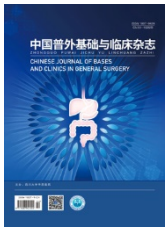

中国普外基础与临床杂志
*Chinese Journal of Bases and Clinics in General Surgery*
ISSN 1007-9424,CN 51-1505/R

# 《中国普外基础与临床杂志》网络首发论文

题目：基于 AI 的诊断准确性和预后研究报告规范：TRIPOD+AI 声明解读
作者：粟文，赖泽鹏，蒋昊林，曾子倩，陈卫中
收稿日期：2024-11-25
网络首发日期：2025-01-02
引用格式：粟文，赖泽鹏，蒋昊林，曾子倩，陈卫中．基于 AI 的诊断准确性和预后研究报告规范：TRIPOD+AI 声明解读[J/OL]．中国普外基础与临床杂志.
https://link.cnki.net/urlid/51.1505.R.20241230.1604.008

·指南解读·

# 基于 AI 的诊断准确性和预后研究报告规范：TRIPOD+AI 声明解读

粟文，赖泽鹏，蒋昊林，曾子倩，陈卫中

成都医学院公共卫生学院流行病与卫生统计学教研室（成都 610500）

陈卫中

教授，硕士研究生导师，成都医学院流行病与卫生统计学教研室主任。任四川省课程思政教育指导委员会委员、国家医师资格考试试题开发专家委员会委员、中国卫生信息学会卫生统计学教育专业委员会常务委员、健康统计学专业委员会委员，《中国妇幼卫生杂志》、《南方医科大学学报》、《成都医学院学报》编委。四川省一流课程、课程思政课程负责人，参与多部国家级规划教材的编写。主要从事医学临床试验和诊断试验统计学方法研究，先后主持或参与临床试验设计及统计分析项目 20 余项；以第一或通信作者发表 SCI、中文核心等学术论文 30 余篇。

**【摘要】** 随着临床和生物大数据的极大丰富，机器学习技术通过结合多方面的信息以预测个体的健康结局，在科研及学术论文中应用日益广泛，但关键信息报告的不足也逐渐显现，包括数据偏倚、模型对不同群体的公平性、数据质量和适用性问题，以及在真实临床环境中保持预测准确性和可解释性的难度等，增加了将预测模型安全有效地应用于临床实践的复杂性。针对这些问题，多变量预测模型个体预后或诊断的透明报告（transparent reporting of a multivariable prediction model for individual prognosis or diagnosis，TRIPOD）+人工智能（artificial intelligence，AI）声明在 TRIPOD 的基础上提出了针对机器学习模型的报告规范，以提升模型的透明性、可重复性和健康公平性，从而改善机器学习模型的应用质量。当前，国内基于机器学习技术的预测模型研究日益增多。为帮助国内读者更好地理解和应用 TRIPOD+AI，笔者结合实例对其进行了解读，希望为研究人员报告质量提升提供支持。

**【关键词】** TRIPOD+AI；机器学习；机器学习模型；报告规范；预测模型

## AI-based diagnostic accuracy and prognosis research reporting guideline: interpretation of the TRIPOD+AI Statement

SU Wen, LAI Zepeng, JIANG Haolin, ZENG Ziqian, CHEN Weizhong

*Department of Epidemiology and Statistics, School of Public Health, Chengdu Medical College, Chengdu 610500, P. R. China*

*Corresponding author: CHEN Weizhong, Email: wejone@126.com*

**【Abstract】** With the increasing availability of clinical and biomedical big data, machine learning is being widely used in scientific research and academic papers. It integrates various types of information to predict individual health outcomes. However, deficiencies in reporting key information have gradually emerged. These include issues like data bias, model fairness across different groups, and problems with data quality and applicability. Maintaining predictive accuracy and interpretability in real-world clinical settings is also a challenge. This increases the complexity of safely and effectively applying predictive models to clinical practice. To address these problems, TRIPOD+AI (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis+artificial intelligence) introduces a reporting standard for machine learning models. It is based on TRIPOD and aims to improve transparency, reproducibility, and

DOI：10.7507/1007-9424.202411127

基金项目：2024 年四川省研究生教育教学改革项目（项目编号：YJGXM24-C159）；2024 年校级研究生教育教学改革项目（项目编号：YJG202410）

通信作者：陈卫中，Email：wejone@126.com

粟文和赖泽鹏为共同第一作者

health equity. These improvements enhance the quality of machine learning model applications. Currently, research on prediction models based on machine learning is rapidly increasing. To help domestic readers better understand and apply TRIPOD+AI, we provide examples and interpretations. We hope this will support researchers in improving the quality of their reports.

【**Keywords**】 TRIPOD+AI; machine learning; machine learning model; reporting guideline; predictive model

随着全球科技革命 4.0、生命科学革命 3.0 时代的到来，在精准医学理念下，医学的数据化、精准化、智能化特征越来越明显，医学和生物科学领域数据得到极大丰富。由数据驱动的人工智能（artificial intelligence，AI）及机器学习（machine learning）技术的发展，以及在医学领域的应用，使大规模、高维度、动态性的医学大数据得以快速整合，其产生的判别或分类模型（discriminative model）和预测模型（predictive model）在医学领域中被用来判别/区分不同的疾病状态，预测发病风险或预后，展现出在疾病诊断、预后预测和治疗决策支持中的巨大潜力[1-2]。但是，机器学习模型的应用也带来了诸如数据偏倚、模型透明性、结果可重复性等挑战[3]，因此规范化的报告标准显得尤为重要。为应对这一需求，构建起数据驱动的诊断、预后研究标准化体系刻不容缓。为此，Gary 等发布了基于机器学习的多变量预测模型个体预后或诊断的透明报告（transparent reporting of a multivariable prediction model for individual prognosis or diagnosis，TRIPOD）+AI 声明。为帮助学者更好地理解和应用 TRIPOD+AI 声明，笔者结合实例对其进行了解读，希望为研究人员提升报告质量提供支持。

