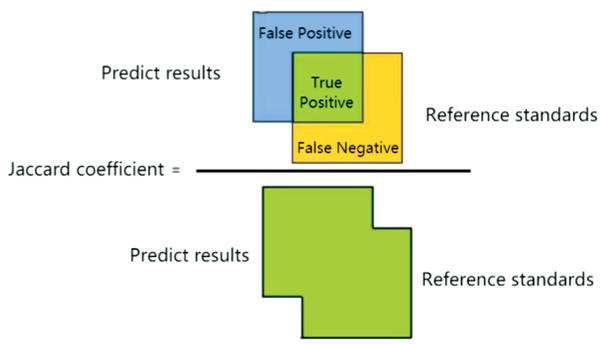

Figure 3 Jaccard coefficient calculation.

Design of experiments: The design of the clinical trial should be suitable for answering the questions of the clinical trial, including the type of trial, prospective or retrospective, single-center or multi-center, and superior design, non-inferior design, or single group target value design. For example, clinical trials targeting intervention models need to ensure sufficient follow-up of participants to ensure that the intervention is safe and effective for a certain period of time. For the AI medical device clinical trial of the medical imaging diagnostic models, in order to avoid the influence of subjective factors, uncertainties, and other factors of doctors, the Multi-Reader Multi-Case design of experiments can be used to ensure a comprehensive evaluation of model performance and reduce errors caused by individual differences of researchers. 2) Research participant group: Clinical trials need to have a clear research participant group that is representative of the studied population. Clinical trials need to select research participants reasonably based on their characteristics and trial objectives, ensuring the representativeness and diversity of the samples. 3) Ethical issues: Clinical trials should comply with ethical principles. Research participants should sign an informed consent form before participating in clinical trials, and clinical trials should obtain approval from the ethics committee[31]. 4) Sample size: Clinical trials should have an appropriate sample size that meets the requirements of statistical analysis to discover meaningful differences between groups. 5) Control and blind design: Participants in intervention clinical trials should be randomly divided into a treatment group and a control group, and a double-blind method should be used to minimize selection bias and ensure the comparability of each group at baseline. The diagnostic or predictive clinical design of experiments should be suitable for answering the questions of clinical trials. Diagnostic clinical trials should use current clinical standard methods as control methods. 6) Test results: The measured results should be clearly defined and related to clinical trial issues, and standardized methods should be used for measurement. 7) Data analysis: The statistical analysis of the data should be appropriate, and the experimental results should be presented in a clear and transparent manner. 8) Adverse events: Clinical trials should report any adverse events that occurred during the trial period and evaluate the safety and tolerability of the clinical trial.

SUMMARY

Ophthalmology is the most active clinical specialty in medical AI. With the continuous increase of clinical research on ophthalmic AI based on ophthalmic imaging and AI technology, we have developed evaluation guidelines for clinical ophthalmic AI research to ensure the quality and reliability of clinical ophthalmic AI research. This guideline summarizes the background and methods of developing the guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology, introduces international guidelines for AI clinical research evaluation, and discusses the evaluation methods of clinical ophthalmic AI research. This guideline introduces general evaluation methods of clinical ophthalmic AI research, evaluation methods of clinical ophthalmic AI models, and commonly-used indices and formulae for clinical ophthalmic AI model evaluation in detail, and amply elaborates the evaluation methods of clinical ophthalmic AI trials. The development of this guideline can help improve the design, implementation, and quality of clinical research protocols, thereby improving the integrity and transparency of research and reducing potential biases. The purpose of this guideline is to provide recommendations for the evaluation of clinical ophthalmic AI research and raise regulatory awareness of ophthalmic clinical research evaluation among relevant researchers. In clinical ophthalmic AI research, researchers can select evaluation indices and formulae according to the research process and model type.

This guideline is the first in the evaluation of clinical ophthalmic AI research. With the gradual introduction of laws, regulations, and policies on the application of AI technology in the medical field, the content of this guideline will be further discussed and updated. Valuable suggestions and opinions on the shortcomings are welcome to continuously update and improve this guideline[39].

