DOI: 10.16781/j.CN31-2187/R.20240775

·论　著·

多阅片者多病例设计在人工智能辅助阅片影像诊断试验评价中的应用

宛慧琴[1]，向　蔓[1]，潘喆敏[1]，秦婴逸[2]，何　倩[2]，贺　佳[1,2*]
1. 同济大学医学院，上海 200092
2. 海军军医大学（第二军医大学）卫生勤务学系军队卫生统计学教研室，上海 200433

［**摘要**］ **目的**　使用多阅片者多病例（MRMC）设计评价人工智能（AI）辅助阅片影像诊断试验的临床效能，以期为影像诊断试验的临床评价提供科学依据。**方法**　采用影像诊断试验中广泛应用的 MRMC 设计，详细阐述了 MRMC 设计中 Obuchowski-Rockette（OR）法的模型构建及其检验方法。实例研究共收集了 3 家医院 200 例受试者的 CT 影像资料，其中 133 例为肋骨骨折患者，68 例为非肋骨骨折患者，由 3 位阅片医师对所有 CT 影像进行判读。分析在 2 种阅片方式（医师＋AI 辅助阅片、医师独立阅片）下肋骨骨折检出的 AUC 值、灵敏度和特异度的差异。**结果**　AI 辅助阅片组的 AUC 值为 0.958，医师独立阅片组的 AUC 值为 0.902，两组 AUC 值差异有统计学意义（$P<0.001$）。AI 辅助阅片组总体的灵敏度为 0.970，特异度为 0.946；医师独立阅片组的灵敏度为 0.838，特异度为 0.966；两组灵敏度差值为 0.131（95% *CI* 0.091～0.171），特异度差值为－0.020（95% *CI* －0.059～0.020），说明 AI 辅助阅片与医师独立阅片的灵敏度差异有统计学意义而特异度差异无统计学意义。两组的阳性似然比均大于 10，阴性似然比均小于 0.2，阳性预测值都接近 1，说明 AI 辅助阅片影像诊断试验的诊断准确性高。**结论**　AI 辅助阅片在提高诊断效能方面有显著优势，不仅可以提高肋骨骨折诊断的准确性和检出率，还能提高医师工作效率，优化医院服务。

［**关键词**］ 人工智能；多阅片者多病例设计；Obuchowski-Rockette 法；肋骨骨折；诊断准确性
［**引用本文**］ 宛慧琴，向蔓，潘喆敏，等. 多阅片者多病例设计在人工智能辅助阅片影像诊断试验评价中的应用［J］. 海军军医大学学报，2025，46（4）：504-510. DOI: 10.16781/j.CN31-2187/R.20240775.

Application of multi-reader multi-case design in evaluating artificial intelligence-assisted imaging diagnostic trials

WAN Huiqin[1], XIANG Man[1], PAN Zhemin[1], QIN Yingyi[2], HE Qian[2], HE Jia[1,2*]
1. School of Medicine, Tongji University, Shanghai 200092, China
2. Department of Military Health Statistics, Faculty of Medical Services, Naval Medical University (Second Military Medical University), Shanghai 200433, China

［**Abstract**］ **Objective**　To evaluate the clinical efficacy of artificial intelligence (AI)-assisted imaging diagnostic trials using multi-reader multi-case (MRMC) design, so as to provide a scientific basis for clinical evaluation of imaging diagnostic trials. **Methods**　The MRMC design, widely used in imaging diagnostic trials, was adopted in this study. The Obuchowski-Rockette (OR) method of MRMC design was detailed, including model construction and test methods. A case study was conducted, collecting imaging data of 200 subjects from 3 hospitals, with 133 cases of rib fractures and 68 cases of non-rib fractures. Three radiologists reviewed all CT images of the subjects. The area under curve (AUC) value, sensitivity and specificity in detecting rib fractures between 2 reading modalities (radiologists with AI assistance vs radiologists reading independently) were compared. **Results**　The AI-assisted reading group had an AUC value of 0.958, while the radiologist-independent reading group had an AUC value of 0.902, showing a significant difference ($P<0.001$). The overall sensitivity and specificity of the AI-assisted reading group were 0.970 and 0.946, respectively; while the sensitivity and specificity of the radiologist-independent reading group were 0.838 and 0.966, respectively. The difference of sensitivity between groups was 0.131 (95% confidence interval［*CI*］0.091-0.171), and the difference of specificity was －0.020 (95% *CI* －0.059-0.020), indicating a significant difference in sensitivity but not in specificity between AI-assisted and radiologist-independent reading groups. Both groups had positive likelihood ratios (＋LR) greater than 10 and negative likelihood ratios (－LR) less than 0.2, with positive predictive values approaching 1, suggesting that the diagnostic accuracy of the AI-assisted imaging diagnostic trials was high. **Conclusion**　The AI-assisted reading method demonstrates a significant advantage in enhancing diagnostic

