Evaluating the Clinical Utility of Artificial Intelligence Assistance and its Explanation on the Glioma Grading Task

Weina Jin*

School of Computing Science, Simon Fraser University, Burnaby, Canada

Mostafa Fatehi†

Division of Neurosurgery, The University of British Columbia, Vancouver, Canada

Ru Guo‡

Division of Neurosurgery, The University of British Columbia, Vancouver, Canada

Ghassan Hamarneh§

School of Computing Science, Simon Fraser University, Burnaby, Canada

Abstract

Clinical evaluation evidence and model explainability are key gatekeepers to ensure the safe, accountable, and effective use of artificial intelligence (AI) in clinical settings. We conducted a clinical user-centered evaluation with 35 neurosurgeons to assess the utility of AI assistance and its explanation on the glioma grading task. Each participant read 25 brain MRI scans of patients with gliomas, and gave their judgment on the glioma grading without and with the assistance of AI prediction and explanation. The AI model was trained on the BraTS dataset with 88.0% accuracy. The AI explanation was generated using the explainable AI algorithm of SmoothGrad, which was selected from 16 algorithms based on the criterion of being truthful to the AI decision process. Results showed that compared to the average accuracy of $82.5 \pm 8.7\%$ when physicians performed the task alone, physicians' task performance increased to $87.7 \pm 7.3\%$ with statistical significance (p-value $= 0.002$) when assisted by AI prediction, and remained at almost the same level of $88.5 \pm 7.0\%$ (p-value $= 0.35$) with the additional assistance of AI explanation. Based on quantitative and qualitative results, the observed improvement in physicians' task performance assisted by AI prediction was mainly because physicians' decision patterns converged to be similar to AI, as physicians only switched their decisions when disagreeing with AI. The insignificant change in physicians' performance with the additional assistance of AI explanation was because the AI explanations did not provide explicit reasons, contexts, or descriptions of clinical features to help doctors discern potentially incorrect AI predictions. The evaluation showed the clinical utility of AI to assist physicians on the glioma grading task, and identified the limitations and clinical usage gaps of existing explainable AI techniques for future improvement.

Keywords: Artificial Intelligence; Neuro-Imaging; Neurosurgery; Explainable Artificial Intelligence; Clinical Study; Human-Centered Artificial Intelligence

Running title Evaluating Clinical Utility of AI and Explanation to Grade Glioma
Funding This study was funded by BC Cancer Foundation–BrainCare Fund. This research was also enabled in part by the computational resources provided by NVIDIA and the Digital Research Alliance of Canada (alliancecan.ca).
Code availability The code to train the AI model, generate the heatmap explanation, and analyze study results is available at: https://github.com/weinajin/multimodal_explanation.
Conflict of Interest All authors declare no financial or non-financial competing interests.
Authorship

WJ: Conceptualization, Study Design, Software Development, Data Analysis, Writing

MF: Conceptualization, Study Design, Recruitment, Writing

RG: Study Design, Recruitment, Writing

GH: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Funding acquisition

*Weina Jin and Mostafa Fatehi are co-first authors.
Corresponding author: weinaj@sfu.ca. 8888 University Dr, Burnaby, BC, Canada. V5A 1S6
†fatehi@alumni.ubc.ca
‡rucheng@student.ubc.ca
§hamarneh@sfu.ca

1 Introduction and Motivation

Artificial intelligence (AI) and machine learning technologies have transformative potential in medicine, as evidenced by the ever-increasing research advances in medical AI in recent years [39, 37]. Using AI in medicine has the potential to support decision-making processes for healthcare providers and improve patient care. This is largely due to the predictive capability of AI to learn to recognize patterns from raw and high-dimensional data, such as medical images, electronic health records, and genomic data. In neuro-oncological settings, AI and machine learning have been studied in a wide range of applications, including identifying the grading and genetic mutations of brain tumors [26, 15, 14], predicting patients' prognosis [53, 33], segmenting tumors based on magnetic resonance imaging (MRI) [40], triaging patients based on computed tomography (CT) scans [47], and discovering radiomics and radiogenomics for brain tumors [44]. Elsewhere, there are emerging cases of deploying AI in routine neuro-radiology practice, such as the AI-based RAPID software to detect vessel occlusion and triage stroke patients based on CT angiography [1].

Despite the above research advances, the widespread implementation of AI faces considerable challenges in translation from bench to bedside [22, 29]. A prerequisite to clinical deployment is bridging the AI evaluation gap between existing algorithmic evaluations and desired clinical evaluations. As with any new medical intervention, AI needs to undergo rigorous evaluation prior to its clinical implementation. Jin et al. [26] previously proposed four phases of clinical utility evaluation for AI in neuro-oncology, analogous to the conventional four phases of clinical trials for drugs or medical devices: **phase I** is the algorithmic evaluation that evaluates the performance of AI model alone on unseen test data; **phase II** evaluates the primary efficacy of AI assistance on the collaborative clinical user-AI task performance, conducted in *experimental* settings on *simulated* tasks; **phase III** further confirms the efficiency of the AI assistance on collaborative clinical user-AI task performance, conducted on a larger scale randomized controlled trial (RCT) in *clinical* settings on *real-world* tasks; and **phase IV** is for post-marketing software support and surveillance. It is worth noting that from phase II and above, all evaluations include clinical users. Existing work in AI in neuro-oncology [14, 15, 53, 33, 40] has generally conducted phase I algorithmic evaluation, which cannot reflect the clinical utility of AI in assisting clinical users in the clinical workflow. Phase II clinical evaluation of AI has been conducted in other specialties such as orthopedics [9], psychiatry [24], and ophthalmology [41]. Phase III RCT clinical evaluation of AI in clinical settings had been conducted in specialties such as gastroenterology [51], and results showed a large variation of AI clinical utility [54, 45]. To the best of our knowledge, there are no phase II (and above) studies on the clinical utility of AI assistance in neuro-oncology, and this is the research gap we aim to bridge in this study.

In addition to the above AI evaluation gap, another significant hurdle for AI clinical implementation is the interpretability or explainability problem of AI. The state-of-the-art AI models, namely deep neural networks, are black-box models, and their decision processes are incomprehensible even to AI engineers. This impedes the clinical use of AI, as clinical users will often require an explanation or justification from AI other than a mere prediction, due to the high-stakes nature of clinical decision-making [49]. Furthermore, explanations may enable physicians to identify potential errors of AI, and potentially to achieve complementary human-AI performance, where the human-AI team outperforms either AI or human alone [48, 52, 13, 6]. Therefore, we leverage the latest technical advances in explainable AI (XAI) as a feature of the AI system in our clinical evaluation study.

In this work, we recruited physicians and conducted a phase II clinical evaluation of AI in an experimental setting on a simulated clinical task based on brain MRI: classifying a glioma case into a glioblastoma (GBM, WHO grade IV), or a WHO grade II or III glioma. This is a clinically relevant question that helps guide subsequent management decisions. Tumor grading is also a routine and ubiquitous task in neuro-oncological settings and is commensurate with our participants' knowledge in neuro-oncology. Ultimately, other tumor genetic characteristics, such as isocitrate dehydrogenase (IDH) mutation status and O6-methylguanine-DNA methyltransferase (MGMT) methylation status, are critical for treatment and prognostication, but are more challenging for clinicians to predict from imaging studies. Thus, in the present proof-of-concept study, we have focused on the task of differentiating GBM from grade II/III diffuse gliomas. Traditionally, clinicians have relied upon patient characteristics, image findings, along with neurological signs and symptoms to decide whether to proceed with an aggressive resection, perform a biopsy, or to continue with watchful waiting. The interpretation of imaging findings is contingent upon the neuro-radiologist's and neurosurgeon's experience, and this likely contributes to some of the heterogeneity seen in practice among clinicians who treat patients with gliomas [19]. A potential AI-based tool that can accurately predict tumor genetics and histologic grading would potentially not only decrease the heterogeneity in management, but also help guide biopsy plans and improve the ability to prognosticate outcomes.

