Journal of the American Medical Informatics Association, 2023, **30(10)**, 1684–1692
https://doi.org/10.1093/jamia/ocad118
Advance access publication 10 August 2023
Research and Applications

Research and Applications

Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias

Ding-Yu Wang[1,2,3], Jia Ding[4], An-Lan Sun[4], Shang-Gui Liu[1,2,3], Dong Jiang[1,2,3], Nan Li[5,*], and Jia-Kuo Yu[1,2,3,*]

[1]Department of Sports Medicine, Peking University Third Hospital, Institute of Sports Medicine of Peking University, Beijing, China
[2]Beijing Key Laboratory of Sports Injuries, Beijing, China
[3]Engineering Research Center of Sports Trauma Treatment Technology and Devices, Ministry of Education, Beijing, China
[4]Beijing Yizhun Medical AI Co., Ltd, Beijing, China
[5]Research Center of Clinical Epidemiology, Peking University Third Hospital, Beijing, China

*Corresponding Authors: Jia-Kuo Yu, MD, PhD, Department of Sports Medicine, Peking University Third Hospital, No.49 North Garden Road, Beijing 100191, China; yujiakuo@126.com; Nan Li, PhD, Research Center of Clinical Epidemiology, Peking University Third Hospital, No.49 North Garden Road, Haidian, Beijing 100191, China; linan917@163.com
Ding-Yu Wang and Jia Ding contributed equally to this study as co-first authors.

ABSTRACT

Background: Incorporating artificial intelligence (AI) into clinics brings the risk of automation bias, which potentially misleads the clinician's decision-making. The purpose of this study was to propose a potential strategy to mitigate automation bias.

Methods: This was a laboratory study with a randomized cross-over design. The diagnosis of anterior cruciate ligament (ACL) rupture, a common injury, on magnetic resonance imaging (MRI) was used as an example. Forty clinicians were invited to diagnose 200 ACLs with and without AI assistance. The AI's correcting and misleading (automation bias) effects on the clinicians' decision-making processes were analyzed. An ordinal logistic regression model was employed to predict the correcting and misleading probabilities of the AI. We further proposed an AI suppression strategy that retracted AI diagnoses with a higher misleading probability and provided AI diagnoses with a higher correcting probability.

Results: The AI significantly increased clinicians' accuracy from 87.2%±13.1% to 96.4%±1.9% ($P < .001$). However, the clinicians' errors in the AI-assisted round were associated with automation bias, accounting for 45.5% of the total mistakes. The automation bias was found to affect clinicians of all levels of expertise. Using a logistic regression model, we identified an AI output zone with higher probability to generate misleading diagnoses. The proposed AI suppression strategy was estimated to decrease clinicians' automation bias by 41.7%.

Conclusion: Although AI improved clinicians' diagnostic performance, automation bias was a serious problem that should be addressed in clinical practice. The proposed AI suppression strategy is a practical method for decreasing automation bias.

Key words: deep learning, automation bias, AI suppression, clinician-AI interaction

INTRODUCTION

In recent years, there have been significant advancements in the application of artificial intelligence (AI) in the medical field, with various algorithms being successfully integrated into clinical practice.[1,2] This has led to a rapid transition towards an AI-driven healthcare system, in which AI augments clinicians' capabilities and contributes to improved diagnostic accuracy and efficiency.[3–7] However, this also brings the risk of automation bias[5,8,9]; that is, the clinician over-accepts the inappropriate advice of an automated system, ignoring contrary data or conflicting human decisions.[10,11] The automation bias problem potentially affects clinician decision-making for millions of patients.[12]

The causes of automation bias are complex and multifaceted, including factors such as the robustness of the automated system, task load and complexity, clinicians' experiences, trust in the system, and awareness of automation bias.[11,13] To mitigate this issue, both upgrading AI performance and improving user practices are necessary. Currently, efforts to address automation bias tend to focus on technological solutions, such as reducing bias in algorithms, while neglecting the importance of user practice.[14] Although algorithms can be optimized to have high area under the receiver-operating characteristic curve (AUC) values, such as 0.95 or 0.99, it is inevitable that AI systems produce errors in real-world clinical environments with unpredictable events, and clinicians are confronted with a potentially inaccurate system output. Thus, it is important to support user practices that retain and promote the clinician's initiative, as the integration of AI algorithms into clinics and the increasing reliance on these tools increases the risk of automation bias.

In this research, we utilized the diagnosis of anterior cruciate ligament (ACL) rupture, a prevalent sports injury,[15] as a case study to evaluate the impact of automation bias on the decision-making of clinicians with varying levels of expertise. By analyzing the clinician-AI interaction, we proposed a potential AI suppression strategy for mitigating automation bias from the user perspective.

Received: 27 August 2022. **Revised:** 30 May 2023. **Editorial Decision:** 11 June 2023. **Accepted:** 19 June 2023

MATERIALS AND METHODS

Our study was conducted with approval from the Peking University Third Hospital (PUTH) ethical committee (IRB00006761-M2020243). Our Review Board waived the requirement for informed consent from patients as this laboratory study solely simulated clinicians' interaction with patient data, did not interfere with actual clinical practice, and no patient-identifying information was collected throughout the study. Additionally, consent from clinicians was also waived by the ethics committee because this was a laboratory study.

ACL rupture detection system

The ACL rupture detection system was developed using a training dataset of 8484 magnetic resonance imaging (MRIs) (Supplementary Table S1) and employed a deep learning architecture incorporating ResNet50[16] and Siamese[17] algorithms. In a preliminary evaluation, the system demonstrated a sensitivity of 90.0%, a specificity of 85.3%, and an area under the receiver-operating characteristic curve of 0.953 when tested on a validation dataset of 2273 prospectively collected knee MRIs (Supplementary Table S1).

