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

Weina Jin[1] (co-first author), **Mostafa Fatehi**[2] (co-first author),

Ru Guo[2], **Ghassan Hamarneh**[1]

1. Medical Imaging Analysis Lab, School of Computing Science, Simon Fraser University
2. Division of Neurosurgery, The University of British Columbia

HAMARNEH
Medical Image Analysis Research Group

SFU SIMON FRASER UNIVERSITY

UBC THE UNIVERSITY OF BRITISH COLUMBIA

Code available at
github.com/weinajin/multimodal_explanation

Outline A phase 2 clinical user study of AI on glioma grading

1. **Study motivation**
2. **Study design**: material, participants, procedure
3. **Result**: doctor+AI > doctor?

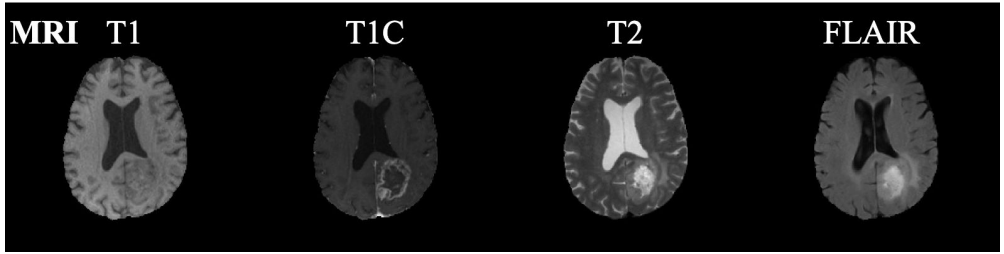
Disclosure statement

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).
All authors declare no financial or non-financial competing interests.

Motivation

Glioma: most common primary tumor of brain and spine

Initial investigation includes CT or MRI for diagnosis, tumor grading, treatment planning, progression tracking, prognosis, etc.

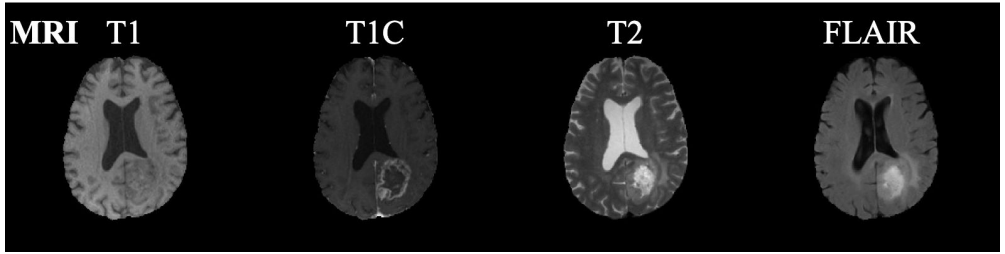


[1] Artificial intelligence in glioma imaging: challenges and advances. J Neural Eng. 2020
[2] Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open. 2020.

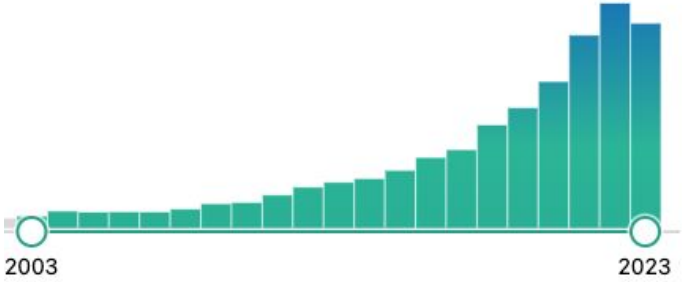
Motivation

Glioma: most common primary tumor of brain and spine

Initial investigation includes CT or MRI for diagnosis, tumor grading, treatment planning, progression tracking, prognosis, etc.



Artificial intelligence (AI) may be a potential tool to assist doctors in glioma imaging tasks, as it is trained to recognize patterns from a large amount of data



PubMed keywords: glioma + AI
https://bit.ly/AI_in_glioma_imaging [1]

[1] Artificial intelligence in glioma imaging: challenges and advances. J Neural Eng. 2020
[2] Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open. 2020.

Motivation

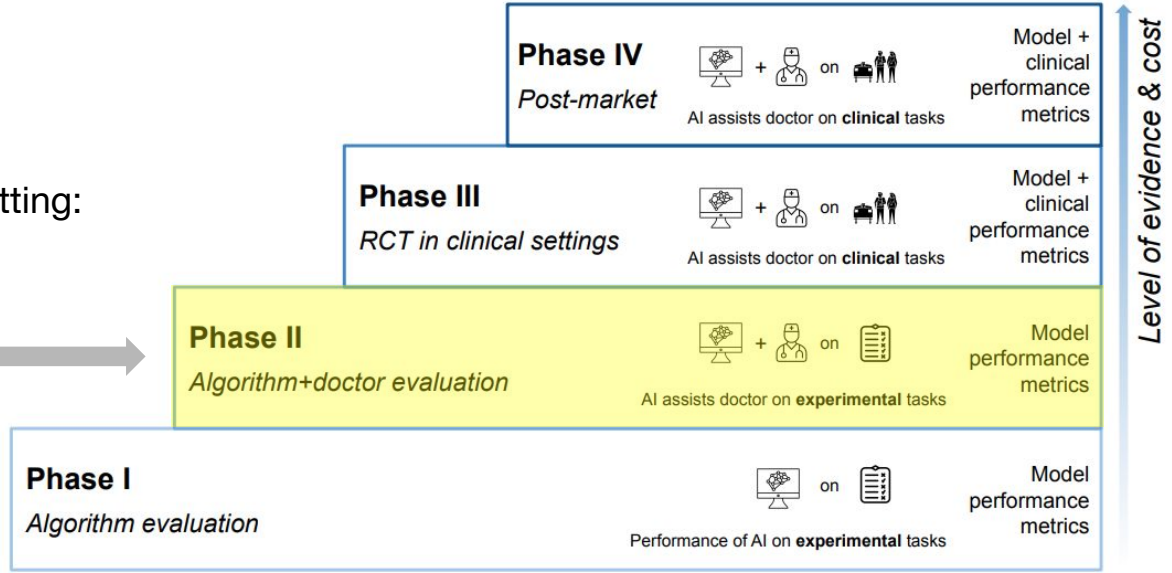
To safeguard the use of AI in clinical setting:

1. **Clinical evaluation**

   Phase 2 clinical study

2. **eXplainable AI (XAI)**

   Being able to explain decisions to clinical users



The four phases of evaluating the clinical utility of AI in glioma imaging [1].