## 1　TRIPOD+AI 声明的制订背景

Moher 等在 2010 年着手进行 TRIPOD 的开发，并在 2015 年发布（https:// www.tripod-statement.org/），旨在为开发或评估预测模型性能的研究提供最低限度的报告建议[4]。但随着机器学习技术的兴起，支持向量机、随机森林、深度学习等机器学习算法给模型带来了革命性的变革，虽然 TRIPOD 声明与建模技术本身关联不大，但制定之时主要针对的是由理论驱动的统计回归模型，其与机器学习技术在建模策略、数据处理、评价目标等方面都有较大差别，对报告的透明度和完整性提出了新的要求，因此急需对声明进行更新。TRIPOD 团队的领导者和合作学者于 2019 年 4 月启动了 TRIPOD+AI 的开发，并于 2022 年 7 月最终确定了 TRIPOD+AI 的条目。其中的“+”表示其是以 TRIPOD 为基础，适用于统计回归模型或机器学习方法开发的预测模型的研究，同时为与现有涉及 AI 的研究报告指南保持一致，使用附加术语“AI”，但实际上支撑模型的为机器学习算法[5]。因此，为了便于阅读，笔者仍然称之为机器学习。

## 2　TRIPOD+AI 声明条目的解读

该声明适用于采用机器学习或传统回归方法，开发和（或）评估预测模型的研究报告，其核查清单涵盖了题目、摘要、前言、方法、开放科学、患者与公众参与、结果、讨论 8 个部分，共 27 个主条目、52 个子条目。另外，专门制定了 13 个条目的摘要核查清单。详细参见补充材料 1 和补充材料 2。

笔者以发表在 *European Journal of Heart Failure* 杂志的论文“Machine learning-based prediction of in-hospital death for patients with Takotsubo syndrome: the InterTAK-ML model”[6]为实例，在解释条目的基础上对例文进行分析。报告条目检查清单详见补充材料 1，以帮助读者更好地理解与应用 TRIPOD+AI 声明。

### 2.1　标题

条目 1：明确研究为开发和（或）评估多变量预测模型，以及适用的目标人群和所预测的结局。

解读：与 TRIPOD 相比，TRIPOD+AI 更强调研究的具体应用背景，包括是否涉及机器学习，同时明确目标人群、结局指标，以及研究是开发模型还是验证模型。因此，标题应使用清晰且专业的术语，使读者一目了然研究的核心内容和机器学习应用特点[7]。

例文的中文标题为“基于机器学习的 Takotsubo 综合征患者住院死亡风险预测：InterTAK-ML 模型”，明确了研究的核心内容，即通过机器学习模型预测 Takotsubo 综合征患者的住院死亡风险。标题清晰地反映了研究的目标人群——Takotsubo 综合征患者，以及预测的结局——患者在住院期间死亡。标题明确提及了机器学习方法，并通过模型名称“InterTAK-ML”进一步突出了研究的核心方法和成果，直接传达了研究的目标和方法。然而，标题未明确区分研究是专注于模型的开发、评估，还

是两者兼顾，但从研究目的和结果来看，应是模型开发研究。整体而言，标题使用了简洁的专业术语，基本符合该条目的要求。

### 2.2 摘要

条目 2：参见 TRIPOD+AI 获取摘要清单（具体见补充材料 2）。摘要是论文的精简概述，其目的是用简洁清晰的语言向读者传递研究过程及其核心发现。清单可以帮助研究人员确保摘要内容的全面性和结构化，主要包括：标题、背景、目的、方法、结果、讨论、资金和注册信息。

解读：首先是背景部分，需要简要说明所研究问题的医学背景和研究进展，以明确构建或评估预测模型的动机，及其在临床诊疗实践中的必要性。该部分报告应简洁充分，解释为什么这项研究至关重要，并为研究目的的引出奠定基础。其次是目的部分，应详细说明研究的具体目标，尤其要明确研究是专注于模型的开发和（或）评估，从而清晰明确地传递研究的性质和研究的重点。第三是方法部分，必须明确报告以下内容：① 数据的来源及纳入排除标准。例如数据的地理来源、样本大小、数据的收集方式等，并报告数据的质量控制措施，以确保预测模型的有效性和外推性；② 尤其是对于预后模型而言，需要简要描述模型预测的时间范围，以明确预测结果的意义和价值；③ 概述模型类型，无论使用的是回归模型、判别模型、深度学习模型，还是贝叶斯模型，均需简要描述[8]；④ 概述建模的关键步骤及内部验证方法（如交叉验证或留一法），以便让读者了解模型的鲁棒性和可靠性[9]。第四是结果部分，应至少包括 3 方面内容：① 研究对象及所研究结局事件的数量；② 最终纳入模型的预测因子，包括人口学变量、临床特征、健康信息等；③ 模型的预测性能，例如 ROC 曲线下面积（area under curve，AUC）值及其置信区间。第五是讨论部分，作者应从整体上解释研究结果的意义，并指出模型的实际应用场景或局限性，有助于帮助读者理解研究的影响以及未来可能的研究方向。第六是研究注册信息，包括注册号和注册数据库的名称，以确保研究的透明性和可追溯性。

例文摘要没有背景部分，因此没有在摘要提出所研究问题的医学背景和研究进展，但在目的、方法、结果及讨论部分，数据来源清晰，模型类型明确，建模步骤、验证方法到位，且研究对象、预测因子、模型性能指标报告完整。但在讨论的部分比较欠缺，没有指出模型可能存在的局限性，也缺乏对更进一步研究的建议，只提到了该模型的优越性。同时摘要部分也没有资金和注册信息。

### 2.3 前言

**2.3.1 背景** 条目 3：3a，阐述研究的医学背景（包括诊断或预后），以及开发或评估预测模型的理由，包括对既有模型的引用或参考；3b，描述目标人群和预测模型在临床路径中的预期用途，以及模型的预期使用者（如医疗保健专业人员、患者、公众）；3c，描述可能存在的健康不平等问题。

解读：相比 TRIPOD，TRIPOD+AI 更关注机器学习技术处理多源性、高维度、动态性医学大数据的优势，要求通过高效且适当的算法妥善解决相关问题，从而提升医疗预测模型的质量和实用性，为医疗领域提供更精准、更个性化的服务。因此，可以在背景中报告机器学习模型在处理本研究真实数据情境下的独特优势或价值[10]。

由于机器学习模型主要是通过“学习”源数据表现出来的特征完成模型构建，其预测性能和有效性很大程度上也取决于其应用的目标人群和具体应用场景与源数据是否适配。因此，必须从人口社会学特征、生活行为方式、疾病特征等方面对目标人群进行准确定义，并对应用场景进行精确划分，比如初次诊断、复发诊断、鉴别诊断，以及短期、长期预后等，以确保读者能够更好地了解模型的适用范围和泛化能力。