Dataset

Two hundred prospectively collected cases, including the patient's medical history, physical examination, and knee MRIs, were collected from the sports medicine and orthopedics clinic of PUTH. All the data used in this study were anonymized and used in the clinician-AI interaction test. Patient identifying information, such as name, was replaced with a code during medical history collection. Any identifying information in the MRI, such as name, was erased using anonymization software. The patient's postsurgical diagnosis is considered the gold standard for reference MRI. In cases where the patient did not undergo surgery, the diagnosis of anterior ACL injury was determined by a panel of 2 senior sports medicine surgeons and 1 senior musculoskeletal (MSK) radiologist with over 10 years of experience in diagnosing ACL injuries. They independently made the diagnosis based on the patient's medical history, physical examination, and MRI. Any inconsistencies in diagnosis were discussed among the panel to arrive at a final diagnosis.

Clinician participants

Forty clinicians from 20 hospitals were invited to participate in the clinician-AI interaction test (Table 1). They were divided into 3 groups: the sports medicine expert group, the sports medicine trainee group, and the nonsports medicine clinician group.

The sports medicine expert group consisted of 9 senior experts from PUTH. All of these clinicians were attending physicians with a sports medicine fellowship and had 5–10 years of experience in diagnosing ACL injuries.

The sports medicine trainee group consisted of 16 clinicians from 12 hospitals (including PUTH). These clinicians were sports medicine fellows undergoing sports medicine training, with some limited exposure to ACL rupture cases.

The nonsports medicine clinician group consisted of 15 clinicians from 9 hospitals (including PUTH). These clinicians were attending physicians with an orthopedics fellowship and had 5–10 years of experience in general orthopedics, and would encounter ACL patients in their daily work but did not receive systematic sports medicine training.

Preparation for the clinician-AI interaction test

Previous research has shown that the end-user performs better, and their trust in a machine is calibrated when an algorithmic decision is presented with a low confidence index.[18] The confidence index is related to the absolute value of the model output, which ranges from 0 to 1. However, it can be difficult for clinicians to decipher the model's direct output into a confidence level. To address this issue, we divided the AI output range into 9 parts and calculated the accuracy (positive/negative predictive value) of the AI diagnosis for each part, using a preliminary validation dataset of 2273 MRIs (Supplementary Table S1), to represent the confidence.

The ACL rupture detection system processed the 200 MRIs, and AI diagnoses with confidence were prepared for the clinician-AI interaction test. The clinician-AI interaction test was conducted on a website. The user interface for the clinician-AI interaction test is shown in Figure 1.

Prior to the test, user education was provided to the clinicians, including information on the model's performance in the preliminary study, the meaning of the model's diagnosis, and the confidence level. Examples of both correct and incorrect AI diagnoses were presented to the clinicians. It was emphasized that they should not completely rely on the AI system, particularly when the confidence level is low.

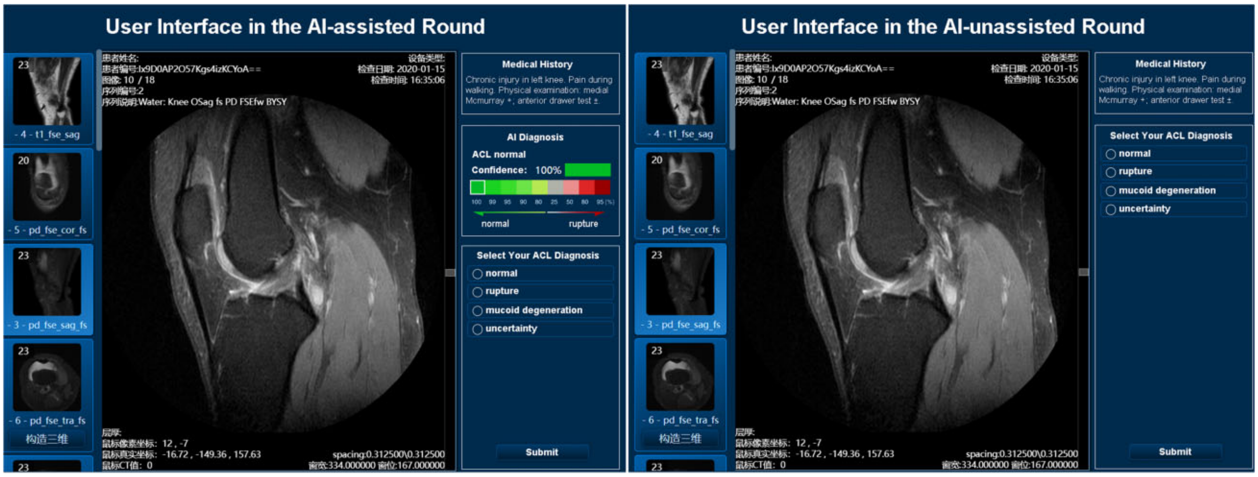
Clinician-AI interaction test

A randomized cross-over design was implemented. Clinicians were assigned to either the AI-assisted group or the AI-unassisted group using block randomization, stratified by physician category. In the first round, clinicians were asked to diagnose 200 MRIs with or without AI assistance. After the completion of the first round, there was a washout period of 14 days, after which the clinicians switched to the AI-assisted or AI-unassisted group and rediagnosed the previous 200 MRIs. In each round, clinicians were asked to select 1 of 4 choices for the status of the ACL: normal, rupture, mucoid degeneration, and uncertainty.

