

· 胸部影像学 ·

人工智能辅助软件可提升疲劳状态下放射科规培医师对肺结节的检测效能

王亮, 许迪, 孙丹丹, 顾俊, 伍建林, 于晶

【摘要】 目的:探讨在人工智能(AI)肺结节检测软件的辅助下能否提升疲劳状态的放射科规培医师对肺结节的检测效能。**方法:**搜集 182 例患者的 1 mm 薄层胸部 CT 图像,有一位放射科规培医师分别在 3 种模式下进行阅片:正常状态下独立阅片(A 组)、疲劳状态下(即一天日常工作满 8 小时以上)独立阅片(B 组)、疲劳状态下使用 AI 软件辅助阅片(C 组),三种阅片模式均间隔洗脱期(2 周),分别记录每次阅片时检出结节的位置、大小和数目。将 3 次肺结节检出结果与金标准(由 2 位从事胸部影像诊断超过 8 年的中级医师结合 AI 筛查结果分别作出诊断,再由 1 位从事胸部影像诊断超过 15 年的高级医师最终审核确定)进行比较,计算敏感度和(患者)人均假阳性(误诊)结节数来评价 3 种模式的检测效能。**结果:**经金标准确认 1281 个肺结节,A 组检出真阳性结节 592 个、假阳结节 297 个,敏感度 46.21%,人均误诊结节数为 1.63;B 组检出真阳性结节 517 个、假阳结节 225 个,敏感度 40.36%,人均误诊结节数为 1.24;C 组检出真阳性结节 995 个、假阳结节 165 个,敏感度 77.67%,人均误诊结节数为 0.91。B 组的敏感度和人均误诊结节数均较 A 组降低,差异均有统计学意义($P < 0.05$);C 组的敏感度较 B 组提高,且人均误诊结节数降低,差异均有统计学意义($P < 0.05$);C 组的敏感度较 A 组提高,人均误诊结节数降低,差异均有统计学意义($P < 0.05$)。**结论:**疲劳显著降低了放射科规培医师对肺结节的检测效能,但在 AI 软件辅助下能明显提高疲劳状态下放射科规培医师对肺结节的检出效能,甚至超过其正常状态下的水平。

【关键词】 肺结节; 人工智能; 计算机辅助检测; 疲劳; 体层摄影术, X 线计算机

【中图分类号】 R814.42; R734.2 **【文献标识码】** A **【文章编号】** 1000-0313(2021)04-0475-06

DOI:10.13609/j.cnki.1000-0313.2021.04.011

开放科学(资源服务)标识码(OSID):



Study on the effect of AI-assisted software on the detection efficiency of pulmonary nodules in fatigued radiological residents in standardized training WANG Liang, XU Di, SUN Dan-dan, et al. Department of Radiology, Zhongshan Hospital, Dalian University, Dalian 116001, China

【Abstract】 Objective: The purpose of this study was to investigate whether the assistance of artificial intelligence (AI)-assisted pulmonary nodule detection software could improve the detection efficiency of pulmonary nodules in fatigued radiological residents with standardized training. **Methods:** A total of 182 patients who underwent chest CT scans was collected. Three readings methods were conducted by a standardization training intern with 4 months chest CT experience: the resident in daily work condition (not fatigued) without AI software (group A), the same fatigued resident (after 8 hours working) without AI-software (group B), and the fatigued resident with AI software (group C). Each of the three readings had a washout period of more than 15 days. The location, size and number of nodules in the three groups were recorded and compared with the gold standard (the diagnosis was made by two intermediate physicians who had been engaged in chest imaging diagnosis for more than 8 years in combination with AI screening results, and was finally confirmed by a senior physician who had been engaged in chest imaging diagnosis for more than 15 years), the number of true positive and false-positive nodules was calculated. The performance of three reading methods regarding lung nodule detection was assessed by sensitivity and the number of false positive nodules (error diagnosis) per pa-