［**收稿日期**］ 2024-11-14　　［**接受日期**］ 2024-12-26
［**基金项目**］ 上海市卫生健康委员会新兴交叉领域研究专项（2022JC011）. Supported by Emerging Interdisciplinary Research Project of Shanghai Municipal Health Commission (2022JC011).
［**作者简介**］ 宛慧琴，硕士生. E-mail: 2849493602@qq.com
*通信作者（Corresponding author）. Tel: 021-81871441, E-mail: hejia63@yeah.net

efficiency, not only improving the diagnostic accuracy and detection rate of rib fractures, but also improving the work efficiency of radiologists and optimizing hospital services.

［**Key words**］ artificial intelligence; multi-reader multi-case design; Obuchowski-Rockette method; rib fractures; diagnostic accuracy

［**Citation**］ WAN H, XIANG M, PAN Z, et al. Application of multi-reader multi-case design in evaluating artificial intelligence-assisted imaging diagnostic trials[J]. Acad J Naval Med Univ, 2025, 46(4): 504-510. DOI: 10.16781/j.CN31-2187/R.20240775.

人工智能（artificial intelligence，AI）在医学各个领域的应用需求日益攀升，尤其是在医学影像学领域，有研究表明AI技术对肺栓塞[1]、皮肤癌[2]等疾病的检测效能可以达到临床医师水平。在传统医学实践中，影像学图片主要依靠医师的个人经验和解释，不可避免会受到医师技术水平和经验认识等多方面因素的影响[3-5]。AI技术的出现可以解决这一难题，其可以在输入影像图片后几乎同时输出报告结果，并且不存在时空和医师压力疲劳[6]等局限，可以降低漏诊的可能性。同时，AI还能精准定位并标注病灶位置，进一步提升了诊断的精度[7-8]。

研究者通常在影像诊断试验中使用多阅片者多病例（multi-reader multi-case，MRMC）设计来评价试验的诊断准确性[9]，这也是美国FDA推荐使用的评价计算机辅助设计临床诊断的有效方法[10-11]。我国食品药品监督管理局也分别在2019年和2021年发布的《深度学习辅助决策医疗器械软件审评要点》[12]和《乳腺X射线系统注册技术审查指导原则》[13]中推荐使用MRMC设计。MRMC设计是在影像诊断试验中由多名阅片者对研究中多名病例的影像图片进行阅读，即每个阅片者会阅读每个病例的每一张影像图片，每例患者会有多个阅片结果。与传统诊断试验方法相比，MRMC设计由于考虑了复杂的相关结构，有更强的检验效能，同时检验方法也有别于传统的t检验而需要专门的统计学分析方法。

本文将详细介绍MRMC设计中的Obuchowski-Rockette（OR）法，包括其原理和计算方法，并结合实际的肋骨骨折案例评估AI辅助阅片系统的性能，旨在为影像诊断试验的评价提供参考。

1　MRMC设计的统计学分析方法

1.1　OR法模型构建　OR法是MRMC设计常用的检验统计量的分析方法。OR法在设计中纳入了MRMC设计的相关性[14]，构建了包括相关结构的两因素混合效应方差分析模型，即

$$\hat{\theta}_{ij}=\mu+\tau_i+R_j+(\tau R)_{ij}+\varepsilon_{ij};\ i=1,\cdots,I;j=1,\cdots,J \quad (1)$$

其中θ_{ij}为第i个诊断方法第j名阅片者的诊断试验准确度估计值，分量$\mu+\tau_i$为诊断方法i的平均准确度，μ为常数，τ_i为诊断方法i的固定效应，R_j为阅片者j的随机效应。R_j、$(\tau R)_{ij}$均为随机变量且服从正态分布，均值为0；方差分别为σ_R^2、$\sigma_{\tau R}^2$，分别反映阅片者之间、阅片者和诊断方法之间的交互作用。误差项ε_{ij}不独立，服从多元正态分布。

