Guidelines and Evaluation of Clinical Explainable AI in Medical Image Analysis

Supplementary Material S1: Physician User Study Findings

Weina Jin Xiaoxiao Li Mostafa Fatehi Ghassan Hamarneh

We present the main findings from the physician user study, including five themes: clinical utility of AI (Section U1) and its explanation (Section U2), clinical requirements of explanation (Section U3) and multi-modal explanations in particular (Section U4), and factors that could influence physicians' quantitative assessment of explanation (Section U5). The user study details and data analysis are presented in Section U6. We number the participants with N1, N2, ..., or O1, O2, where N indicates neurosurgeon participants in the interview, and O indicates open-ended responses in the online survey. Whenever necessary, we included participants' verbatim quotes despite some minor grammatical errors.

Contents

U1 Clinical utility of AI

U1.1 Clinical decision support

The main role of AI mentioned by physicians is to provide clinical decision support, a "second opinion" (N2), or "another level of evidence" (N5). Physicians would not delegate their full clinical tasks to AI at the current stage.

> *"I would do my own interpretation, I would see what AI thought, and I would maybe modify mine. but I wouldn't go 100% on AI. I would just use it, there's more information before I made a decision."* (N3)

> *"With the original plan the ideal AI program would differentiate that it is a glioma prior to attempting to grade the glioma, that's a very lofty goal. It's a very difficult thing I know as this is. This (current AI) would be useful still in operative planning, if you suspect a glioma, to have this confirmatory step, to say that there's another level of evidence suggesting that it's either high or low grade."* (N5)

U1.2 Benefits to junior physicians

The clinical value of AI may be more significant for junior physicians.

> *"The whole idea of AI is that a lowly-trained person can do the same job as a highly-trained person. You don't need 10, 20 years of experience, you don't need to have looked at hundreds of thousands of these (MRI cases)."* (N3)

Specifically, for junior physicians, AI can provide the following support:

1. AI provides decision support for time-sensitive cases, and hard cases.

   > *"If you want to ask a colleague and no one is available, or if you're trying to do something quickly, and want to make a decision yourself initially, you can just check that decision."* (N2)

   > *"This (AI system) would be helpful in asking for a second opinion if someone else is not around, when you're a little bit more uncertain."* (N2)

2. AI improves junior physicians' learning and problem-solving skills: to *"help you reaffirm what you're learning"* (N2).

   > *"I think the use cases would probably end up being, junior learners that are trying to learn how to read these scans, that might want to check with someone. Because if you're looking at a scan, you may have an initial thought about what the diagnosis might be, and you might look up that diagnosis, and look up on Radiopedia what the specific features are. This (AI) would be a way to check your thinking instantly."* (N2)

U2 Clinical utility of explainable AI

U2.1 Heatmap helps to localize important features and easy-to-miss lesions

All five neurosurgeon interviewees utilized the feature localization information of heatmaps to 1) inspect important features for an AI decision; and 2) to identify easy-to-miss small focal lesions.

"I find the color map quite useful for decrypting what the AI is using, its contributory factors." (N5)

"If at first pass (of reading MRI) you miss a small Flair (modality) hyper-intensity that doesn't show up with contrast enhancement, that was helpful (for the heatmap) to help determine." (N4)

"The color maps were just pretty, but they didn't explain anything, except for multi-centricity, like if there were multiple tumors, then (the current heatmap) it's helpful. Here, you can see that there's a second tumor up front, and some people might have missed that. So that way (with the heatmap), the computer is kind of nudging you, and says, 'Hey look, there's a second lesion.' " (N3)

"Take a picture of the (MRI) scan, and it (AI) would then give you another opinion. And then you can either trust what you initially thought, or you know look further into the features that you think are kind of driving your decision, and what the AI is trying to bring your eyes to. That explainability feature was quite cool. Because the other parts of the scan that I had initially missed just based on the fact that I was reviewing the scan pretty quickly, the AI obviously picks it up, where there were tiny little dots of tumor that could have indicated more diffuse spread of the tumor itself, and it can bring your attention to other small foci of tumor that you may have initially missed." (N2)

"I think that the AI (and its explanation) helps with a kind of a scanning bias that we have. As humans where we see the large and obvious lesion and we're immediately focused on that. I think that the AI sometimes seem to be better at noticing secondary satellite lesions that definitely changed the diagnostic and probably the prognostic opinion. So a single large lesion can be distracting and keep your attention, whereas smaller secondary lesions I noticed that they were picked up by the heatmap quite well." (N5)

U2.2 Resolving disagreement

All five neurosurgeons mentioned that explanation is most useful during decision discrepancies between AI and physicians.

