·专家述评·

眼科人工智能临床研究评价指南(2023)

杨卫华[1],邵 毅[2],许言午[3,4],《眼科人工智能临床研究评价指南(2023)》专家组,中国医药教育协会眼科影像与智能医疗分会,中国医药教育协会智能医学专业委员会

引用:杨卫华,邵毅,许言午,等. 眼科人工智能临床研究评价指南(2023). 国际眼科杂志 2023;23(7):1064-1071

基金项目:国家自然科学基金资助项目(No.61906066);深圳市医疗卫生三名工程项目(No.SZSM202011015);深圳市科技计划项目(No.KCXFZ20211020163813019)

作者单位:[1](518040)中国广东省深圳市眼科医院 深圳市眼病防治研究所;[2](330006)中国江西省南昌市,南昌大学第一附属医院眼科;[3](510641)中国广东省广州市,华南理工大学未来技术学院;[4](510320)中国广东省广州市,人工智能与数字经济广东省实验室(广州)

作者简介:杨卫华,博士,主任医师,深圳市眼科医院大数据与人工智能办公室主任,研究方向:眼科人工智能、眼科影像。

通讯作者:邵毅,博士,主任医师,副主任,研究方向:眼科人工智能、角膜病及眼科影像. freebee99@163.com;许言午,博士,教授,研究方向:眼科人工智能、医学大数据处理. ywxu@ieee.org

收稿日期:2023-05-04 修回日期:2023-05-30

摘要

人工智能(AI)技术在医学领域的应用是当前的热点。眼科作为医学领域中的AI应用前沿专业之一,运用机器学习技术应用于诊断、干预和预测眼科疾病方面取得了显著的成果。基于眼科AI临床研究的需求,为契合眼科AI临床诊疗发展的实际情况,中国医药教育协会眼科影像与智能医疗分会和智能医学专业委员会组织专家结合近年来国内外AI临床研究的评价报告,经过多轮讨论和修改,形成了针对眼科AI临床研究的评价指南。该指南包括了眼科AI临床研究评价指南制定的背景和方法、AI临床研究评价的国际指南介绍、眼科AI临床研究评价方法等内容,详细介绍了眼科AI临床研究通用评价方法、眼科AI临床研究模型评价方法、常用眼科AI临床研究模型评价指标和计算公式,并详细阐述了眼科AI临床试验评价方法。该指南的制定旨在为眼科AI临床研究人员提供指导和规范,并推动眼科AI临床研究的评价向着规范化和标准化方向发展,进一步提高眼科AI临床研究评价的整体水平。

关键词:人工智能;眼科;评价;临床研究;机器学习;深度学习

DOI:10.3980/j.issn.1672-5123.2023.7.03

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 Special Committee of Chinese Medicine Education Association

Foundation items: 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)

[1]Shenzhen Eye Institute; Shenzhen Eye Hospital, Shenzhen 518040, Guangdong Province, China; [2]Department of Ophthalmology, 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; Pazhou Lab, Guangzhou 510320, Guangdong Province, China. ywxu@ieee.org

Received:2023-05-04 Accepted:2023-05-30

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 Special 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, introduction to international guidelines on the clinical research evaluation of AI, and the evaluation methods of ophthalmic AI models. This guideline introduces general evaluation methods of clinical ophthalmic AI research, evaluation methods of clinical AI models, and common indices and formulae for clinical AI model evaluation in detail, and amply elaborates the evaluation method 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

Citation: Yang WH, Shao Y, Xu YW, *et al*. Guidelines on Clinical Research Evaluation of Artificial Intelligence in Ophthalmology (2023). *Guoji Yanke Zazhi* (*Int Eye Sci*) 2023; 23(7):1064-1071

0 引言

人工智能(artificial intelligence, AI)是计算机科学的一个分支,旨在开发智能机器,使它们能够像人类一样进行学习、推理、判断和决策。AI 包含很多子领域和技术,如自然语言处理、计算机视觉[1]、机器学习[2]、深度学习网络[3]等。AI 被广泛应用于医疗保健、金融、交通运输、制造等领域[4]。随着计算机技术和数据处理能力的不断提升,AI 的发展和应用也越来越广泛和深入。眼科疾病是影响全球人口健康的重要疾病之一,包括白内障、青光眼、糖尿病视网膜病变、年龄相关性黄斑变性、病理性近视等。临床研究对于了解疾病的病理生理机制、发展预防和治疗策略、提高患者生活质量以及降低医疗成本等方面都具有重要意义。AI 在眼科临床研究领域的应用主要包括眼科疾病的预测和诊断[5-6]、治疗和干预、预防和管理等[7-8]。其中,基于眼科影像和 AI 技术的眼科疾病的早期筛查系统,如糖尿病视网膜病变眼底图像辅助诊断软件[9]、眼底病变眼底图像辅助诊断软件(适用于慢性青光眼样视神经病变、糖尿病视网膜病变)[10-11]、慢性青光眼样视神经病变眼底图像辅助诊断软件等产品均通过了中国国家药品监督管理局三类医疗器械注册证的注册审批。