假设阅片者和诊断方法之间的误差相等，在不同的阅片者和诊断方法的情况下等协方差，则产生以下协方差

$$Cov(\varepsilon_{ij},\varepsilon_{i'j'})=\begin{cases}\sigma_\varepsilon^2 & i=i',j=j'\\ Cov_1 & i\neq i',j=j'\\ Cov_2 & i=i',j\neq j'\\ Cov_3 & i\neq i',j\neq j'\end{cases} \quad (2)$$

其中Cov_1表示使用不同诊断方法时相同阅片者的诊断准确性指标的协方差，Cov_2表示使用相同诊断方法时不同阅片者的诊断准确性指标的协方差，Cov_3表示使用不同诊断方法时不同阅片者的诊断准确性指标的协方差。Obuchowski和Rockette[14]建议对协方差排序为$Cov_1\geqslant Cov_2\geqslant Cov_3\geqslant 0$，用总体相关性进行表示则可得到$r_i=Cov_i/\sigma_\varepsilon^2$，$i=1,2,3$。由于在实际临床试验中通常只进行1次试验，无法直接计算协方差矩阵，可以使用Bootstrap[15]、Jackknife[16]和DeLong方法[17]对数据进行重抽样，从而估计协方差矩阵。使用这3种协方差估计方法在计算诊断准确性指标时得到的估计值一致，但估计值的置信区间存在细微的差异。

1.2　OR法假设检验　OR法的零假设（null hypothesis）是各诊断方法的效应相等，即H_0:$\tau_1=\cdots\tau_i$，备择假设为H_1:$\tau_1\neq\cdots\tau_i$。Pavur和Nath[18]于1984年提出的校正F统计量方法，为MRMC设计OR法的数据分析提供了重要的校正机制。他们校正了F统计量的系数，纳入了数据中的相关结构，

保证了统计分析的准确性。同时，优化后 F 统计量的自由度与优化前 F 统计量的自由度保持一致，即

$$F = \frac{MS(T)}{MS(T*R)+J(Cov_2-Cov_3)} \quad (3)$$

其中 $MS(T) = \frac{J}{I-1}\sum_{i=1}^{I}(\hat{\theta}_{i\cdot} - \hat{\theta}_{\cdot\cdot})^2$；

$MS(T*R)=\frac{1}{(I-1)(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\theta}_{ij} - \hat{\theta}_{i\cdot} - \hat{\theta}_{\cdot j} + \hat{\theta}_{\cdot\cdot})^2$；

下标被点替换表示是缺失下标的平均值。由于在实践中，一般无法获得 Cov_2 和 Cov_3 的准确值，用估计值或预测值来代替，式（3）则表示为

$$F^* = \frac{MS(T)}{MS(T*R)+J(\widehat{Cov}_2-\widehat{Cov}_3)} \quad (4)$$

1.3　OR 法的优化　由于原始的 OR 法存在诸如过于保守、误差项不独立等问题，Hillis 等[19]对 OR 法的检验统计量和自由度都进行了修正，得到

$$F_{OR}=\frac{MS(T)}{MS(T*R)+Jmax(\widehat{Cov}_2-\widehat{Cov}_3, 0)} \quad (5)$$

$$ddf_H=\frac{[MS(T*R)+Jmax(\widehat{Cov}_2-\widehat{Cov}_3, 0)]^2}{[MS(T*R)]^2/[(I-1)(J-1)]} \quad (6)$$

其中 $MS(T)$ 是诊断方法的均方，$MS(T*R)$ 是诊断方法和阅片者交互作用的均方，$\widehat{Cov}_2$ 和 $\widehat{Cov}_3$ 是 Cov_2 和 Cov_3 的估计值。考虑到 $\widehat{Cov}_2-\widehat{Cov}_3$ 的结果可能为负值，Hillis 等[19]添加了当其为负数时设置为 0 的 max 项，避免了检验统计量的分母可能为 0 的情况，也使检验统计量变大，相应的 P 值将会变小，使得 OR 法克服了过于保守的可能。

用 θ_i 表示诊断方法 i 的预期准确度指标，则当只有两组进行比较时，诊断方法的组间差值 $\theta_i-\theta_{i'}$（$i\neq i'$）的近似 100（1−α）置信区间为

$$\hat{\theta}_{i\cdot} - \hat{\theta}_{i'\cdot} \pm t_{\alpha/2}; ddf_H\sqrt{\frac{2}{J}MS_{den_{OR}}} \quad (7)$$