ACKNOWLEDGEMENTS

Member of the Formation Guideline Expert Group

Writing expert

Wei-Hua Yang	Shenzhen Eye Hospital; Shenzhen Eye Institute
Yan-Wu Xu	School of Future Technology, South China University of Technology, China; Pazhou Lab. Guangzhou
Hui-Hui Fang	Pazhou Lab. Guangzhou
Yi Shao	The First Affiliated Hospital of Nanchang University
Shao-Chong Zhang	Shenzhen Eye Hospital; Shenzhen Eye Institute
Yong-Yue Wei	Center for Public Health and Epidemic Preparedness & Response, Peking University
Zu-Guo Liu	Eye Institute of Xiamen University
Ji-Yin Zhou	The Second Affiliated Hospital of Army Medical University
Yong-Jin Zhou	School of Biomedical Engineering, Health Science Center, Shenzhen University

Experts involved in drafting

Sunee Chansangpetch	Department of Ophthalmology, King Chulalongkorn Memorial Hospital
Hao Chen	Eye Hospital, Wenzhou Medical University
Jie Chen	Pengcheng Laboratory
Yu-Zhong Chen	Beijing Airdoc Technology Co., Ltd.
Hong-Guang Cui	First Affiliated Hospital, School of Medicine, Zhejiang University
Qi Dai	Eye Hospital, Wenzhou Medical University
Wei-Wei Dai	Aier Institute of Digital Ophthalmology & Visual Science
Ai-Jun Deng	The Affiliated Hospital of Weifang Medical University
Lin Ding	People's Hospital of Xinjiang Uygur Autonomous Region
Li-Xin Duan	Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
Hua-Zhu Fu	Institute of High Performance Computing, A*STAR, Singapore
Zong-Yuan Ge	Beijing Airdoc Technology Co., Ltd.
Wei Han	Second Affiliated Hospital, School of Medicine, Zhejiang University
Hou-Bin Huang	Senior Department of Ophthalmology, PLA General Hospital; Hainan Hospital of PLA General Hospital
Qin Jiang	Eye Hospital, Nanjing Medical University
Bai-Ying Lei	School of Biomedical Engineering, Health Science Center, ShenZhen University
Gen-Jie Ke	AnHui Provincial Hospital
Hu Liu	The First Affiliated Hospital with Nanjing Medical University
Shi-Ying Li	Xiang'an Hospital of Xiamen University and Medical Center of Xiamen University
Wen Li	Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
Xiao-Meng Li	Hong Kong University of Science and Technology
Xiao-Qing Liu	Visionary Intelligence Ltd.
Yan Lou	School of Intelligent Medicine, China Medical University
Pei-Rong Lu	The First Affiliated Hospital of Soochow University
Zong-Ming Song	Henan Provincial People's Hospital; Henan Eye Hospital
Bin Sun	Shanxi Eye Hospital
Ming-Kui Tan	School of Software Engineering, South China University of Technology
Li-Ming Tao	The Second Affiliated Hospital of Anhui Medical University
Cheng Wan	Nanjing University of Aeronautics and Astronautics
Rui-Li Wei	Shanghai Changzheng Hospital, Naval Medical University
Jian Wu	Second Affiliated Hospital School of Medicine; School of Public Health, Zhejiang University
Xuan Xiao	Renmin Hospital of Wuhan University
Jie Xu	Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University
Wen Xu	Second Affiliated Hospital, School of Medicine, Zhejiang University
Fan Xu	The People's Hospital of Guangxi Zhuang Autonomous Region
Jing-Jing Xu	Visionary Intelligence Ltd.
Yong-Sheng Yang	Eye Hospital, China Academy of Chinese Medical Sciences
Jin Yao	Eye Hospital, Nanjing Medical University
Juan Ye	Second Affiliated Hospital, School of Medicine, Zhejiang University
Li-Jing Yue	Guangdong Second Traditional Chinese Medicine Hospital
Dong-Dong Zhang	Beijing Zhenhealth Technology Co., Ltd.
Guang-Hua Zhang	Big data Intelligent Diagnosis and Treatment Industry College of Taiyuan University
Guo-Ming Zhang	Shenzhen Eye Hospital; Shenzhen Eye Institute
Hong Zhang	Eye Hospital, Harbin Medical University
Zhi-Chang Zhang	School of Intelligent Medicine, China Medical University
Yi-Tian Zhao	Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences
Bo Zheng	School of Information Engineering, Huzhou University
Hui-Fang Zhou	Shanghai Ninth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine

Guidelines Statement: This guideline was drafted by the expert workgroup of Guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology (2023), some experts from the Ophthalmic Imaging and Intelligent Medical Branch and the Intelligent Medicine Professional Committee of the China Medical Education Association of the China Medical Education Association. All the experts involved in the development of this guideline declare that they maintain an objective stance and base their recommendations on professional knowledge, global research data, and clinical research experience. After thorough discussions, the guideline is formed with unanimous agreement from all the experts.