To provide evidence for the safe, accountable, and effective use of AI in clinical settings, we conducted a nationwide clinical study in Canada on the glioma grading task. We recruited 35 neurosurgeons, each reading a set of 25 brain MRIs without and with AI assistance on glioma grade prediction and explanation. The AI model was trained on the same task on the publicly-available Multimodal Brain Tumor Segmentation (BraTS) dataset with an accuracy of 88.0%. The AI explanation was a color map overlaid on the MRI to highlight important regions for AI prediction. We selected the SmoothGrad XAI algorithm [46] to generate color maps for the trained AI model. SmoothGrad was selected from 16 commonly-used XAI algorithms based on the computational evaluation of how truthful the XAI algorithm reflects the AI decision process. Results showed that physicians' average task accuracy improved from $82.5 \pm 8.7\%$ without AI assistance to $87.7 \pm 7.3\%$ with the assistance of AI prediction (p-value=0.002). The additional assistance of AI explanation did not change the accuracy, which was $88.5 \pm 7.0\%$ (p-value=0.35). Doctors assisted by either AI prediction alone or AI prediction and explanation combined did not achieve complementary doctor-AI task performance. The study confirmed the effect of AI on enhancing physicians' clinical task performance in a simulated clinical setting. The study also identified the limitations and possible failure reasons of existing explainable AI techniques for future improvement. This is the first study in neuro-oncology to evaluate the clinical utility of AI assistance.

The article is organized as follows: we describe the AI model, XAI algorithm, and study design in Section 2. We present the study results in Section 3, discuss the clinical utilities of AI and its explanation in Section 4, and analyze limitations and future work in Section 5. In Supplemental S1-S3, we provide the qualitative results (S1), additional methods and quantitative results (S2), and the study survey content (S3).

2 Materials and Methods

2.1 Study Material

2.1.1 MRI data

We used the publicly-available Multimodal Brain Tumor Segmentation (BraTS) 2020 dataset [35, 5] in the glioma grading clinical study, as well as to train the AI model. The BraTS dataset contains routine clinically-acquired, pre-operative brain MRI scans from patients with glioma. The brain MRIs in BraTS dataset were obtained with different clinical protocols and various scanners from 19 institutions, including the publicly-available TCGA/TCIA repositories [4, 3]. Each MRI scan consists of four MRI pulse sequences of T1-weighted, T1-weighted contrast enhancing (T1C), T2-weighted, and T2 Fluid Attenuated Inversion Recovery (FLAIR). MRIs in the BraTS dataset were pre-processed images with the pre-processing steps of co-registration to the same T1 anatomic template, resampling to $1mm^3$ voxel resolution, and skull-stripping. The original pre-processing methods are detailed in Bakas et al. [5]. The data underwent additional pre-processing for data augmentation during model training, including random flipping, random rotation, sequence-wise intensity normalization, and resizing from the original data dimension of $240 \times 240 \times 155$ to $128 \times 128 \times 128$ in width, height, and depth. Each MRI scan is associated with a tumor grade label of a GBM or grade II/III glioma, which was pathologically confirmed. The total number of MRI cases is 369, including 76 cases of grade II/III glioma, and 293 cases of GBM.

2.1.2 AI model and algorithmic evaluation on glioma grading task

We trained an AI model using the BraTS 2020 dataset to grade glioma MRIs. The AI model receives an MRI input and outputs a glioma grade of either a GBM or a grade II/III glioma. The model architecture is a VGG-like [43] three-dimensional (3D) convolutional neural network (CNN), with six 3D CNN layers connected to two fully connected layers. We stratified split the BraTS dataset into 65% training (239 cases), 15% validation (56 cases), and 20% (74 cases) hold-out test set by keeping the same grade II/III : GBM ratio in each set. There were no patient ID overlapping across the three datasets. To handle the imbalanced data, we used a weighted sampler that equalizes the data sampling probability of each class during training [11]. The training, validation, and test accuracies of the AI model were 80.28%, 92.86%, and 90.54%, respectively. The fine-grained model performance metrics are in Supplemental S2 Fig. 1, which was also shown to participants in the clinical study.

From the test set, we sampled a subset of 25 MRIs as the clinical test subset used in the glioma grading clinical study. We sampled the subset by keeping the same ratio of the correctly/incorrectly predicted grade II/III glioma or GBM as the confusion matrix of model performance in Supplemental S2 Fig. 1. This is to keep an equivalent performance of the AI model on the test set and the clinical test subset. In the clinical test subset, there were 7 cases of grade II/III glioma, and 18 cases of GBM. The AI model had an accuracy of 88.00% on the clinical test subset.

2.1.3 Generating and selecting the optimal AI explanation

The AI model we trained to grade glioma is a black-box CNN model. To explain the model decisions to physicians, we applied post-hoc XAI algorithms that act as a surrogate model to approximate the black-box AI model by probing the model parameters and/or input-output pairs [27]. From 16 post-hoc XAI algorithms that can generate a feature attribution map or heatmap (named as color map in the study) to explain AI prediction using the important image regions, we selected SmoothGrad [46], which was the most truthful to the AI model decision process [28]. Fig. 1 shows four examples of SmoothGrad explanation used in the study [1]. The evaluation method and result on XAI truthfulness are detailed in Supplemental S2.

2.2 Participant

We recruited physicians to evaluate their clinical task performance without and with AI assistance. The inclusion criteria for the study participants were: the participant must hold an MD degree or equivalent; and must be a consultant neurosurgeon, radiologist, or

[1] We have a note for Fig. 1 panel (D) in Supplemental S2 to raise an issue on its MRI ground truth label.

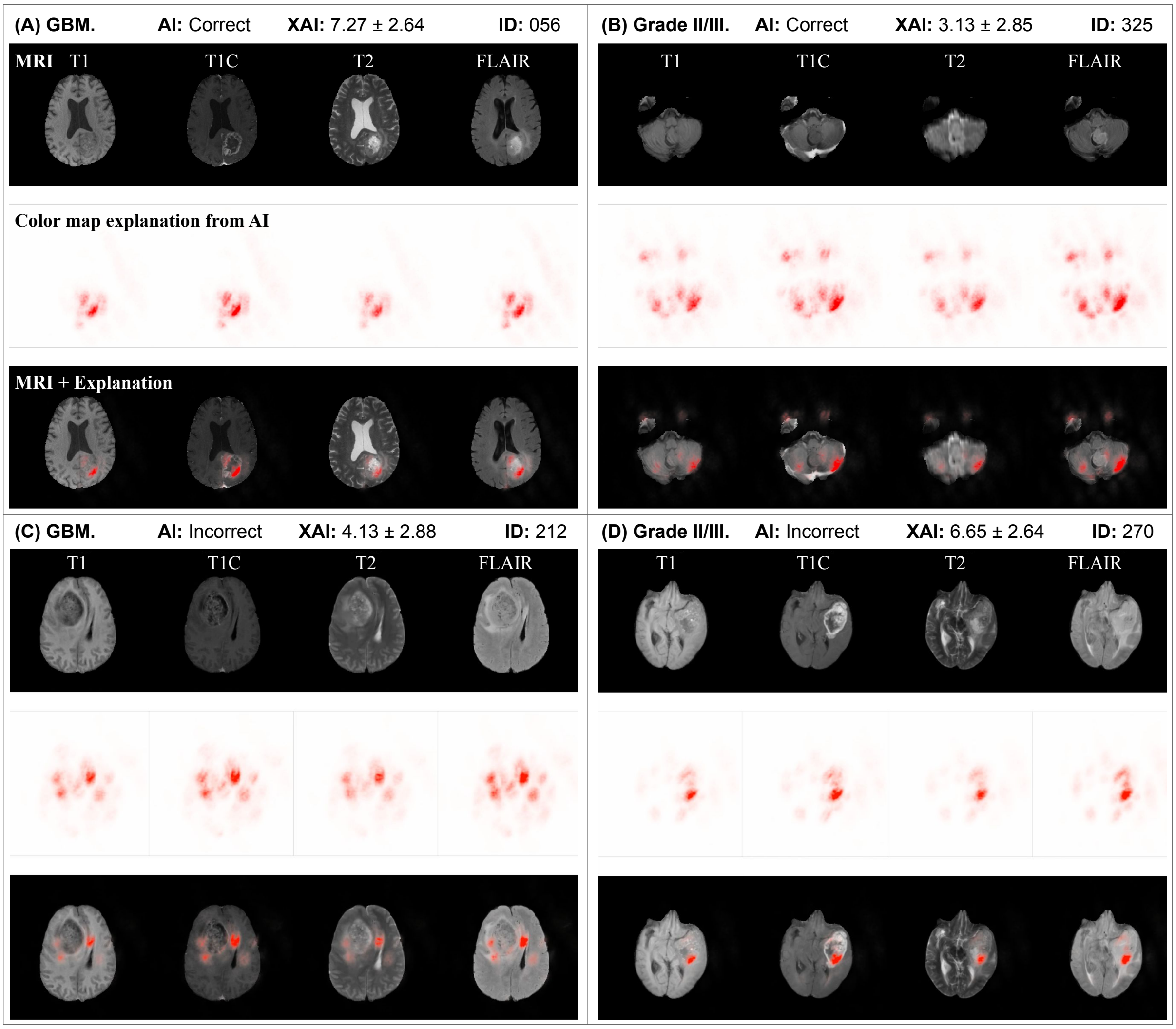

Figure 1: Visualization of four MRI scans and their corresponding color map explanations used in the study. We include two GBM and two Grade II/III cases where AI predicts correctly or incorrectly with plausible or implausible explanations (indicated as the XAI score of participants' mean rating on explanation quality on a 0–10 scale). ID is the BraTS dataset ID. Within each panel, each column is an MRI pulse sequence of T1, T1C, T2, and FLAIR. Row 1 is the MRI image, Row 2 is the color map showing the important image regions for AI prediction, and Row 3 is the color map overlaid on the MRI. The original MRIs and color maps are both 3D images, and we visualize one 2D image in axial view.