作者单位: 116001 大连, 大连大学附属中山医院放射科(王亮、许迪、孙丹丹、伍建林、于晶); 100025 北京, 北京推想科技有限公司(顾俊)

作者简介: 王亮(1981—), 男, 山东文登人, 主治医师, 主要从事胸部影像诊断及人工智能辅助诊断临床应用工作。

通信作者: 于晶, E-mail: yujing1969@163.com

tient (ED/P). **Results:** A total of 1281 nodules was confirmed to be gold standard nodules. Group A detected 592 true positive nodules and 297 false-positive nodules with a sensitivity of 46.21%, and the number of false positive nodules was 1.63 per CT scan. Group B detected 517 true positive nodules and 225 false-positive nodules with a sensitivity of 40.36%, and ED/P was 1.24. Group C detected 995 true positive nodules and 167 false-positive nodules with a sensitivity of 77.67%, and ED/P was 0.91. The sensitivity and ED/P in group B were lower than those in group A, and the differences were statistically significant ($P < 0.05$); the sensitivity of group C was higher than that of group B, and ED/P was reduced, and the differences were both statistically significant ($P < 0.05$); the sensitivity of group C was higher than that of group A, ED/P was reduced, and the differences were statistically significant ($P < 0.05$). **Conclusion:** Fatigue can reduce the detection efficiency of pulmonary nodules for radiologists. However, fatigued intern aided with an AI-assisted lung nodule detection system could achieve a higher detection rate for pulmonary nodules, even exceeding the results obtained in their routine working state without fatigue.

【Key words】 Artificial intelligence; Computed aided diagnosis; Pulmonary nodules; Fatigue; Tomography, X-ray computed

近几年来,国内外学者越来越关注医疗差错、疲劳等因素在医学诊断和治疗等方面产生的影响,尤其是在放射领域。在放射医师中,回顾性分析发现首诊错误率(假阳性和假阴性)高达约 30%^[1],其中疲劳所致的诊断错误占有相当高的比例。随着 CT 薄层图像技术的广泛应用,图像负荷骤然增加,导致放射科医师工作量及工作时长日益增长。近年来人工智能技术发展迅速,基于人工神经网络的智能计算机辅助诊断系统已逐渐嵌入临床工作流程之中^[2],尤其以基于深度学习的肺结节检测软件的表现更为抢眼。人工智能辅助肺结节检测系统作为放射科医师的“密友”,有助于提高对肺结节的检出率和准确性^[3]。国内外众多学者对不同的类似软件及模型在辅助临床工作中的价值和限度等进行了深入研究和分析,但少有学者研究其对疲劳状态下放射科医师工作效能的影响。目前放射科处于一线诊断工作的医师绝大多数为规培医师。本文旨在探究 AI 辅助软件能否提高疲劳状态下放射科规培医师在临床工作中对肺结节的诊断效能。

材料与方法

1. 临床资料

前瞻性将 2019 年 1 月 1 日—1 月 10 日在本院使用 2 代炫速双源 CT 行胸部 CT 检查的 200 例患者纳入观察。图像纳入标准:①深吸气后屏气状态下连续扫描图像;②无金属伪影、呼吸伪影等原因导致的 CT 图像质量不佳;③无肺炎、肺不张和胸腔积液等病变而影响病灶的确认及标注者;④排除肺内无结节或结节数量过多(结节总数大于 20 个)的患者;⑤排除肺叶或肺段切除术后患者。最终共纳入符合条件的 182 例患者,女 103 例,男 79 例,年龄 21~77 岁,平均(49.5 ±

14.2)岁。

2. 图像采集和重建

使用 Somatom Definition Flash CT 机,常规行 CT 平扫,扫描范围自胸廓人口至后肋膈角尖端水平,于吸气末单次屏气内完成扫描。标准扫描方案:管电流采用自动曝光控制技术,120 kV,准直宽度 128×0.6 mm,螺距 1.2,常规重建层厚 1.0 mm、层距 0.7 mm。图像采集完成后分别采用高分辨算法(B70f)和标准算法(B30f)自动重建原始扫描数据,并将所有图像传输至 PACS 和人工智能服务器(InferRead CT Lung Research, Infervision, Beijing, China)进行肺结节自动检测,获得检测结果(包括结节数目、长径及位置等)。