The clinicians' sensitivity, specificity, accuracy, and reading time in the 2 rounds were calculated and compared. The

Table 1. Demographics of clinician participants

	Overall	Sports medicine expert	Sports medicine trainee	Nonsports medicine clinician
Number	40	9	16	15
Age	36 (28–45)	38 (33–42)	34 (28–40)	37 (31–45)
Sex (male: female)	34:6	5:4	15:1	14:1
Sports medicine training	–	Completed	Undergoing	None
Year of expertise (beyond resident)	–	5–10	1–3	5–10
Centers	20 hospitals	PUTH only	12 hospitals	9 hospitals



User Interface in the AI-assisted Round

患者姓名:
患者编号:lx9D0AP2O57Kgs4izKCYoA==
图像: 10 / 18
序列编号:2
序列说明:Water: Knee OSag fs PD FSEfw BYSY
设备类型:
检查日期: 2020-01-15
检查时间: 16:35:06
层厚:
鼠标像素坐标: 12 , -7
鼠标真实坐标: -16.72 , -149.36 , 157.63
鼠标CT值: 0
spacing:0.312500\0.312500
窗宽:334.000000 窗位:167.000000

Medical History

Chronic injury in left knee. Pain during walking. Physical examination: medial Mcmurray +; anterior drawer test ±.

AI Diagnosis

ACL normal

Confidence: 100%

Select Your ACL Diagnosis

○ normal

○ rupture

○ mucoid degeneration

○ uncertainty

Submit

User Interface in the AI-unassisted Round

患者姓名:
患者编号:lx9D0AP2O57Kgs4izKCYoA==
图像: 10 / 18
序列编号:2
序列说明:Water: Knee OSag fs PD FSEfw BYSY
设备类型:
检查日期: 2020-01-15
检查时间: 16:35:06
层厚:
鼠标像素坐标: 12 , -7
鼠标真实坐标: -16.72 , -149.36 , 157.63
鼠标CT值: 0
spacing:0.312500\0.312500
窗宽:334.000000 窗位:167.000000

Medical History

Chronic injury in left knee. Pain during walking. Physical examination: medial Mcmurray +; anterior drawer test ±.

Select Your ACL Diagnosis

○ normal

○ rupture

○ mucoid degeneration

○ uncertainty

Submit

Figure 1. The user interface of the clinician-AI interaction test. The MRI images with patient's medical history were displayed. The AI diagnosis part with both the AI diagnosis and the confidence was shown in the AI-assisted round. An indicator color was introduced to offer the clinicians perceptual intuition of the AI diagnosis and its confidence. In the AI-unassisted round, the AI diagnosis part was hidden while other parts remained unchanged.

reading time for each MRI was defined as the period from when the clinician opened the MRI to when they submitted their diagnosis, and was recorded in seconds.

The clinician's internal diagnosing threshold in terms of probability was calculated using a binary logistic regression model, as described by Plasencia *et al.*[19] Briefly, the following relationship existed between the probability that physicians gave an ACL rupture diagnosis and the risk of ACL rupture calculated by the AI:

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta r, \qquad (1)$$

where P indicates the clinician's ACL rupture diagnosis probability, r indicates the risk of ACL rupture, α is the intercept coefficient, and β is the regression coefficient for ACL rupture risk. The clinician's internal diagnosing threshold was defined as the value of risk at which clinicians are equally likely to diagnose normal ACL or ACL rupture. In such cases, $P = .5$ and $r = -\alpha/\beta$, according to Equation (1).

We also conducted a semistructured interview with 20 randomly selected clinicians following the test (Supplementary Table S3). Six open-ended questions were asked to investigate the clinicians' perceptions of AI and their strategies for utilizing it (see Supplementary Materials—Qualitative study). The interviews were analyzed using the grounded theory method. A total of 15 nominal codes were identified, and 5 subcategories were obtained.

Clinician-AI interaction analysis

We defined a clinician-AI interaction as one clinician diagnosing one MRI with AI assistance. In total, there were 8000 clinician-AI interactions (200 MRIs×40 clinicians). Based on the correctness of the clinician's AI-unassisted and AI-assisted diagnosis, each clinician-AI interaction had 3 outcomes: a correcting event, a misleading event (automation bias), and a null event. A misleading event (automation bias) was defined as the clinician's original correct diagnosis being misled by a wrong AI diagnosis. A correcting event was defined as the clinician's incorrect diagnosis being corrected by a correct AI diagnosis. A null event was defined as the clinician's diagnosis not being corrected or misled by the AI diagnosis.

The correcting, misleading, and null events among the 8000 clinician-AI interactions in the clinician-AI interaction test were identified. We further calculated the proportion of correcting and misleading events in the clinician-AI interactions in 9 ranges with different AI diagnoses and powers in Supplementary Table S2.

Correcting proportion=

$$\frac{\text{Correcting event number}}{\text{Clinician} - \text{AI interaction number}} \times 100\%.$$

Misleading proportion=

$$\frac{\text{Misleading event number}}{\text{Clinician} - \text{AI interaction number}} \times 100\%.$$

An ordinal logistic regression model was used to predict the probability of correcting and misleading events with a given model output value. We used data from the clinician-AI interaction test to build the logistic regression model. The independent variables were the AI output value (X, continuous), clinician group (G, categorical), and degree of difficulty of each MRI (D, clinician average error rate of certain MRIs without AI assistance, continuous). The difficulty of the MRI was calculated as $D = \frac{\text{Wrong diagnosis number from 40 clinicians}}{\text{Total diagnosis number from 40 clinicians}}$. The dependent variables were misleading events ($Y = 0$), null events ($Y = 1$), and correcting events ($Y = 2$), which were assigned as ordinal variables because the outcomes of the misleading, null, and correcting events varied from negative to positive. As the association between the AI correcting/misleading effect and the model output probability was potentially nonlinear, we used cubic splines with 3 degrees of freedom to fit.

