

DOI: 10.16781/j.CN31-2187/R.20240467

• 短篇论著 •

影像诊断试验中多阅片者多病例设计的样本量估计研究

向 蔓¹, 宛慧琴¹, 潘喆敏¹, 何 倩², 秦婴逸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** 实现影像诊断试验中多阅片者多病例设计下的样本量估计, 探索阅片者和病例均随机、阅片者固定病例随机、阅片者随机病例固定 3 种情形下阅片者和受试者的数量组合。**方法** 采用 Van Dyke 数据集中 114 例受试者(45 例被金标准诊断为主动脉夹层)的影像, 并由 5 名影像科医生在自旋回波(Spin-echo)和电影(Cine)2 种序列 MRI 下的判读结果为预试验数据, 使用 Obuchowski-Rockette 法估计正式试验中所需的样本量。**结果** 医生在 Spin-echo MRI、Cine MRI 序列下判别主动脉夹层的平均 AUC 值分别为 0.941 (95% CI 0.899~0.983)、0.897 (95% CI 0.837~0.957)。当效应量取两者 AUC 差值 0.044、阅片者数量为 5 时, 阅片者和病例均随机、阅片者固定病例随机、阅片者随机病例固定 3 种情形下分别需要受试者 337、162、282 例。**结论** 多阅片者多病例设计在进行样本量估计时需考虑阅片者和受试者两者数量, 相同情况下, 阅片者人数越多所需的受试者数量越少; 3 种情形中, 阅片者(≥ 5 人)和病例均随机时所需样本量最多。

[关键词] 影像诊断试验; 多阅片者多病例; Obuchowski-Rockette 法; 样本量

[引用本文] 向蔓, 宛慧琴, 潘喆敏, 等. 影像诊断试验中多阅片者多病例设计的样本量估计研究[J]. 海军军医大学学报, 2025, 46(5): 693-698. DOI: 10.16781/j.CN31-2187/R.20240467.

Sample size estimation of radiological diagnostic tests in multireader multicase design

XIANG Man¹, WAN Huiqin¹, PAN Zhemin¹, HE Qian², QIN Yingyi², 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 estimate the sample size of radiological diagnostic tests in multireader multicase (MRMC) design, and to explore the numbers of readers and cases under 3 different inference situations: random-reader random-case, fixed-reader random-case, and random-reader fixed-case. **Methods** The images of 114 participants (45 cases diagnosed as aortic coarctation by the gold standard) in the Van Dyke dataset were used, and 5 radiologists read these images under 2 different magnetic resonance imaging (MRI) sequences (Spin-echo and Cine MRI) to obtain the pre-experiment data. Then Obuchowski-Rockette method was used to estimate sample size. **Results** The mean area under curve (AUC) of aortic coarctation determined by radiologists was 0.941 (95% confidence interval [CI] 0.899-0.983) with the Spin-echo MRI sequence, and 0.897 (95% CI 0.837-0.957) with the Cine MRI sequence. When the effect size was 0.044 and the number of readers was 5, we needed 337 participants for random-reader random-case, 162 participants for fixed-reader random-case, and 282 participants for random-reader fixed-case. **Conclusion** In MRMC design, we need both the number of readers and cases; the larger the number of readers, the smaller the cases required. We need more samples under the situation of random-reader random-case, when the number of readers is ≥ 5 .

[Key words] radiological diagnostic test; multireader multicase; Obuchowski-Rockette method; sample size

[Citation] XIANG M, WAN H, PAN Z, et al. Sample size estimation of radiological diagnostic tests in multireader multicase design[J]. Acad J Naval Med Univ, 2025, 46(5): 693-698. DOI: 10.16781/j.CN31-2187/R.20240467.

[收稿日期] 2024-06-30 **[接受日期]** 2024-10-08

[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), 上海市产业协同创新项目(2021-cyxt1-kj10)。Supported by Emerging Interdisciplinary Research Project of Shanghai Municipal Health Commission (2022JC011) and Shanghai Industry Collaborative Innovation Project (2021-cyxt1-kj10).