[1] Artificial intelligence in glioma imaging: challenges and advances. J Neural Eng. 2020
[2] Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open. 2020.

Research question

Doctor + AI > Doctor ?

Research question

Doctor + AI > Doctor ?

If so,

Doctor + AI > max(Doctor, AI) ?

Complementary doctor-AI performance

Research question

Doctor + AI > Doctor ?

Suggestion Explanation

If so,

Doctor + AI > max(Doctor, AI) ?

Suggestion Explanation

→ Complementary doctor-AI performance

Outline A phase 2 clinical user study of AI on glioma grading

1. **Study motivation**
2. **Study design**: material, participants, procedure
3. **Result**: doctor+AI > doctor?

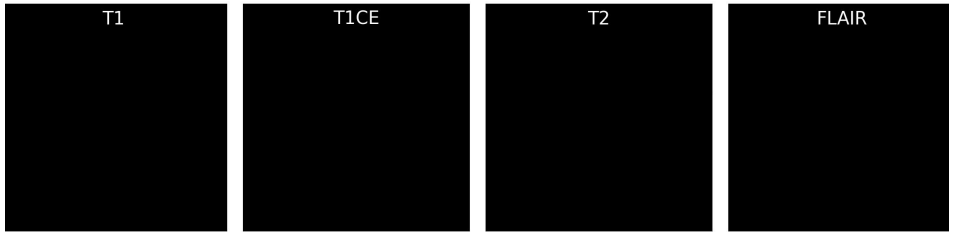
Glioma grading on MRI Data

BraTS 20' Dataset [1]

n=369

Grade 2-3 glioma

n=76



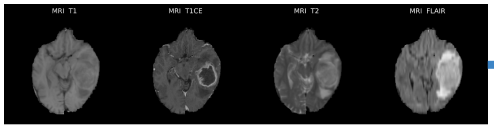
Tumor mask contour

Grade 4 glioblastoma

n=293

[1] The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). Menze, et al., IEEE TMI 2015.

Glioma grading on MRI AI assistance: a suggestion + an explanation



Input MRI

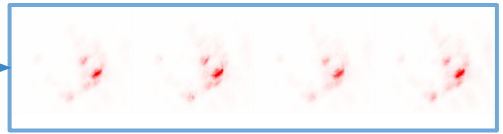
AI model

3D convolutional neural network

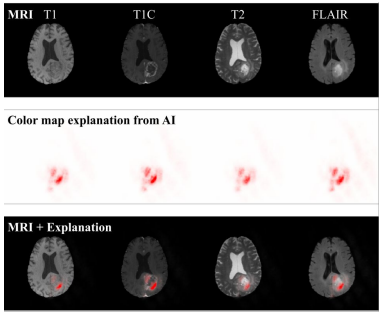
AI's suggestion:
Grade 4 glioblastoma

Accuracy: 88%

AI's explanation:

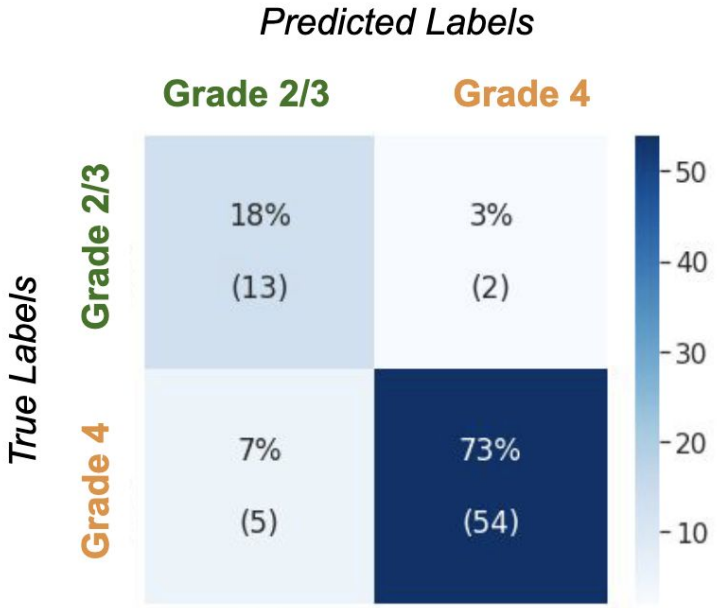


SmoothGrad XAI algorithm [1]



[1] Smoothgrad: removing noise by adding noise, 2017. Arxiv: 1706.03825

Glioma grading on MRI AI model performance



	Sensitivity	Specificity	Positive Predictive Value	F1 score	Number of class in the test set
Grade 2 or 3 glioma	0.87	0.92	0.72	0.79	*15*
Grade 4 glioblastoma	0.92	0.87	0.96	0.94	*59*