neuro-radiologist, or a trainee in neurosurgery, radiology, or neuro-radiology. Since the study was conducted anonymously as an online survey, two stages of eligibility screening were conducted: one was conducted at the beginning of the online survey, where participants were filtered by their answers to the questions about their roles in medical practice and their medical specialty; The other was conducted using a post-survey screening process to filter out responses that did not meet the inclusion criteria due to random guess or lack of required expertise in neuro-oncological MRI interpretation. We did so by only including participants whose task accuracy when performing the grading task alone was above 0.55. The accuracy threshold was set to be slightly higher than the random guess accuracy of 0.5. We used convenience sampling and recruited participants by directly contacting the researchers' national-wide clinical research network. The recruitment period was from October 2021 to February 2022.

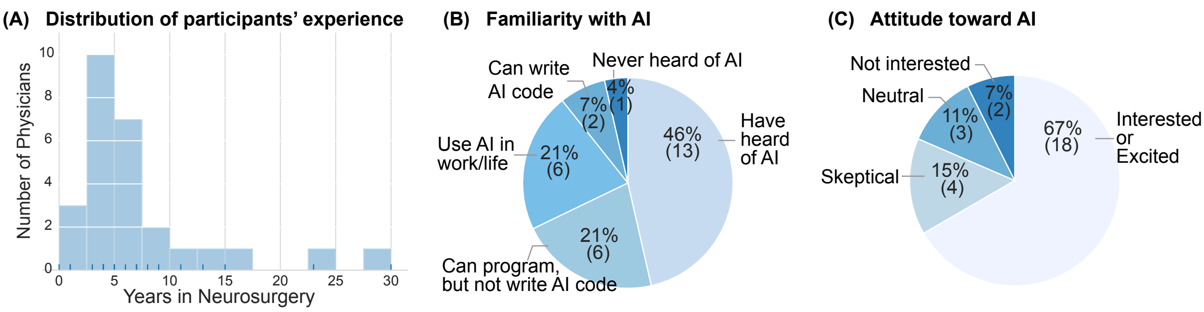

Figure 2: Participants' demographic information. Panel (A) shows the distribution of participants' years of neurosurgical practice; The sticks on the x-axis show each participant's years of practice. Panel (B) shows participants' familiarity with AI, and panel (C) shows their attitude towards AI, with the percentage and number of participants (in parentheses) indicated for each category.

A total of 35 participants met the inclusion criteria and were enrolled in the study. The recruitment rate was 14.8% (35 out of 236 eligible participants contacted). Among them, 29 participants completed the survey, while 6 participants dropped out without completing the survey (their numbers of completed MRI interpretation cases were: 2, 2, 2, 3, 8, and 17, respectively). In addition, five participants took the interview to provide qualitative comments on the AI system in the survey. The self-report demographic data of the participants were[2]: female: male = 7:19; age: 34.7 ± 8.2 (mean $\pm$ std of the 26 reported participants); all participants were from the neurosurgery specialty, and their positions were: 12 attending neurosurgeons, 2 neurosurgical fellows, and 21 neurosurgical residents. Their years of practicing medicine were 9.8 ± 9.1, and their years of practicing neurosurgery were 7.1 ± 6.5. Figure 2 summarizes their familiarity with AI and their attitude towards AI. Most participants (96%) were familiar with AI technologies, e.g. they had heard of AI, or had used it at work or in their daily lives. Over 2/3 of the

[2]Since the questions on age and gender were not set to be mandatory, we report the collected data on age and gender, which may not include all participants.

participants had a positive attitude towards AI, whereas the rest had a skeptical or neutral attitude. The detailed demographics are listed in Supplemental S2 Table 1.

2.3 Study design and procedure

We designed a pre-post clinical study to examine the clinical utility of AI system regarding its benefit to physicians' task performance. The study consisted of an online glioma grading survey (30-40 minutes) and an optional remote interview (20-30 minutes). Participants provided separate informed consents for the survey and the interview.

The online survey is the main part of the study, where participants read a set of 25 MRIs without and with the assistance of AI prediction and its explanation. The MRIs were sampled from the BraTS 2020 dataset as described in Section 2.1. The sequence in which MRIs were shown was randomized for each participant to avoid bias due to MRI reading order. At the beginning of the survey, participants were instructed to read the consent form and provide their electronic consent by clicking the checkbox "I agree to participate in the study". In the survey, participants were first introduced to the AI system and its performance on the test set. Then they began the MRI reading task. For each MRI, the participants first gave their own judgment. Then the AI prediction was revealed to them, and they were asked to give their current and possibly updated judgment of the glioma grade. The AI prediction was shown only after participants gave their own judgment. Next, participants were asked about their willingness to check AI explanation of how it arrived at the prediction, and were shown a color map explanation from AI with important regions of prediction highlighted (examples are in Fig. 1). The MRI and color map explanation were both 3D images shown in video format, in which participants could control the video play to view different MRI slices and their corresponding color map explanation. After that, participants were asked to provide their final (again, possibly modified) judgment on the glioma grade, and evaluate the agreement between their own clinical judgment and color map on an 11-point scale from 0 to 10 (Table 1).

We also asked participants to rate, on an 11-point scale, their trust level in the AI system and willingness to incorporate this AI suggestion into routine clinical practice. The two questions were asked at three time points: at the beginning of the survey before exposure to any information on the AI system as a baseline, after viewing AI performance metrics (Supplemental S2 Fig. 1) on the test set, and after using AI with its prediction and explanation assistance for the 25 MRIs. The survey ends with the question to select and rank possible goal(s) of checking AI explanations, and a short demographic questionnaire on the participant's medical experience, familiarity with AI, attitude towards AI, age, and gender. The full survey content is in Supplemental S3.

After completing the online survey, participants were given monetary compensation ($50 CAD gift card) as appreciation for their time and effort. The participant could choose to participate in an optional remote interview. In it, participants talked about their user experience and commented on the AI system. Participants provided additional verbal consent prior to the interview. The remote interview sessions were video- or audio-recorded for qualitative data analysis. The study was approved by the Research Ethics Board of Simon Fraser University (Ethics number: H20-03588).

Variable name	Survey question	Survey question options
DR	What grade of glioma would you predict this MRI to be?	• Grade II/III glioma • Grade IV glioma
DR+AI	After viewing AI's suggestion, what is your current judgment on the tumor grade?	ditto
DR+XAI	After viewing AI's explanation, what is your final judgment on the tumor grade?	ditto
Need explanation	Would you like to check the explanation from AI for this MRI?	Yes/No option
Explanation quality	How closely does the highlighted area of the color map match with your clinical judgment?	[0-10 scale] • 0, Not close at all • 5, Somewhat close • 10, Very close
Trust	What is your trust level in this AI model?	[Scale from -5 to 5] • -5, Totally distrust the AI • 0, Neutral, neither distrust nor trust • 5, Totally trust the AI
Willingness to use AI	How likely will you incorporate this AI's suggestions into your routine clinical practices, such as diagnosis, prognosis, and medical management?	[0-10 scale] • 0, Not likely • 5, Somewhat likely • 10, Very likely
Explanation goal	When are you most likely to check those color map explanations from AI?	Select and rank from a set of 15 predefined options and a self-filled option

Table 1: List of variables collected in the survey, and their corresponding survey questions. For the trust scale, we post-processed the responses by adding 5 to all responses, so that the scale range is from 0 to 10. In the following text, we underline the variable names listed in the table. Variables above the double horizontal line were asked for each MRI case, and the ones below were only asked once or several times at different time points.