Disclaimer: The content of this guideline only represents the suggestions and guidance from involved experts on clinical research evaluation methods. This guideline is for reference only and does not represent any laws or regulations. Despite extensive consultation and discussion among experts, some contents are inevitably not comprehensive. The suggestions provided in this guideline are not mandatory opinions, and practices that are inconsistent with this guideline do not imply errors or impropriety. There are still many issues that need to be explored in clinical practice, and ongoing and future clinical studies will provide further evidence. With the accumulation of clinical experience and the emergence of new treatment methods, this guideline needs to be regularly revised and updated in the future to bring more clinical benefits to patients.

Foundations: Supported by National Natural Science Foundation of China (No.61906066); the San Ming Project of Medicine in Shenzhen (No.SZSM202011015); Shenzhen Science and Technology Program (No. KCXFZ20211020163813019).

Conflicts of Interest: Yang WH, None; **Shao Y,** None; **Xu YW,** None; **Expert Workgroup of Guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology (2023),** None; **Ophthalmic Imaging and Intelligent Medicine Branch of Chinese Medicine Education Association,** None; **Intelligent Medicine Committee of Chinese Medicine Education Association,** None.

REFERENCES

1 Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, Liu Y, Topol E, Dean J, Socher R. Deep learning-enabled medical computer vision. *NPJ Digit Med* 2021;4(1):5.

2 Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015;349(6245):255-260.

3 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553): 436-444.

4 Davenport TH, Ronanki R. Artificial intelligence for the real world. *Harvard Business Review* 2018;96(1):108-116.

5 Ao DH, Tian XR, Ma MX, Zhang B, Chen M, Peng YL. Intelligent diagnostic model of keratoconus based on deep learning algorithm. *Guoji Yanke Zazhi (Int Eye Sci)* 2023;23(2):299-304.

6 Escamez CSF, Martinez SP, Fernandez NT. High interpretable machine learning classifier for early glaucoma diagnosis. *Int J Ophthalmol* 2021;14(3):393-398.

7 Ruan S, Liu Y, Hu WT, Jia HX, Wang SS, Song ML, Shen MX, Luo DW, Ye T, Wang FH. A new handheld fundus camera combined with visual artificial intelligence facilitates diabetic retinopathy screening. *Int J Ophthalmol* 2022;15(4):620-627.

8 Savoy M. IDx-DR for diabetic retinopathy screening. *Am Fam Physician* 2020;101(5):307-308.

9 He J, Cao T, Xu F, Wang S, Tao H, Wu T, Sun L, Chen J. Artificial intelligence-based screening for diabetic retinopathy at community hospital. *Eye (Lond)* 2020;34(3):572-576.

10 Li F, Pan J, Yang D, Wu J, Ou Y, Li H, Huang J, Xie H, Ou D, Wu X, Wu B, Sun Q, Fang H, Yang Y, Xu Y, Luo Y, Zhang X. A multicenter clinical study of the automated fundus screening algorithm. *Transl Vis Sci Technol* 2022;11(7):22.

11 Han R, Cheng G, Zhang B, Yang J, Yuan M, Yang D, Wu J, Liu J, Zhao C, Chen Y, Xu Y. Validating automated eye disease screening AI algorithm in community and in-hospital scenarios. *Front Public Health* 2022;10:944967.

12 Yang WH, Zheng B, Wu MN, Zhu SJ, Fei FQ, Weng M, Zhang X, Lu PR. An evaluation system of fundus photograph-based intelligent diagnostic technology for diabetic retinopathy and applicability for research. *Diabetes Ther* 2019;10(5):1811-1822.

13 Zheng B, Yang WH, Wu MN, Zhu SJ, Weng M, Zhang X, Zhang MJ. Establishment and application of diabetic retinopathy intelligent assisted diagnostic technology evaluation system based on fundus photography. *Chin J Exp Ophthalmol* 2019;37(8):674-679.

14 Guidelines for artificial Intelligent Medicine Special Committee of China Medicine Education Association, National Key Research and Development Program of China "Development and Application of Ophthalmic Multimodal Imaging and Artificial Intelligence Diagnosis and Treatment System" Project Team. Intelligent diabetic retinopathy screening system based on fundus photography. *Chin J Exp Ophthalmol* 2019,37(8):593-598.

15 Glaucomatology Group of Chinese Ophthalmological Society, Artificial Intelligence Group of China Association of Medical Equipment. Guidelines on standardized design and application of artificial intelligent assisted glaucoma screening system based on fundus photography (2020). *Chin J Ophthalmol* 2020;56(6):423-432.

16 Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health* 2020;2(10):e549-e560.

17 Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health* 2020;2(10):e537-e548.

18 Sounderajah V, Ashrafian H, Golub RM, *et al*. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11(6):e047709.

19 Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, Logullo P, Beam AL, Peng L, Van Calster B, van Smeden M, Riley RD, Moons KG. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11(7):e048008.