基于眼科影像和 AI 技术的眼科 AI 临床研究如火如荼,随着眼科 AI 临床研究的不断增多,确保其质量和可靠性的评价指南变得尤为必要。这不仅可以确保研究数据的准确性和有效性,而且能提高研究的可重复性和可比性。此外,对 AI 算法和模型的验证和认证也非常关键,以确保其在真实临床环境中的有效性和可靠性[12-13]。因此,中国医药教育协会眼科影像与智能医疗分会和智能医学专业委员会组织成立了《眼科人工智能临床研究评价指南(2023)》专家组,制定适用于眼科 AI 临床研究评价的指南。本指南主要针对基于眼科影像和 AI 技术[14-15]的眼科 AI 临床研究,旨在全面总结眼科 AI 临床研究评价的方法,可以保障眼科 AI 临床研究的质量和可靠性,促进眼科 AI 临床研究的透明度和规范性,同时保护研究参与者隐私和数据安全,平稳推动眼科 AI 临床研究和应用的发展。

1《眼科人工智能临床研究评价指南(2023)》制定方法

基于目前眼科 AI 临床研究评价问题,中国医药教育协会眼科影像与智能医疗分会、智能医学专业委员会组织眼科 AI 专家、眼科临床研究专家、眼科医学伦理专家和眼科 AI 产品研发科学家于 2022-07 成立眼科 AI 临床研究评价指南专家组,于 2022-07-25 开始对眼科 AI 临床研究的相关专家进行访谈,收集并整理相关领域中涉及的眼科 AI 临床研究评价问题及在相关 AI 技术临床研究中面临的困难。由于眼科 AI 临床研究评价尚未形成统一的可遵守的指南,本指南专家组在认真学习国内外眼科 AI 临床研究文献、研究文献的基础上,结合眼科 AI 临床研究的实践经验,召开线下和线上会议,针对收集的眼科 AI 临床研究评价问题进行充分讨论和论证。由执笔专家组成员撰写指南初稿,初稿形成后通过电子邮件和微信方式由各位专家独立阅读并提出修改意见,分别提交指南撰写组核心成员,修改意见经过整理并通过微信、邮件方式和线上会议进行讨论和归纳。指南在修改期间充分接受参与专家的建议和指导意见,最终达成指南终稿,旨在指导眼科 AI 临床研究评价。本指南制定过程历时近 1a。

2 AI 临床研究评价的国际指南介绍

目前,国际上还没有针对眼科 AI 临床研究的评价指南。然而,有一些通用的规范 AI 临床研究或临床试验的指南可以参考。例如 2020 年发布的干预性临床试验的建议-AI 扩展版(Standard Protocol Items: Recommendations for Interventional Trials - Artificial Intelligence, SPIRIT-AI)[16]和临床试验报告统一标准-AI 扩展版(Consolidated Standards of Reporting Trials - Artificial Intelligence, CONSORT-AI)[17],2021 年发布的诊断准确性研究报告标准-AI 扩展版(Standards for Reporting of Diagnostic accuracy studies-Artificial Intelligence, STARD-AI)[18]和个体预后或诊断的多变量预测模型的透明报告-AI 扩展版(Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis - Artificial Intelligence, TRIPOD-AI)[19]。其中,SPIRIT-AI 是涉及 AI 的干预措施临床试验的规范性指南,针对 AI 临床试验方案应报告的特定信息,应与 SPIRIT 2013 和其他 SPIRIT 扩展指南一同使用,目的是促进 AI 临床试验设计和方法的透明度,以促进理解、解释和同行评审[16]。类似地,CONSORT-AI 用于规范涉及 AI 的干预措施临床试验报告,建议提供对 AI 干预措施的清晰描述,包括使用所需的指导和技能、AI 干预集成的环境、AI 干预的输入和输出处理、AI 与人类的交互以及错误案例分析,促进 AI 干预措施临床试验报告的透明度和完整性[17]。STARD-AI 是用于规范以 AI 为核心的诊断测试准确性研究报告的指南,提出需对数据预处理方法、AI 测试开发方法(如数据集划分、模型校准、训练时停止准则、使用外部验证集)、公平度量指标、非标准性能指标、可解释性以及人与 AI 测试的交互等内容进行报告,旨在提高 AI 诊断测试准确性研究的透明度和公平性[18]。TRIPOD-AI 是针对多变量 AI 预测模型研究报告的指南,以帮助研究者透明地报告研究内容,并帮助查阅者理解研究方法和结果,从而减少研究浪费[19]。

3 眼科 AI 临床研究评价方法

眼科 AI 临床研究的环节包括眼科检查数据采集和管理、模型开发、临床试验、临床应用 4 个关键环节。本指南将针对这些关键环节介绍评价方法。值得注意的是,眼科 AI 临床研究模型可按照临床应用的任务分为干预模型、诊断模型、预测模型[20-21]3 种。具体地,眼科 AI 干预模型可作为独立干预措施或联合常规干预措施用于对特定疾病或症状的治疗、预防或管理等;眼科 AI 诊断模型用于确定是否存在某种疾病或病变及其分类、分级;眼科 AI 预测模型用于根据研究参与者的特征预测未来疾病的风险或治疗的效果。因此,对模型评价方法的介绍将按这 3 种眼科 AI 临床研究模型分别展开。此外,由于临床试验是医疗器械在国内和国外上市的必要条件[22-23],本指南将在第 4 节单独介绍眼科 AI 临床试验的评价方法。