The primary model was as follows:

$$\ln\left(\frac{P(Y \le i)}{1 - P(Y \le i)}\right) = \beta_{0i} - \left(\beta_1 D + \beta_{2j} G_j + f(X)\right) + \varepsilon, \qquad (2)$$

where $i = 0$, 1, and 2 indicate a misleading, null, and correcting event, respectively. $j = 0$, 1, and 2 indicate the sports medicine expert group, sports medicine trainee group, and

Table 2. Clinicians' performances with and without AI assistance

	Overall	Sports medicine expert	Sports medicine trainee	Nonsports medicine clinician
Accuracy (%)				
AI-unassisted	87.2 ± 13.1	96.5 ± 1.1	89.8 ± 9.0	79.0 ± 15.7
AI-assisted	96.4 ± 1.9	98.2 ± 0.9	96.2 ± 1.8	95.5 ± 1.7
P value	<.001	.006	.002	.001
Sensitivity (%)				
AI-unassisted	90.9 ± 12.1	92.5 ± 2.1	92.9 ± 6.0	87.7 ± 18.3
AI-assisted	93.8 ± 4.2	96.1 ± 1.9	93.6 ± 3.7	92.6 ± 5.1
P value	.135	.015	.844	.683
Specificity (%)				
AI-unassisted	86.8 ± 15.5	97.6 ± 1.4	89.4 ± 11.6	77.4 ± 18.1
AI-assisted	97.4 ± 2.1	98.8 ± 1.2	97.2 ± 2.2	96.8 ± 2.0
P value	<.001	.011	<.001	.002
Interpretation time (in seconds)				
AI-unassisted	14.0 ± 5.3	8.8 ± 2.0	14.9 ± 3.2	16.1 ± 6.2
AI-assisted	9.4 ± 4.6	8.4 ± 2.6	10.3 ± 4.3	8.9 ± 5.6
P value	<.001	.722	.011	.003

Data are shown as the mean±SD. Accuracy, sensitivity, and specificity: Wilcoxon matched-pair signed-rank test.

nonsports medicine clinician group, respectively. $f(X)$ represents a cubic spline function of the output probability of the model with 3 degrees of freedom. β_0 represents a fixed intercept. ε represents a random intercept. The model was conducted using R (version 4.1.0; R Development Core Team) with the packages "rms" and "segmented."

In the logistic regression model, we first calculated the probability of Y switching from 1 to 2 (correcting event, represented by a green point) and 1 to 0 (misleading event, represented by a red point). The cubic splines were used to fit the correcting and misleading probability (green and red lines). By calculating the intersection point between the 2 lines, we were able to identify the model output range where the predicted misleading event probability exceeded the correcting probability.

Based on the logistic regression model, we proposed an AI suppression strategy. Under this strategy, AI diagnoses with a higher misleading probability would be retracted, and clinicians would be allowed to make their own decisions in order to decrease automation bias, while AI diagnoses with a higher correcting probability would be retained to augment clinicians' capabilities.

We simulated a third-round clinician-AI interaction test with this AI suppression strategy to examine whether it would decrease automation bias. It was assumed that the clinicians' diagnoses would switch to their AI-unassisted diagnoses when the AI suppression strategy was used. The third-round test results were simulated by combining the clinician's AI-assisted diagnosis when the AI diagnosis had a higher correcting probability and the clinician's AI-unassisted diagnosis when the AI diagnosis had a higher misleading probability in the 2 rounds above.

Statistics

A paired t test was used to compare the clinicians' accuracy, sensitivity, specificity, and interpretation time between the AI-assisted and AI-unassisted rounds. The logistic regression model of the clinician's internal diagnosing threshold and clinician-AI interaction was described earlier. A post hoc analysis was used to calculate the power of the subgroup sample size in terms of accuracy (sports medicine expert, $1-\beta=0.98$; sports medicine trainee, $1-\beta=0.92$; nonsports medicine clinician, $1-\beta=0.99$).

RESULTS

AI increased clinicians' overall accuracy

The application of AI significantly increased the clinicians' accuracy from 87.2%±13.1% to 96.4%±1.9% ($P<.001$) (Table 2). The clinicians' sensitivity and specificity were improved, which increased the utility of the MRI in the ACL diagnosis.[20] The assistance of AI changed the clinicians' internal diagnostic threshold from 0.531 to 0.563.

Although the AI increased the overall accuracy of the clinicians, this benefit usually occurred at both ends of the AI output, with the clinicians' accuracy reaching 100% (Figure 2). However, when the AI output probability was approximately 3.4×10^{-3} to 9.9×10^{-1}, their AI-assisted accuracy dropped, and it was even below the AI-unassisted accuracy of approximately 5.3×10^{-2} to 5.3×10^{-1} (Figure 2, Overall).

AI provided both correcting and misleading effects

We identified 833 (10.4%) correcting events and 132 (1.7%) misleading events among the 8000 clinician-AI interactions in the test (Table 3). The clinicians' mistakes in the AI-assisted round were associated with misleading events, accounting for 45.5% (132/290) of the total mistakes, indicating widespread automation bias. However, all clinicians denied that the AI had any misleading effect on themselves in the interviews (eg, "The mistakes of the AI are obvious. I don't think the AI misleads me."). They only reported the positive aspects of the AI in the interview (Supplementary Table S2).

The distributions of correcting and misleading events were different (Figure 3). More correcting events occurred below the probability of 6.9×10^{-4} and over the probability of 9.9×10^{-1}, while more misleading events occurred in the range of 5.3×10^{-2} to 5.3×10^{-1}, where the AI diagnosis confidence was 25%. Although we warned that the AI diagnosis might be unreliable when the confidence was low before the test, clinicians still made more mistakes under these circumstances.