其中 $MS_{den_{OR}}$ 是式（5）右侧的分母，则式（7）写为

$$\hat{\theta}_{i\cdot} - \hat{\theta}_{i'\cdot} \pm t_{\alpha/2};$$

$$ddf_H\sqrt{\frac{2}{J}[MS(T*R)+Jmax(\widehat{Cov}_2-\widehat{Cov}_3, 0)]} \quad (8)$$

如果不满足零假设成立的条件，则 F_{OR} 服从 $F_{t-1,\ ddf;\ \lambda}$ 的非参数分布，非中心参数 λ 表示为

$$\lambda=\frac{J\sum_{i=1}^{I}(\theta_i - \theta_{\cdot})^2}{\sigma_{tR}^2+\sigma_{\varepsilon}^2-Cov_1+(J-1)(Cov_2-Cov_3)} \quad (9)$$

2　实例研究

2.1　资料和方法

2.1.1　一般资料　本研究为回顾性设计，筛选并纳入了 3 家医院（北京医院、南京市第一医院、广州医科大学附属第一医院）收集到的 200 例受试者 2021 年 4−12 月拍摄的胸部 CT 影像资料，其中阳性受试者 132 例，阴性受试者 68 例；男 107 例，女 93 例，年龄为 18～92 岁，平均（56.83±16.59）岁。CT 影像的纳入标准：（1）患者年龄≥18 岁，性别不限；（2）影像重建层厚＜2 mm；（3）符合医学影像 DICOM 3.0 标准连续采集序列，且未经压缩、格式转换等处理的胸部影像。影像的排除标准：（1）胸部或背部有金属植入物的影像；（2）患者有肋骨病理性骨折、自发性骨折或肋骨其他非骨折性病变；（3）研究者认为不适合入组的影像。

2.1.2　研究方法　本研究依据权威专家组的评审结果确立了肋骨骨折诊断的金标准。专家组由 3 名职称为主治医师及以上的医师组成，均具有 10 年及以上的 CT 骨折影像学诊断经验。评审过程中，2 位医师对肋骨骨折进行了独立评估，并在识别出肋骨骨折时于影像上进行标注。评估结束后，对这 2 位医师的评估结果进行了比对，若两者结果一致，则将其定为金标准；若 2 位医师的评估结果存在分歧，则由第 3 位医师进行仲裁，以确立金标准。

3 名具有骨折阅片经验的低年资研究医师先随机安排，再确定是先进行 AI 辅助阅片还是先进行独立阅片，但在 AI 辅助阅片或医师独立阅片后都需要 30 d 的洗脱期。每位医师在 2 个阶段进行阅片的影像顺序不同，3 名医师在阅读同一批影像图片时的影像顺序也不同，以此达到降低偏倚的目的。为避免专家组、研究医师和 AI 阅片结果的相互影响，采用盲态阅片方式并去除患者敏感信息后，再将影像交与专家组、研究医师、AI 阅片。3 名阅片医师分别在 AI 辅助阅片和独立阅片 2 种模态下阅读所有 CT 影像，得到的阅片结果为二分类（阳性或阴性）。主要评价指标为优效下的 AUC 值，次要评价指标为优效下的灵敏度和非劣效下的特异度，优效界值为 0，非劣效界值为−0.1。同时使用似然比和预测值作为判定试验诊断准确性的辅助指标。

2.2　统计学处理　采用R 4.4.1 软件进行数据分析，使用优化后的OR 法进行统计分析，协方差矩阵分别使用Jackknife、Bootstrap 和DeLong 方法进行估计，AUC 值使用RJafroc 包计算，病例层面的灵敏度和特异度估计使用dplyr 包和binom 包计算。检验水准（α）为 0.05。

2.3　研究结果

2.3.1　肋骨骨折检出的ROC 曲线分析　AI 辅助阅片组和医师独立阅片组肋骨骨折检出的ROC 曲线AUC 值分别为 0.958 和 0.902，两组AUC 差值（AI 辅助阅片－医师独立阅片）为 0.056，且差异有统计学意义（P＜0.001）。OR 法使用 3 种协方差估计方法获得的置信区间略有不同，但 P 均小于 0.001，可以认为 2 种阅片方式对肋骨骨折诊断的准确性差异有统计学意义，AI 辅助阅片的诊断性能优于医师独立阅片。见表 1。