3.1 眼科 AI 临床研究通用评价方法

3.1.1 数据采集和管理的评价 针对眼科 AI 临床研究中数据采集和管理环节的评价是为了确保研究数据的数量、质量、完整性、安全性以及可靠性[24]。具体的评价方法建议覆盖以下几个方面:(1)数据数量评价:评价收集数据的数量,确保其符合临床研究中模型的开发、性能的验证等要求。(2)数据质量评价:评价数据的质量[25-26],包括数据的完整性、准确性、逻辑性、一致性和可用性等,确保数据的质量符合要求[27]。(3)数据清洗评价:评价数据清洗过程是否保持脱敏、是否符合逻辑、是否有效等。(4)数据标签评价:评价数据标签,即参考标准[28]的构建过程和标签质量,确保数据标签可靠。对于依赖人工标注而生成的标签,需评价标注流程的规范性、标注人员和设备、标注过程以及标注质量[29]。(5)数据存储评价:评价数据的存储质量,确保数据的存储安全且符合要求。常用的方法包括检查数据的存储位置、存储介质和存储方式等。(6)数据管理评价:评价数据的管理质量,确保数据的管理安全且符合要求。常用的方法包括检查数据的管理过程和数据管理人员的能力等[30]。(7)数据使用评价:评价数据的使用质量,确保数据的使用以及共享过程安全且符合要求。常用的方法包括检查数据使用的目的、范围、伦理性[31]、合法性,以及数据共享的政策、共享方式和目的等。

3.1.2 眼科 AI 模型开发的评价 针对眼科 AI 临床研究中模型开发环节的评价是为了确保研究开发的模型具有高质量、可靠性以及稳定性。具体的评价方法建议覆盖以下几个方面:(1)开发数据集的评价:评价开发 AI 模型所使用的数据集的质量、数量、均衡性是否足够,数据集的代表性如何,训练集、验证集、测试集的划分是否合理;评价标签的定义方法是否有充分的临床依据。(2)特征选择和提取评价:若需要人工选择特征,则评价选择的特征是否能够对模型的性能产生重要影响,同时评价特征提取方法是否合适。(3)眼科 AI 模型性能评价:使用常见的指标评价模型的性能,确保模型能够准确地预测目标变量,详见 3.2 小节。(4)交叉验证:使用交叉验证方法(如 k 折交叉验证)来评价模型的泛化能力,确保模型能够在新数据上进行准确预测。(5)模型解释性评价:评价模型的解释性,确保模型的预测结果可被临床解释和理解。(6)模型稳定性评价:评价模型对数据噪声和随机性的稳定性,确保模型在面对不同数据集时能够产生一致的结果。(7)模型适应性评价:评价模型在不同群体和不同环境下的适应性,确保模型能够在实际应用中产生准确的结果。

3.1.3 眼科 AI 模型临床应用的评价 针对眼科 AI 模型临床应用的评价是为了确保临床应用的安全、有效以及可重复性。具体的评价方法建议覆盖以下几个方面:(1)安全性评价:评价临床应用过程是否存在数据隐私和安全性等方面的问题,以保护研究参与者的隐私权和个人信息。(2)内部有效性评价:评价研究结果的准确性、可信度和适用性。内部有效性的高低取决于研究设计的合理性、研究组和对照组的选取和分配、盲法设计、研究过程中的控制和管理以及数据分析的可靠性等因素。(3)外部有效性评价:评价研究结果的推广能力和普适性。外部有效性的高低取决于研究样本的代表性、试验环境的真实性、研究方法的通用性和研究结果的适用性等因素。(4)可重复性评价:评价研究结果是否能被重复验证,即评价 AI 模型在不同数据集上的性能是否稳定、性能波动范围是否可接受,在不同设备上的表现是否一致,在同一数据多次输入情况下的预测结果是否一致。可重复性的高低取决于模型开发阶段数据的代表性、研究过程的透明度、研究方法的清晰度、数据的公开性和分析的可重复性等因素。(5)应用效果评价:评价临床应用中的效果,包括对患者诊断和治疗的指导和改善程度。(6)卫生经济学分析评价:评价在临床应用中的卫生经济学价值,包括成本效果分析、成本效用分析、成本效益分析等,成本包括人力、物力和经济成本等,产出指标包括实际应用过程中产生的临床效果、质量调整生命年和节约的医疗费用等。

3.2 眼科 AI 临床研究模型评价方法

3.2.1 眼科 AI 干预模型的评价 眼科 AI 干预模型可作为独立干预措施或联合常规干预措施用于对特定疾病或症状的治疗、预防或管理等。为证明眼科 AI 干预模型对治疗目标病症有效,眼科 AI 干预模型临床研究的评价主要指标是干预过程评价和干预效果评价两方面。干预过程的评价可通过与常规干预措施直接比较,从干预过程的时长、安全性和有效性、卫生经济学等方面开展评价,根据指标数据的类型选择适合的统计学方法进行比较[32-34]。干预效果的评价通常使用临床结局指标来衡量,如死亡率、疾病复发率、生存期等,可以通过干预后症状减轻、疾病进展或生存率等结果来评价,详见 3.3.1 小节。

3.2.2 眼科 AI 诊断模型的评价 诊断模型是用于确定是否存在某种疾病或病变的模型。评价诊断模型的主要目标是考察其诊断准确性,可使用的评价指标可包括灵敏度、特异度、准确率和 Kappa 一致性系数等,详见 3.3.2 小节。