3. 金标准及阅片方式

由 2 位从事胸部影像诊断工作超过 8 年的中级医师结合 AI 辅助软件的检测结果逐层阅片并记录结节数目、长径及位置,两人检测结果交由第 3 位从事胸部影像诊断工作超过 15 年的主任医师进行审核,做出最终诊断作为金标准。

受试者为影像专业规范化培训医师(从事胸部影像诊断工作 4~12 个月),测试前将肺结节标注标准告知受试者并进行培训(依据 2018 年《胸部 CT 肺结节数据标注与质量控制专家共识》进行培训,达到能够在正常工作状态下准确识别各类型结节)。受试者按以下 3 种模式下进行阅片:模式 A 为不借助 AI 辅助软件,受试者在正常状态下(即一天工作之前)独立阅读每个病例的 CT 图像;模式 B 为两周过后(洗脱期),受试者在疲劳状态下(即一天工作满 8 小时后)不借助 AI 辅助软件再次阅片;模式 C 为再次洗脱期过后,受试者在疲劳状态下结合 AI 辅助软件的检测结果再次

阅片。三组阅片结果分别以 Excel 表格形式记录结节的位置、大小和数量。每次阅片之后,将肺结节的检测结果与金标准进行比对,并由受试者回顾分析每次漏检结节的原因并记录。

4. 统计学分析

分别计算 3 组中检出肺结节的敏感度、(患者)人均误诊结节数。敏感度计算方法为每组真阳性结节总数/总结节数;人均误诊结节数计算方法为每组假阳性结节数(误诊结节总数)/CT 检查患者例数。比较三种阅片模式对肺结节的检测能力。进一步将结节按照大小分为 3 类,分别为长径<4 mm、4 mm≤长径<10 mm和 10 mm≤长径<30 mm^[4],分别计算 B、C 两组中对 3 类肺结节的检出敏感度和人均误诊结节数。利用 SPSS 20.0 统计学软件进行统计学分析,对正态分布的计量资料采用配对样本 *t* 检验,非正态分布的

计量资料的组间比较采用(配对样本)Wilcoxon 秩检验,以 *P*<0.05为差异具有统计学意义。

结 果

按照金标准,200 例患者共检出 1281 个肺结节。A、B 和 C 三种阅片模式对肺结节的检出情况及比较结果见表 1。A、B 模式之间敏感度和人均误诊结节数的差异均具有统计学意义(*P*<0.05),疲劳状态下规培医师对肺结节的检测敏感度下降,但人均误诊结节数也有所有下降。应用 AI 辅助软件后,疲劳状态下规培医师对肺结节的检测效能明显提升,人均误诊结节数下降(图 1),差异均具有统计学意义(*P*<0.05)。应用 AI 辅助肺结节检测软件后,疲劳状态下的规培医师对肺结节的检测效能已经超过其正常状态的水平(图 2),人均误诊结节数下降程度更为明显,差异均具

表 1 正常状态和疲劳状态下对肺结节检测情况的比较

指标	A 组	B 组	C 组	模式 A vs. 模式 B		模式 B vs. 模式 C		模式 A vs. 模式 C	
				Z 值	P 值	Z 值	P 值	Z 值	P 值
真阳性结节数	592	517	995	—	—	—	—	—	—
假阳性结节数	297	225	165	—	—	—	—	—	—
敏感度	46.21%	40.36%	77.67%	-2.582	0.010	-10.399	0.000	-9.681	0.000
人均误诊结节数	1.63	1.24	0.91	-3.263	0.001	-2.039	0.041	-5.458	0.000