[作者简介] 向 蔓, 硕士生. E-mail: 1923681995@qq.com

* 通信作者(Corresponding author). E-mail: hejia63@yeah.net

在医疗领域,许多疾病的诊断需要借助影像数据(如X线、CT、超声、MRI等),其中90%的影像判读由影像科医生完成^[1]。随着深度学习技术的快速发展,人工智能(artificial intelligence, AI)技术在医疗领域得到了广泛应用, AI辅助阅片产品在影像诊断中表现良好,可帮助医生提高诊断的灵敏度、缓解阅片压力、提升阅片效率,已有多款产品完成注册上市(如肿瘤、肺结节等辅助识别软件)^[2]。在评价产品诊断效果的试验中(医生单独阅片 vs 医生借助AI辅助阅片产品阅片),考虑到使用场景和对象的差异性,我国国家药品监督管理局和美国FDA均推荐采用多阅片者多病例(multireader multicase, MRMC)设计^[3-5]。

MRMC设计是指多名阅片者使用不同方法对病例影像进行判读的一种设计,常用于比较2种或多种诊断方法的效果,其中全交叉MRMC设计最为常见,广泛应用于计算机辅助诊断、AI辅助影像诊断等领域的临床评价中^[6-7],可有效避免医生和使用场景等因素带来的偏倚。近年来已有学者运用MRMC方差分析模型比较不同诊断技术(2D和3D断层成像)诊断乳腺肿瘤患者的差异^[8],其后又在MRMC设计下的胸部CT影像AI辅助诊断系统临床试验中比较不同方式下肺结节的检出与诊断效果,同时在不同阅片相关性条件下进行病例样本量估计的探索性分析,总结影响样本量的因素^[9]。样本量一直是临床试验中的重要话题,鉴于MRMC设计的复杂性,样本量估计具有一定难度。本研究在国内外相关研究基础上将MRMC设计中的方差分析方法用于样本量估计,实现不同情形下的样本量估计,探索合适的阅片者和病例数量组合。

1 材料和方法

1.1 MRMC设计分析方法

1.1.1 Obuchowski-Rockette (OR) 法模型 在MRMC设计中,数据间会产生复杂的相关结构,阅片者会使用不同方法对同一张影像进行阅片,而不同阅片者也会对同一张影像阅片,同一张影像会有多个阅片结果,因此在进行分析时需要对方法进行调整,OR法是常用方法之一。OR法由Obuchowski和Rockette^[10]在1995年提出,其将诊断方式作为固定效应,阅片者作为随机效应,直接使用原始观察值生成的准确性指标构建方差分析模型。考虑

N 次重复试验情况,模型如下:

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

其中 θ 指 n 次重复试验下第 i 种诊断方法第 j 名阅片者的准确度指标估计值(如AUC估计值)。 μ 为固定常数,指所有诊断方法、阅片者以及重复阅片情况下准确度指标的均值; τ_i 为第 i 种诊断方法的固定效应; R_j 为第 j 名阅片者的随机效应; $(\tau R)_{ij}$ 为第 i 种诊断方法与第 j 名阅片者交互的随机效应; ε_{ij} 为随机误差项。同时该模型假定 R_j 独立服从均值为0、方差为 σ_r^2 的正态分布,即 $R_j \sim N(0, \sigma_r^2)$; $(\tau R)_{ij}$ 独立服从均值为0、方差为 $\sigma_{\tau r}^2$ 的正态分布,即 $(\tau R)_{ij} \sim N(0, \sigma_{\tau r}^2)$ 。