Candidates of 16 post-hoc heatmap explanation methods on the glioma task

Gradient based

Grad-CAM

Gradient

Input x Gradient

SmoothGrad

Deconvolution

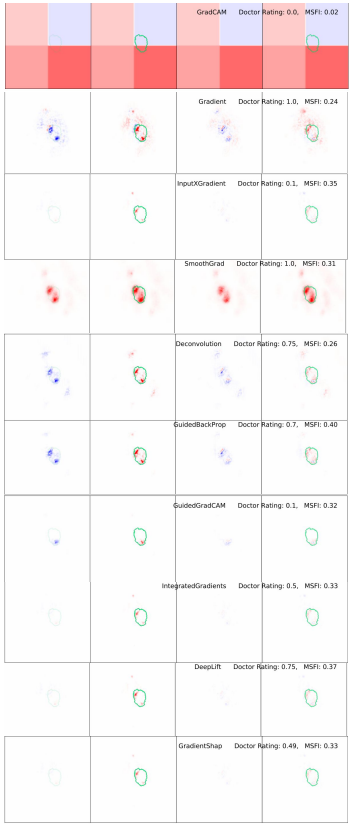
Guided Backpropagation

Guided Grad-CAM

Integrated Gradient

DeepLIFT

Gradient SHAP



Perturbation based

Occlusion

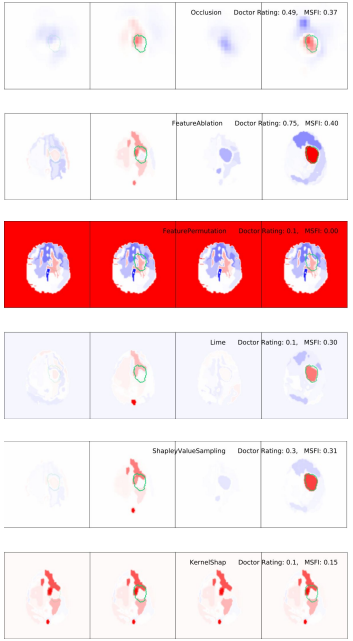
Feature Ablation

Feature Permutation

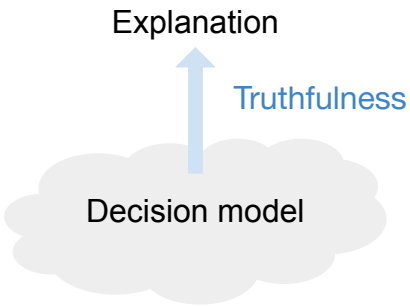
LIME

Shapley Value Sampling

Kernel SHAP

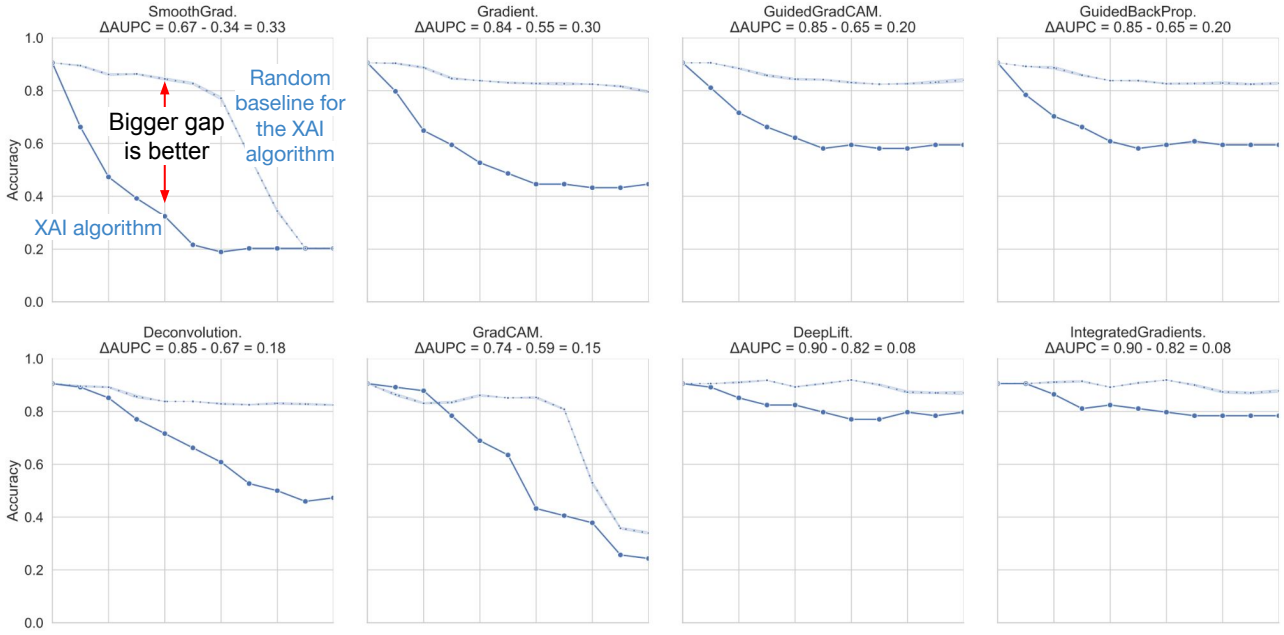


The AI explanation method was selected for being relatively truthful to AI decision-making



Assumption:
Removing important features will cause AI performance to drop.

Gradual feature removal experiment



.....

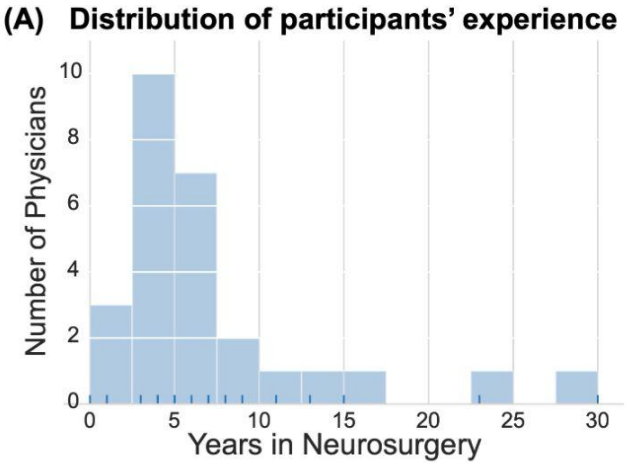
Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis, 2023

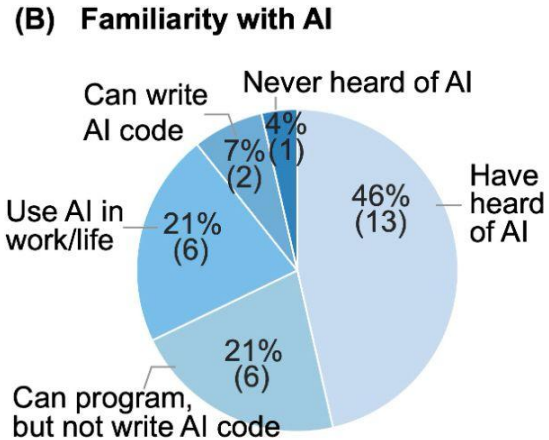
Participants

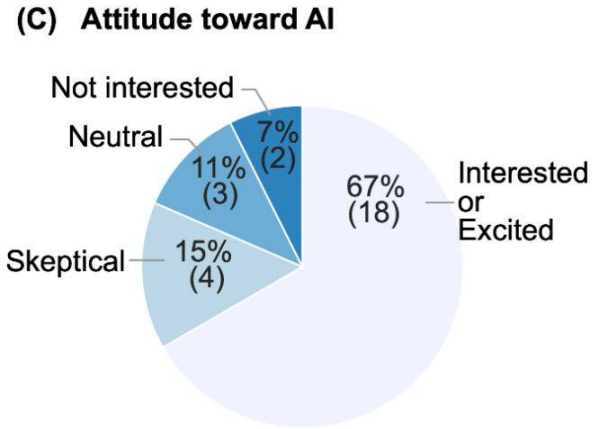
- **Inclusion criteria**
  - Neurosurgeon, radiologist, or neuro-radiologist;
  - Attending, fellow, or resident
- **Eligibility screening** by:
  - survey screening questions, and
  - **DR** task accuracy>0.55