2.4 Statistical Analysis

We conduct statistical analysis to test the following null hypotheses on the utility of AI in AI-supported task performance, in achieving complementary doctor-AI task performance, and in physicians' trust and willingness to use AI.

1. There are no differences in physicians' accuracies on the glioma grading task among the three conditions: 1) Physician performing the task alone, denoted as DR; 2) Physician performing the task with the assistance of AI prediction, denoted as DR+AI; 3) Physician performing the task with the assistance of AI prediction and explanation, denoted as DR+XAI.

2. The accuracy of collaborative doctor-AI team performance of DR+AI or DR+XAI is no higher than the best performance of AI alone or doctor alone (DR).

3. There are no differences in physicians' trust level among the three time points: 1) Initial baseline without knowing any information from AI; 2) After viewing AI performance metrics; and 3) After using AI with its prediction and explanation assistance for the 25 MRIs.

4. There are no differences in the physicians' willingness to use AI among the three time points: 1) Initial baseline without knowing any information from AI; 2) After viewing AI performance metrics; and 3) After using AI with its prediction and explanation assistance for the 25 MRIs.

To test the hypotheses, a one-way analysis of variance (ANOVA) with repeated measures [36] is performed when data fulfill the assumptions of normality and sphericity. We use Shapiro-Wilk test of normality [42] and Mauchly's test for sphericity [34] to test the assumptions for ANOVA. If the null hypothesis is rejected, a post-hoc analysis is conducted using Tukey's HSD (honestly significant difference) test when data meet the assumption of homogeneity of variances. Otherwise, if assumptions for ANOVA are violated, we use the non-parametric Friedman test, and a post-hoc analysis of Wilcoxon signed-rank test with Bonferroni correction. For hypothesis 2, to compare the accuracy of doctor-AI team with the best accuracy of doctor or AI alone, a one-sided t-test is performed if the data fulfill the assumption of normality. Otherwise, a non-parametric test is used.

Additionally, the Spearman correlation coefficient is used to measure the association between two continuous variables; and the chi-square test of independence is conducted to test the association between two categorical variables. Unless otherwise stated, we use a significance level $\alpha = 0.05$. The statistical analysis was performed using Python statistical package SciPy[3] and Pingouin[4].

[3]http://scipy.org/

[4]http://pingouin-stats.org/index.html

2.4.1 A pilot study to estimate sample size

Before launching the formal national study, we conducted a pilot study to iterate the survey content and estimate the sample size. Six neurosurgical residents were recruited in the pilot study. With a two-sided test size of 5% and a power of 90%, based on the effect size of 1.3 between DR and DR+AI, the estimated sample size is 13.

3 Results

In the article, we report the quantitative analysis of the survey data, and provide the full results of the qualitative data analysis in Supplemental S1. The qualitative data are from the interview and free-text input in the survey. We discuss findings from both quantitative and qualitative data in the Discussion Section 4. We number the participants with N1, N2, ... or O1, O2, ... when directly quoting their words.

3.1 Physicians' task performance in three decision-support conditions

A total of 2279 glioma grading decisions were collected for the three decision-support conditions of 1) DR: physician performing the task alone (761 decisions); 2) DR+AI: physician performing the task with AI prediction assistance (759 decisions); and 3) DR+XAI: physician performing the task with AI prediction and explanation assistance (759 decisions). Participants' average task accuracies for the three conditions were: DR: $82.49 \pm 8.69\%$ (mean±std), DR+AI: $87.70 \pm 7.33\%$, and DR+XAI: $88.52 \pm 7.02\%$. The descriptive statistics of participants' task accuracy for the three conditions are in Table 2.

Condition	N	M±SD	Min	25% Q	Mdn	75% Q	Max
DR	35	82.49 ± 8.69	60.00	80.00	84.00	88.00	100.00
DR+AI	35	87.70 ± 7.33	68.00	84.00	88.00	92.00	100.00
DR+XAI	35	88.52 ± 7.02	72.00	84.00	88.00	92.00	100.00

Table 2: Descriptive statistics for all participants' task performance accuracy (%). N - number of participants, M - mean, SD - standard deviation, Q - quantile, Mdn - median.

All data passed the sphericity assumption test for ANOVA, but the data on DR condition did not pass the normality assumption. Therefore, we used the non-parametric Friedman test instead, and results showed a statistically significant difference in task accuracies among the three conditions, $\chi^2_F(2) = 23.53, p < 0.001$. We then conducted post-hoc analysis using Wilcoxon signed-rank tests with Bonferroni correction. Results showed that the DR+AI condition had a statistically higher accuracy compared to the DR condition ($Z = 9.0, p = 0.002$); similarly, the DR+XAI condition had a statistically higher accuracy compared to the DR condition ($Z = 3.0, p = 0.0004$). However, the accuracies between DR+AI and DR+XAI conditions did not show statistically significant difference ($Z = 15.5, p = 0.35$) (Fig. 3). We also calculated the effect size using common language effect size, and results showed a physician has a probability of 67.2% of having a higher accuracy when assisted by AI prediction (DR+AI) than performing the task alone (DR), a probability of 71.0% of having a higher accuracy when assisted by AI prediction and explanation (DR+XAI) than performing the task alone (DR), but only a probability of 53.6% of having a higher accuracy when assisted by AI prediction and explanation (DR+XAI) than assisted by AI prediction alone (DR+AI). In addition to the above result using the accuracy metric, in Supplemental S2 Table 3, we also report results and statistical tests using other performance metrics, including sensitivity, specificity, and F1 score, and the results showed a similar trend.

In addition to the change of performance in the three conditions, we also tested whether complementary doctor-AI task accuracy was achieved, which indicates doctors assisted by AI outperform either AI or doctors alone. Since AI had a higher accuracy than the mean accuracy of doctors alone, we compare the doctor-AI team accuracy with the AI accuracy of 88.00% using a one-sample, one-sided t-test. When physicians were assisted by AI prediction only, the accuracy of the collaborative doctor-AI team (DR+AI) was no higher than the AI accuracy of 88.00%, $t(34) = -0.239$, p-value =0.59; Similarly, when physicians were assisted by AI prediction and explanation, the accuracy of the collaborative doctor-AI team (DR+XAI) was no higher than the AI accuracy of 88.00%, $t(34) = 0.436$, p-value =0.33.