The level of expertise of the clinician also influenced the likelihood of being corrected or misled. Clinicians with less

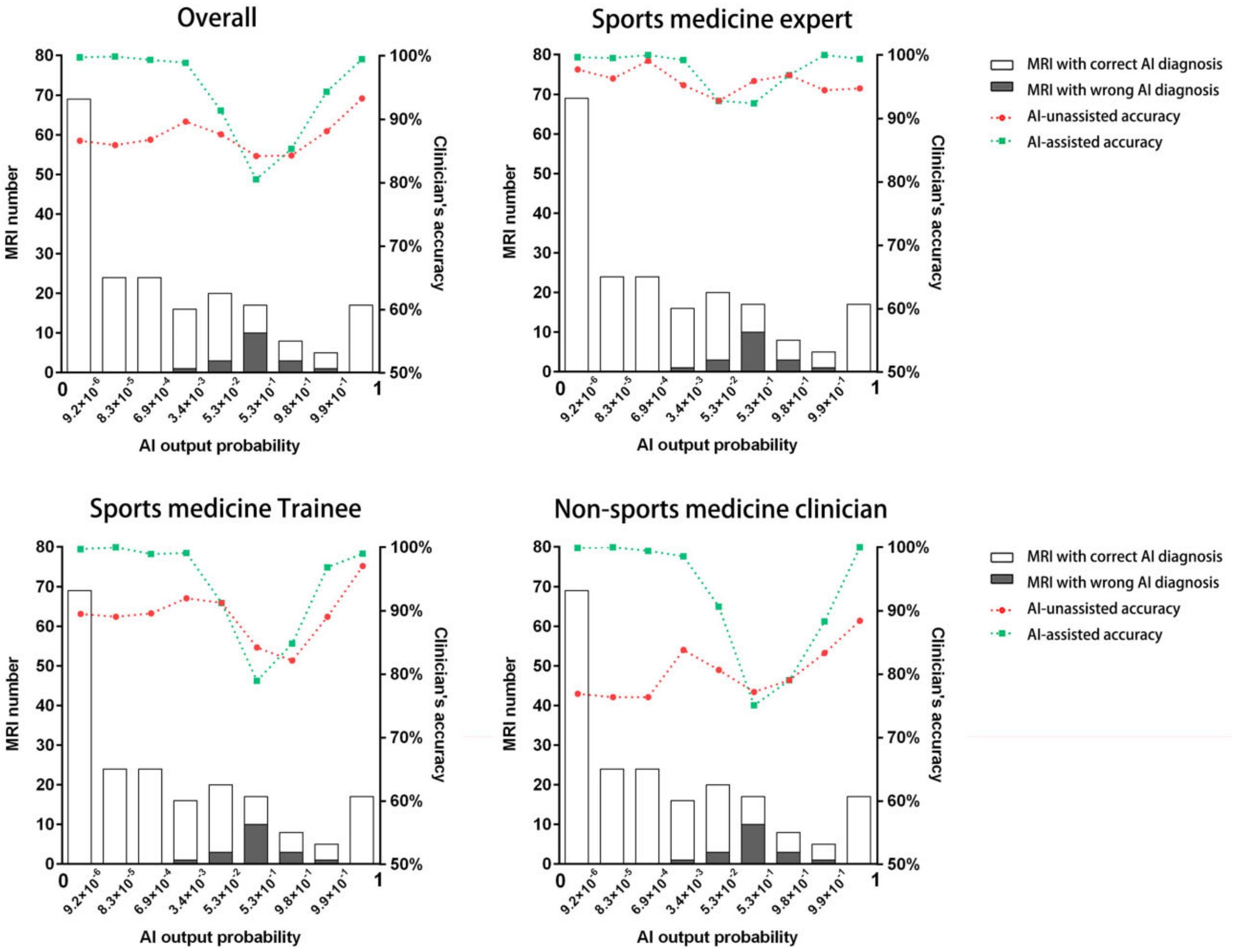


Figure 2. The AI's and clinician's accuracy in different AI output probability ranges. The clinician's AI-assisted accuracy was highly associated with the AI output range. Clinicians' AI-assisted accuracy might drop below the AI-unassisted accuracy in a certain range.

Table 3. List of the requirements and proportions for correcting, misleading, and null events in the clinician-AI test

AI diagnosis	Clinician's AI-unassisted diagnosis	Clinician's AI-assisted diagnosis	Event	Proportion (%)
Right	Right	Right	Null	79.9
Right	Wrong	Wrong	Null	0.3
Right	Right	Wrong	Null	0.3
Right	Wrong	Right	Correcting	10.4
Wrong	Right	Right	Null	5.3
Wrong	Wrong	Wrong	Null	1.3
Wrong	Right	Wrong	Misleading (automation bias)	1.7
Wrong	Wrong	Right	Null	0.8

experience tended to rely more on the AI and were more vulnerable to AI mistakes (sports medicine expert baseline odds ratio [OR]: 1; sports medicine trainee OR: 1.92, 95% CI: 1.37–2.71; and nonsports medicine clinician OR: 4.73, 95% CI: 3.36–6.67).

The AI suppression strategy reduced the automation bias

Given that the correcting and misleading events were highly associated with the AI output, we built an ordinary logistic regression model to predict the correcting and misleading probability of the AI output (green and red points in Figure 4). The regression curves of the predicted correcting (green line) and misleading (red line) probabilities had 2 intersection points ($x = 0.1019$ and 0.8843) that divided the output range into the "benefit" zone with higher correcting probability and the "risk" zone with higher misleading probability. In addition, we found that the "risk" zone varied among the clinician groups (Table 4 and Supplementary Figure S1). The AI suppression strategy could be customized according to the features of the users and the working environment.