Participants

- 35 participants (recruitment rate 15%) in neurosurgery
- 12 attending, 2 fellows, 21 residents
- Years of practicing neurosurgery: 7.1 ± 6.5
- Female: male = 7:19; age: 34.7±8.2



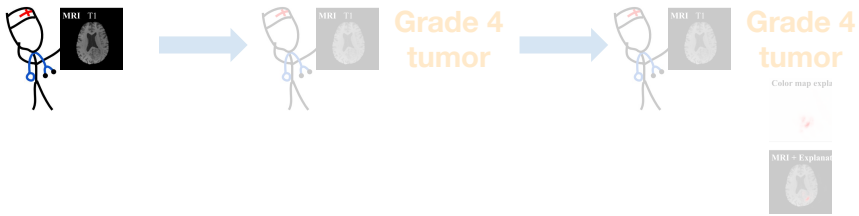




Study design

- National online **survey**, 25 MRI, 30-40 min.

  Participants gave judgment at three conditions:



1. **DR**

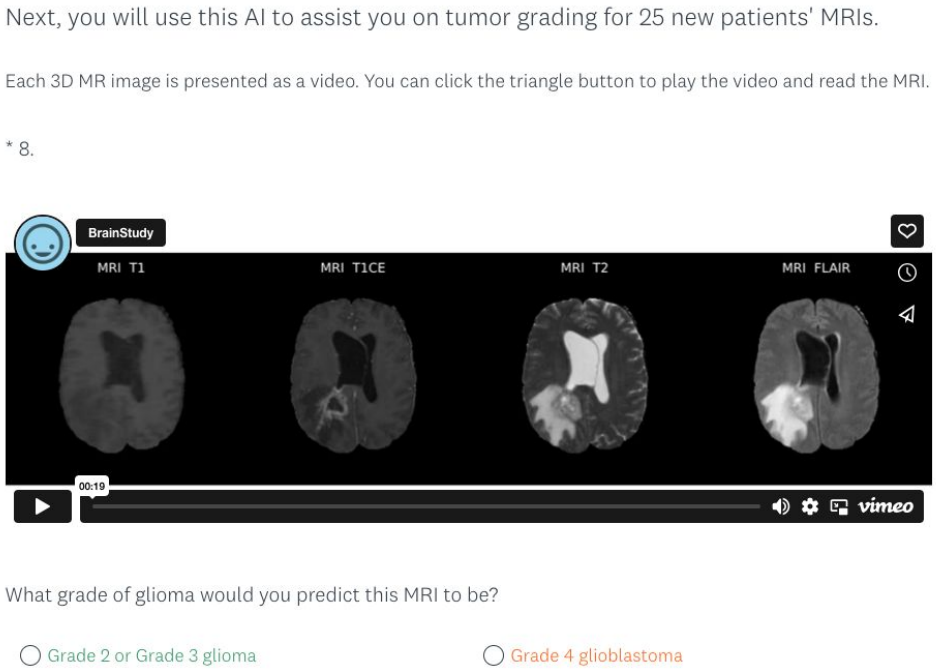
Doctor only

2. **DR+AI**

Doctor assisted by AI **prediction**

3. **DR+XAI**

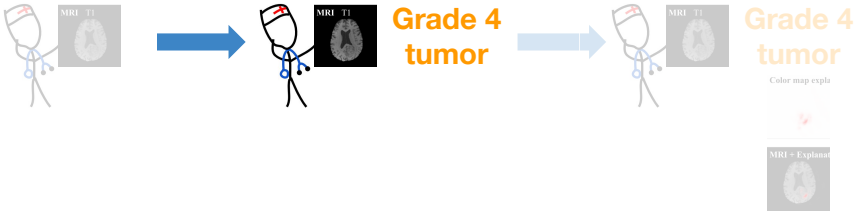
Doctor assisted by AI prediction & **explanation**



Study design

- National online **survey**, 25 MRI, 30-40 min.

  Participants gave judgment at three conditions:



1. DR

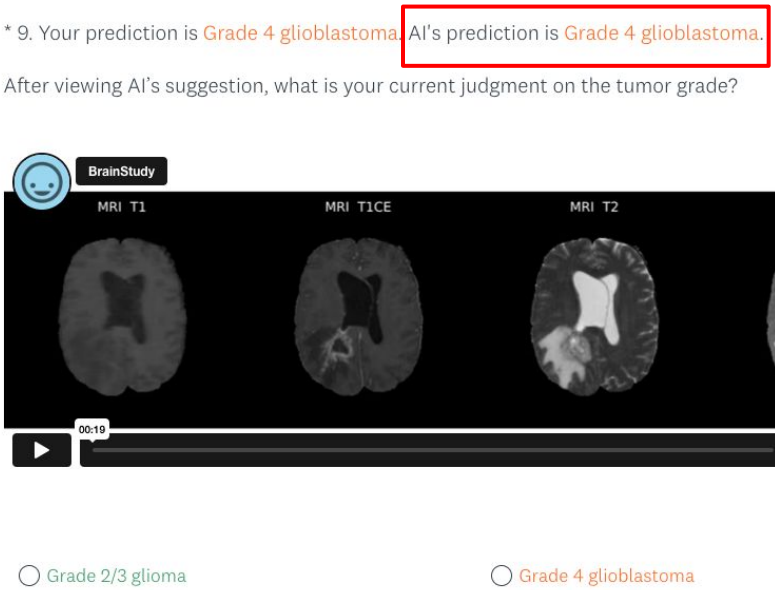
Doctor only

2. **DR+AI**

Doctor assisted by AI **prediction**

3. DR+XAI

Doctor assisted by AI prediction & explanation

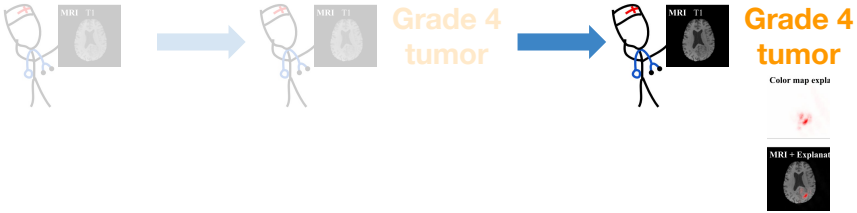


Study Ethics No.: H20-03588

Study design

- National online **survey**, 25 MRI, 30-40 min.