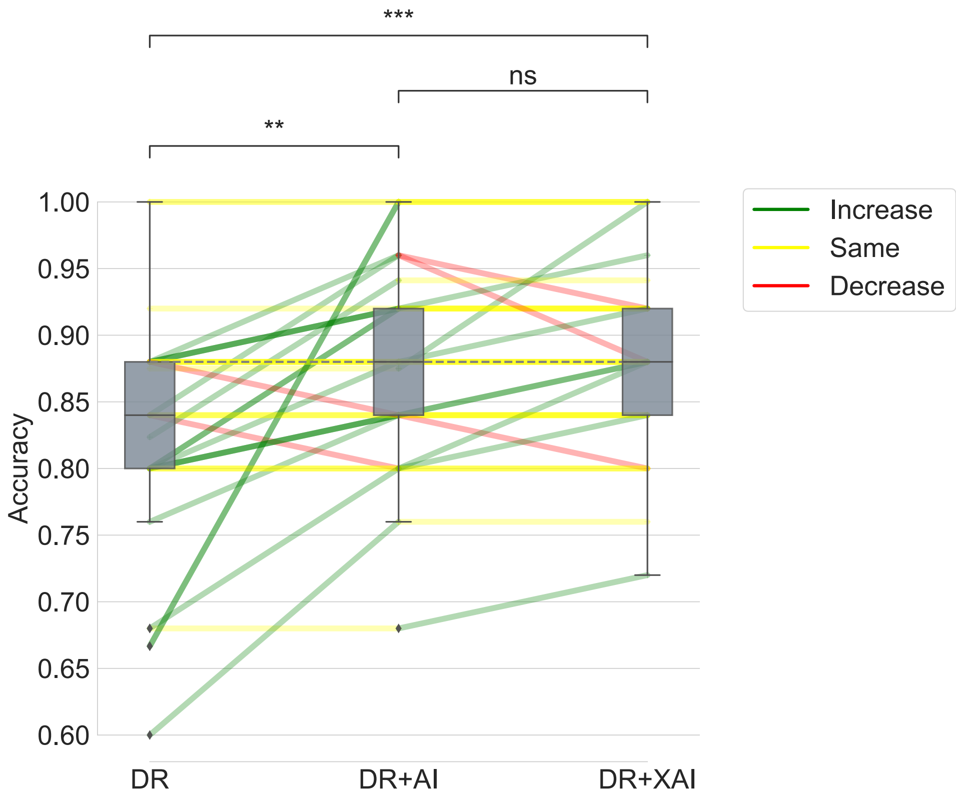

Figure 3: Participants' task accuracies on glioma grading in three conditions: 1) DR: Physicians performing the task alone; 2) DR+AI: Physicians performing the task with AI assistance (with predictions from AI); 3) DR+XAI: Physician performing the task with XAI assistance (with predictions and explanations from AI). We show box plots for the three conditions. The colored lines between boxes show each participant's accuracy change between the conditions, with green lines indicating an accuracy increment, red indicating a decrement, and yellow indicating no change. The darkness of colored lines encodes the frequency of such a change. The dashed line indicates the accuracy of the AI model of 88.0%. ns: $p > 0.05$, $** : 0.001 \leq p \leq 0.01$, $*** : 0.0001 \leq p \leq 0.001$.

3.2 Decision agreement and decision change patterns

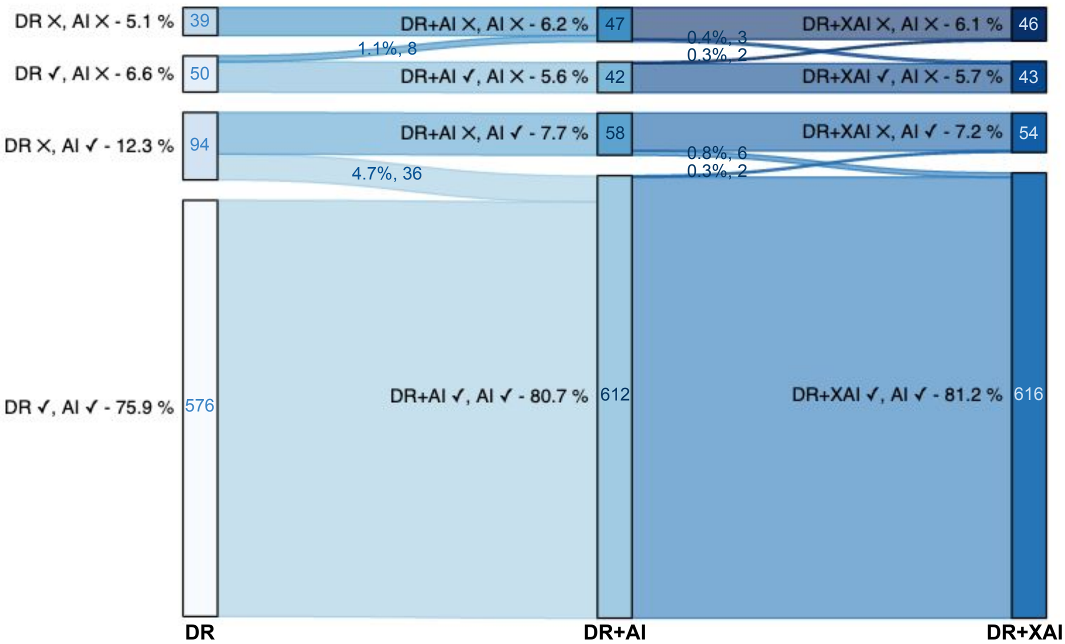

Figure 4: Participants' decision change stream plot for each error category of all participants. The three columns represent the three conditions of DR, DR+AI, and DR+XAI, respectively. The four rectangles within each column record the number and percentages of cases when the doctors' and the AI's decisions were correct (✓) or not (×), e.g., the decision agreement tally is reported in the first and fourth rectangles, reflecting when the doctor and AI made the same decisions (where both were incorrect or correct). The total number of decisions was 759 for each column.

We analyzed the fine-grained decision agreement and decision change in the three conditions of DR, DR+AI, and DR+XAI, and visualized such patterns as a decision change stream plot in Fig. 4. The subgroup analysis for attending and resident + fellow physician subgroups showed similar patterns (Supplemental S2 Fig.6).

For the decision agreement pattern, as shown in Fig. 4, as a baseline when physicians performed the task alone (DR), physicians' and AI decisions agreed with each other in 81.0% (615/759) decisions. The decision agreement increased to 86.8% (659/759) when physicians were assisted by AI prediction (DR+AI), and further increased to 87.2% (662/759) when physicians were assisted by both AI prediction and explanation (DR+XAI).

For physicians' decision change pattern during AI assistance, as shown in Fig. 4, with the assistance of AI prediction (DR+AI condition), physicians changed 5.8% (44/759) of their original decision to AI prediction, and such decision change occurred only during decision disagreement between AI prediction and physicians' decision. Among these decision change cases, 81.8% (36/44) were correct changes, i.e., resulted in a corrected decision that matched the ground truth diagnosis, and the remaining 18.2% (8/44) were incorrect changes, i.e., leading to an erroneous decision. With further assistance from AI explanation (DR+XAI condition), physicians changed 1.7% (13/759) of their decisions, and such decision change occurred during both decision agreement (0.7%, 5/759) and disagreement (1.1%, 8/759) between AI prediction and physicians' decision. Among them, 69.2% (9/13) changed correctly, and 30.8% (4/13) changed incorrectly.

3.3 Trust and willingness to use AI

We tested whether participants would calibrate their level of trust in the tested AI system and their willingness to use AI with the exposure to AI performance metrics and AI usage experience. Participants' level of trust in AI and willingness to use AI were recorded at three time points: 1) the initial baseline without knowing any information from AI; 2) after viewing AI performance metrics; and 3) after using AI predictions and explanations for the 25 MRIs. The descriptive statistics of the two variables at three time points are listed in Table 3. In addition, the two variables trust in AI and willingness to use AI are highly correlated, with a Spearman correlation coefficient of 0.70 ($p < 0.001$).

	Time point	N	M±SD	Min	25% Q	Mdn	75% Q	Max
Trust	Bsl	29	5.31 ± 2.04	0.00	5.00	5.00	7.00	9.00
	Pfm	29	6.72 ± 1.67	3.00	5.00	7.00	8.00	9.00
	Use	29	6.62 ± 2.68	0.00	6.00	8.00	8.00	9.00
Willingness to use AI	Bsl	29	4.10 ± 2.79	0.00	2.00	5.00	5.00	10.00
	Pfm	29	5.07 ± 2.45	0.00	3.00	5.00	7.00	10.00
	Use	29	4.59 ± 3.16	0.00	1.00	5.00	8.00	9.00

Table 3: Descriptive statistics for participants' trust and willingness to use AI. N - number of participants, M - mean, SD - standard deviation, Q - quantile, Mdn - median. The three time points are: 1) Bsl: the initial baseline without knowing any information from AI; 2) Pfm: after viewing AI performance metrics, and 3) Use: after using AI predictions and explanations for the 25 MRIs.