20 Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606.

21 Choi S, Park J, Park S, Byon I, Choi HY. Establishment of a prediction tool for ocular trauma patients with machine learning algorithm. *Int J Ophthalmol* 2021;14(12):1941-1949.

22 National Medical Products Administration, National Health Commission. Quality Management Standards for Clinical Trials of Medical Devices. Published March 24, 2022. Accessed on June 25, 2023. https://www.nmpa.gov.cn/xxgk/fgwj/xzhgfxwj/20220331144903101.html

23 National Medical Products Administration. Technical Guidelines for Clinical Evaluation of Medical Devices. Published September 18, 2021. Accessed on June 25, 2023. https://www.nmpa.gov.cn/ylqx/ylqxggtg/20210928170338138.html

24 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;20(1):144-151.

25 China Association for Quality Inspection. Annotation and quality control specifications for fundus color photographs (T/CAQI 166-2020). *Chin J Exp Ophthalmol* 2021;39(9):761-768.

26 China Association for Quality Inspection. Annotation and quality control specifications for fundus color photograph. *Intelligent Medicine* 2021;1(2):80-87.

27 National Medical Products Administration. Artificial intelligence Medical Device Quality Requirements and Evaluation Part 2: General Requirements for Datasets [YY/T 1833.2—2022]. Accessed on June 25, 2023. http://app.nifdc.org.cn/biaogzx/qxqwk.do?formAction=view&id=2c9048d881c8ca520181d69cd9e149ae

28 National Medical Products Administration. Artificial intelligence Medical Device Quality Requirements and Evaluation Part 1: Terminology [YY/T 1833.1—2022]. Accessed on June 25, 2023. http://app.nifdc.org.cn/biaogzx/qxqwk.do?formAction=view&id=2c9048d881c8ca520181d69c35d44998

29 National Medical Products Administration. Artificial intelligence Medical Device Quality Requirements and Evaluation Part 3: General Requirements for Data Annotation [YY/T 1833.3—2022]. Accessed on June 25, 2023. http://app.nifdc.org.cn/biaogzx/qxqwk.do?formAction=view&id=2c9048d882c4f3180182c936a8b124f9

30 IEEE Engineering in Medicine and Biology Society. IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence. *IEEE* 2022:2801-2022.

31 Expert Workgroup of Expert Consensus for Ethics of Clinical Application of Artificial Intelligence in Ophthalmology, Digital Imaging and Intelligent Medicine Branch of China Medical Education Association, Intelligent Medicine Committee of China Medical Education Association. Expert consensus for ethics of clinical application of artificial intelligence in ophthalmology (2023). *Chin J Exp Ophthalmol* 2023;41(1):1-7.

32 Salmi LR, Saillour-Glénisson F, Alla F, Boussat B. Evaluation and research on public health interventions. *Rev Epidemiol Sante Publique* 2023;71(2):101836.

33 Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, Boyd KA, Craig N, French DP, McIntosh E, Petticrew M, Rycroft-Malone J, White M, Moore L. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ* 2021;374:n2061.

34 Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, Moore L, O'Cathain A, Tinati T, Wight D, Baird J. Process evaluation of complex interventions: Medical Research Council guidance. *BMJ* 2015;350:h1258.

35 Ocular Fundus Diseases Group of Chinese Ophthalmological Society, Expert Group for Artificial Intelligence Research, Development, and Application. The standardized design and application guidelines: a primary-oriented artificial intelligence screening system of the lesion sign in the macular region based on fundus color photography. *Chin J Ocul Fundus Dis* 2022;38(9):711-728.

36 National Health Commission. Guideline for evaluation of qualitative test performance [WS/T 505-2017]. Accessed on June 25, 2023. http://www.nhc.gov.cn/wjw/s9492/201710/057f0376b9ed473fb3bd39c70318ad3a.shtml

37 Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, McGinn T, Guyatt G. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* 2017;318(14):1377-1384.

38 Nattino G, Pennell ML, Lemeshow S. Assessing the goodness of fit of logistic regression models in large samples: a modification of the Hosmer-Lemeshow test. *Biometrics* 2020;76(2):549-560.

39 Yang WH, Shao Y, Xu YW; Expert Workgroup of Guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology (2023); Ophthalmic Imaging and Intelligent Medicine Branch of Chinese Medicine Education Association, Intelligent Medicine Committee of Chinese Medicine Education Association. Guidelines on clinical research evaluation of artificial intelligence in ophthalmology(2023). *Guoji Yanke Zazhi (Int Eye Sci)* 2023;23(7):1064-1071.