  Participants gave judgment at three conditions:



1. **DR**

Doctor only

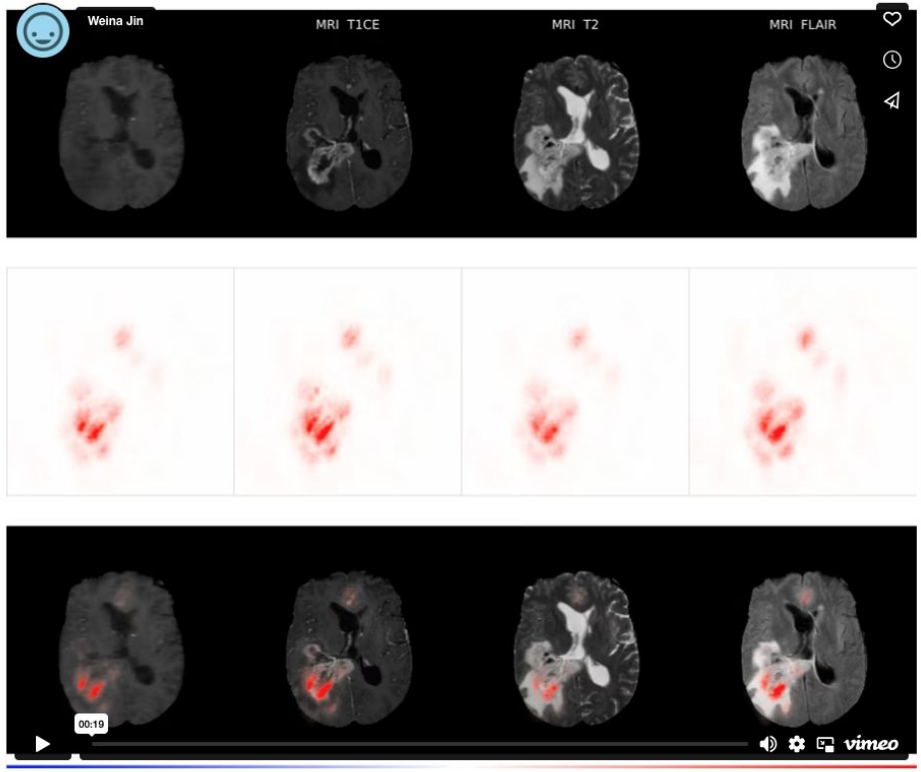
2. **DR+AI**

Doctor assisted by AI **prediction**

3. **DR+XAI**

Doctor assisted by AI prediction & **explanation**

* 11. Please click the triangle "play" button to play the video and read the color map explanation.



After viewing AI's explanation, what is your final judgment on the tumor grade?

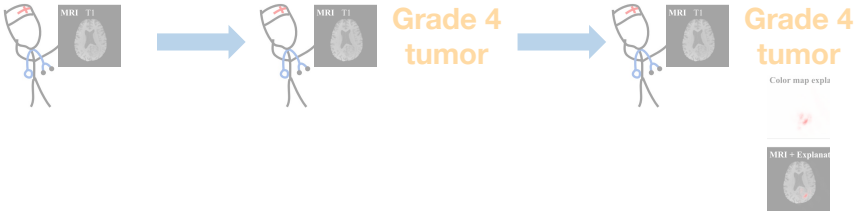
◯ Grade 2/3 glioma ◯ Grade 4 glioblastoma

Study Ethics No.: H20-03588

Study design

- National online **survey**, 25 MRI, 30-40 min.

  Participants gave judgment at three conditions:



1. **DR**

Doctor only

2. **DR+AI**

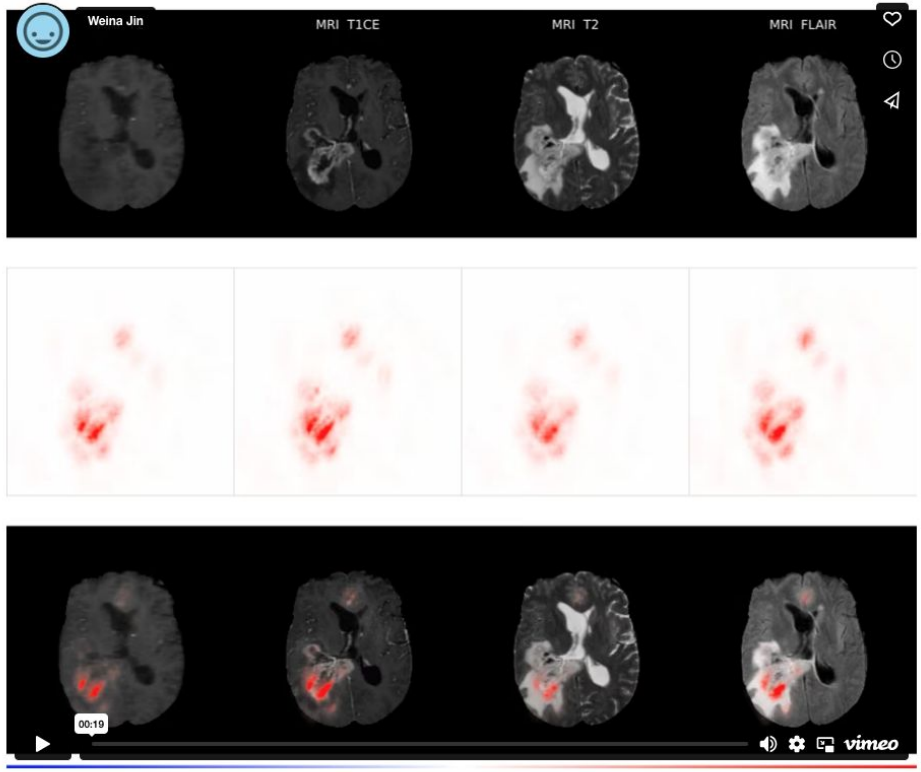
Doctor assisted by AI **prediction**

3. **DR+XAI**

Doctor assisted by AI prediction & **explanation**

- Post-survey, one-to-one **interview**, 20-30 min

* 11. Please click the triangle "play" button to play the video and read the color map explanation.



After viewing AI's explanation, what is your final judgment on the tumor grade?