由于受到数据来源、质量、预测因子选择等诸多因素影响，常常导致应用于预测模型的源数据出现“偏向”特定群体的情况，而通用的模型预测性能评价指标往往也掩盖了模型在不同群体中的表现，相比于传统模型机器学习方法更容易出现健康不公平问题，因此要求作者要在模型的开发和评估阶段，增加偏倚检测与公平性评估的步骤，以避免模型对某些群体产生系统性误差[11]。

例文在背景部分，详细说明了 Takotsubo 综合征的疾病特性及其严重性与 InterTAK-ML 模型开发理由，并明确目标人群与模型的预期用途，包括模型的作用和预期使用者，模型是预测短期预后，且相比于以往常用的德国和意大利压力性心肌病（GEIST）评分系统，例文提出的新模型可以提供更精确的预测，减少传统方法的局限性，但是在该部分没有明确说明健康不平等问题。

**2.3.2 目的** 条目 4：明确研究的目标，并说明研究是否涉及模型的开发、评估，或者两者兼有。

解读：与 TRIPOD 在前言目的部分写作要求一致，TRIPOD+AI 需要作者清晰地陈述研究的具体目标是什么，是为了开发一个新模型，还是评估

既有模型的性能，还是同时进行模型的开发和评估。

例文在背景中明确地提到是进行模型的开发，模型评估未直接提及。

### 2.4 方法

**2.4.1** *数据*　条目 5：5a，分别描述用于模型开发和评估数据集的数据来源（例如随机试验、队列、常规治疗或注册研究数据）、使用这些数据的理由以及数据的代表性；5b，明确收集数据的关键日期，包括对象招募的开始、结束日期，以及随访结束日期（如果适用）。

解读：TRIPOD+AI 与 TRIPOD 一致，都要求详细描述数据的来源和收集时间，并解释选择利用这些数据的理由和数据的代表性，从而提供更多信息让读者能够判断模型时代背景，以及推广到更广泛的人群或临床实践中的可能性[11]。对于用于诊断的判别模型数据一般不涉及随访，而对于预后模型，随访时长无疑会对模型的开发和应用产生重要影响，必须对随访时长和预测的时间范围做明确界定。同时，需要强调的是，在机器学习中为了保证模型的预测准确性和泛化能力，模型开发和评估应在不同的数据集上进行，因此应分别进行说明。

例文中明确说明了使用了两个数据集，训练和内部验证队列均来自国际 Takotsubo 登记研究（International Takotsubo Registry，InterTAK），具有较强的国际代表性，招募时间为 2011 年到 2021 年。外部验证队列来源于 Takotsubo Italian Network，入组了 2007 年至 2018 年的患者，为独立的来源数据。没有提到随访的结束日期。

**2.4.2** *研究对象*　条目 6：6a，说明研究现场的基本特征（例如初级医疗机构、二级医疗机构、社区人群等），以及所选研究中心的数量和位置；6b，描述纳入研究对象的纳入和排除标准；6c，提供研究对象接受所有治疗或其他医学干预的详细情况，且需说明在模型开发或评估期间如何处理干预特征。

解读：TRIPOD+AI 在研究对象的描述上延续了 TRIPOD 的要求，包括明确研究对象来源人群，以及研究对象的纳入和排除标准，以便于读者评估数据的质量、对象代表性与结果的外推性[12]。同时，TRIPOD+AI 声明特别强调了对于治疗或其他干预措施的报告，这不仅是完整准确地报告数据特征的需要，以确保模型的可重复性和可比性。更重要的原因在于，干预可能通过改变数据特征，影响特征工程、挑战模型结构机器参数等多个方面，进一步影响模型的准确性和泛化能力。因此，在模型开发或评估期间治疗干预特征就显得尤为重要，是必须报告的内容。

例文说明了数据来源于 17 个国家、58 个心血管中心，但是未说明医疗机构层级和具体人群来源，提到以 InterTAK 诊断标准作为纳入依据，但没有列出具体的纳入排除标准。只提到了收集干预相关数据，但是没有明确说明具体干预特征，也未说明是否纳入模型或进行控制。

**2.4.3** *数据准备*　条目 7：详细描述所有数据预处理和质量核查的内容，并说明其在不同社会人口学特征群体中的一致性。

解读：TRIPOD+AI 要求详细说明数据预处理和质量检查流程，这一要求与机器学习模型的特点密切相关。机器学习数据质量的依赖度极高，数据噪声、缺失值或信息偏倚会直接影响模型的准确性和泛化能力[13]。因此，研究者需在模型开发前进行数据核查，包括验证数据是否合格（符合纳入和排除标准），缺失值、异常值、逻辑错误的识别与处理等。在合并不同来源的数据时，应确保数据完整性，包括数据格式标准化、键值匹配、映射规则等的明确定义。此外，数据的质量问题可能因人群特征不同而出现不同的表现，在数据准备阶段应保持核查策略和措施的一致性，以提高模型在不同人群的公平性和适用性[14]。

例文描述了缺失值处理、多重共线性分析等过程，提到数据通过标准化表格和临床记录审查收集，并剔除了高缺失变量，通过敏感性分析验证了模型在不同群体中的性能一致性。

**2.4.4** *结局指标*　条目 8：8a，明确定义模型预测的结局指标和时间范围，包括如何以及何时评估、选择该指标的理由，并解释评估方法在不同人群是否一致；8b，如果结局指标的测量需要主观评估，应描述评估者的资质和人口学特征；8c，报告实现盲法评估的所有措施。

解读：本条目中 8a 和 8c 延续了 TRIPOD 结局指标的要求，通过明确结局指标的定义、测量时间或时间窗口、测量方法等内容，以准确提供模型预测内容信息，如疾病状态、是否复发、手术需求或治疗效果等，研究者还需要解释评估方法在不同群体中的一致性[15]。以上措施共同保障结局指标评估的科学性和公平性，并通过盲法测量减少人为干扰。此外，TRIPOD+AI 新增了 8b，类似于病理诊断、影像分析等结局指标，其结果依赖主观判断时，应详细说明评估人员的专业背景和人口特征，包括学历、临床经验及专业领域，以确保他们有能力进行准确评估，并帮助读者判断不同特征评估者