3.2.3 眼科 AI 预测模型的评价 预测模型用于根据研究参与者的特征预测疾病的风险、生理结构的变化,或预测治疗效果。评价预测模型可包含评价疾病未来发生与否的分类结果,评价未来生理结构测量参数的回归结果等。在有明确预测标签(参考标准)的情况下,可使用的评价指标可包括均方根误差、平均绝对误差、灵敏度、特异度等;在没有明确预测标签(参考标准)的情况下,可使用的评价指标可包括与其他优秀方法获得结果的阳性符合率、阴性符合率、总符合率等,详见 3.3.3 小节。

3.3 常用眼科 AI 临床研究模型评价指标和计算公式 本指南提供了常用的眼科 AI 模型评价指标及其计算公

式[28,35],不同模型的临床研究应根据实现的任务选择不同的指标进行评价。

3.3.1 眼科 AI 干预模型的常用结局评价指标和计算公式

(1)干预模型死亡率,指研究参与者在干预后死亡的比例:

$$死亡率=\frac{干预后死亡人数}{干预人数}\times 100\% \quad ①$$

(2)干预模型疾病复发率,指研究参与者在干预后疾病再次发作的比例:

$$疾病复发率=\frac{干预后疾病复发人数}{干预人数}\times 100\% \quad ②$$

(3)干预模型生存期,指研究参与者从干预开始到死亡或失访之间的天数。

3.3.2 眼科 AI 诊断模型的常用评价指标和计算公式

(1)混淆矩阵,一种特殊的、具有两个维度的可视化矩阵,可用于监督学习评价时比较分类结果和实际测得值。混淆矩阵的每一行代表了预测类别,每一行的数据总数表示预测为该类别的数据的数目;每一列代表了数据的真实归属类别,每一列的数据总数表示该类别的数据数目;每一元素中的数值表示对应真实类别数据被预测某类的数目(表 1)。

表 1 混淆矩阵示意

混淆矩阵		参考标准		合计
		阳性	阴性	
预测值	阳性	真阳性(TP)	假阳性(FP)	R_1
	阴性	假阴性(FN)	真阴性(TN)	R_2
合计		C_1	C_2	N

注:TP:实际为阳性的样本被正确地判别为阳性的数量;FP:实际为阴性的样本被错误地判别为阳性的数量;FN:实际为阳性的样本被错误地判别为阴性的数量;TN:实际为阴性的样本被正确地判别为阴性的数量;R_1:真阳性和假阳性例数的总和,R_2:假阴性和真阴性例数的总和,C_1:真阳性和假阴性例数的总和,C_2:假阳性和真阴性例数的总和,N:样本总数量。

(2)灵敏度(Sensitivity, Sen),又可称召回率(Recall, R)、查全率,是真阳性样本占全体阳性样本的比例:

$$Sen=\frac{TP}{TP+FN} \quad ③$$

(3)特异度(Specificity, Spe),真阴性样本占全体阴性样本的比例:

$$Spe=\frac{TN}{TN+FP} \quad ④$$

(4)似然比(Likelihood Ratio, LR),同时反映敏感度和特异度的复合指标,即患病者中得出某一筛检研究结果的概率与未患病者得出这一概率的比值。

阳性似然比(Positive Likelihood Ratio, +LR),筛检结果的真阳性率与假阳性率之比,比值越大,研究结果阳性时为真阳性的概率越大:

$$+LR=\frac{Sen}{1-Spe} \quad ⑤$$

阴性似然比(Negative Likelihood Ratio, -LR),筛检结果的假阴性率与真阴性率之比,其比值越小,研究结果阴性时为真阴性的可能性越大:

$$-LR=\frac{1-Sen}{Spe} \quad ⑥$$

(5)准确率(Accuracy, Acc),算法诊断正确的样本占全体样本的比例:

$$Acc=\frac{TP+TN}{N} \quad ⑦$$

(6)精确率(Precision, Pre),又称阳性预测值(Positive Prediction Value, PPV)、查准率,是真阳性样本占算法判为阳性样本的比例:

$$Pre=\frac{TP}{TP+FP} \quad ⑧$$

(7)阴性预测值(Negative Prediction Value, NPV),真阴性样本占被算法判为阴性样本的比例:

$$NPV=\frac{TN}{TN+FN} \quad ⑨$$

(8)漏检率(Miss Rate, MR),也称为漏报率、漏诊率、漏警率、假阴性率,即检测中未发现的阳性样本占全体阳性样本的比例:

$$MR=1-\frac{TP}{TP+FN}=1-Sen \quad ⑩$$

(9)误检率(False Alarm Rate, FA),也称为误报率、误诊率、虚警率、假阳性率,即全体阴性样本中被错误地预测为阳性样本的比例:

$$FA=1-\frac{TN}{TN+FP}=1-Spe \quad ⑪$$

(10)F_1 分数(F_1 Score),召回率和精确率的调和平均数:

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

式中,P 表示精确率;R 表示召回率。

(11)约登指数(Youden Index, YI),也称正确指数,假设假阴性(漏诊率)和假阳性(误诊率)危害同等意义,约登指数为灵敏度与特异度之和减去 1,指数越大说明筛查效果越好。

$$YI=Sen+Spe-1 \quad ⑬$$

(12)Kappa 系数(Kappa Value),用于评价筛查系统与参考标注诊断结果一致性的指标:

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

式中$p_o=(TP+TN)/N$, $p_e=(R_1C_1+R_2C_2)/N\times N$,即:

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

(13)受试者操作特征(Receiver operating characteristic,ROC)曲线下面积(area under curve,AUC):ROC 是通过在一组(一系列)预设阈值下估计的筛查系统在测试集上的灵敏度和特异度,从而产生一组(1-特异度,灵敏度)操作点,将这些操作点依次连接形成的曲线,AUC 即为该曲线和 X 轴所围成的面积(图 1),可用于度量分类模型的性能,取值范围一般为 0.5~1,且值越大代表模型分类效果越好。

(14)精确率-召回率(Precision-Recall,PR)曲线:PR 曲线与 ROC 类似,是通过在一组(一系列)预设阈值下估计的筛查系统在测试集上的精确率和召回率,从而产生一组(召回率,精确率)操作点,将这些点依次连接形成的曲线(图 1)。

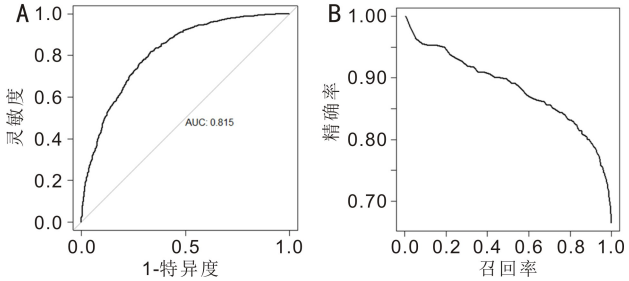

图 1　ROC 曲线及 AUC 指标与 PR 曲线示意图　A:ROC 曲线;B:PR 曲线。

3.3.3 眼科 AI 预测模型的常用评价指标和计算公式

预测模型若输出分类类别结果,则可使用 3.3.2 小节提供的评价指标和计算公式进行评价;若输出为连续数值结果,则可使用如下的评价指标和计算公式:

(1)均方根误差(Root Mean Square Error, RMSE)可以衡量预测值和真值之间的偏差,能够反映出测量的精确度。均方根误差越接近于 0,表明模型对于目标值预测的效果越好:

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \quad ⑯$$

式中,N 为总样本数,y_i 为第 i 个样本的真值,$\hat{y}_i$ 为第 i 个样本的预测值。

(2)平均绝对误差(Mean Absolute Error, MAE),是各个测量值与参考标准的偏差绝对值的平均值。平均绝对误差可避免误差相互抵消的问题,准确地反映实际预测误差的大小:

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i-y_i| \quad ⑰$$

(3)平均绝对百分比误差(Mean Absolute Percentage Error, MAPE),是一种相对度量,相较于 MAE,MAPE 计算了预测值和参考标准偏差相对参考标准的百分比:

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

MAPE 的范围是[0,+∞),值为 0 代表完美模型,值大于 100%代表劣质模型。注意当参考标注值为 0 时公式不可用。

(4)对称平均绝对百分比误差(Symmetric Mean Absolute Percentage Error,SMAPE),与 MAPE 相比,计算公式分母中的参考标准绝对值被替换为参考标准绝对值和预测值绝对值的中值:

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

SMAPE 的取值范围为[0,200%],当参考标准和预测值同时为 0 时公式不可用。

(5)R^2,也叫决定系数,是回归预测值和标定值之间拟合程度的统计系数。R^2 值介于 0~1 之间,越接近 0,表明模型的预测结果越接近随机;越接近 1,表明模型回归预测目标值的拟合效果越好:

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

式中,N 为总样本数,y_i 为第 i 个样本的真值,$\hat{y}_i$ 为第 i 个样本的预测值,$\bar{y}$ 为所有样本真值的平均值。

(6)当预测模型输出的参考标准未知时,可将待评价方法结果与其他方法获得的结果进行符合率的计算,如阳性符合率、阴性符合率、总符合率,如表 2[36] 和公式所示:

表 2　参考标准未知的 2×2 表

待评价方法	比较方法		合计
	阳性	阴性	
阳性	a	b	a+b
阴性	c	d	c+d
合计	a+c	b+d	n

$$\text{阳性符合率}=\frac{a}{a+c}\times 100\% \quad ㉑$$

$$\text{阴性符合率}=\frac{d}{b+d}\times 100\% \quad ㉒$$

$$\text{总符合率}=\frac{a+d}{n}\times 100\% \quad ㉓$$

(7)对预测模型除了评价其准确性,对其校准或拟合优度的考察也十分重要。校准或拟合优度被认为是预测模型最重要的属性之一,它反映了预测模型正确估计绝对风险的程度,校准不当的预测模型会低估或高估目标结果[37]。校准或拟合优度的评价方法通常使用 Hosmer-Lemeshow 拟合度检验和校准曲线。

Hosmer-Lemeshow 拟合优度检验(HL 检验)[37],用于判断预测值与真实值之间的差异情况。若 $P\leqslant 0.05$,表示预测值与真实值之间的差异具有统计学意义,说明模型拟合度较差;若 $P>0.05$,则提示通过 HL 检验,说明预测值与真实值之间无明显差异[38]。

校准曲线(Calibration Curve)[37]用于辅助观察模型的预测概率是否接近于真实概率,是实际发生率-预测发生率的散点图,本质上是拟合优度检验的结果可视化。