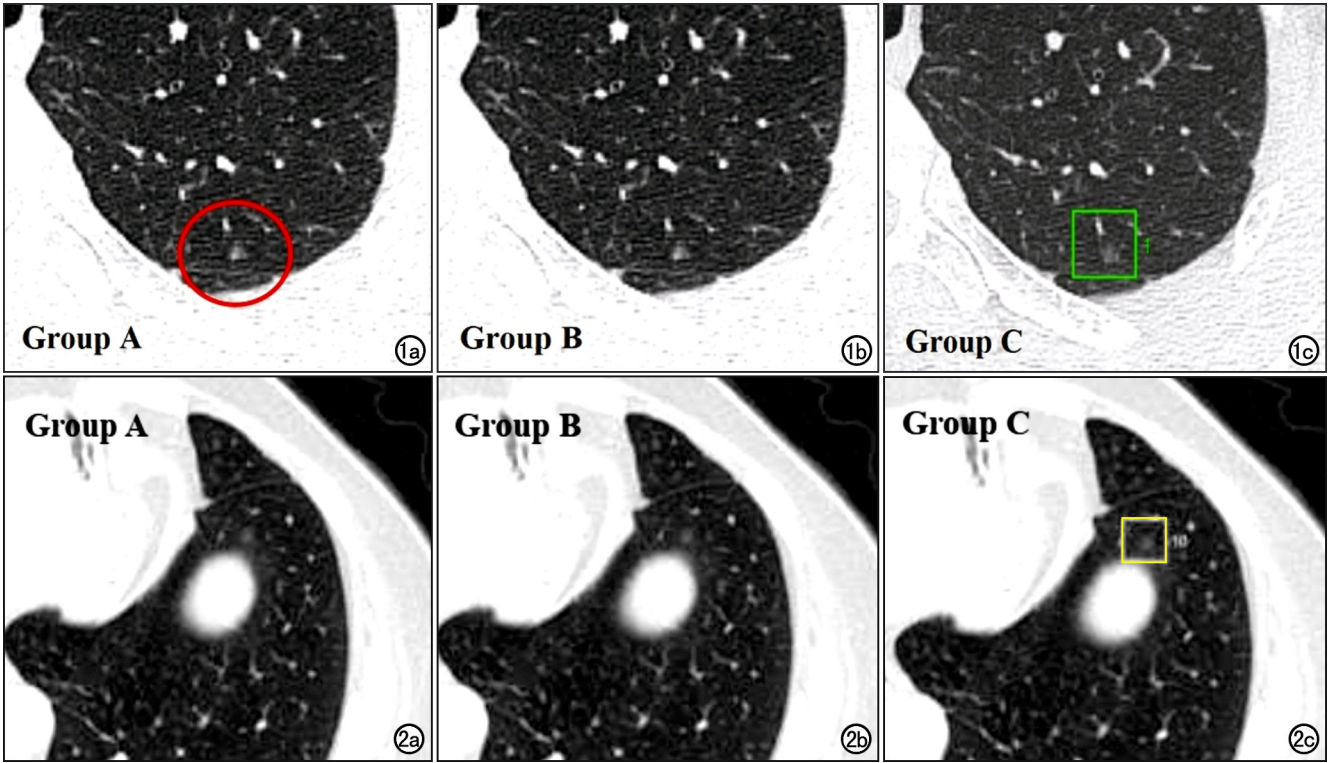


图 1 患者,男,56 岁。金标准确认左肺上叶有一磨玻璃结节。a)正常状态下(模式 A)该结节被检出(红色圆圈);b)在疲劳状态下(模式 B)独立阅片时未检出;c)AI 软件辅助下疲劳状态(C 模式)的规培医师准确检出结节(绿色方框)。图 2 患者,男,57 岁。金标准确认左肺下叶前内基底段有一磨玻璃结节。a)正常状态下(模式 A)该结节未被检出;b)在疲劳状态下(模式 B)独立阅片时未检出此结节;c)AI 软件辅助下疲劳状态(C 模式)的规培医师准确检出此结节(黄色方框)。