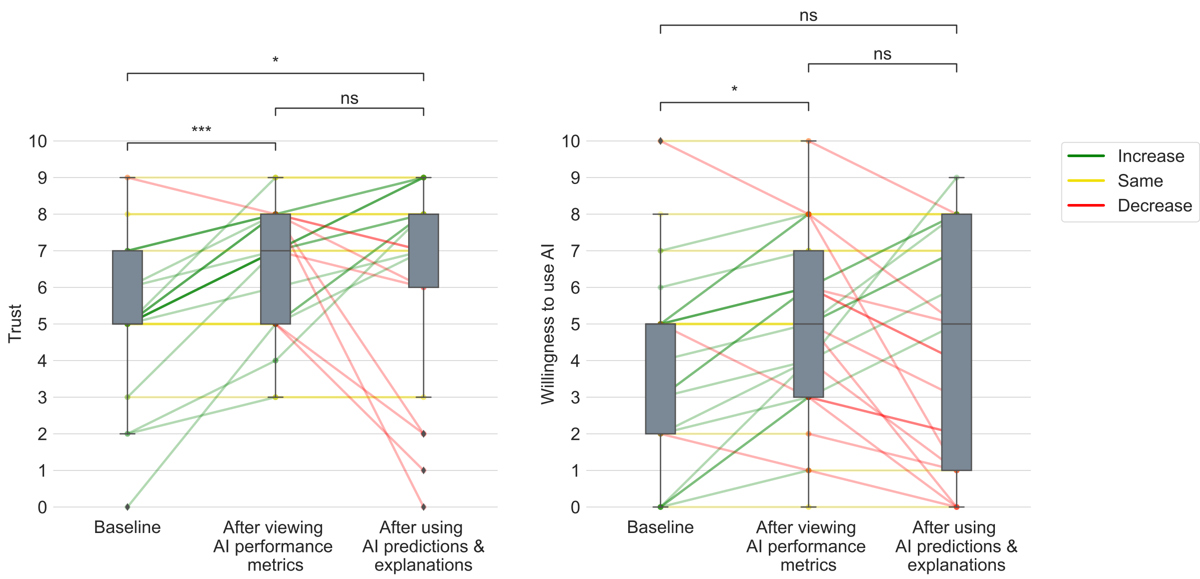

Figure 5: Box plots and changes of participants' trust in the AI (left), and willingness to use AI (right) at the initial baseline, after viewing AI performance metrics, and after using AI predictions and explanations. Both dependent variables (y axis) are reported on a 0-10 point scale. The colored lines in between indicate a change in the ratings of each participant, with green indicating an increment, red indicating a decrement, and yellow indicating no change. The darkness of colored lines encodes the frequency of such a change. ns: $p > 0.05$, $* : 0.01 \leq p \leq 0.05$, $** : 0.001 \leq p \leq 0.01$, $*** : 0.0001 \leq p \leq 0.001$.

Part of the trust and willingness to use AI data did not pass the sphericity and normality assumption test for ANOVA. Therefore, we used the non-parametric Friedman test instead. Results showed a statistically significant difference among the three time points for both trust in AI ($\chi^2_F(2) = 16.97, p = .0002$), and willingness to use AI ($\chi^2_F(2) = 8.09, p = .018$). We conducted post-hoc analysis using Wilcoxon signed-rank tests with Bonferroni correction to identify the statistically different pairs. For the level of trust in the AI system, participants rated a statistically higher trust after viewing AI performance metrics compared with the initial baseline ($Z = 4.0, p = .0004$); and a statistically higher trust after using AI predictions and explanations for the 25 MRIs compared with the initial baseline ($Z = 58.0, p = .025$); but there was no significant difference between the trust level after viewing AI performance metrics, and after using AI predictions and explanations for the 25 MRIs. For the level of willingness to use AI, participants only rated a statistically higher willingness to use AI after viewing AI performance metrics compared with the initial baseline ($Z = 29.5, p = .012$); and the rest pairwise tests did not show statistically significant differences. The statistical test results are visualized in Fig. 5.

3.4 Clinical usage scenarios for AI explanation

Physicians' behavior of seeking an explanation is usually to support their subsequent clinical sub-tasks. We summarized such potential explanation goals from literature [25], and asked participants to select and rank them. The results are shown in Table 4. The top-rated explanation goals were related to the critical nature of clinical tasks, and AI explanation was useful mainly to safeguard clinical decision-making.

Explanation goal	**Selected times**	**Rankings**
To build and calibrate my trust in this AI	18	**1** (7), **2** (4), **3** (3), **4** (3), **5** (1),
When I doubt about the prediction from AI	15	**1** (8), **2** (1), **3** (3), **4** (3),
To verify AI's decisions	16	**1** (2), **2** (6), **3** (3), **4** (2), **5** (1), **6** (2),
To ensure the safety use of the AI	11	**1** (2), **2** (2), **3** (4), **4** (3),
For a difficult case, when I am not certain	10	**1** (1), **2** (3), **3** (1), **4** (2), **5** (2), **7** (1),
To learn from AI	7	**2** (1), **3** (1), **4** (2), **6** (2), **8** (1),
To improve my patients' outcomes	5	**1** (3), **4** (1), **5** (1),
To ensure fairness and no biases in the AI model	5	**2** (1), **3** (2), **5** (1), **9** (1),
When I am trading off among multiple objectives for my patient	2	**2** (1), **3** (1),
To meet the ethical requirements	2	**2** (1), **5** (1),
To make Differential Diagnosis	3	**3** (1), **5** (1), **10** (1),
To make new medical discovery	3	**5** (1), **6** (1), **7** (1),
Before discussion with my colleagues	1	**5** (1),
To meet the legal requirements	1	**6** (1),
To generate report or patient chart	0	

Table 4: Ranking of explanation goals. The "Selected times" are the number of times each goal was selected by participants, and "Rankings" shows participants' ranking (in bold) and the frequency of such ranking (in parentheses).

In addition to the above general explanation goal in a clinical setting, for each MRI case, we asked a yes/no question on whether the participant needs to check the AI explanation (need explanation). We calculated each participant's need explanation degree (0.0: do not need explanation for any cases, and 1.0: need explanation for all cases) by the ratio of "yes" answers out of all the recorded responses among the 25 MRI cases. The trend of need explanation shows polarization: 66% (23/35) of the participants had a need explanation degree of less than 0.3, and 20% (7/35) had a need explanation degree of above 0.7. In particular, 14% (5/35) of the participants completely did not need any explanation for all cases (need explanation degree = 0.0), and 9% (3/35) of the participants needed explanations for every case (need explanation degree = 1.0). We further conducted a chi-square test of independence on whether need explanation was associated with decision agreement, and the relationship between the two variables was significant, $\chi^2(1, N = 746) = 62.7, p < 0.001$. When there was a decision disagreement between AI and physicians' initial judgment, physicians were more likely to check the explanation.

For the quality of the specific explanation content, we collected 744 ratings on the color map explanation quality on a 0-10 point scale, and obtained an average quality rating of 6.12±2.92 (mean±std). Each color map explanation received ratings with large standard deviations ranging from 2.39 to 3.04. The rating for each of the 25 AI explanations is listed in Supplemental S2 Table 2, and we visualize four explanations with high or low ratings in Fig. 1.

4 Discussion

4.1 The clinical utility of AI prediction

In our study on the glioma grading clinical task, physicians' initial task performance was lower than AI performance. With AI assistance, physicians' task performance significantly increased. However, the improvement did not exceed AI performance. The result aligns with prior phase II clinical evaluation of AI on medical image analysis tasks to diagnose knee lesions [9], diabetic retinopathy [41], and pulmonary adenocarcinoma [31], where physicians exhibited better task performance with the assistance of a superior AI (AI that outperformed physicians); But it diverges from the similar phase II study in psychiatry on medical record data [24], where a superior AI assistance did not improve clinicians' accuracy in treatment selection. These divergent results indicate the variability in the effect of AI assistance in clinical settings, and suggest the importance of conducting clinical studies to validate the clinical utility of AI assistance on specific clinical tasks.

Our clinical study validates the clinical utility of AI as a physician performance booster in the glioma grading task. The quantitative result echoes the qualitative result (Supplemental S1 Section 1) wherein physicians regarded AI as a "*second opinion*" (N2), or "*another level of evidence*" (N5). Quantitative results showed that such performance improvement was more prominent for junior physicians (residents and fellows, Supplemental S2 Section 2.2). Qualitative results further showed junior physicians can potentially benefit from AI in time-sensitive cases and hard cases, and AI can potentially improve junior physicians' learning and problem-solving skills by "*reaffirming what you're learning*" (N2).

By further analyzing the decision change pattern, the observed physician performance improvement with the assistance of AI prediction was mainly due to the fact that physicians' decision patterns converged to be similar to AI, as physicians only switched their decisions during decision disagreement (Fig. 4). The decision disagreement between physicians and AI caused physicians "*to pause and then go through the images ... to understand the disagreement*" (N5). But since in this condition, physicians were only assisted by AI prediction alone, they could not access more information from AI decision process to resolve disagreement, and one of the expected utilities of AI explanation is to provide additional information to facilitate physicians' decision-making.