3.3.4 眼科 AI 临床研究中其他常用评价指标和计算公式

(1)数据有效使用率,是数据收集和处理过程中,最终被有效使用的数据占总数据量的比例:

$$\text{数据有效使用率}=\frac{\text{有效数据量}}{\text{总数据量}}\times 100\% \quad ㉔$$

(2)样本量估算公式,可根据眼科 AI 模型的预期效果,推导测试集中各类别数据需要的数量:

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

式中,Z 为置信水平的 Z 统计量,Δ 为允许误差,P 为预期的准确率、灵敏度、特义度等评价指标,N 为所需样本量。通常设定参数估计双侧可信区间的可信度为 95%(即Ⅰ类错误 α 为 0.05,双侧),则 $Z_{1-\alpha/2}=1.96$,预期评价指标估计精度(可信区间半宽度)Δ 通常设置为 5%。

(3)评价多类别分类眼科 AI 研究任务时,对于多分类眼科 AI 研究任务,若多个类别互相独立,则可将多类别的评价转化为多个二分类问题的评价,每一类的阴性样本定义为总样本中除了该类别为阳性的样本之外的所有样本。可计算的评价指标包括 Micro/Macro F_1 值,Micro/Macro AUC 和 Kappa 值。

其中,Macro F_1 和 Macro AUC 值是先分别计算每一类预测的 F_1 值和 AUC 值,然后将各个类别的 F_1 值和 AUC 值取平均:

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

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

式中,C 为分类任务的总类别数。

Micro F_1 和 Micro AUC 值则是先计算总体样本的真阳性、假阳性、真阴性和假阴性样本数,再根据 F_1 和 AUC 定义进行计算,即:

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

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

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

Micro AUC 依赖全局的混淆矩阵,在绘制全局 ROC 曲线时,横纵坐标点分别代表全局的 1-特异度和灵敏度,即 $\left(1 - \frac{\sum_{i=1}^{C} TN_i}{\sum_{i=1}^{C} TN_i + \sum_{i=1}^{C} FP_i}, \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + \sum_{i=1}^{C} FN_i}\right)$。 (31)

Micro/Macro F1, Micro/Macro AUC 均为 0~1 之间的数值,值越接近 1 表示多分类模型的效果越好。

Kappa 一致性系数在评价多分类任务时:

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

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

(4)眼科 AI 临床研究中结构区域分割评价时,评价结构(生理结构、病灶等)区域分割结果是否准确的评价指标通常有 DICE 系数和 Jaccard 系数:

DICE 系数(Dice Coefficient),结构区域分割轮廓与参考标准轮廓的交集占分割轮廓与参考标准轮廓平均值的比例(图 2):

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

其中 $|X\cap Y|$ 是 X 和 Y 之间的交集,$|X|$ 和 $|Y|$ 分表表示 X 和 Y 的元素的个数。

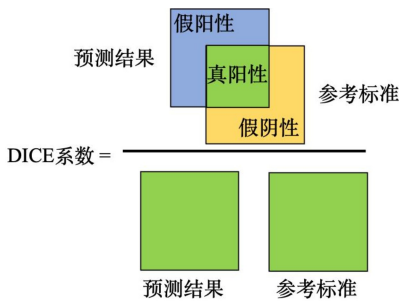

图 2 DICE 系数计算示意。

Jaccard 系数(Jaccard Coefficient),结构区域分割轮廓与参考标准轮廓的交集占分割轮廓与目标轮廓并集的比例(图 3),又称交并比(Intersection over Union, IoU):

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

4 眼科 AI 临床试验评价方法

临床试验是临床研究的重要组成部分,用于验证药物或医疗器械的安全性和有效性。对于眼科 AI 临床试验的的评价方法建议覆盖以下几个方面:试验设计、研究参与者群体、伦理问题、样本量、对照和盲法设计、试验结果、数据分析、不良事件等。(1)试验设计:临床试验的设计应适合于回答临床试验的问题,包括试验类型、前瞻性还是回顾性、单中心还是多中心、优效性设计或非劣性设计还是单组目标值设计等。例如,针对干预模型的临床试验需保证对参与者进行足够时间的随访,确保干预在一定时期内是安全有效的。针对医学影像诊断模型的 AI 医疗器械临床试验,为避免医生的主观因素和不确定性等因素的影响,可采用多阅片者多病例(Multi-reader Multi-case, MRMC)试验设计,确保全面评价模型性能,减小因研究者个体差异造成的误差。(2)研究参与者群体:临床试验需有一个明确的研究参与者群体,该研究参与者群体是被研究人群的代表。临床试验需要根据研究参与者的特点和试验目的,合理选择研究参与者,保证样本的代表性和多样性。(3)伦理问题:临床试验应符合伦理原则,研究参与者在参加临床试验前应签署知情同意书,且临床试验应获得伦理委员会的批准[31]。(4)样本量:临床试验应具有合适的样本量,满足统计分析的要求,以发现组间有意义的差异。(5)对照和盲法设计:干预性临床试验研究参与者应随机分为治疗组和对照组,且应采用双盲方法,以尽量减少选择偏差,并确保各组在基线时具有可比性。诊断性或预测性临床试验设计应适合于回答临床试验的问题。诊断性临床试验应以目前临床上标准的方法作为对照方法。(6)试验结果:被测量的结果应明确定义并与临床试验问题相关,并应使用标准化方法进行测量。(7)数据分析:数据的统计分析应适当,试验结果应以清晰透明的方式呈现。(8)不良事件:临床试验应报告在试验期间发生的任何不良事件,并应评价临床试验的安全性和耐受性。