1.1.2 误差项协方差矩阵 在OR法模型中误差项 ε_{ij} 并不独立,而是均数为0、长度为 $(I \times J \times N)$ 的向量,且服从多元正态分布。该法假定不同阅片者和诊断方法之间的误差相等且在不同阅片者间和不同诊断方法间协方差一致,由此产生的协方差矩阵如下:

$$\text{Cov}(\varepsilon_{ijn}, \varepsilon_{ij'n}) = \begin{cases} \sigma_e^2 & i=i', j=j' \text{ 相同诊断方法, 相同阅片者} \\ \text{Cov}_1 & i \neq i', j=j' \text{ 不同诊断方法, 相同阅片者} \\ \text{Cov}_2 & i=i', j \neq j' \text{ 相同诊断方法, 不同阅片者} \\ \text{Cov}_3 & i \neq i', j \neq j' \text{ 不同诊断方法, 不同阅片者} \end{cases}$$

Obuchowski和Rockette认为临床实际中 $\text{Cov}_1 \geq \text{Cov}_2 \geq \text{Cov}_3 \geq 0$ 。在实际情况中,考虑时间与成本往往不会进行重复试验,即试验次数 $n=1$,因此无法直接估计协方差矩阵,可采用Bootstrap^[11]、Jackknife^[12]和DeLong^[13]方法,先对数据进行重抽样,形成多个数据集,再进行估计。

1.1.3 OR法统计量 OR法的假设检验中原假设为 $H_0: \tau_1 = \dots = \tau_i$,即各种诊断方式的效应相同;备择假设为 $H_1: \tau_1 \neq \dots \neq \tau_i$,即各种诊断方式的效应不同。模型中的误差项并不独立,原始 F 值可能产生不准确的结果,因此Obuchowski和Rockette基于Pavur和Nath^[14]的研究提出了校正的统计量 F 。原始 F 值($F_{\text{未校正}}$)和校正 F 值(F_{OR})如下:

$$F_{\text{未校正}} = \frac{MS_t}{MS_r} \sim F_{I-1, (I-1)(J-1)}$$

$$F_{\text{OR}} = \frac{MS_t}{MS_r + J(\text{Cov}_2 - \text{Cov}_3)} \sim F_{I-1, (I-1)(J-1)}$$

其中, I 为诊断方式数量, J 为阅片者数量, MS_t 为诊断方式的均方值, MS_r 为诊断方式与阅片者交互

项的均方值, F_{OR} 值中分子自由度 df_1 为 $(I-1)$, 分母自由度 df_2 为 $(I-1)(J-1)$ 。原始 OR 法中存在 F_{OR} 值分母为负以及阅片者数量较小时结果过于保守的问题。针对这些问题, Hillis 等^[15-16] 对 F_{OR} 值和分母自由度进行了优化。优化后的 F_{OR} 值 (F_{ORH})、分母自由度 ($df_{2\ ORH}$) 如下:

$$F_{ORH} = \frac{MS_t}{MS_{tr} + J[\max(\widehat{Cov}_2 - \widehat{Cov}_3, 0)]} \sim F_{I-1, df_{2\ ORH}}$$

$$df_{2\ ORH} = \frac{\{MS_{tr} + J[\max(\widehat{Cov}_2 - \widehat{Cov}_3, 0)]\}^2}{\frac{(MS_{tr})^2}{(I-1)(J-1)}}$$

1.2 MRMC 设计样本量计算

1.2.1 阅片者与病例情况 在影像诊断试验 MRMC 设计中, 计算样本量时需要考虑阅片者和受试者两方面的数量, 根据来源可将两者归为固定或随机效应。若阅片者/受试者数量充足, 且可在一定范围内抽样获得, 则可视为随机效应; 若阅片者/受试者来自固定群体, 数量有限且无法通过抽样获得, 则可视为固定效应。对于随机效应, 研究结论可推广至所抽样的群体, 而在固定效应中结论仅适用于参与试验的个体。根据两者的来源可分为 3 种情形: 阅片者和病例均随机 (random-reader random-case, RRRC)、阅片者固定病例随机 (fixed-reader random-case, FRRC)、阅片者随机病例固定 (random-reader fixed-case, RRFC)。