是否存在潜在偏见的问题，提高研究的透明度和可靠性[16]。

例文明确说明模型预测的主要结局是住院死亡，时间范围是患者住院期间的死亡事件，也提到了选择该指标的理由，并讨论了指标在不同人群中的一致性。而“住院死亡”是客观指标，无需主观评估。数据来源于注册研究，流程较为标准化，有一定的盲法评估效果。

**2.4.5** 预测因子 条目 9：9a，描述初始预测因子的选择原因（例如参考相关文献、既往模型、数据的可用性），及选择过程；9b，明确定义所有预测因子，包括其测量方式和测量时间（以及实现盲法评估的所有措施）；9c，如果预测因子的测量需要主观评估，请说明预测因子评估者的资质和人口学特征。

解读：虽然机器学习理论上能够处理高维数据，但在实际应用中，如何从海量信息中筛选出关键特征，减少噪音特征干扰，同时缩短训练时间、减少过拟合风险，从而提高模型的性能和可解释性，进行预测因子选择是机器学习必要的步骤。其中，初始预测因子的选择可以基于文献或既往模型研究、专家意见，以及行业惯例等依据，也需要综合考虑数据获取、质量保证的难易程度等现实情况，即数据的可用性。确定最终纳入模型的预测因子阶段，在机器学习中称为特征选择，可以基于数据探索、统计学方法、过滤技术等，比如选择与结局变量相关系数更大的指标，或基于 LASSO 回归等筛选预测因子。另外，一些高级的机器学习算法（如随机森林、梯度提升树等）本身具有特征重要性评估功能，可以利用其结果选择重要的特征重新拟合模型。根据 TRIPOD+AI 的要求，以上内容均需要报告，以提高机器学习模型的透明性和可解释性。9b、9c 内容与结局指标报告要求基本一致，这里不再赘述。

例文中提到初始变量选择基于临床相关性、文献和数据可用性，并结合岭回归进行筛选；定义了变量及其测量方式，明确了采集时间；但预测因子大多为客观数据，主观评估需求较少，因此没有说明评估者的资质和人口学特征。

**2.4.6** 样本量 条目 10：解释研究样本量是如何确定的（分别针对模型开发和评估），包括所有样本量计算的细节，并论证研究中的样本量是否足以回答研究问题。

解读：相比于 TRIPOD 的要求，TRIPOD+AI 提出了关于样本量更详细的要求。这主要是因为在机器学习中，通常都需要大样本量支持，以保证模型的性能、泛化能力，并可以一定程度上避免模型过拟合，以及训练和评估模型时数据分布差异的问题。但样本量过大势必会增加数据清洗和预处理的难度，同时也会使模型训练时间显著增加，对计算资源需求大增。因此，在实际研究中仍需要根据研究目标、数据特征、模型复杂度，以及对模型性能的要求，合理确定样本量的大小。有关的样本量估计方法包括经验法、基于分布理论的统计估计方法，以及数学模拟等方法[17-18]。但目前大部分有关机器学习的文献，对样本量提及均较少。

例文中并没有明确说明样本量是如何确定的，也未提供计算细节。因此，样本量是否能充分回答研究问题存在一定的不确定性。

**2.4.7** 缺失值 条目 11：说明缺失数据处理方法，以及数据剔除的原因。

解读：TRIPOD+AI 未对缺失值处理方法提出特定要求，允许研究者根据具体情况选择适当方法。这主要是因为一些机器学习模型在处理缺失值方面更加灵活，比如决策树及其集成算法（如随机森林、梯度提升树等）、神经网络等，允许不对缺失值进行事先填补，而是在模型训练过程中基于数据增强或迁移学习等数据驱动策略，自动适应缺失值的存在，根据其他完整信息以及缺失值本身的分布特点拟合模型，也同时避免了缺失值处理方式不当造成模型拟合错误的问题，显示出比传统插补方法更有效的特点。同时，如果在分析过程中忽略或剔除了某些数据，需解释原因，以评估其合理性和公平性。

例文中提到对缺失值超过 30% 的变量直接剔除；在交叉验证过程中，使用计量资料的中位数和计数资料的众数进行插补缺失值，但未深入讨论剔除数据或变量的特性及其对模型的影响。

**2.4.8** 统计分析方法 条目 12：12a，描述数据的分析目的（如用于模型开发和性能评估），包括是否进行了数据集划分，并考虑样本量要求；12b，根据模型类型，描述预测因子在分析中的处理方式（如函数形式、重缩放、转换、标准化等）；12c，明确模型类型，解释模型选择理由，描述所有的模型构建步骤，包括超参数调整优化和内部验证方法；12d，描述不同来源（如医院、国家）的数据之间是否存在模型参数估计和性能评价中的异质性，并报告识别和处理方法，参考 TRIPOD-Cluster 声明的特别注意事项[19-20]；12e，明确定义研究中用于模型性能（如区分度、校准度、临床效用等）评价的指标和图表（以及选择理由），明确模型选择过程（如果

适用）；12f，描述在既有模型评估过程中是否进行了更新（如重新校准），包括模型整体层面的更新，或适用人口社会学群体或环境层面的更新；12g，对于模型评估，描述模型预测值是如何获得的［如公式、代码、对象、应用程序编程接口（application programming interface，API）等］。

解读：条目 12 包括 7 个子条目，其中 a、b、c 只针对模型开发研究；f、g 只针对模型评估研究；d、e 同时适用于两种不同目的的研究。研究者需要准确报告相关内容，以提高研究的透明度，为其他研究者复现模型奠定基础。

相比于传统预测模型的构建，机器学习模型在开发、评估时，特别注重模型的预测准确性（内部有效性），并避免过拟合以提升模型的应用泛化能力（外部有效性），因此扩展和细化了许多要求，尤其对于深度学习，数据集划分是标准步骤。根据分析目标不同，一般将数据集定义为了训练集（training set）和测试集（testing set）。其中，用于模型开发或训练的数据集称为训练集，在训练过程中用于对模型结构、超参数等进行调整优化的数据集为验证集（validation set）。而测试集是在模型开发完成后，用于最终评估模型性能的数据集。比如，可以采用静态留出法划分数据集，将原始数据集按照随机抽取的方式将 60% ~ 80% 的数据作为训练集，10% ~ 20% 的数据作为验证集，10% ~ 20% 的数据作为测试集。但该方法对数据的划分方式比较敏感，不同的划分方式可能得到不同结果，在模型训练阶段可以采取交叉验证法（cross validation）进行，包括留一法、K 折交叉验证等，其基本思想是通过多次动态划分训练集和验证集，将结果进行综合作为训练结果，尤其对于小样本数据而言，能够充分利用有限的数据进行学习和验证，以保证模型的训练结果和泛化能力，有效减低过拟合风险。