5 总结

眼科是医学 AI 最为活跃的临床专科,随着基于眼科影像和 AI 技术的眼科 AI 临床研究的不断增多,为保障眼科 AI 临床研究的质量和可靠性,我们制定了眼科 AI 临床

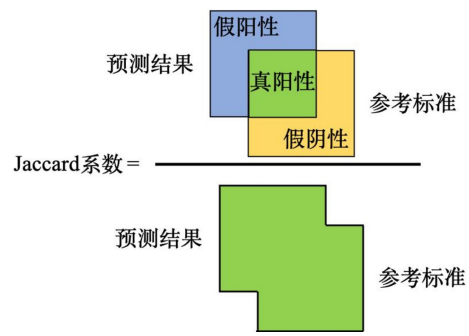

图 3 Jaccard 系数计算示意。

研究评价指南。本指南总结了眼科 AI 临床研究评价指南制定的背景和方法、介绍了 AI 临床研究评价的国际指南、并讨论了眼科 AI 临床研究评价方法。详细介绍了眼科 AI 临床研究通用评价方法、眼科 AI 临床研究模型评价方法、常用眼科 AI 临床研究模型评价指标和计算公式,并详细阐述了眼科 AI 临床试验评价方法。眼科 AI 临床研究评价指南的制定,有助于改进临床研究方案的设计、实施和研究质量,从而提高研究的完整性和透明度,减少潜在的偏倚。本指南的目的是提出眼科 AI 临床研究评价的建议,从而提高相关人员对眼科 AI 临床研究评价的规范意识。眼科 AI 临床研究中,研究者可根据研究的环节、模型的类型来选用相对应的评价指标和计算公式。

本指南是第一部关于眼科 AI 临床研究评价的指南,随着医学领域中 AI 技术应用方面法律法规政策方针的逐步出台,本指南内容将得到进一步的讨论和更新。欢迎对本指南存在的不足提出宝贵的建议和意见,使得本指南能够不断更新和完善。

形成指南专家组成员

执笔专家:

杨卫华 深圳市眼科医院 深圳市眼病防治研究所
许言午 华南理工大学未来技术学院 人工智能与数字经济广东省实验室(广州)
方慧卉 人工智能与数字经济广东省实验室(广州)
邵 毅 南昌大学第一附属医院
张少冲 深圳市眼科医院 深圳市眼病防治研究所
魏永越 北京大学公众健康与重大疫情防控战略研究中心
刘祖国 厦门大学眼科研究所
周吉银 陆军军医大学第二附属医院
周永进 深圳大学医学部生物医学工程学院

参与起草的专家(按姓氏拼音排列):

Sunee Chansangpetch 泰国朱拉隆功国王纪念医院眼科
陈 浩 温州医科大学附属眼视光医院
陈 杰 鹏城实验室
陈羽中 北京鹰瞳科技发展股份有限公司
崔红光 浙江大学医学院附属第一医院
戴 琦 温州医科大学附属眼视光医院
戴伟伟 爱尔数字眼科研究所
邓爱军 潍坊医学院附属医院
丁 琳 新疆维吾尔自治区人民医院
段立新 电子科技大学(深圳)高等研究院
付华柱 新加坡科技研究局高性能计算研究所
戈宗元 北京鹰瞳科技发展股份有限公司
韩 伟 浙江大学医学院附属第二医院
黄厚斌 解放军总医院眼科医学部 解放军总医院海南医院
蒋 沁 南京医科大学附属眼科医院
雷柏英 深圳大学医学部生物医学工程学院
柯根杰 安徽省立医院
刘 虎 南京医科大学第一附属医院
李世迎 厦门大学附属翔安医院暨厦门大学医学中心
李 文 电子科技大学(深圳)高等研究院
李小萌 香港科技大学
刘小晴 北京致远慧图科技有限公司
娄 岩 中国医科大学智能医学学院
陆培荣 苏州大学附属第一医院
宋宗明 河南省立眼科医院 河南省人民医院
孙 斌 山西省眼科医院
谭明奎 华南理工大学软件学院
陶黎明 安徽医科大学第二附属医院
万 程 南京航空航天大学
魏锐利 海军军医大学上海长征医院
吴 健 浙江大学医学院附属第二医院 浙江大学公共卫生学院
肖 璇 武汉大学人民医院
徐 捷 首都医科大学附属北京同仁医院 北京市眼科研究所
徐 雯 浙江大学医学院附属第二医院
徐 帆 广西壮族自治区人民医院
许晶晶 北京致远慧图科技有限公司
杨永升 中国中医科学院眼科医院
姚 进 南京医科大学附属眼科医院
叶 娟 浙江大学医学院附属第二医院
岳丽菁 广东省第二中医院
张冬冬 北京至真互联网技术有限公司
张光华 太原学院大数据智能诊疗产业学院
张国明 深圳市眼科医院 深圳市眼病防治研究所
张 弘 哈尔滨医科大学附属第一医院眼科医院
张志常 中国医科大学智能医学学院
赵一天 中国科学院慈溪生物医学工程研究所
郑 博 湖州师范学院信息工程学院
周慧芳 上海交通大学医学院附属第九人民医院