1.2.2 样本量估算考虑 在实际的影像诊断试验中, 估计样本量时需重点考虑 5 个方面的内容^[17]:

(1) 诊断试验设计 (优效、非劣效)、MRMC 设计类型、检验水准 α 、II 类错误 β 、病患比例; (2) 准确度指标、效应量; (3) OR 法中相关参数; (4) 统计量 F 分布的非中心参数和分母自由度;

(5) 实际功效。

对于试验设计, 我国医疗器械技术审评中心建议: “优先选择同品种产品或临床参考标准 (即临床金标准) 进行非劣效对照设计, 若无同品种产品且难以获取临床参考标准可选择替代方法, 如选择用户结合软件联合决策与用户单独决策进行优效对照设计”^[4]。全交叉设计 (所有阅片者采用所有阅片方式对所有病例影像进行阅片) 是 MRMC 中最常见的一类。阳性和阴性病例的比例一般为 1 : 1, 或者与预试验数据一致^[18]。

在影像诊断试验中, AUC 是常见的准确度评价指标^[19], 灵敏度、特异度及相关衍生指标也可作为评价指标。通常 AUC 和灵敏度采用优效设计, 特异度采用非劣效设计。其中 AUC 计算结果为受试者总数, 灵敏度计算结果为阳性病例数, 特异度计算结果为阴性病例数。效应量 (d) 为不同诊断方法准确度指标的差值, 一般来自预试验 (d_{obs}) 或同类型试验, 也可使用临床上确定的差异值。由于抽样误差的存在, 预试验和实际情况中的效应量可能存在差异, 可取 2 倍标准差来表示离散程度和变异范围, 在本研究中取 $d = |d_{obs}|$ 和 $d = |d_{obs}| + 2\sigma$ (σ 为标准差)^[20] 来估计样本量。

使用 OR 法时, 可先通过预试验数据求解 σ_{tr}^2 、 σ_e^2 、 Cov_1 、 Cov_2 、 Cov_3 的值, 再进行样本量估计。计算非中心参数 (λ) 和分母自由度 (df_2) 时, 两者估计值见表 1。当原假设 (2 种阅片方法诊断准确度无差别) 不成立时^[21], 在 RRRC、RRFC 两种情形下 F_{OR} 值将服从近似非中心 F 分布 $F_{I-1, df_2; \lambda}$; 在 FRRC 情形下, F_{OR} 值将服从近似非中心 χ^2 分布 $\chi^2_{I-1; \lambda}$ 。

表 1 MRMC 设计中非中心参数和分母自由度情况

不同情形	非中心参数估计值 ($\hat{\lambda}$)	分母自由度估计值 ($\hat{df}_2$)
RRRC	$\frac{rd^2/2}{\sigma_{tr}^2 + \frac{c^*}{c} [\sigma_e^2 - \widehat{Cov}_1 + (r-1)\max(\widehat{Cov}_2 - \widehat{Cov}_3, 0)]}$	$\frac{\{\sigma_{tr}^2 + \frac{c^*}{c} [\sigma_e^2 - \widehat{Cov}_1 + (r-1)\max(\widehat{Cov}_2 - \widehat{Cov}_3, 0)]\}^2}{\{\sigma_{tr}^2 + \frac{c^*}{c} [\sigma_e^2 - \widehat{Cov}_1 - \max(\widehat{Cov}_2 - \widehat{Cov}_3, 0)]\}^2 / [(t-1)(r-1)]}$
FRRC	$\frac{rd^2/2}{\frac{c^*}{c} [\sigma_e^2 - \widehat{Cov}_1 + (r-1)\max(\widehat{Cov}_2 - \widehat{Cov}_3, 0)]}$	NA
RRFC	$\frac{rd^2/2}{\sigma_{tr}^2 + \frac{c^*}{c} [\sigma_e^2 - \widehat{Cov}_1 - \max(\widehat{Cov}_2 - \widehat{Cov}_3, 0)]}$	$(t-1)(r-1)$