表 1　3 种协方差估计方法下肋骨骨折的 ROC 曲线分析

Tab 1　ROC curve analysis of rib fracture detection using 3 covariance estimation methods

Method	$AUC_{AI+reader}$ (95% *CI*)	AUC_{reader} (95% *CI*)	Difference (95% *CI*)	*F* value	*P* value
$OR_{Jackknife}$	0.958 (0.935, 0.981)	0.902 (0.878, 0.926)	0.056 (0.032, 0.080)	20.983	＜0.001
$OR_{Bootstrap}$	0.958 (0.935, 0.981)	0.902 (0.879, 0.925)	0.056 (0.030, 0.081)	18.656	＜0.001
OR_{DeLong}	0.958 (0.935, 0.981)	0.902 (0.878, 0.926)	0.056 (0.032, 0.080)	21.070	＜0.001

ROC: Receiver operating characteristic; OR: Obuchowski-Rockette; $OR_{Jackknife}$: OR-based AUC with Jackknife covariance; $OR_{Bootstrap}$: OR-based AUC with Bootstrap covariance; OR_{DeLong}: OR-based AUC with DeLong covariance; AUC: Area under curve; AI: Artificial intelligence; *CI*: Confidence interval.

2.3.2　肋骨骨折检出的灵敏度和特异度　在 200 例受试者中，根据金标准判定 132 例为肋骨骨折，68 例无肋骨骨折。AI 辅助阅片组肋骨骨折检出的灵敏度和特异度分别为 0.970 和 0.946，而医师单独阅片组肋骨骨折检出的灵敏度和特异度分别为 0.838 和 0.966。两组肋骨骨折检出的灵敏度的差值为 0.131（95% *CI* 0.091～0.171），特异度差值为－0.020（95% *CI* －0.059～0.020），说明 AI 辅助阅片的灵敏度优于医师独立阅片，而 AI 辅助阅片的特异度提高虽不显著，但非劣于医师独立阅片的特异度（优效界值为 0，非劣效界值为－0.1）。见表 2。

表 2　各阅片医师在 2 种模态下检出肋骨骨折的灵敏度和特异度

Tab 2　Sensitivity and specificity of rib fracture detection by radiologists with 2 different modalities

Reader	Sensitivity (95% *CI*)			Specificity (95% *CI*)		
	AI＋reader	Reader	Difference	AI＋reader	Reader	Difference
Reader 1	0.970 (0.924, 0.992)	0.841 (0.767, 0.899)	0.129 (0.060, 0.198)	0.926 (0.837, 0.976)	0.956 (0.876, 0.991)	－0.029 (－0.108, 0.050)
Reader 2	0.970 (0.924, 0.992)	0.848 (0.776, 0.905)	0.121 (0.053, 0.189)	0.971 (0.898, 0.996)	0.971 (0.898, 0.996)	0.000 (－0.057, 0.057)
Reader 3	0.970 (0.924, 0.992)	0.826 (0.750, 0.886)	0.144 (0.073, 0.215)	0.941 (0.856, 0.984)	0.971 (0.898, 0.996)	－0.029 (－0.098, 0.039)
Overall reader	0.970 (0.948, 0.984)	0.838 (0.798, 0.873)	0.131 (0.091, 0.171)	0.946 (0.906, 0.973)	0.966 (0.931, 0.986)	－0.020 (－0.059, 0.020)

AI: Artificial intelligence; *CI*: Confidence interval.

2.3.3　肋骨骨折检出的似然比和预测值　在 AI 辅助阅片和医师独立阅片 2 种情境下肋骨骨折检出的阳性似然比均＞10，阴性似然比均＜0.2，阳性预测值均趋近于 1，而阴性预测值不高，这与本研究的主要研究指标相吻合。见表 3。

表 3　各阅片医师在 2 种模态下检出肋骨骨折的似然比和预测值

Tab 3　Likelihood ratios and predictive values of rib fracture detection by radiologists with 2 different modalities

Reader	＋LR		－LR		PPV		NPV	
	AI＋reader	Reader	AI＋reader	Reader	AI＋reader	Reader	AI＋reader	Reader
Reader 1	13.187	19.061	0.033	0.166	0.962	0.974	0.940	0.756
Reader 2	32.970	28.848	0.031	0.156	0.985	0.982	0.943	0.767
Reader 3	16.485	28.076	0.032	0.180	0.970	0.982	0.941	0.742
Overall reader	17.983	24.433	0.032	0.167	0.972	0.979	0.941	0.755

＋LR: Positive likelihood ratio; －LR: Negative likelihood ratio; PPV: Positive predictive value; NPV: Negative predictive value; AI: Artificial intelligence.