在分析中，预测因子的处理方式对预测模型开发至关重要，直接影响模型对数据的理解和利用能力，进而影响模型的预测性能。需要根据探索性分析的结果，明确预测因子纳入模型的函数形式。例如，数据中预测因子与目标变量之间存在非线性关系，选择非线性函数形式无疑更为恰当。当然，也可以通过数据转换将非线性关系转化为线性关系，以满足特定算法的需求、提高数据的可学习性，避免欠拟合问题。同时，为了统一不同特征的尺度，提升模型收敛速度，并便于数据理解与比较，一般需要对因子进行归一化、标准化等处理，这些都需要在方法部分进行报告。

无论是新模型开发还是既有模型的评估，预测模评估都是十分重要的内容，它不仅可以指导模型优化与选择，而且可以了解模型性能的优劣，以确保模型的泛化能力和实际应用中的可信度。若根据应用目标、模型类型不同，开发出了多种指标供选择，报告中应对相关指标进行明确定义，并说明选择依据。

传统模型通常没有分层评估的需求，而在机器学习中，特别关注了不同群体或来源的数据间模型参数和性能估计结果的异质性问题，需要研究者说明对其的识别过程和处理方式，以确保模型在不同人群、不同医院或地区等多样环境中的泛化能力。

针对既有模型的评估，要求报告预测值获取方法，需要提供有关预测值计算的公式、代码、API 等，以确保其透明性和评价的有效性[21]。

例文所选模型开发和内部验证集来源于 InterTAK Registry，样本按照 75%：25% 随机分配为训练集和内部验证集；外部测试集来自独立的 Takotsubo Italian Network。数据预处理采用重缩放和缺失值处理，未提到是否进行了转换或标准化。为了处理高维数据，且多变量之间存在共线性的问题，选择岭回归进行特征选择。例文详细描述了模型构建步骤及内部验证方法，也列出了模型性能评估的指标和选择理由，但未直接提到对数据异质性的处理方法。例文的研究未涉及既有模型的更新，因此不需要描述更新内容。明确了预测值生成基于岭回归或 logistic 回归模型，对输入变量和来源进行了详细描述。

**2.4.9　类别不平衡**　条目 13：如果使用了解决类别不平衡方法，应说明理由和具体方法，以及后续重新校准模型或预测结果的方法。

解读：类别不平衡是机器学习中的常见问题。例如，对于罕见病，当训练数据集中患者和非患者人数相差非常悬殊，造成患者数量过少，即出现了类别不平衡。如果不进行特殊处理，多数类样本会主导模型的学习优化过程，进而使模型在预测时更倾向于给出多数类的结果。因此 TRIPOD+AI 专门增加了对类别不平衡处理方法的要求，在数据层面可以采用过采样、欠采样，在算法层面采用加权损失函数或生成对抗网络等进行处理[22]。处理类别不平衡后，可能需要对模型进行校准，以确保不同类别的预测概率真实反映实际情况，常用的方法有 Platt scaling 或 Isotonic regression 等[23]。

在例文中并没有提到类别不平衡问题，因此也没有对应方法的应用。

**2.4.10** 公平性 条目 14：描述用于解决模型公平性问题的方法及其原理。

解读：如前所述，与传统预测模型相比，机器学习方法对数据本身的依赖性增强，在构建过程中更为复杂，因此保证模型在不同人群适用时的公平性就显得尤为重要。除了前面在数据收集、预处理、特征选择等方面的体现外，模型训练过程中，可以通过对不平衡群体进行特殊调整、加权损失函数等方法进行处理。在模型评估中，可以使用特定的公平性指标，如均衡误差率、差异影响、统计公平性等，或通过分层交叉验证亚组独立分析，比较不同群体的预测准确性、敏感性、特异性等指标，确保模型在不同特征群体表现的一致性或公平性[24]。

在例文中，并没有明确提到模型公平性问题及其解决方法。

**2.4.11** 模型输出结果 条目 15：明确预测模型结果形式（如分类或分类概率），提供分类的详细信息、分类依据，以及分类阈值的确定方法。

解读：当预测模型用于状态预测或分类任务时，模型一般是根据预测因子的取值，获得属于预定义分类类别的概率，进一步根据概率阈值标准，判断出相应的类别，作为模型预测结果。比如，根据对象特征利用模型获得有病的概率为 58.2%，根据概率>50% 判定为有病的阈值标准，则将对象判定为有病。条目要求报告确定分类阈值的依据，一般是根据应用场景，通过 ROC 曲线结合临床意义，通过最大化敏感性和（或）特异性来设置阈值[25]。

例文中以住院死亡作为二分类任务的目标变量，模型性能通过 AUC 指标进行评估。

**2.4.12** 训练与评估 条目 16：识别模型开发与评估数据集在医疗环境、入选标准、结果和预测因子方面的任何差异。

解读：机器学习模型与传统预测模型都要求详细描述模型开发过程中使用的数据，验证或评估时所使用数据之间的差异，并分析其对模型性能的潜在影响，以确保模型在实际应用中的可靠性和适用性。TRIPOD+AI 声明还增加了对医疗环境和入选标准方面的要求。主要是因为开发数据和评估数据可能是不同来源的数据，比如不同医院或不同的入选标准，如未充分考虑以上差异，可能会导致模型的预测效果显著下降，或者模型无法很好地适应新的数据集[26-27]。

例文中提到开发数据集与外部验证数据集来源不同，这说明研究开发的模型可以评估在不同地理和医疗环境下的泛化能力。两组数据集均为 InterTAK 诊断标准纳入患者，但未深入讨论两个数据集在预测因子分布和结局发生率上的差异及其影响。

**2.4.13** 伦理批准 条目 17：列出批准本研究的机构研究委员会或伦理委员会，并说明是否已获得研究对象的知情同意，或是否获得了伦理委员会的豁免许可。

解读：机器学习通常需要大量的患者数据来训练和验证模型，可能增大患者隐私和数据安全的敏感性。因此，TRIPOD+AI 强调对数据使用的伦理审批和知情同意，以保护数据来源的合法性和道德性，确保研究在合法合规的前提下进行。