MRMC: 多阅片者多病例; RRRC: 阅片者和病例均随机; FRRC: 阅片者固定病例随机; RRFC: 阅片者随机病例固定; r : 待估计阅片者数量; NA: 不适用。 d : 诊断方法准确度指标的差值; c^* : 预试验中受试者数量; c : 待估计的受试者数量; t : 诊断方法数量; σ_{tr}^2 : 阅片者与诊断方法交互项的方差; σ_e^2 : 误差项的方差。

结合表 1 中的公式,可得到 2 种诊断方法准确性具有差异的实际功效 (Power) ^[18]:

(1) 对于 RRRC、RRFC 两种情况

$$Power=Prb(F_{1,\widehat{df}_2;\widehat{\lambda}}>F_{1-\alpha;1,\widehat{df}_2})$$

(2) 对于 FRRC

$$Power=Prb(\chi^2_{1;\widehat{\lambda}}>\chi^2_{1-\alpha;1})$$

结合Ⅱ类错误β值的限制,最终可推导出所需受试者数量(c)和阅片者数量(r)的样本量组合。

1.3 数据来源 本研究中准确度指标AUC、效应量以及 OR 法中的相关参数均来自预试验,使用公开的 Van Dyke 数据集^[22]作为预试验结果,估计 MRMC 设计下所需阅片者和病例样本数量。该数据集中共有 114 例受试者(45 例为主动脉夹层患者),由 5 名影像科医生分别对受试者在自旋回波 (Spin-echo) 和电影 (Cine) 2 种序列 MRI 下的影像进行阅片,比较不同序列检测主动脉夹层的表现。医生使用 5 分等级量表对影像进行判读: 1=肯定无主动脉夹层, 2=可能无主动脉夹层, 3=不确定主动脉夹层, 4=可能为主动脉夹层, 5=肯定为主动脉夹层。

1.4 统计学处理 本研究采用 MRMC 全交叉设计,使用 R 4.2.3 软件和 R Jafroc 2.1.2 包进行样本量估计。采用双侧检验,α=0.05,β=0.2。

2 结 果

在预试验 Van Dyke 数据集中,5 名影像科

医生在 Spin-echo MRI 阅片方式下 AUC 值区间为 0.905~0.999,平均 AUC 值为 0.941 (95% CI 0.899~0.983); 在 Cine MRI 序列下 AUC 值区间为 0.830~0.973,平均 AUC 值为 0.897 (95% CI 0.837~0.957)。见表 2。

表 2 两种诊断方法下阅片者的 AUC 结果

阅片者	Spin-echo MRI (95% CI)	Cine MRI (95% CI)
阅片者 1	0.948 (0.905, 0.991)	0.920 (0.861, 0.987)
阅片者 2	0.905 (0.847, 0.963)	0.859 (0.788, 0.930)
阅片者 3	0.922 (0.864, 0.980)	0.904 (0.849, 0.959)
阅片者 4	0.999 (0.998, 1.000)	0.973 (0.939, 1.000)
阅片者 5	0.930 (0.879, 0.981)	0.830 (0.749, 0.911)
平均值	0.941 (0.899, 0.983)	0.897 (0.837, 0.957)

AUC:曲线下面积;Spin-echo MRI:自旋回波序列磁共振成像;Cine MRI:磁共振电影成像;CI:置信区间。

估计正式试验样本量时以平均 AUC 为准确度指标,采用优效、MRMC 全交叉设计,α=0.05,β=0.2。根据预试验数据,计算 OR 法中所需参数值为 σ_r²=0.000 20、σ_e²=0.000 80、Cov₁=0.000 35、Cov₂=0.000 34、Cov₃=0.000 24,效应量为 2 种诊断方法下平均 AUC 的差值,即 d_{obs}=0.044,σ=0.022,计算 ① d=|d_{obs}|=0.044 和 ② d=|d_{obs}|+2σ=0.088 在不同情况下所需样本数量,结果见表 3。