◯ Grade 2/3 glioma ◯ Grade 4 glioblastoma

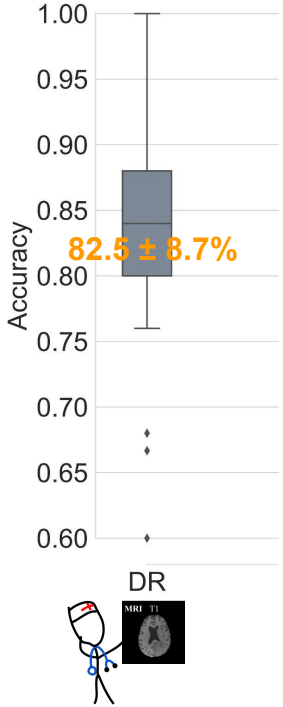
Study Ethics No.: H20-03588

Outline A phase 2 clinical user study of AI on glioma grading

1. **Study motivation**
2. **Study design**: material, participants, procedure
3. **Result**: doctor+AI > doctor?

Result

Is doctor+AI better than doctor only?



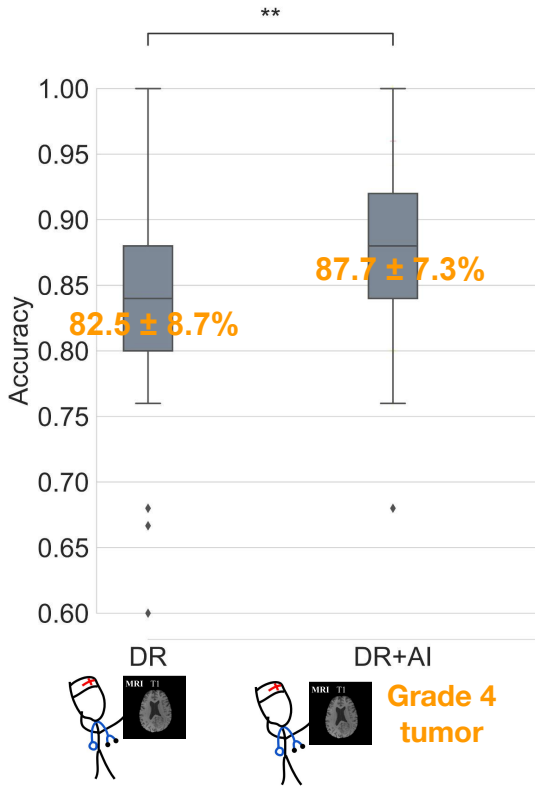
Result

Is doctor+AI better than doctor only?

Doctor + AI > Doctor ✅

On improving doctors' task performance:

- AI prediction (**DR+AI**) is helpful



Wilcoxon signed-rank tests with Bonferroni correction

Result

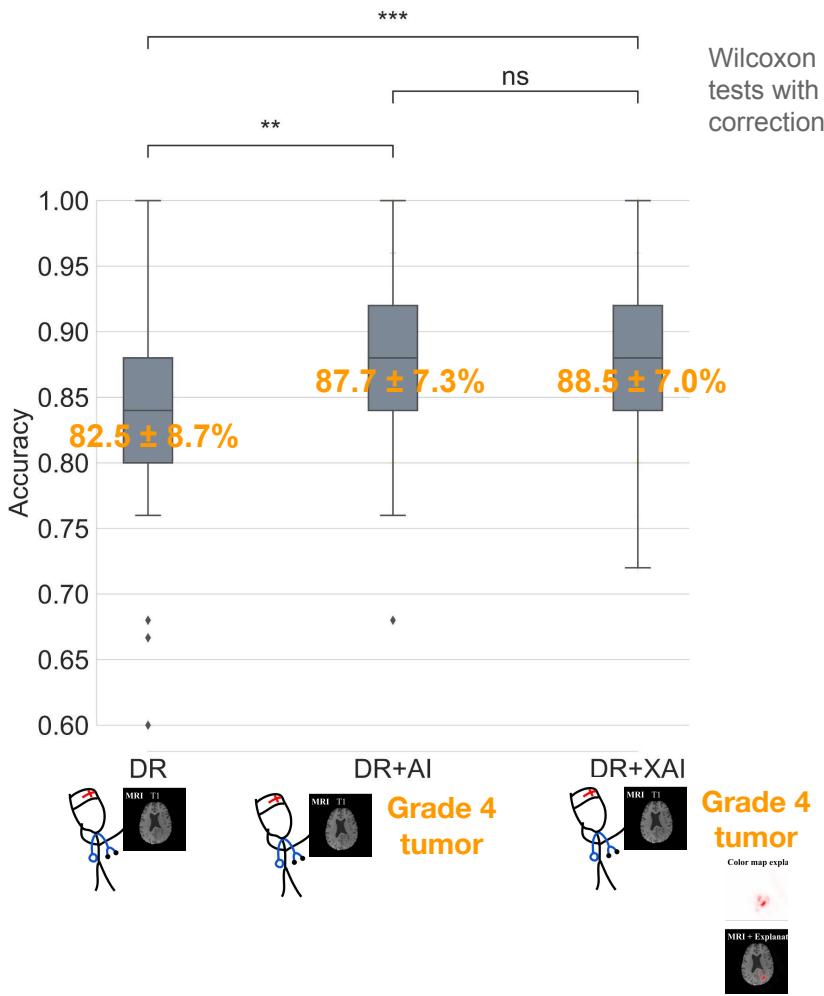
Is doctor+AI better than doctor only?

Doctor + AI > Doctor ✅

Suggestion Explanation

On improving doctors' task performance:

- AI prediction (**DR+AI**) is helpful
- AI explanation (**DR+XAI**) not show additional value



Wilcoxon signed-rank tests with Bonferroni correction

Result

Is doctor+AI better than doctor only?