例文中，在方法部分并没有提到知情同意及伦理审批。

### 2.5 开放科学

条目 18：18a，提供本研究的资金来源及资助方在本研究中的角色；18b，声明所有作者的利益冲突及财务披露情况；18c，说明研究方案的获取途径，或声明未制定研究方案；18d，提供研究的注册信息，包括注册名称和注册编号，或声明本研究未注册；18e，提供获取研究数据的详细信息；18f，提供获取分析代码的详细信息。

解读：条目 18 的 6 个子条目主要规定了需要公开的信息内容。相比传统预测模型，TRIPOD+AI 特别强调公开透明化，委员会鼓励研究者分享数据，以便其他研究者能够验证和重现结果。

在例文“提供资金”部分，作者明确说明了资金来源，并明确了资助方在研究中的角色，声明了无利益冲突，但没有提供研究注册信息，也未声明未注册。同时也没有提供代码获取的途径，虽然研究使用了国际性数据集，但未明确说明数据是否公开，也没有提供数据获取的方式。

### 2.6 患者和公众参与

条目 19：提供在研究设计、实施、报告、解释或传播过程中，患者和公众参与情况的详细信息，或声明无相关参与。

解读：该条目体现了研究是否充分考虑到患者和公众的观点、需求和期望，即从对象的角度，考虑模型的实用性、可操作性和可读性，从而提高医学研究的质量和影响力。这是对传统 TRIPOD 标准的一个重要补充。

例文中未提及患者和公众的参与情况。

### 2.7 结果

**2.7.1** 研究对象 条目 20：20a，描述研究过程中研究对象的变动情况，包括出现和未出结局事件的

人数。对于随访性研究，还需提供随访时间的概要。使用图表形式可能会使表达更清晰。20b，报告对象的整体特征，如果可能应报告不同来源或现场对象的特征，包括关键日期、关键预测因子（含人口学特征）、接受治疗、样本大小、结局事件数量、随访时间和数据缺失量。建议以表格形式报告。报告不同关键人口学特征对象间的差异。20c，在模型评估中，展示与开发数据中关键预测相关变量（人口学特征、预测因子和结局指标）分布的比较结果。

解读：声明特别建议研究者以变动流程图的形式总结报告对象招募、排除、剔除的情况，并同时报告出现与不出现结局事件的人数。采用表格形式报告对象或数据的分布情况，并建议对于不同来源、医院或研究中心的数据，进行分别报告，以及报告不同特征人群在预测因子方面的差异，主要目的仍然是增加研究的透明性，清晰地呈现出数据全貌，从而衡量数据的代表性和质量状况，也提示研究者和读者可能存在的类别不平衡等问题，为后续模型构建策略和评估奠定基础，进而提高模型的公平性和可靠性。

对于模型评估而言，通过比较模型开发数据集和评估（或测试）数据集的分布，可以了解模型在从开发数据集到评估（或测试）数据集的转换过程中，是否能够适应不同的数据分布情况，从而评估模型的泛化能力。如果两个数据集的分布差异过大，可能意味着模型在新的数据分布下性能会受到影响，无法很好地对未知数据进行准确预测。

例文未涉及长期随访，因此没有相关的随访描述，但详细列出了出现和未出现结局事件的人数。文章通过表格全面报告了患者的整体特征，包括人口学信息、预测因子和治疗情况，明确了结局事件人数、总体患者数量以及住院死亡率。文章还提到剔除了缺失率较高的变量，但未在表格中单独列出缺失情况。此外，未区分开发集和外部验证集患者的详细特征对比。研究还提供了外部验证的性能评估，并指出预测因子的来源一致。

**2.7.2　模型开发**　条目 21：明确说明各分析任务（如模型开发、超参数调整、模型评估）中研究对象和结局事件的数量。

解读：该条目是 TRIPOD+AI 的一个独有条目，其与对象中对样本量的要求是基本一致的。研究对象和结局事件的数量直接关系到模型所能学习到的信息丰富程度。一般来说，较大的样本量能让模型接触到更多的数据模式和特征关系，有助于提高模型的准确性和稳定性。在超参数调整过程中，研究对象的数量会影响超参数的选择和调整策略。例如，在一个小样本的机器学习任务中，对于决策树模型的树深度这一超参数，由于样本量不足，可能无法准确判断不同树深度设置下模型的真实性能差异，从而难以选出最佳的超参数值。在模型评估阶段，样本量或结局数量过小可能导致评价结果不稳定，会影响对模型可靠性和泛化能力的判断。因此，要求作者完整详细地报告各分析阶段的样本量，可以更全面地理解模型开发的背景，帮助读者评估模型的表现和合理性，也为其他研究者进行验证和进一步研究提供了基础信息[28]。

例文对模型开发阶段的样本量和结局事件信息有一定描述，提到使用交叉验证方法进行了超参数调整；提供了外部验证数据集和模型性能评估指标。

**2.7.3　模型定义**　条目 22：提供完整预测模型的详细信息（例如公式、代码、对象、API），以便进行新个体预测和第三方评估、使用，包括关于获取或重复使用的限制条件（例如可免费获取、专有等）。

解读：机器学习模型常伴随有复杂的算法和大量的参数，其公式、代码、对象、应用程序接口是理解和应用模型的基础，如果只提供简要描述，很难全面理解其内部机制。公开模型细节，如代码和 API，可以让研究者和用户更透明地了解模型的设计和预测流程，也促进了模型的传播和应用。这也和前面的条目 18 呼应。

例文没有提供完整的预测模型详细信息（如公式、代码或 API），也未说明获取模型的方式或限制条件。

**2.7.4　模型性能**　条目 23：23a，报告模型性能评价指标的估计值及其置信区间，包括在关键亚组（如社会人口学特征）中的表现。可以考虑采用图表形式进行展示。23b，如有评估，报告模型在不同人群间性能的差异，参见 TRIPOD-Cluster[20]。

解读：如方法中叙及的，模型的性能评价无论对于模型开发和评估都是十分重要的内容，应详细报告模型性能评价指标的估计值及其置信区间。作者可参考 TRIPOD-Cluster 报告规范针对不同亚型人群进行模型性能评价，并进行适当的假设检验，如 *t* 检验、ANOVA 等，比较不同亚组模型性能的差异是否具有统计学意义[29]。基于统计结果，解释和讨论模型性能的差异来源，说明可能的异质性原因及其对模型应用的影响，如数据质量、群体特征和预测因子与结局之间的关系，并提出改进模型