有统计学意义($P < 0.05$)。

两种阅片模式对不同大小肺结节的检出情况见表 2。应用 AI 辅助软件后规培医师对 < 4 mm 结节的检出敏感度提升了 40.28%，对 4~10 mm 结节检出敏感度提升了 27.08%。

回顾性分析 191 个漏检结节的特征和原因：64 个结节的长径为 0.9~3.0 mm；28 个结节位于血管旁或与血管走行相延续，短径与血管相仿(图 3)；28 枚结节位于胸膜下；28 个结节为密度对比较低的纯磨玻璃结节；另有 43 个结节被规培医生成功发现，但又将其错误地排除。显示经金标准确认的右肺中叶实性血管旁结节，短径与血管管径相近，在日常工作前和在人工智能辅助下的疲劳状态下的规培医生准确检出，在疲劳状态下独立阅片时虽被观察到，但未被确认为结节，最终漏检该结节。

讨 论

Waite 等^[5]对疲劳在放射学中的影响进行了综述，从本质上讲，它是一种精神疲倦和体能消耗，可以表现在身体和认知两方面^[5]。对于放射科医生来说，长时间的集中注意力工作、全面地观察不同负荷的图像、使用三维图像观察复杂的解剖结构，以及快速阅读的时效要求等诸多方面，都是导致其产生疲劳的因素。

众多因素中，眼睛疲劳和视力模糊的严重程度随着阅读图像的数量和总阅读时间增加而增加，并且保持聚焦变得越来越困难^[6-7]。瑞典职业疲劳量表提供了一个感知疲劳的模型，它通过 5 个方面(缺乏能量、体力消耗、身体不适、缺乏动力和嗜睡)的评分来评价疲劳程度^[8]。Krupinski 等^[9-10]应用瑞典职业疲劳量表进一步研究发现，放射科医师在日常工作中疲劳感是不断蓄积的过程，在日常的临床图像判读工作满 8 小时后，其 SOFI 评分更差，并且阅片诊断的错误率增加、准确性降低。因此，本研究中以日常工作满 8 小时作为疲劳状态的定义。临床工作中，由于疲劳所造成的错误最常见的是遗漏错误和假阳性错误，由此引发的医患双方对簿公堂的案例屡见不鲜。如何破解疲劳对临床工作带来的不利影响，众多学者提出的各种理论和方案尚未达成共识。而人工智能辅助软件的临床应用及其优异表现，给解决临床医师的疲劳问题带来了新的思路。

疲劳状态可导致放射科医师对肺结节的检出率降低、漏诊结节增多及准确性下降^[10]，而这一表现在缺乏工作经验的规培医师中体现得更为明显^[11]。本研究得到了类似结果，B 模式比 A 模式的检出敏感度下降了 5.85%，多漏检 191 个结节，这对于本身诊断经验不足的规培医师来说较正常状态下降了近 12.66%；

表 2 疲劳状态独立阅片与疲劳状态 AI 辅助阅片对不同大小结节检测情况的比较

阅片模式	真阳性 结节数	假阳性 结节数	敏感度	Z 值	P 值	人均误诊 结节数	Z 值	P 值
直径 < 4 mm				-10.201	0.000		-4.001	0.000
B 模式	363	163	36.56%			0.94		
C 模式	763	84	76.84%			0.47		
直径 4~10 mm				-5.466	0.000		-1.408	0.159
B 模式	148	62	53.43%			0.54		
C 模式	223	81	80.51%			0.70		
直径 10~30 mm				-1.732	0.083		0.000	1.000
B 模式	6	0	54.55%			0		
C 模式	9	0	81.82%			0		