Doctor + AI > Doctor ✅

Suggestion Explanation

On improving doctors' task performance:

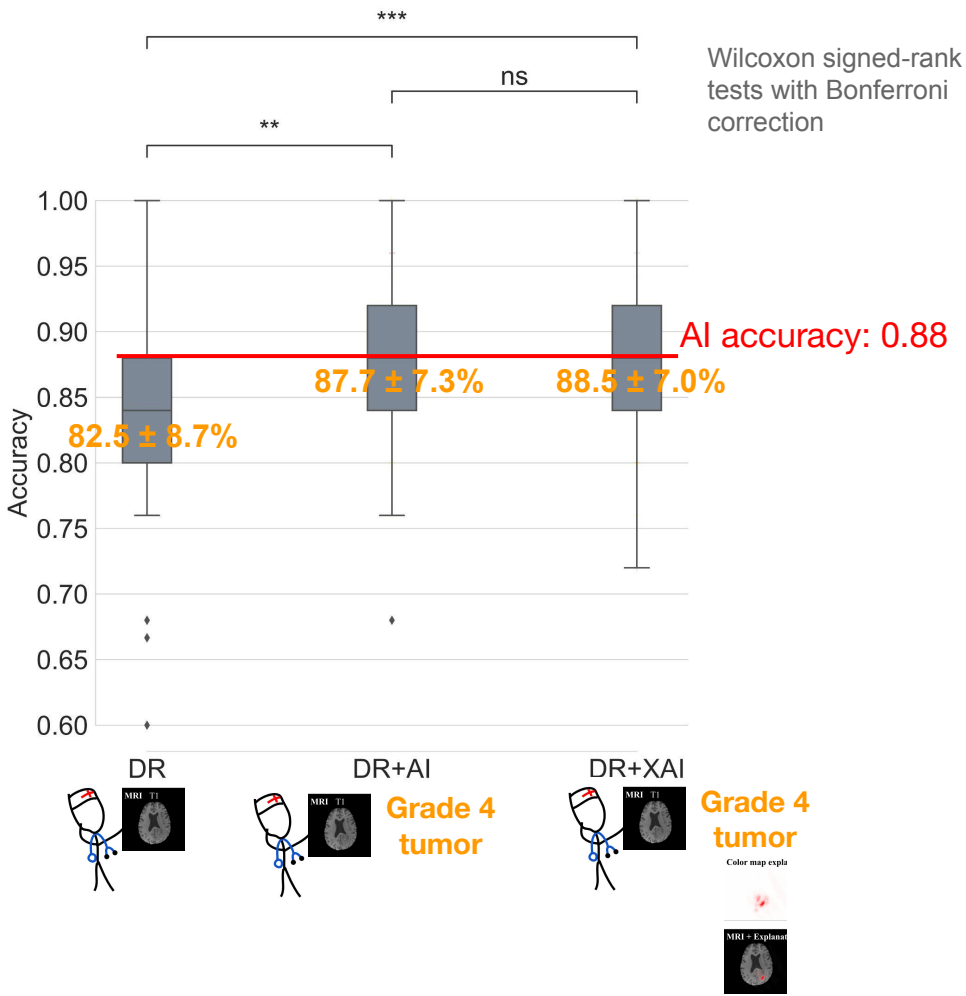
- AI prediction (**DR+AI**) is helpful
- AI explanation (**DR+XAI**) not show additional value

If so,

Doctor + AI > max(Doctor, AI) ❌

Suggestion Explanation

Not achieve complementary doctor-AI performance in **DR+AI** or **DR+XAI**

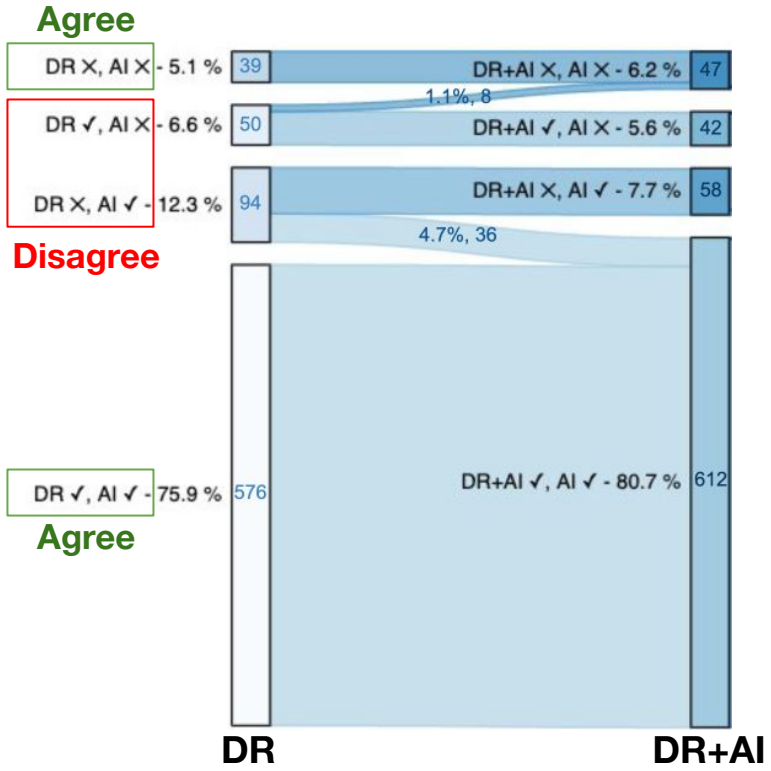


Wilcoxon signed-rank tests with Bonferroni correction

Result Why doctors improved their performance with AI prediction?

Because physicians' decision patterns converged to be more similar to AI decisions, as they **only switched decisions** when disagreeing with AI.