In addition, our study also inspected factors that influence physicians' adoption of AI for decision support, including their trust and willingness to use AI. We noticed that after viewing AI performance metrics, physicians' trust and willingness to use AI increased significantly compared to the baseline. However, after using the AI on 25 cases with the assistance of AI prediction and explanation, physicians' trust and willingness to use AI diverged, and the average ratings of trust and willingness to use AI did not show a significant difference compared with the ones after only viewing AI performance metrics. This indicates physicians may perceive different messages from AI prediction and explanation information, and construct different mental models of AI [7] accordingly to calibrate their trust and willingness to use AI. Such a hypothesis is evidenced by physicians' positive and negative comments for AI explanation from the qualitative data (detailed in Section 4.2). Furthermore, qualitative results showed that to establish trust, some physicians request more information beyond the AI performance metrics alone, such as information on prediction confidence and dataset (Supplemental S1 Section 3.3).

4.2 The clinical utility of AI explanation

In the study, with the additional assistance of AI explanation, physicians' task performance did not show a significant difference compared to the performance with AI prediction assistance only, and did not achieve complementary doctor-AI team performance. The finding aligns with a similar phase II clinical study involving AI heatmap explanation in diabetic retinopathy [41], and other similar AI-supported decision-making experiments involving laypersons on an age prediction task [16], and on a criminal justice decision support [2], where presenting AI prediction alone would improve human accuracy, but there was no additional performance boost with the assistance of feature attribution (i.e. heatmap) explanation. This indicates that the existing AI explanation failed to indicate for physicians when to rely on AI recommendations, and when not to. Otherwise, the doctor-AI team performance would have been higher than either AI or physician alone. Indeed, by looking into their fine-grained decision change pattern, physicians had initiated both correct and incorrect decision changes that were relatively equivalent in amount, which explains the source of the statistically insignificant difference in accuracies between the assistance of AI prediction and the additional explanation. Prior human-subject studies in the human-computer interaction field observed a similar effect of AI explanation in decision support: the AI explanation tends to only increase the chances of humans accepting AI suggestions regardless of AI correctness [8, 24, 30]. This indicates that additional strategies [17, 38, 32, 21] are needed to carefully craft the design of explainable AI algorithms and interfaces to reduce such overreliance risk on AI [10, 12] and achieve its desired clinical impact, such as complementary doctor-AI task performance.

Qualitative results revealed reasons for the failure of AI explanation to boost physician-AI team performance. Physicians had a mixed view of the clinical utility of AI explanation: they saw the heatmap explanation as a useful tool to help them localize important features and easy-to-miss lesions. They also used explanation as a "cross-check" (N3) tool to verify AI decisions, calibrate their trust in AI, and ensure the safe use of AI (N1), especially during decision disagreement (Supplemental S1 Section 2). However, when physicians were using AI explanation to verify AI decisions, they found that the heatmap explanation only provided limited information on the location of important features, but failed to give explicit reasons, contexts, or descriptions of the highlighted features (Supplemental S1 Section 3.1). For example, doctors would request reasons why the heatmap explanation only highlighted part of the relevant features: "*What does that (color map region) mean? ... Was it central necrosis? But it couldn't be the central necrosis, because there's more central necrosis in the temporal lobe, and that area didn't get highlighted. So anyway, I don't know, it's just confusing.*" (N3) This phenomenon can be shown in panels A, C, D in Fig. 1, where the heatmap only highlighted partial regions of the contrast-enhancing tumor edge without a pathological description of the features or why it was important. Similarly, doctors would also request reasons why some image features outside the tumor were highlighted: "*There are quite a few areas that tend to 'light up' on the prediction map that are false positives. And it would be good to know where/why these false positive areas are interpreted as such.*" (O1) This phenomenon can be illustrated in panel B in Fig. 1. This finding echoes similar feedback from pathologists in a user study using heatmap explanations [18]. By comparing the heatmap explanation information with the clinical decision-making process (Supplemental S1 Section 3.1), we further identified that the existing heatmap explanation was missing critical information to construct a clinically relevant explanation [28]: the heatmap explanation neither provided descriptive information on the pathology of the highlighted image features, nor justified why and how the highlighted regions lead to the AI decision [20]. Such information is indispensable for physicians to construct a complete chain of reasoning with AI explanation, and future work is needed to develop new XAI techniques that incorporate more clinically relevant information into the form of AI explanation, such as combining text description with feature localization.

The ranking of explanation goals revealed a wide range of clinical usage scenarios to use AI explanation in a clinical context. In addition, physicians' need for explanation showed polarization and varied from person to person, which indicates the use of explanation could be an individualized choice, and the contents of explanation could be personalized or presented on demand [23, 25].

5 Limitations and future work

Despite the national-scale study, the total number of participants was 35, which was relatively small. Despite our best recruitment effort, we did not get any enrollment of radiologists. These factors limit further statistical analysis, such as multivariate regression analysis to identify variables associated with physicians' performance improvement. The study was a phase II evaluation using a simulated clinical task on retrospective MRI data only, and participants neither had access to patients nor their information other than MRI. To fully assess the clinical utility of AI and its explanation, future work can conduct phase II studies on the newly improved AI systems. If the results are promising, phase III randomized controlled clinical user studies are needed on tasks in real-world clinical settings on retrospective clinical data or prospective cases. In addition to assessing efficacy, more work is needed to evaluate the safety and side effects when using AI in clinical settings to get a balanced view of the strengths and limitations of AI. The study design on the sequence of introducing prediction and explanation in doctors' decision-making process may introduce bias, because we cannot distinguish whether the decision after introducing AI explanation was due to the effect of users' interpretation of explanation, or the anchoring effect [50] from users' previous judgment and inspection of AI prediction, where users were reluctant to change their decisions. Further work may design studies to mitigate the bias introduced by the sequence of exposure, such as comparing two scenarios of showing explanation or prediction first, and testing differences in the outcomes.

Future work may improve the existing XAI methods in the following ways: 1) in the technical development phase, XAI developers and researchers should seek more clinical input to understand the clinical reasoning and clinical users' requirements when incorporating AI assistance in their decision-making process, so that the XAI techniques can potentially achieve complementary human-AI performance; 2) In the clinical deployment phase, additional training or tutorial sessions may be developed to enable clinical users to understand the capability of AI, and incorporate the additional cognitive strategies while interpreting the AI explanation.

In AI and explanation-assisted clinical decision settings, there can be different ways of human-AI collaboration, across a spectrum from AI passively involved in decision-making by providing a second opinion and explanation when needed, to AI actively participating in the decision-making and requiring humans' review of its decisions and explanations. In clinical settings, users may choose or switch among different ways of collaboration with AI. Our study investigated one setting similar to the second opinion one, where users form their own judgment before checking AI prediction and its explanation. With the newly improved XAI methods, future study design can explore other ways of human-AI collaboration and its influence on the clinical utility of AI and explanation.

6 Conclusion

As a fast-advancing technology, AI has the potential to transform neuro-oncological practice and assist physicians in a variety of clinical tasks such as tumor segmentation and disease prediction. To overcome the clinical translational gap of moving AI from bench to bedside and accumulate evidence for the safe and effective use of AI, we conducted a phase II evaluation on the clinical utility of AI and its explanation, which is analogous to the phase II clinical trial on the primary efficacy of a new intervention on a small-scale population.

The Canada-wide online survey study recruited 35 neurosurgeons. Each participant read 25 brain MRIs from patients with gliomas, and gave their judgment on the glioma grading without and with the assistance of AI prediction and explanation. Results showed that compared to physicians performing the task alone, when assisted by AI prediction, physicians' task performance increased significantly to be equivalent to AI performance. But the additional assistance of AI explanation did not further boost physicians' performance. Complementary doctor-AI team performance was not achieved. In addition, physicians' trust in the AI system and willingness to use AI increased after viewing AI performance metrics compared to the baseline, and such levels did not change after using AI predictions and explanations. The study showed the clinical utility of AI assistance in improving physicians' task performance, and revealed limitations of existing AI explanation techniques for future improvement.

Acknowledgements

We thank all the physician participants for their time and valuable input in the study. We thank Ben Cardoen, Hanene Ben Yedder, and Kumar Abhishek for helping us review the manuscript. We thank the anonymous reviewers for the helpful comments. This study was funded by BC Cancer Foundation–BrainCare Fund. This research was also enabled in part by the computational resources provided by NVIDIA and the Digital Research Alliance of Canada (alliancecan.ca).

References

[1] Julie Adhya, Charles Li, Laura Eisenmenger, Russell Cerejo, Ashis Tayal, Michael Goldberg, and Warren Chang. Positive predictive value and stroke workflow outcomes using automated vessel density (rapid-cta) in stroke patients: One year experience. *The Neuroradiology Journal*, 34(5):476–481, 2021. PMID: 33906499.