性能或应对异质性的建议。

例文提供了模型的主要性能指标（AUC、敏感性、特异性）及其置信区间，并通过外部验证评估了模型在不同人群中的泛化能力。

**2.7.5** 模型更新 条目 24：如果模型有更新，请报告所有更新结果，包括更新后的模型及其性能。

解读：随着新数据的积累，原有模型在某些特定人群或情境下的预测性能可能下降，因此模型必须不断更新调整以提高自身性能。机器学习模型与传统模型一样，必须报告每次模型更新的结果，包括更新后的模型和后续性能。作者应提供更新后模型的详细信息，包括模型的输入、输出、所有中间层和连接，以及任何新添加的预测因子或调整后的各项参数。作者还应报告更新后模型的性能表现，包括区分度、校准度以及其他相关性能指标，并指出未来研究的方向和局限性。

例文中没有提到模型更新，因此并没有提到模型更新的内容。

## 2.8 讨论

**2.8.1** 解读 条目 25：对主要结果进行整体解读，包括本研究目的，及在已有研究的回顾中讨论公平性问题。

解读：TRIPOD+AI 要求研究者在讨论部分对主要结果进行总体解读，相比于传统模型，机器学习模型必须考虑公平性问题。作者需要阐述主要的研究发现并引用关键数据、图表或统计结果以提供支持，讨论研究结果是否达到了预期的研究目的，将本研究的结果与以往相关研究进行对比，分析本研究的独特贡献和创新点。解读结果时，特别需要关注公平性问题，通过探讨样本是否出现选择偏倚、研究结果是否适用于所有相关人群等问题，从而提高研究的普适性和应用价值[30]。

例文在讨论部分，提到文章缺乏全面的种族数据，该模型在欧洲和亚洲以外人群中的适用性有限。同时对纳入研究的变量进行了简化，可能会影响不同群体间的公平性。

**2.8.2** 局限性 条目 26：讨论该研究所存在的局限性（如样本缺乏代表性、样本大小、过拟合、缺失数据）及其所引起的偏倚、统计不确定性和外推性的影响。

解读：TRIPOD+AI 鼓励研究者像构建传统预测模型一样，详细讨论研究中的各种局限性，并分析这些局限性可能对研究结果带来的偏倚、不确定性以及对可推广性的影响。机器学习模型的局限性主要在于数据质量与偏倚、模型解释性、外部适应性、临床整合难度、伦理和隐私问题[31-32]。作者应客观、全面地探讨这些局限性，以帮助读者评估研究结果的可信度、适用性和可推广性。

例文“讨论”部分提到了研究设计的固有限制、种族多样性的局限性、变量选择的局限性、时间因素的局限等。

**2.8.3** 模型适用性 条目 27：27a，描述在应用预测模型时，如何评估和处理低质量或不可得数据（如预测因子数据）；27b，明确用户在处理输入数据或使用模型时是否需要进行交流合作，以及需要具备的专业知识水平；27c，讨论下一步研究的方向和计划，重点关注模型的适用性和可推广性。

解读：相比传统预测模型，机器学习模型在实际应用中对实施指导的要求更为严格。由于现实情况下输入数据可能因各种原因而质量差或不可用，因此 TRIPOD+AI 要求研究者在报告中详细描述如何评估和处理这类数据，以确保模型的准确性和可靠性。作者应具体说明评估输入数据质量所用的标准和方法，以及对质量差或缺失数据的处理策略，如数据插补、数据修正或删除，并讨论这些方法可能对模型拟合带来的影响[33]。此外，TRIPOD+AI 要求研究者明确用户在使用模型时是否需要进行数据处理，并指出所需的专业知识水平，以确保模型被正确、有效地应用。

在例文中，模型采用了简化设计，仅依赖 10 个最相关变量，从而在数据不完整的情况下仍能保持较高的实用性和可靠性。此外，论文提到，InterTAK-ML 模型被开发为一种用户友好的在线工具，用户可以通过输入简单的变量实现操作，降低了使用门槛。对于未来的研究方向，论文提出了进一步验证模型适用性和推广性的必要性，同时强调探索模型在不同医疗环境中的表现，以确保其在广泛临床场景中的实用价值。

## 2.9 TRIPOD+AI 清单的使用

TRIPOD+AI 声明作为 TRIPOD 的更新版，其检查清单将完全取代 TRIPOD（2015）。项目组极力倡导研究者在论文撰写伊始便运用该清单，以明晰相关内容，并着手准备相关细节。

研究人员需要下载 TRIPOD+AI 清单文件，可从官方站点（https:// www.tripod-statement.org/）下载或从补充材料 1 获取，并对照清单中的每一条目逐一进行检查，并进行标记，以便于编辑或同行评审快速定位，提高评审效率。如果报告中有对应条目的内容，应标明页码或具体位置。如果条目不适用于本研究，需在清单中填写“NA”（not applicable,

不适用），并解释不适用或无法提供的原因。如果由于文章篇幅限制，难以在正文中全部呈现，比如有关模型性能的指标或图表、数据分析代码或数据集说明等内容，可放入补充材料，并在主文中引用。此外，建议提供开放科学声明，明确数据和代码的共享情况（如通过公开链接访问），以支持透明性和可重复性。完成报告后，可将清单作为附录提交。

TRIPOD+AI 项目组一再强调，该声明仅用于提升预测模型研究报告的透明性和质量，不是质量评估工具。同时，TRIPOD+AI 检查表中的大部分条目虽然均考虑论文的自然顺序，但有些条目是基于其特殊考虑进行的顺序安排。因此，该清单并不推荐结构化格式，具体顺序应取决于预测模型及目标刊物的格式要求。

## 3　小结

TRIPOD+AI 作为针对机器学习预测模型的报告规范，系统涵盖了从模型开发到评估的全过程，要求研究者明确数据来源、变量选择依据、数据预处理步骤，并对模型性能进行全面评估。它并不是一个质量评估工具，而是为研究者提供最低限度报告要求的规范，旨在确保预测模型研究的透明性和完整性。相比传统预测模型的报告要求，TRIPOD+AI 增加了对公平性分析、偏差校正和结果透明化的细化要求，特别是在评估模型泛化能力时，需明确外部验证的流程和结果，并分析不同人群中的表现差异。但 TRIPOD+AI 条目众多，涵盖了模型开发、评估、预测因子选择、结果分析等多个方面，直接应用可能对研究者存在一定挑战。为帮助国内研究者更好地理解和使用这一指南，笔者翻译了相关条目，并结合具体文章进行解读。这一工作将促进国内医学 AI 研究的规范化发展，并为临床实践提供更可靠、更透明的支持工具。