Sixteen MRIs (8%) had a high-risk AI diagnosis in this study. If we had retracted the AI diagnosis of these 16 MRIs while providing AI diagnosis of the rest as the AI suppression strategy suggested, the misleading events were estimated to be decreased by 41.7%, while the correcting events only

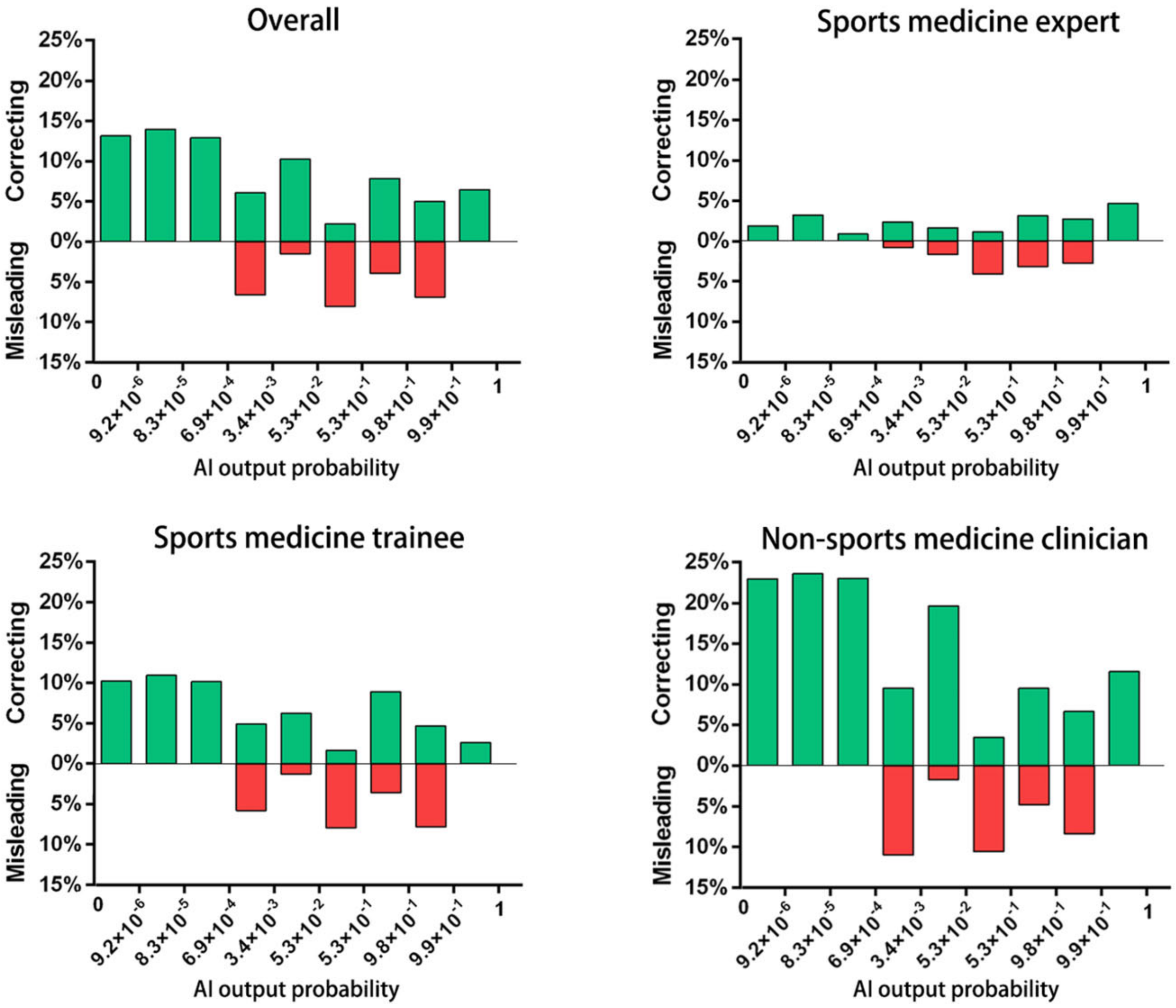


Figure 3. The correcting and misleading proportion of the AI output range. The correcting effect of the AI was accompanied by the misleading effect.

decreased by 3.2%. The clinician's AI-assisted mistakes were estimated to be decreased by 8.6% (Table 4).

DISCUSSION

In summary, our study highlights the importance of considering the potential for automation bias when incorporating AI into clinical practice. The use of AI can greatly improve clinicians' diagnostic accuracy and efficiency and the utility of MRI,[20] but it is crucial to also address the potential for automation bias by implementing strategies such as identifying high-risk zones for misleading AI diagnoses and allowing clinicians to make their own decisions for these cases. Our findings suggest that the AI suppression strategy can decrease automation bias and promote effective interaction between clinicians and AI systems.

Widespread automation bias has been well recognized since clinical decision support systems were introduced.[11] In the era of AI, AI technology provides a higher form of automation than other technologies, implying that AI can reinforce the risks of automation bias. This phenomenon is being realized by researchers and clinicians.[5,8,9] Automation bias causes clinicians to have higher diagnostic error,[5,21] and the entire spectrum of clinicians, including experts, could be the victims.[8] In this study, we found that clinicians of all expertise levels were vulnerable to automation bias, even though AI improved their overall diagnostic accuracy and efficiency. Nonsports medicine clinicians with less experience were found to be 4.73 times more likely to be influenced by the AI, highlighting the importance of addressing automation bias for these individuals who also benefited the most from AI assistance. Considering the potential risk of automation bias in the clinician-AI interaction, simply evaluating the performance of the AI algorithm is not sufficient to ensure patient safety. It is important to evaluate the clinician-AI interaction in order to ensure that the algorithm is being used to benefit patients.