[2] Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6618–6626, May 2021.

[3] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels for the pre-operative scans of the tcga-gbm collection. 2017.

[4] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels for the pre-operative scans of the tcga-lgg collection. 2017.

[5] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1), September 2017.

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11405–11414, May 2021.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2429–2437, Jul. 2019.

[8] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[9] Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F.

Beaulieu, Geoffrey M. Riley, Russell J. Stewart, Francis G. Blankenberg, David B. Larson, Ricky H. Jones, Curtis P. Langlotz, Andrew Y. Ng, and Matthew P. Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Medicine*, 15(11):e1002699, nov 2018.

[10] Zana Buccinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.

[11] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[12] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169, 2015.

[13] Shan Carter and Michael Nielsen. Using artificial intelligence to augment human intelligence. *Distill*, 2(12), December 2017.

[14] P Chang, J Grinband, B D Weinberg, M Bardis, M Khy, G Cadena, M.-Y. Su, S Cha, C G Filippi, D Bota, P Baldi, L M Poisson, R Jain, and D Chow. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *American Journal of Neuroradiology*, 39(7):1201–1207, 2018.

[15] Kyu Sung Choi, Seung Hong Choi, and Bumseok Jeong. Prediction of IDH genotype in gliomas with dynamic susceptibility contrast perfusion MR imaging using an explainable recurrent neural network. *Neuro-Oncology*, 21(9):1197–1209, sep 2019.

[16] Eric Chu, Deb Roy, and Jacob Andreas. Are visual explanations useful? A case study in model-in-the-loop prediction. *CoRR*, abs/2007.12248, 2020.

[17] Pat Croskerry. Cognitive forcing strategies in clinical decisionmaking. *Annals of Emergency Medicine*, 41(1):110–120, 2003.

[18] Theodore Evans, Carl Orge Retzlaff, Christian Geißler, Michaela Kargl, Markus Plass, Heimo Müller, Tim-Rasmus Kiehl, Norman Zerbe, and Andreas Holzinger. The explainability paradox: Challenges for xai in digital pathology. *Future Generation Computer Systems*, 133:281–296, 2022.

[19] Mostafa Fatehi, Leeor S. Yefet, Swetha Prakash, Brian D. Toyota, and Peter A. Gooderham. Current trends in neurosurgical management of adult diffuse low-grade gliomas in canada. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*, page 1–4, 2022.

[20] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, nov 2021.

[21] Mark L. Graber, Stephanie Kissam, Velma L. Payne, Ashley N.D. Meyer, Asta Sorensen, Nancy Lenfestey, Elizabeth Tant, Kerm Henriksen, Kenneth LaBresh, and Hardeep Singh. Cognitive interventions to reduce diagnostic error: a narrative review. *BMJ quality & safety*, 21(7):535–557, jul 2012.

[22] Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1):30–36, jan 2019.

[23] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, may 2021.

[24] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry 2021 11:1*, 11(1):1–9, feb 2021.

[25] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. EUCA: the end-user-centered explainable AI framework. 2021.

[26] Weina Jin, Mostafa Fatehi, Kumar Abhishek, Mayur Mallya, Brian Toyota, and Ghassan Hamarneh. Artificial Intelligence in Glioma Imaging: Challenges and Advances. *Journal of Neural Engineering*, 17(2):21002, April 2020.

[27] Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks. *MethodsX*, 10:102009, 2023.

[28] Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical Image Analysis*, 84:102684, 2023.

[29] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), October 2019.

[30] Himabindu Lakkaraju and Osbert Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 79–85, New York, NY, USA, 2020. Association for Computing Machinery.

[31] Jiaoyang Li, Lingxiao Zhou, Yi Zhan, Haifeng Xu, Cheng Zhang, Fei Shan, and Lei Liu. How does the artificial intelligence-based image-assisted technique help physicians in diagnosis of pulmonary adenocarcinoma? A randomized controlled experiment of multicenter physicians in China. *Journal of the American Medical Informatics Association*, 10 2022. ocac179.

[32] Geoffrey K. Lighthall and Cristina Vazquez-Guillamet. Understanding Decision Making in Critical Care. *Clinical Medicine & Research*, 13(3-4):156–168, dec 2015.

[33] Luke Macyszyn, Hamed Akbari, Jared M. Pisapia, Xiao Da, Mark Attiah, Vadim Pigrish, Yingtao Bi, Sharmistha Pal, Ramana V. Davuluri, Laura Roccograndi, Nadia Dahmane, Maria Martinez-Lage, George Biros, Ronald L. Wolf, Michel Bilello, Donald M. O'Rourke, and Christos Davatzikos. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-Oncology*, 18(3):417–425, mar 2016.

[34] John W. Mauchly. Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, 11(2):204–209, 1940.

[35] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, October 2015.

[36] David S Moore. *Introduction to the Practice of Statistics*. WH Freeman and company, 2009.

[37] Urs J. Muehlematter, Paola Daniore, and Kerstin N. Vokinger. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *The Lancet Digital Health*, 3(3):e195–e203, mar 2021.

[38] Geoffrey R. Norman, Sandra D. Monteiro, Jonathan Sherbino, Jonathan S. Ilgen, Henk G. Schmidt, and Silvia Mamede. The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. *Academic Medicine*, 92(1):23–30, jan 2017.

[39] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. AI in health and medicine. *Nature Medicine*, 28(1):31–38, 2022.

[40] Ramin Ranjbarzadeh, Abbas Bagherian Kasgari, Saeid Jafarzadeh Ghoushchi, Shokofeh Anari, Maryam Naseri, and Malika Bendechache. Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Scientific Reports*, 11(1):10930, 2021.

[41] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Scott Barb, Anthony Joseph, Michael Shumski, Jesse Smith, Arjun B. Sood, Greg S. Corrado, Lily Peng, and Dale R. Webster. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology*, 126(4):552–564, apr 2019.

[42] S. S. SHAPIRO and M. B. WILK. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611, 12 1965.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[44] Gagandeep Singh, Sunil Manjila, Nicole Sakla, Alan True, Amr H Wardeh, Niha Beig, Anatoliy Vaysberg, John Matthews, Prateek Prasanna, and Vadim Spektor. Radiomics and radiogenomics in gliomas: a contemporary update. *British Journal of Cancer*, 125(5):641–657, 2021.

[45] George C M Siontis, Romy Sweda, Peter A Noseworthy, Paul A Friedman, Konstantinos C Siontis, and Chirag J Patel. Development and validation pathways of

artificial intelligence tools evaluated in randomised clinical trials. *BMJ Health & Care Informatics*, 28(1), 2021.

[46] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

[47] Joseph J. Titano, Marcus Badgeley, Javin Schefflein, Margaret Pain, Andres Su, Michael Cai, Nathaniel Swinburne, John Zech, Jun Kim, Joshua Bederson, J. Mocco, Burton Drayer, Joseph Lehar, Samuel Cho, Anthony Costa, and Eric K. Oermann. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature Medicine*, 24(9):1337–1341, sep 2018.

[48] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

[49] Stefano Triberti, Ilaria Durosini, Giuseppe Curigliano, and Gabriella Pravettoni. Is explanation a marketing problem? the quest for trust in artificial intelligence and two conflicting solutions. *Public Health Genomics*, 23(1-2):2–5, 2020.

[50] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[51] Pu Wang, Xiaogang Liu, Tyler M Berzin, Jeremy R Glissen Brown, Peixi Liu, Chao Zhou, Lei Lei, Liangping Li, Zhenzhen Guo, Shan Lei, Fei Xiong, Han Wang, Yan Song, Yan Pan, and Guanyu Zhou. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *The Lancet Gastroenterology &amp Hepatology*, 5(4):343–351, April 2020.

[52] Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6):70–79, may 2019.

[53] Hao Zhou, Martin Vallières, Harrison X. Bai, Chang Su, Haiyun Tang, Derek Oldridge, Zishu Zhang, Bo Xiao, Weihua Liao, Yongguang Tao, Jianhua Zhou, Paul Zhang, and Li Yang. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro-Oncology*, 19(6):862–870, jun 2017.

[54] Qian Zhou, Zhi-hang Chen, Yi-heng Cao, and Sui Peng. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *npj Digital Medicine*, 4(1):154, 2021.