3 讨　论

AI 技术的快速发展促使各种 AI 辅助产品大量出现，如何评价这些 AI 产品的性能成为非常重要的研究课题。MRMC 设计非常适用于影像诊断试验的评价，加之我国食品药品监督管理局也推荐使用 MRMC 设计对医药器械进行评价，因而明确 MRMC 设计评价所使用的方法尤为重要。本文介绍了 MRMC 设计中 OR 法的原理与检验过程并通过实例进行统计分析，意在将研究中包括的相关结构尤其是阅片者变异纳入统计检验，进一步提高诊断试验的准确性与精确度。若直接忽视阅片者变异，使用 MRMC 设计中常规的配对 *t* 检验或未校正的方差分析，极有可能产生错误的研究结果[9]。

MRMC 设计常用的另一种方法是 Dorfman-Berbaum-Metz（DBM）法，该方法通过伪值构建三向混合效应方差分析模型。与 OR 法提供了一个可接受的概念框架而受到重视相比，DBM 法因使用伪值缺乏固有解释力而未能提供一个具有实际意义的概念模型，被认为是一种“工作模型”[19]。OR 法和 DBM 法在模型构建和方差分解等方面各有优势，但也存在等价的情况。具体而言，在以下 3 种情况下 2 种方法可实现等价计算：（1）DBM 法使用标准化伪值，OR 法用 Jackknife 法估计协方差矩阵；（2）DBM 法使用原始伪值，OR 法依然采用 Jackknife 法估计协方差矩阵；（3）DBM 法应用类伪值（quasi pseudovalue）。但 DBM 法存在仅限于精度估计及适度保守等缺点，Hillis 等[19-21]针对 DBM 法存在的问题进行了优化，提出了校正后的检验统计量和分母自由度，并证明了 DBM 法的计算等价于可被接受的 OR 法，从而增强了模型检验的可靠性和实用性，使 DBM 法在实际临床诊断试验中的应用范围进一步扩大，研究者可以在开展影像诊断试验时考虑选择和使用这 2 种统计分析方法。

AI 在影像学诊断中应用广泛，众多研究成果证实，AI 技术在辅助医师进行疾病诊断时能够显著提升工作效率和准确性，如 AI 在肝脏、前列腺、乳腺及心血管疾病等的诊断中准确性较高[22-25]，并推动疾病诊断向更精准的方向发展。李坤华[26]将头颈部 CT 血管造影自动后处理与手动后处理进行比较，发现使用 AI 系统自动后处理获得的图像质量更高、动脉狭窄的诊断性能更好、后处理时间更短且辐射剂量更少，证明 AI 技术相较于传统的手动后处理技术具有显著优势。一项将 AI 辅助系统用于 CT 检测新鲜骨折的研究同样发现，AI 辅助 CT 系统对各级别肋骨新鲜骨折的诊断灵敏度均高于影像科主任医师[27]。本研究中 AI 辅助阅片下肋骨骨折检出的 AUC 值为 0.958，医师独立阅片肋骨骨折检出的 AUC 值为 0.902，两组差异有统计学意义，与上述研究结论一致。在 2 种诊断方式下肋骨骨折检出的阳性似然比均大于 10，凸显了 AI 技术在提升诊断试验效能方面的重要作用；阴性似然比均低于 0.2，进一步证实了 AI 辅助阅片在有效排除非肋骨骨折方面的能力，从而保证了诊断的准确性；同时，阳性预测值非常接近于 1，这表明 AI 辅助阅片在降低误诊率方面具有显著优势。这些结果都说明 AI 辅助阅片影像诊断试验的诊断准确性较高，有较好的临床应用价值。