表 3 不同情形下 MRMC 全交叉设计的样本量

效应量	阅片者, n	RRRC		FRRC		RRFC	
		受试者, n	实际功效	受试者, n	实际功效	受试者, n	实际功效
0.044	3	>2 000	NA	206	0.802	>2 000	NA
	4	1 009	0.800	179	0.802	>2 000	NA
	5	337	0.800	162	0.800	282	0.800
	6	247	0.800	152	0.802	149	0.801
	7	208	0.800	144	0.802	101	0.801
	8	186	0.801	138	0.801	77	0.803
	9	171	0.800	134	0.803	62	0.802
	10	161	0.801	130	0.801	52	0.803
	3	102	0.802	52	0.805	246	0.800
	4	62	0.804	45	0.804	62	0.804
0.088	5	50	0.805	41	0.805	35	0.807
	6	44	0.805	38	0.802	24	0.803
	7	40	0.801	36	0.802	<20	NA
	8	38	0.806	35	0.807	<20	NA
	9	36	0.803	34	0.808	<20	NA
	10	35	0.807	33	0.807	<20	NA

RRRC:阅片者和病例均随机;FRRC:阅片者固定病例随机;RRFC:阅片者随机病例固定;NA:不适用。

预试验 Van Dyke 数据集中阳性和阴性病例比例为 1 : 1.53 (45 : 69),正式试验中若参照该比例

确定受试者人数,当 d=0.044、r=4 时,在 RRRC 时共需 1 009 例受试者,其中需要阳性病例 399 例、

阴性病例 610 例;若保持阳性与阴性病例比例为 1 : 1, 则 1 009 例受试者中一半可为阳性病例。

相同情况下,阅片者数量和病例数量相互限制,即阅片者数量越多所需的受试者人数越少,如在 $d=0.044$ 、 $r=5$ 时需要受试者 337 例,而当 $r=6$ 时仅需受试者 247 例。

3 种情形中 RRRC 所需样本量最多,当 $d=0.044$ 、 $r=5$ 时,RRRC、FRRC、RRFC 下分别需要受试者 337、162、282 例。同时,效应量越大所需阅片者和受试者人数越少,如在 RRRC 情形下,若 $r=6$ 、 $d=0.044$ 时需要 247 例受试者,而 $d=0.088$ 时仅需 44 例受试者。

3 讨 论

样本量一直是影像诊断试验中的重要内容,随着 OR 法的逐渐成熟,MRMC 设计的样本量研究取得了许多进展。2004 年 Eng^[23] 在归纳样本量影响因素时认为 MRMC 设计中至少需要 4 名阅片者,并提供了效应量、预期 AUC 以及阅片者间变异的参考值。2009 年 Obuchowski^[24] 发现,相比全交叉设计,混合裂区设计需要的阅片者数量更少,但对金标准的要求更高。2011 年,Obuchowski 和 Hillis^[25] 基于 2 个实际案例(肺结节、结肠息肉计算机辅助诊断)对多病灶数据(如患者有多个肺结节)的样本量研究做了探索,估计了不同病灶比例下(所有患者有 1 个病灶,其中 25%、50% 的患者有 2 个病灶)的样本量。2012 年 Chen 等^[26] 对 AUC 值在非劣效设计中的样本量估计做了调整。2018 年 Hillis 和 Schartz^[21] 对 OR 法在不同 MRMC 设计类型下的使用做了详细的方法学介绍。2022 年尚美霞等^[9] 使用病灶层面的 AUC 数据对样本量进行了估计。除 OR 法外,DBM 法也可用于样本量估计,但需将原始值转化为 Jackknife 伪值^[27]。在 OR 法中通过 Jackknife 法估计协方差矩阵或者在 DBM 法中使用类伪值(quasi pseudovalues)进行样本量估计时^[15],两种方法的结果可以相互转换。