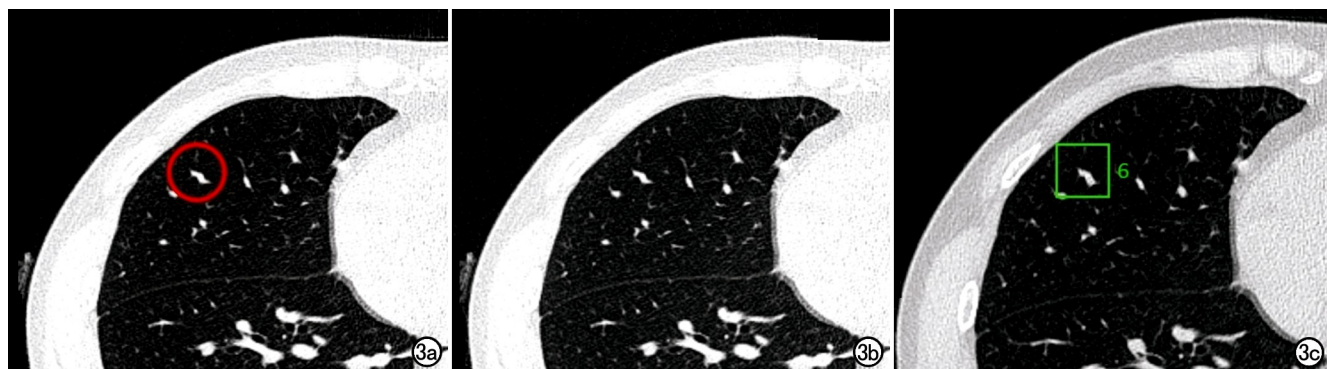


图 3 患者，男，23 岁。金标准确认右肺中叶有一实性结节位于血管旁。a) 正常状态下(模式 A)此结节被准确检出(红色圆圈)；b) 在疲劳状态下(模式 B)独立阅片时虽观察到此结节但未确认，最终导致漏检此结节；c) AI 软件辅助下疲劳状态(C 模式)的规培医师准确检出此结节(绿色方框)。

尽管,人均误诊结节数下降了 0.39,但其主要原因并非单纯误诊结节数的减少,而是与相对应的患者检查例数减少,进而导致检出结节总数(A 组 889 个、B 组 742 个)的下降有一定关系。准确检出直径过小、密度对比差及位置隐匿的结节对规培医师来说即使在正常状态下独立阅片时也是一件不容易的事,此类结节在疲劳状态下易出现漏检是在意料之中的,这与类似 CAD 辅助检测结节的研究结果相符^[11-12]。我们推测规培医师之所以会错误的排除结节,可能与规培医师的诊断经验不足、以及疲劳后认知能力、记忆力和注意力下降等原因有一定关系^[5]。

使用 AI 辅助软件后,疲劳状态下漏检的 191 个结节中有 154 个结节被正确检出,约占漏检结节数的 80.6%。不仅如此,敏感度也有近乎成倍的提升,从 40.36% 提升至 3 组最高的 77.67%,提升率达 37.31%,且人均误诊结节数降至 3 组最低的 0.91。由此可见,疲劳在肺结节检测中所引起的错误(假阴性、假阳性)绝大部分被纠正,同时带来了准确性的大幅度提升。但在 AI 软件辅助下仍有漏检结节的存在,其中包括规培医师与 AI 软件共同漏检的结节,以及由于经验不足或者认知能力下降所致的漏检结节。如果随着规培医师临床诊断经验的不断积累和提升,对此类结节的漏检概率会显著降低。

进一步对不同大小肺结节的检测情况进行分析,放射科规培医师在疲劳状态下应用 AI 辅助软件后,<4mm 结节组中检出敏感度大幅度提升,程度超过 110%,人均误诊结节数下降也较为明显。一方面说明越小的结节越容易被漏诊,尤其是在放射科医师在长时间工作产生疲劳之后,这符合临床上对肺结节检测的普遍经验;另一方面说明 AI 人工智能辅助肺结节检测软件能够“不知疲倦”的达到稳定的检出效能;同时软件检出的结节需要与医师互动,进行人工再判断,这对于诊断经验有限的规培医师,人均误诊结节数的下降表现也是符合预期的。假阳性结节数的减少可以让临床医师无需对这些无临床价值的结节耗费一天中有限的精力,同时让呈现此类结节表现的患者避免不必要的紧张和焦虑。我们注意到,4~10 mm 结节组中 B 模式与 C 模式之间检出敏感度的差异具有统计学意义($P<0.05$),但在人均误诊结节数方面的差异并无统计学意义($P>0.05$)。相对于<4 mm 结节来说,4~10 mm 结节的形态学特征更为复杂,这对于规培医师来说,正确分析判断具有一定的难度,人均误诊结节数略微升高情有可原。