但本研究仍然存在一些局限性。第一，研究的样本量较小，可能会一定程度上影响试验结果。样本量估计是 MRMC 设计中非常重要的内容，由于同时需要估计阅片者和病例的数量，与常规的样本量估计方法有所区别，这将会在后续的研究中进一步探讨。但本研究得到主要评价指标 AUC 值表现出明显差异，可以认为对研究结果产生的影响不大。第二，本研究的实例研究没有考虑到不同年资的阅片医师对研究结果的影响，医师的经验、资历一定程度上会影响诊断准确性。第三，由于肋骨数量多且不规则，容易受呼吸影响而产生伪影，影像的对比度和分辨率等也会影响 AI 辅助阅片系统的诊断性能[28-30]。基于本研究中出现的问题，建议研究者在使用 MRMC 设计评价影像诊断试验时进一步提升影像的质量，增加病例和阅片医师的数量，并对阅片医师的资历进行分组比较，以获得更加全面、准确的结果。

在实际临床应用中，AI 辅助阅片软件能够帮助医师筛选出可疑的病变区域，或在医师初步读片后进行再次验证，从而最大程度降低漏诊和误诊率，进而快速筛选出需要优先治疗的患者，并优化患者的诊治过程。AI 技术不仅提升了医疗服务的质量，还能在一定程度上缓解医疗资源紧张，尤其是在基层医疗机构中，AI 辅助阅片软件的应用可以作为年轻医师的辅助指导[6]，提升他们的诊断能力。然而，当前必须认识到 AI 辅助阅片产品被

设计为影像科医师的辅助工具，而非替代工具[31]，未来仍需更多研究来探索 AI 技术的新发展。

[参 考 文 献]

[1] WEIKERT T, WINKEL D J, BREMERICH J, et al. Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm[J]. Eur Radiol, 2020, 30(12): 6545-6553. DOI: 10.1007/s00330-020-06998-0.

[2] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115-118. DOI: 10.1038/nature21056.

[3] CHAKRABORTY D P. Analysis of location specific observer performance data: validated extensions of the Jackknife free-response (JAFROC) method[J]. Acad Radiol, 2006, 13(10): 1187-1193. DOI: 10.1016/j.acra.2006.06.016.

[4] BEAM C A, LAYDE P M, SULLIVAN D C. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample[J]. Arch Intern Med, 1996, 156(2): 209-213.

[5] WAGNER R F, METZ C E, CAMPBELL G. Assessment of medical imaging systems and computer aids: a tutorial review[J]. Acad Radiol, 2007, 14(6): 723-748. DOI: 10.1016/j.acra.2007.03.001.

[6] VAN DEN BROEK M C L, BUIJS J H, SCHMITZ L F M, et al. Diagnostic performance of artificial intelligence in rib fracture detection: systematic review and meta-analysis[J]. Surgeries, 2024, 5(1): 24-36. DOI: 10.3390/surgeries5010005.

[7] THRALL J H, LI X, LI Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success[J]. J Am Coll Radiol, 2018, 15(3 Pt B): 504-508. DOI: 10.1016/j.jacr.2017.12.026.

[8] TOPOL E J. High-performance medicine: the convergence of human and artificial intelligence[J]. Nat Med, 2019, 25(1): 44-56. DOI: 10.1038/s41591-018-0300-7.

[9] ZHOU X H, OBUCHOWSKI N A, MCCLISH D K. Statistical methods in diagnostic medicine[M]. 2nd ed. New York: John Wiley & Sons, Inc., 2011: 310.

[10] WANG L, WANG H, XIA C, et al. Toward standardized premarket evaluation of computer aided diagnosis/detection products: insights from FDA-approved products[J]. Expert Rev Med Devices, 2020, 17(9): 899-918. DOI: 10.1080/17434440.2020.1813566.

[11] FDA. Clinical performance assessment: considerations for computer-assisted detection devices applied to radiology images and radiology device data in premarket notification (510(k)) submissions—guidance for Industry and Food and Drug Administration Staff[Z/OL]. (2022-09-28)[2023-12-20]. *https://www.fda.gov/media/77642/download.*

[12] 国家药品监督管理局医疗器械技术审评中心. 关于发布深度学习辅助决策医疗器械软件审评要点的通告(2019 年第 7 号)[EB/OL]. (2019-07-03)[2024-11-14]. *https://www.cmde.org.cn//xwdt/shpgzgg/gztg/20190703141714991.html.*

[13] 国家药品监督管理局医疗器械技术审评中心. 国家药监局关于发布视力筛查仪和乳腺X射线系统2项注册技术审查指导原则的通告(2021 年第 42 号)[EB/OL]. (2021-06-24)[2024-11-14]. *https://www.cmde.org.cn//flfg/zdyz/fbg/fbgyy/20210701102800157.html.*

[14] OBUCHOWSKI N A Jr, ROCKETTE H E Jr. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations[J]. Commun Stat Simul Comput, 1995, 24(2): 285-308. DOI: 10.1080/03610919508813243.