利益冲突:

所有作者均声明不存在利益冲突。本指南的制定未接受任何企业的赞助。

指南声明:

本指南为《眼科人工智能临床研究评价指南(2023)》专家组、中国医药教育协会眼科影像与智能医疗分会和中国医药教育协会智能医学专业委员会部分专家起草。所有参与本指南制定的专家均声明,坚持客观的立场,以专业知识、全球研究数据和临床研究经验为依据,经过充分讨论,全体专家一致同意后形成本指南。

免责声明:

本指南的内容仅代表参与制定的专家对临床研究评价方法的建议指导意见,供临床医师参考;本指南的内容不代表任何的法律法规。尽管专家们进行了广泛

的意见征询和讨论,但仍有不全面之处。本指南所提供的建议并非强制性意见,与本指南不一致的做法并不意味着错误或不当。临床实践中仍存在诸多问题需要探索,正在进行和未来开展的临床研究将提供进一步的证据。随着临床经验的积累和新的治疗方法的涌现,未来需要对本指南定期修订、更新,为患者带来更多临床获益。

参考文献

1 Esteva A, Chou K, Yeung S, *et al*. 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 敖弟华, 田熙睿, 马明勋, 等. 基于机器深度学习算法的圆锥角膜智能化诊断模型研究. 国际眼科杂志 2023;23(2):299-304

6 Fernandez Escamez CS, Martin Giral E, Perucho Martinez S, *et al*. High interpretable machine learning classifier for early glaucoma diagnosis. *Int J Ophthalmol* 2021;14(3):393-398

7 Ruan S, Liu Y, Hu WT, *et al*. 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 TY, Xu FP, *et al*. Artificial intelligence-based screening for diabetic retinopathy at community hospital. *Eye* (*Lond*) 2020;34(3):572-576

10 Li F, Pan JY, Yang DL, *et al*. A multicenter clinical study of the automated fundus screening algorithm. *Transl Vis Sci Technol* 2022;11(7):22

11 Han RA, Cheng GW, Zhang BL, *et al*. 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, *et al*. 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 郑博, 杨卫华, 吴茂念, 等. 基于眼底照相的糖尿病视网膜病变智能辅助诊断技术评价体系的建立及应用. 中华实验眼科杂志 2019;37(8):674-679

14 袁进, 雷博, 张明, 等. 基于眼底照相的糖尿病视网膜病变人工智能筛查系统应用指南. 中华实验眼科杂志 2019;37(8):593-598

15 中华医学会眼科学分会青光眼学组, 中国医学装备协会眼科人工智能学组. 中国基于眼底照相的人工智能青光眼辅助筛查系统规范化设计及应用指南(2020年). 中华眼科杂志 2020;56(6):423-432

16 Cruz Rivera S, Liu XX, Chan AW, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health* 2020;2(10):e549-e560

17 Liu XX, Cruz Rivera S, Moher D, *et al*. 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, *et al*. 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, *et al*. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606

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

22 国家药品监督管理局, 国家卫生健康委员会. 医疗器械临床试验质量管理规范[EB/OL].(2022-03-24)[2023-05-14]. https://www.nmpa.gov.cn/xxgk/fgwj/xzhgfxwj/20220331144903101.html

23 国家药品监督管理局. 医疗器械临床评价技术指导原则[EB/OL].(2021-09-18)[2023-05-14]. https://www.nmpa.gov.cn/ylqx/ylqxggtg/20210928170338138.html

24 Weiskopf NG, Weng CH. 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 中国质量检验协会. 眼底彩照标注与质量控制规范(T/CAQI 166-2020). 中华实验眼科杂志 2021;39(9):761-768

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

27 人工智能医疗器械质量要求和评价. 第2部分:数据集通用要求: YY/T 1774—2021

28 人工智能医疗器械质量要求和评价. 第1部分:术语: YY/T 1774—2021

29 人工智能医疗器械质量要求和评价. 第3部分:数据标注通用要求: YY/T 1774—2021

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

31《眼科人工智能临床应用伦理专家共识》专家组, 中国医药教育协会数字影像与智能医疗分会, 中国医药教育协会智能医学专业委员会, 等. 眼科人工智能临床应用伦理专家共识(2023). 中华实验眼科杂志 2023;41(1):1-7

32 Salmi LR, Saillour-Glénisson F, Alla F, *et al*. L'évaluation et la recherche sur les interventions en santé publique [Evaluation and research on public health interventions]. *Rev Epidemiol Sante Publique* 2023;71(2):101836

33 Skivington K, Matthews L, Simpson SA, *et al*. 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, *et al*. Process evaluation of complex interventions: medical Research Council guidance. *BMJ* 2015;350:h1258

35 中华医学会眼科学分会眼底病学组, 人工智能研发应用专家指导组. 面向基层的人工智能眼底彩色照相黄斑区域病变体征筛查系统规范化设计及应用指南. 中华眼底病杂志 2022;38(9):711-728

36 中华人民共和国卫生行业标准. 定性测定性能评价指南. WS/T 505-2017[S/OL]

37 Alba AC, Agoritsas T, Walsh M, *et al*. 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