在选取准确度指标时可使用 AUC、灵敏度、特异度,也可用三者的衍生指标。以特异度为评价指标进行非劣效设计时,非劣效界值的选择须有明确依据。AUC 值的大小依赖于置信度评分情况,评分的分布会影响 ROC 曲线的拟合方式,对于非正态分布的置信分数需要选择非参数方法进

行拟合。同时对病灶水平的样本量进行估计时,可使用病灶层面的 ROC 曲线计算 AUC 值,如定位 ROC 曲线(location ROC curve, LROC)、自由响应 ROC 曲线(free-response ROC curve, FROC)以及在此基础上改进的可替代自由响应 ROC 曲线(alternative free-response ROC curve, AFROC)^[28]。

在 MRMC 设计的样本量估计中需要考虑阅片者间以及病例间的变异,分别估计阅片者和受试者两方面的样本数量。本研究结果发现,在相同情况下,增加阅片者的数量可在一定程度上减少所需受试者数量,因此可根据实际情况选择两者合适的数量组合。在 RRRC、FRRC 和 RRFC 这 3 种情形中,阅片者(≥ 5 人)和病例均为随机因素时所需的样本量最多,相比其他情况这时阅片者和病例的来源更加广泛,产生的变异也更大,若要获得相同的功效需要更多的阅片者和受试者,最终结论也可推广至群体层面。影像诊断试验中效应量的选择一般可取预试验结果或临床推荐值,但由于抽样误差,预试验和实际研究仍存在一定差异,可根据阅片者的波动和试验情况来确定效应量。效应量越小,方法间诊断效果差异所需的样本量越多,试验成本也越高。

本文结合国内外相关研究结果,介绍了 MRMC 全交叉设计下使用 OR 法进行样本量估计的相关步骤和参数,实现了阅片者和病例不同情形下的数量估计。本研究也存在一定的局限性:在分析方法上仅使用 OR 法,缺少 DBM 法在样本量估计上的探索;在 MRMC 设计类型上,局限于全交叉设计下不同场景的样本量估计,暂未扩展至 MRMC 其他设计类型(如嵌套设计、混合裂区设计)的样本量计算;在影像诊断试验的准确度评价指标上局限于 ROC 指标,未探索其他 ROC(如 LROC、FROC、AFROC)指标以及灵敏度、特异度在样本量估计中的应用。在实际开展 MRMC 设计类型的影像诊断试验中,需结合多方因素确定相关指标,以估计阅片者和受试者的样本量,从而确保试验顺利进行。

[参 考 文 献]

- [1] 贾凯丽,王雪梅.医学影像人工智能新进展[J].国际放射医学核医学杂志,2020(1):27-31. DOI: 10.3760/cma.j.issn.1673-4114.2020.01.007.
- [2] 国家药品监督管理局.肺结节CT影像辅助检测软件获批上市[EB/OL].(2020-12-10)[2023-10-20]. <https://www.nmpa.gov.cn/yaowen/ypjgyw/>