Many factors contribute to automation bias by AI. The primary cause of automation bias is the algorithm generating faulty conclusions. Inappropriate human-AI interaction would retain the errors of the algorithm and pass them to the user side. Efforts to address automation bias have often focused on technological solutions, such as reducing bias in the algorithm, while neglecting the role of user practice.[14] Most medical AI algorithms are built without involving the end user and without considering how the end user will use them. If we are using AI to augment clinician abilities and not

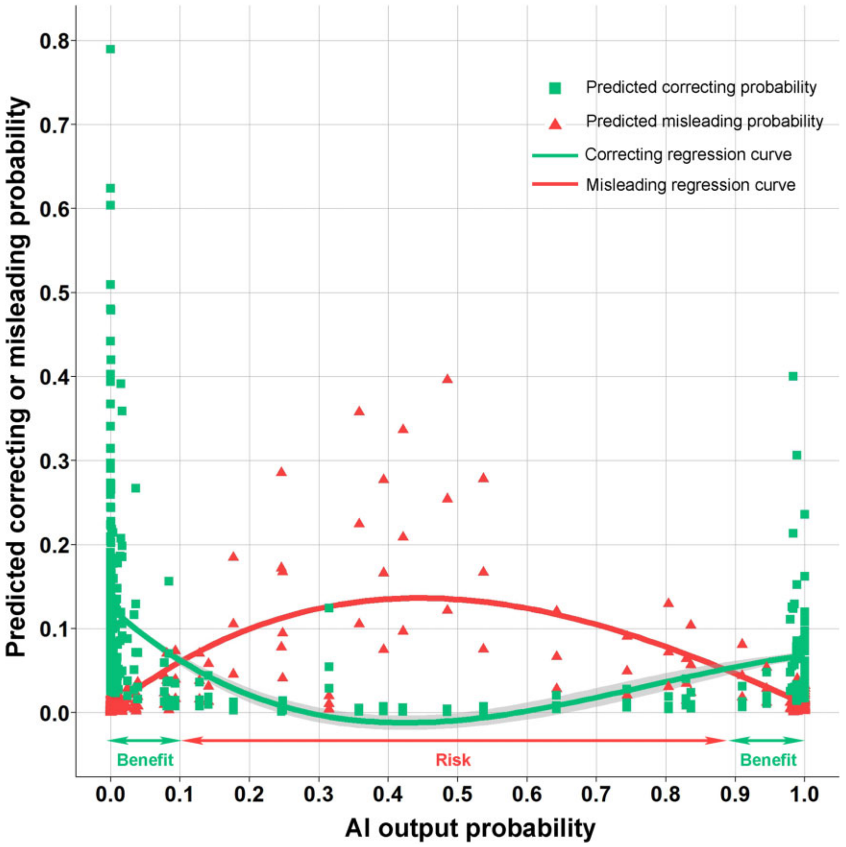


Figure 4. An ordinary logistic regression model was used to predict the correcting and misleading probability of the AI output (green and red points). The 2 intersection points of the regression curves divided the output range into the "benefit" zone with higher correcting probability and the "risk" zone with higher misleading probability.

Table 4. The effect of the AI suppression strategy on automation bias

	Overall ($n = 40$)	Sports medicine expert ($n = 9$)	Sports medicine trainee ($n = 16$)	Nonsports medicine clinician ($n = 15$)
AI "risk" zone	0.1019–0.8843	0.0345–0.7889	0.0804–0.8968	0.1299–0.8949
Clinician's mistake (per person) in AI-assisted round				
Without AI suppression strategy	290 (7.3)	33 (3.7)	123 (7.7)	134 (8.9)
With AI suppression strategy	265 (6.6)	29 (3.2)	109 (6.8)	127 (8.5)
Decrease by	8.6%	12.1%	11.3%	5.2%
Correcting event (per person)				
Without AI suppression strategy	833 (20.8)	40 (4.4)	251 (15.7)	542 (36.1)
With AI suppression strategy	806 (20.1)	37 (4.1)	240 (15.0)	529 (35.3)
Decrease by	3.2%	7%	4.4%	2.3%
Misleading event (per person)				
Without AI suppression strategy	132 (3.3)	14 (1.6)	50 (3.1)	68 (4.5)
With AI suppression strategy	77 (1.9)	7 (0.8)	29 (1.8)	41 (2.7)
Decrease by	41.7%	50.0%	42.0%	39.7%

replace them, human-AI interaction should be considered in the product's design. Another possible reason is that clinicians overlook the automation bias risk. In this study, we found that 45.5% (132/290) of clinicians' mistakes in the AI-assisted round were related to the misleading effect. However, all the clinicians in the interview held an optimistic attitude toward their AI-assisted performances. They denied that the AI had a misleading effect on themselves, even though they were aware that AI would generate faulty diagnoses. This lack of recognition of the problem may lead to a failure to balance self-confidence and trust in AI. Therefore, particular emphasis should be placed on user training to familiarize clinicians with the reliability of AI and the risks of accepting incorrect information.

In this study, we designed a user interface including information on AI diagnosis and confidence to help clinicians better calibrate their trust in AI.[18] An indicating color bar was added to warn of the low certainty index. We also provided user education before the test and emphasized that the AI diagnosis with 25% and 50% confidence was highly unreliable. However, these efforts were insufficient to completely solve the automation bias. Clinicians were vulnerable to faulty AI conclusions once they built up the trust that was necessary to benefit from AI support. Additionally, it was found that faulty AI diagnoses with low confidence can still cause misleading events, even when the probability of the AI diagnosis was far below the clinician's internal threshold for diagnosing ACL rupture. The study suggests that the problem of automation bias may become more severe when the AI provides a faulty diagnosis with high confidence, as humans tend to show a strong preference for options with certainty, according to prospect theory.[22] Further efforts are needed to decrease automation bias and promote clinician-AI interaction.