[15] EFRON B, TIBSHIRANI R J. An introduction to the Bootstrap[J]. J R Stat Soc A Stat, 1993, 43(4): 600. DOI: 10.2307/2348146.

[16] DORFMAN D D, BERBAUM K S, METZ C E. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the Jackknife method[J]. Invest Radiol, 1992, 27(9): 723-731.

[17] DELONG E R, DELONG D M, CLARKE-PEARSON D L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach[J]. Biometrics, 1988, 44(3): 837-845.

[18] PAVUR R, NATH R. Exact F tests in an ANOVA procedure for dependent observations[J]. Multivariate Behav Res, 1984, 19(4): 408-420. DOI: 10.1207/s15327906mbr1904_3.

[19] HILLIS S L, OBUCHOWSKI N A, SCHARTZ K M, et al. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data[J]. Stat Med, 2005, 24(10): 1579-1607. DOI: 10.1002/sim.2024.

[20] HILLIS S L, BERBAUM K S, METZ C E. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis[J]. Acad Radiol, 2008, 15(5): 647-661. DOI: 10.1016/j.acra.2007.12.015.

[21] HILLIS S L. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis[J]. Stat Med, 2007, 26(3): 596-619. DOI: 10.1002/sim.2532.

[22] YASAKA K, AKAI H, KUNIMATSU A, et al. Deep learning for staging liver fibrosis on CT: a pilot study[J].

Eur Radiol, 2018, 28(11): 4578-4585. DOI: 10.1007/s00330-018-5499-7.

［23］ HA R, MUTASA S, KARCICH J, et al. Predicting breast cancer molecular subtype with MRI dataset utilizing convolutional neural network algorithm[J]. J Digit Imaging, 2019, 32(2): 276-282. DOI: 10.1007/s10278-019-00179-2.

［24］ BERNARD A, COMBY P O, LEMOGNE B, et al. Deep learning reconstruction versus iterative reconstruction for cardiac CT angiography in a stroke imaging protocol: reduced radiation dose and improved image quality[J]. Quant Imaging Med Surg, 2021, 11(1): 392-401. DOI: 10.21037/qims-20-626.

［25］ ZHU L, GAO G, LIU Y, et al. Feasibility of integrating computer-aided diagnosis with structured reports of prostate multiparametric MRI[J]. Clin Imaging, 2020, 60(1): 123-130. DOI: 10.1016/j.clinimag.2019.12.010.

［26］ 李坤华. 基于人工智能的多模式CT自动分析在头颈部动脉狭窄患者中的应用研究[D]. 重庆：重庆医科大学，2024.

［27］ 梁洁，孙金磊，李梦远，等. 基于深度学习人工智能辅助系统用于CT检出肋骨新鲜骨折[J]. 中国介入影像与治疗学，2023，20(9)：555-560. DOI: 10.13929/j.issn.1672-8475.2023.09.010.

［28］ BIZIMUNGU R, ALVAREZ S, BAUMANN B M, et al. Thoracic spine fracture in the panscan era[J]. Ann Emerg Med, 2020, 76(2): 143-148. DOI: 10.1016/j.annemergmed.2019.11.017.

［29］ DENNIS B M, BELLISTER S A, GUILLAMONDEGUI O D. Thoracic trauma[J]. Surg Clin North Am, 2017, 97(5): 1047-1064. DOI: 10.1016/j.suc.2017.06.009.

［30］ LIN F C, LI R Y, TUNG Y W, et al. Morbidity, mortality, associated injuries, and management of traumatic rib fractures[J]. J Chin Med Assoc, 2016, 79(6): 329-334. DOI: 10.1016/j.jcma.2016.01.006.

［31］ BLUM A, GILLET R, URBANEJA A, et al. Automatic detection of rib fractures: are we there yet?[J]. EBioMedicine, 2021, 63: 103158. DOI: 10.1016/j.ebiom.2020.103158.

［本文编辑］ 杨亚红