- ylqxyw/20201201110537199.html.
- [3] 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知[EB/OL]. (2019-07-03)[2023-10-20]. <https://www.cmde.org.cn/CL0050/19360.html>.
 - [4] 国家药品监督管理局医疗器械技术审评中心. 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[EB/OL]. (2021-06-24)[2023-10-20]. <https://www.nmpa.gov.cn/ylqx/ylqxgggt/20210629161924195.html>.
 - [5] FDA. Considerations for computer-assisted detection devices applied to radiology images and radiology device data in premarket notification[510(k)] submissions [EB/OL]. (2022-09-28)[2023-10-20]. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology>.
 - [6] GALLAS B D, CHAN H P, D'ORSI C J, et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA[J]. Acad Radiol, 2012, 19(4): 463-477. DOI: 10.1016/j.acra.2011.12.016.
 - [7] 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.
 - [8] 尚美霞,姚晨,康晓平,等. MRMC 方差分析在影像诊断试验多阅片者多病例研究设计中的应用[J]. 中国卫生统计, 2017, 34(5): 705-709, 712.
 - [9] 尚美霞,阎小妍,李雪迎,等. 采用多阅片者多病例设计评估AI辅助医疗产品临床试验的样本量估算和应用[J]. 中国卫生统计, 2022, 39(1): 14-18. DOI: 10.3969/j.issn.1002-3674.2022.01.003.
 - [10] 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.
 - [11] EFRON B, TIBSHIRANI R J. An introduction to the Bootstrap [M]. Boca Raton: Chapman and Hall/CRC, 1994: 45.
 - [12] QUENOUILLE M H. Problems in plane sampling[J]. Ann Math Statist, 1949, 20(3): 355-375. DOI: 10.1214/aoms/1177729989.
 - [13] 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. DOI: 10.2307/2531595.
 - [14] 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.
 - [15] 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.
 - [16] 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.
 - [17] OBUCHOWSKI N A, BULLEN J. Multireader diagnostic accuracy imaging studies: fundamentals of design and analysis[J]. Radiology, 2022, 303(1): 26-34. DOI: 10.1148/radiol.211593.
 - [18] HILLIS S L, OBUCHOWSKI N A, BERBAUM K S. Power estimation for multireader ROC methods an updated and unified approach[J]. Acad Radiol, 2011, 18(2): 129-142. DOI: 10.1016/j.acra.2010.09.007.
 - [19] OBUCHOWSKI N A. Receiver operating characteristic curves and their use in radiology[J]. Radiology, 2003, 229(1): 3-8. DOI: 10.1148/radiol.2291010898.
 - [20] CHAKRABORTY D P. Observer performance methods for diagnostic imaging: foundations, modeling, and applications with R-based examples[M]. Boca Raton: Chapman and Hall/CRC, 2018: 238.
 - [21] HILLIS S L, SCHARTZ K M. Multireader sample size program for diagnostic studies: demonstration and methodology[J]. J Med Imaging (Bellingham), 2018, 5(4): 045503. DOI: 10.1117/1.JMI.5.4.045503.
 - [22] VAN DYKE C, WHITE R, OBUCHOWSKI N, et al. Cine MRI in the diagnosis of thoracic aortic dissection. 79th RSNA meetings[C]. Chicago, IL, 1993, 28. DOI: 10.1148/radiology.188.2.573.
 - [23] ENG J. Sample size estimation: a glimpse beyond simple formulas[J]. Radiology, 2004, 230(3): 606-612. DOI: 10.1148/radiol.2303030297.
 - [24] OBUCHOWSKI N A. Reducing the number of reader interpretations in MRMC studies[J]. Acad Radiol, 2009, 16(2): 209-217. DOI: 10.1016/j.acra.2008.05.014.
 - [25] OBUCHOWSKI N A, HILLIS S L. Sample size tables for computer-aided detection studies[J]. AJR Am J Roentgenol, 2011, 197(5): W821-W828. DOI: 10.2214/AJR.11.6764.
 - [26] CHEN W, PETRICK N A, SAHINER B. Hypothesis testing in noninferiority and equivalence MRMC ROC studies[J]. Acad Radiol, 2012, 19(9): 1158-1165. DOI: 10.1016/j.acra.2012.04.011.
 - [27] 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.
 - [28] HE X, FREY E. ROC, LROC, FROC, AFROC: an alphabet soup[J]. J Am Coll Radiol, 2009, 6(9): 652-655. DOI: 10.1016/j.jacr.2009.06.001.