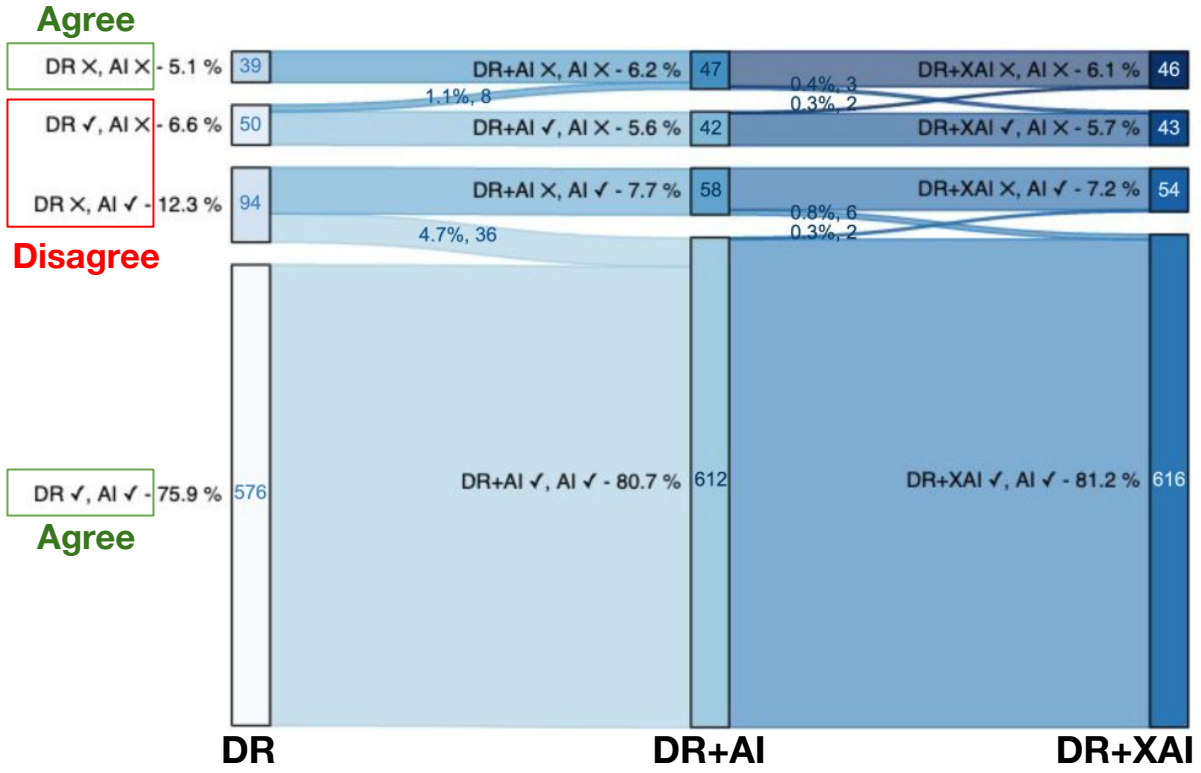
I.e.: doctors' improved performance is due to **overreliance on AI**



Result Why did not AI explanation help?

Because physicians had both **changed decisions correctly and incorrectly** with basically **the same amount**.

I.e.: explanation cannot help doctors to discern potentially questionable decisions of AI.



What does that (color map region) mean? Like hey, which part of my car gets my car moving? It should say press the accelerator. But yours would just show a dashboard of the car, and show that this button had some red, that button had some red, but it's not an explanation. Let's go to an example, and you'll see, what about the red areas under MRI T1CE (modality)? **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.

…These color maps were totally useless **without text, without any context or explanation**, like those details. The color maps were just pretty, but they didn't explain anything.

– Neurosurgeon #3

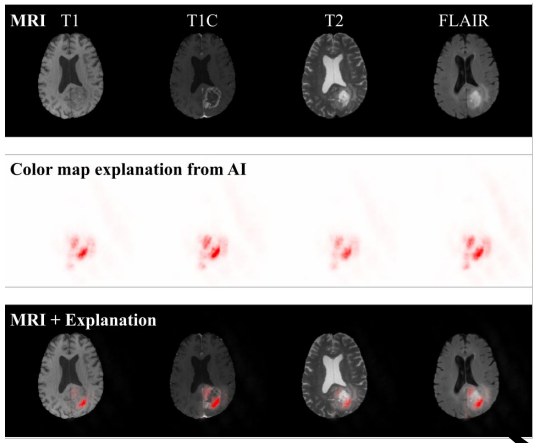
Though the color map is drawing your eyes to many different spots, but I feel like I didn't understand why my eyes were being driven to those spots, like **why were these very specific components important**? And I think that's where all my confusion was.

– Neurosurgeon #2

Qualitative results

Why AI explanation did not help?

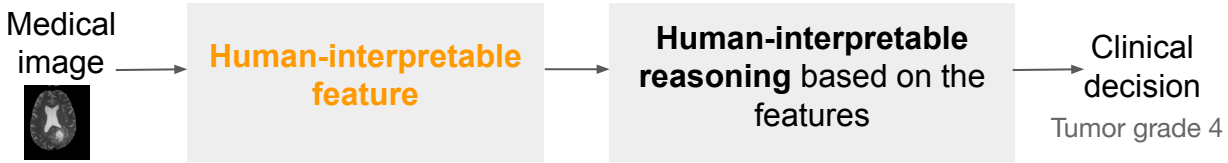
Because the existing AI explanations do not explain in a clinically relevant way



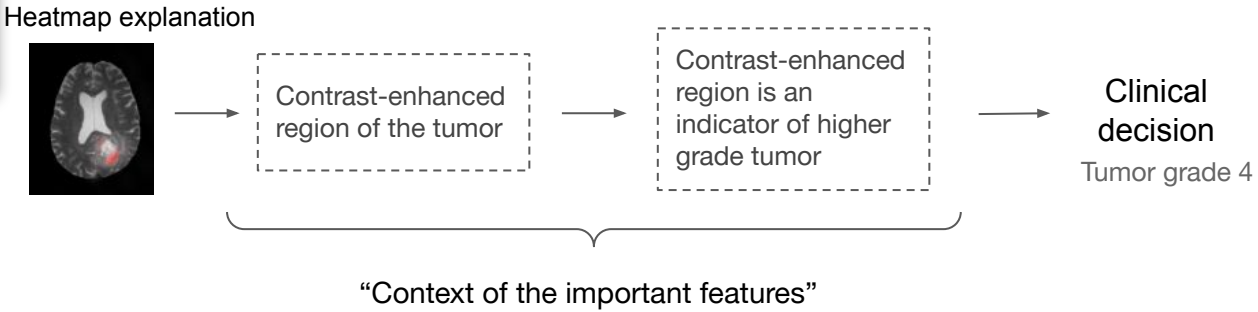
Existing AI explanations do not speak clinical language to explain

> What (explanation) we get currently, when a radiologist read it, they **point out the significant features**, and then they **integrate those knowledge**, and say, to my best guess, this is a glioblastoma. And I have the same expectations of AI (explanation).
>
> – Neurosurgeon #3

Physicians' clinical image interpretation process:



Physicians' interpretation process of AI explanation:



Takeaway A phase 2 clinical user study of AI on glioma grading

- A national clinical study with 35 neurosurgeons on glioma grading task using MRI
- With **AI prediction** assistance, doctors significantly improved task performance
  - Due to overreliance on AI suggestions
- Additional **AI explanation** did not change doctors' performance
  - Due to lack of clinically relevant information in AI explanation

---

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

Weina Jin (co-first author), Mostafa Fatehi (co-first author), Ru Guo, Ghassan Hamarneh

weinaj@sfu.ca

Thank you!

HAMARNEH
Medical Image Analysis Research Group

SFU SIMON FRASER UNIVERSITY

UBC THE UNIVERSITY OF BRITISH COLUMBIA

Code available at
github.com/weinajin/multimodal_explanation