本研究的局限性:首先,金标准没有利用外部参考标准(如组织学证据),而是采用两位从事胸部影像诊

断的中级医师及一位主任医师(工作时间超过 15 年)的 3 次阅片的共识作为内部参照标准,这一局限性在评价 CAD 效能的研究中普遍存在;其次,本研究中并没有对两次疲劳程度进行量化,我们考虑到造成电脑终端操作者视觉疲劳的原因包括众多不可控的客观因素及复杂的主观因素,不可能精确控制或重复,而且本研究中 10~30 mm 的结节数量较少,对这类结节的检测评价需要更大数据量的研究来进一步分析探讨;最后,本研究结果仅代表一个特定算法的在特定参数时 CAD 的表现,并不能完全等同于其它类似软件。本研究的重点是阅片者和 CAD 之间的交互所能产生的提升效果,并非单纯对软件性能的评价。

综上所述,疲劳确实能让放射科规培医师对肺结节诊断效能下降,应用 AI 辅助肺结节检测软件能帮助其提升诊断效能,甚至超越其正常状态的诊断水平,对直径<4 mm 结节检出率的提升尤为显著。

参考文献:

- [1] Makary MA, Daniel Mi. Medical error: the third leading cause of death in the US[J/OL]. BMJ, 2016, 353: e2139. DOI: 10.1136/bmj.i2139.
- [2] 金征宇. 人工智能医学影像应用: 现实与挑战[J]. 放射学实践, 2018, 33(10): 989-991.
- [3] 萧毅, 刘士远. 肺结节影像人工智能技术现状与思考[J]. 肿瘤影像学, 2018, 27(4): 8-11.
- [4] 蔡强, 吴宁, 周丽娜, 等. 计算机辅助检测系统在低剂量 CT 筛查者非钙化肺结节中的应用研究[J]. 中华放射学杂志, 2014, 48(1): 63-66.
- [5] Waite S, Kolla S, Jeudy J, et al. Tired in the reading room: the influence of fatigue in radiology[J]. JACR, 2017, 14(2): 191-197.
- [6] Krupinski EA, Kallergi M. Choosing a radiology workstation: technical and clinical considerations[J]. Radiology, 2007, 242(3): 671-682.
- [7] Krupinski EA. Current perspectives in medical image perception[J]. Atten Percept Psycho, 2010, 72(5): 1205-1217.
- [8] Ahsberg E, Gamberale F, Gustafsson K, et al. Perceived fatigue after mental work: an experimental evaluation of a fatigue inventory[J]. Ergonomics, 2000, 43(2): 252-268.
- [9] Krupinski EA, Berbaum KS, Caldwell RT, et al. Do long radiology workdays affect nodule detection in dynamic CT interpretation[J]. J Am Coll Radiol, 2012, 9(3): 191-198.
- [10] Krupinski EA, Berbaum KS, Caldwell RT, et al. Long radiology workdays reduce detection and accommodation accuracy[J]. J Am Coll Radiol, 2010, 7(9): 698-704.
- [11] 胡琼洁, 夏黎明, 陈冲, 等. 实习医师以共同阅片和第二阅片者模式使用计算机辅助检测系统在低剂量 CT 中的应用研究[J]. 放射学实践, 2018, 33(10): 1022-1028.
- [12] 蔡雅倩, 张正华, 韩丹, 等. AI 对肺磨玻璃结节筛查及定性的临床应用研究[J]. 放射学实践, 2019, 34(9): 958-962.

(收稿日期: 2020-05-611 修回日期: 2020-11-15)