Inspired by the handover strategy in the autopilot and the driver of an autonomous vehicle,[12] we designed an AI suppression strategy to switch the clinician-AI-assisted model to the clinician-alone working model to decrease the automation bias. The correcting and misleading probability of the AI diagnosis was predicted based on the clinician-AI interaction test and the AI suppression was triggered when the misleading probability was higher than the correcting probability. With this strategy, the misleading events were estimated to be decreased by 41.7%, while the correcting events only decreased by 3.2%. This AI suppression strategy was established by sampling and statistics. The disease distributions, clinician characteristics, and clinician-AI interactions should represent the real working environment. Then, the high-risk misleading range could be representative and applied to the clinical practice without post hoc analysis. Compared to investing in improving the algorithm, making adjustments to user practice is more economical and efficient. This study provides a framework for future studies and AI products.

This study has several strengths: (1) A diverse group of 40 clinicians with varying specialties and levels of experience from different hospitals participated in the study, providing a comprehensive examination of the clinician-AI interaction. (2) We used mixed methods to investigate the impact of AI on clinician decision-making and to understand the mechanisms underlying this impact. (3) The study conducted a systematic evaluation of clinician-AI interaction and proposed a strategy to enhance this interaction, which can be applied to other AI models in the future.

The study also has some limitations: (1) This study was not conducted in a real clinical setting, where a doctor would perform a physical examination and be responsible for making an accurate diagnosis. Although real patient documents including medical history and physical examination results from the clinic were used, no patients participated in the test. The test settings, which involved reading 200 MRIs and diagnosing only ACL, do not reflect real-world practice. (2) To avoid causing a heavy burden to the clinicians and decreasing their concentration, we controlled the MRI sample size to 200, which was a relatively small sample size. (3) The third-round clinician-AI interaction test was a simulation and the reading session was not repeated. Instead, the clinician's AI-unassisted diagnosis was used to estimate the AI suppression strategy.

CONCLUSION

Although AI improved clinicians' diagnostic performance, automation bias was a serious problem that should be addressed in clinical practice. The proposed AI suppression strategy is a practical method for decreasing automation bias.

FUNDING

The work was supported by the National Natural Science Foundation of China (U22A20283), Beijing-Tianjin-Hebei Basic Research Cooperation Project (V1640331014004), and the Fundamental Research Funds for the Central Universities (PKU2023XGK002).

AUTHOR CONTRIBUTIONS

WDY: designing the study, evaluating the AI algorithm, managing the clinician test and analyzing the results, and writing the manuscript. LN: evaluating the AI algorithm, managing the clinician test, and analyzing the results. DJ: developing the deep learning model. SAL: developing the deep learning model. LSG, JD, and YJK: managing the clinician test and analyzing the results.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank all the clinicians who spared their time and participated in this study.

CONFLICT OF INTEREST STATEMENT

DJ and SAL are employees of Beijing Yizhun Medical AI Co., Ltd. YJK, WDY, and DJ have a patent for the deep learning model related to this work. LSG, JD, and LN have no competing interests.

DATA AVAILABLITY

The deep learning model can be accessed through the online handbook's described method at https://github.com/KingfishD/userbook-for-the-knee-AI-system.git.

REFERENCES

1. U.S. Food & Drug Administration. (2021). Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed May 30, 2023.
2. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res* 2021; 23 (4): e25759.
3. McBee MP, Awan OA, Colucci AT, *et al.* Deep learning in radiology. *Acad Radiol* 2018; 25 (11): 1472–80.
4. de Siqueira VS, Borges MM, Furtado RG, *et al.* Artificial intelligence applied to support medical decisions for the automatic analysis of echocardiogram images: a systematic review. *Artif Intell Med* 2021; 120: 102165.
5. Eng DK, Khandwala NB, Long J, *et al.* Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. *Radiology* 2021; 301 (3): 692–9.

6. Lindsey R, Daluiski A, Chopra S, *et al.* Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A* 2018; 115 (45): 11591–6.
7. Mori Y, Kudo SE, Misawa M, *et al.* Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. *Ann Intern Med* 2018; 169 (6): 357–66.
8. Tschandl P, Rinner C, Apalla Z, *et al.* Human–computer collaboration for skin cancer recognition. *Nat Med* 2020; 26 (8): 1229–34.
9. Bond RR, Novotny T, Andrsova I, *et al.* Automation bias in medicine: the influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *J Electrocardiol* 2018; 51 (6S): S6–11.
10. Cummings ML. Automation bias in intelligent time critical decision support systems. In: AIAA 1st Intelligent Systems Technical Conference, Illinois, Chicago; 2004: 289–294.
11. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012; 19 (1): 121–7.
12. Sujan M, Furniss D, Grundy K, *et al.* Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform* 2019; 26 (1): e100081.
13. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017; 24 (2): 423–31.
14. Strauß S. Deep automation bias: how to tackle a wicked problem of AI? *Big Data Cogn Comput* 2021; 5 (2): 18.
15. Sanders TL, Maradit KH, Bryan AJ, *et al.* Incidence of anterior cruciate ligament tears and reconstruction: a 21-year population-based study. *Am J Sports Med* 2016; 44 (6): 1502–7.
16. Xie S, Girshick R, Dollár P, *et al.* Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Honolulu, HI: IEEE; 2016.
17. Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); San Diego, CA: IEEE; 2005.
18. Knoery CR, Bond R, Iftikhar A, *et al.* SPICED-ACS: study of the potential impact of a computer-generated ECG diagnostic algorithmic certainty index in STEMI diagnosis: towards transparent AI. *J Electrocardiol* 2019; 57S: S86–91.
19. Plasencia CM, Alderman BW, Barón AE, *et al.* A method to describe physician decision thresholds and its application in examining the diagnosis of coronary artery disease based on exercise treadmill testing. *Med Decis Making* 1992; 12 (3): 204–12.
20. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980; 302 (20): 1109–17.
21. Kiani A, Uyumazturk B, Rajpurkar P, *et al.* Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit Med* 2020; 3: 23.
22. Kahneman D, Tversky A. *Prospect Theory: An Analysis of Decision under Risk*. Hackensack, NJ and Singapore: World Scientific; 2013: 99–127.