**重要声明**

利益冲突声明：本文全体作者阅读并理解了《中国普外基础与临床杂志》的政策声明，我们没有相互竞争的利益。

作者贡献声明：粟文和赖泽鹏共同负责原始英文论文的翻译与解读，撰写初稿，并对文中的关键内容进行深入分析与阐释。两位作者对本研究的整体框架设计和具体内容贡献均等。蒋昊林对翻译内容和解读的准确性进行了全面核对，并对文章结构和语言表达进行了优化，为最终稿的完成提供了重要支持。陈卫中和曾子倩负责全程指导，包括研究方向的选择、学术规范的把控，以及论文最终版本的审阅与修改。

**参考文献**

1　Au EH, Francis A, Bernier-Jean A, *et al*. Prediction modeling-part 1: regression modeling. Kidney Int, 2020, 97(5): 877-884.
2　Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 542(7639): 115-118.
3　Deo RC. Machine learning in medicine. Circulation, 2015, 132(20): 1920-1930.
4　Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ, 2015, 350: g7594. doi: 10.1136/bmj.g7594.
5　Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 2024, 385: e078378. doi: 10.1136/bmj-2023-078378.
6　De Filippo O, Cammann VL, Pancotti C, *et al*. Machine learning-based prediction of in-hospital death for patients with takotsubo syndrome: the InterTAK-ML model. Eur J Heart Fail, 2023, 25(12): 2299-2311.
7　曹煜隆, 单娇, 龚志忠, 等. 个体预后与诊断预测模型研究报告规范——TRIPOD 声明解读. 中国循证医学杂志, 2020, 20(4): 492-496.
8　Van Calster B, McLernon DJ, van Smeden M, *et al*. Calibration: the Achilles heel of predictive analytics. BMC Med, 2019, 17(1): 230. doi: 10.1186/s12916-019-1466-7.
9　Jamarani A, Haddadi S, Sarvizadeh R, *et al*. Big data and predictive analytics: a systematic review of applications. Artificial Intelligence Review, 2024, 57(7):.
10　Adler-Milstein J, Chen JH, Dhaliwal G. Next-generation artificial intelligence for diagnosis: from predicting diagnostic labels to "wayfinding". JAMA, 2021, 326(24): 2467-2468.
11　Smith BT, Smith PM, Harper S, *et al*. Reducing social inequalities in health: the role of simulation modelling in chronic disease epidemiology to evaluate the impact of population health interventions. J Epidemiol Community Health, 2014, 68(4): 384-389.
12　Yang C, Kors JA, Ioannou S, *et al*. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc, 2022, 29(5): 983-989.
13　Bhandari N, Walambe R, Kotecha K, *et al*. A comprehensive survey on computational learning methods for analysis of gene expression data. Front Mol Biosci, 2022, 9: 907150. doi: 10.3389/fmolb.2022.907150.
14　Chen P, Wu L, Lei Wang. AI fairness in data management and analytics: a review on challenges, methodologies and applications. Appl Sci, 2023, 13(18): 10258. doi: 10.3390/app131810258.
15　Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol, 2022, 22(1): 287. doi: 10.1186/s12874-022-01768-6.
16　Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. Sci, 2024, 6(1): 3. doi: 10.3390/sci6010003.
17　Christodoulou E, van Smeden M, Edlinger M, *et al*. Adaptive sample size determination for the development of clinical prediction models. Diagn Progn Res, 2021, 5(1): 6. doi: 10.1186/s41512-021-00096-5.

18 Ng W, Minasny B, Mendes WDS, *et al*. The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. SOIL, 6: 565-578. https://doi.org/10.5194/soil-6-565-2020, 2020.
19 Debray TPA, Collins GS, Riley RD, *et al*. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. BMJ, 2023, 380: e071058. doi: 10.1136/bmj-2022-071058.
20 陶立元, 刘珏. 基于多源数据的个体预后或诊断多因素预测模型报告规范 (TRIPOD-Cluster) 解读. 中华医学杂志, 2023, 103(36): 2893-2897.
21 Gerds TA, Kattan MW. Medical risk prediction. New York: Chapman and Hall/CRC, 2021: 312.
22 Cusworth S, Gkoutos GV, Acharjee A. A novel generative adversarial networks modelling for the class imbalance problem in high dimensional omics data. BMC Med Inform Decis Mak, 2024, 24(1): 90. doi: 10.1186/s12911-024-02487-2.
23 Rajaraman S, Ganesan P, Antani S. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS One, 2022, 17(1): e0262838. doi: 10.1371/journal.pone.0262838.
24 Liu S, Vicente LN. Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. Comput Manag Sci, 2022, 19: 513-537.
25 Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation, 2007, 115(5): 654-657.
26 Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med, 2016, 375(13): 1216-1219.
27 Finlayson SG, Subbaswamy A, Singh K, *et al*. The clinician and dataset shift in artificial intelligence. N Engl J Med, 2021, 385(3): 283-286.
28 Probst P, Wright M, Boulesteix AL. Hyperparameters and tuning strategies for random forest. WIREs Data Mining Knowl Discov, 2018, 9.
29 Talic S, Shah S, Wild H, *et al*. Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. BMJ, 2021, 375: e068302.
30 Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat Med, 2020, 26(1): 16-17.
31 Mazzolenis ME, Bulat E, Schatman ME, *et al*. The ethical stewardship of artificial intelligence in chronic pain and headache: a narrative review. Curr Pain Headache Rep, 2024, 28(8): 785-792.
32 Gil-Fuster E, Eisert J, Bravo-Prieto C. Understanding quantum machine learning also requires rethinking generalization. Nat Commun, 2024, 15(1): 2277. doi: 10.1038/s41467-024-45882-z.
33 de Jong J, Emon MA, Wu P, *et al*. Deep learning for clustering of multivariate clinical patient trajectories with missing values. Gigascience, 2019, 8(11): giz134. doi: 10.1093/gigascience/giz134.

收稿日期：2024-11-25 修回日期：2024-12-23
本文编辑：罗云梅