"It's (heatmap) very helpful when there's discordance between (me and AI). If it's affirming the diagnosis that I've made based on the (MRI) scan, then it's not as relevant. But I would scrutinize it, if it disagreed with me, to determine whether I would change my mind or not. So I think it's not helpful when you're agreeing, but it is helpful to more carefully consider when you disagree." (N4)

"I think the use case where it (heatmap) is very useful is when there's a discordance. So when the algorithm and my opinion disagree, understanding that disagreement is where the heatmap is very useful." (N5)

The way doctors resolve decision disagreement is to reassess the explanation together with the inputs by themselves, to see whether the explanation is reasonable according to their clinical prior knowledge. A reasonable explanation that aligns with clinical evidence is more likely to persuade doctors to take AI's suggestion, whereas an explanation that deviates from doctors' own judgment on the contributing features may alert doctors to carefully inspect the original input image and the highlighted areas on the heatmap, especially to inspect for the easy-to-miss lesions potentially highlighted by the heatmap. During the inspection, if doctors did not identify any clinically meaningful features from heatmap localization, doctors will probably negate the validity of the explanation, may not take AI's suggestion despite its high accuracy and potentially correct prediction, and stick to their own judgment for the final decision. The next subsection (§U2.3) presents more findings on using explanations to verify AI decisions.

"That's where the AI's explanation would be valuable (during decision disagreement). Because then I could say, 'Well no, that's a low grade.' And the AI say 'No, it's a GBM.' But if the AI said, 'Well, based on multicentricity and mass effect.' And I'm like, "Oh I missed, oh I didn't see that the AI picked up that there was a second lesion.' Then I would say, 'Oh sorry, I got it wrong. AI got it right."' (N3)
— A reasonable explanation persuaded doctors to take AI's suggestion.

"Like the two times we disagreed, I would still stick with mine over AI. But that's after seeing those red (heat) maps. Like I said, the red maps were useless. They didn't help at all." (N3)
— An unreasonable explanation did not persuade doctors to take AI's suggestion.

"There's only one that I had discordance (with AI prediction), and I think I did change my final judgment depending on looking at the areas that are highlighted. I think I just reevaluated there were areas that they (the heatmap) had picked up that I had not taken into account, when I scrolled through the (MR) images myself, and the heatmap would cause me to scrutinize those same areas on the source imaging, so I would look at that heatmap be, 'okay, what's different about that area that it thinks it's important?' So the highlighted areas of clinical relevance that I missed. I would scrutinize that area a little bit more to see if there was something different about it to sway my decision-making." (N4)
— Doctor reevaluated his own judgment based on heatmap localization.

"I think that it (heatmap) helped me understand in cases where there was a discordance between what AI was suggesting, why those disagreements might exist. So in cases where there was very little overlap between what I thought were important factors, and what the AI was interpreting based on the heatmap, I was more confident in disagreeing with the AI. If it had pointed out additional factors that I had not previously noticed, then I was more confident in changing my opinion based on the new points identified. So the disagreement caused me to pause and then go through the images again with the aid of the heatmap, to see if there was something that I missed. ... So on each individual one, I looked at the components (features) that I thought were important for my diagnosis. I chose my diagnosis, and after that when there was disagreement with the AI, I would go through the (MRI) scan again and the heatmap, first looking for anything I had missed, and second looking at the components of the heatmap that the AI felt were important. If the heatmap aligned with the components that I thought were important, and there was simply a disagreement on grade, based on that, I typically stuck with my own opinion over the opinion of the AI. And if the heatmap identified factors that I had not seen on my first look at the scan, then I reconsidered, and often changed my opinion." (N5)
— Doctor's detailed process to identify reasons for decision disagreement based on the alignment of doctor's and AI's important features.

Despite the extra time and effort spent in re-inspecting input image and explanation, a participant still sees the explanation useful to identify doctors' potential errors.

"Although it does add work, the work that it adds is only in cases where I may have otherwise made an error. So I find it (explanation) very useful as an error checking modality." (N5)

U2.3 Verifying AI decision, and calibrating trust

Since explanation reveals the AI model's decision process, physicians tend to assess the plausibility of explanation based on its agreement with clinical prior knowledge, and use such explanation quality information as a proxy or *"internal validation"* (N1) to judge AI's decision credibility for a given case.

"This color map didn't work as well as the previous one, and so that's kind of an internal validation to me. I want to see scenarios where it (AI) doesn't work, and I can tell that. So this is reassuring to me that, like they (AI) can make a mistake, and I can call it out, I can determine

the mistake. ... Presumably this AI system won't do very well, and then you can show me why it didn't do well, like when you do the explanation, you can say it didn't do well, because it looked at the wrong places." (N1)

"It's using relevant points to make that decision, and that in itself would increase my confidence in it. I don't know exactly what it's doing, but I know that it's looking in the same areas and as looking at the patterns of enhancement, and the patterns of high signal within the areas that I'm looking as well, so that intuitively allows me to have some confidence in it. But when you start to see stuff that is in my opinion irrelevant, it makes you question to some extent exactly what it's taking into account in its algorithmic decision-making process." (N4)

"It (the heatmap) might change your confidence level when you're looking at areas which don't appear to have clinical relevance that are being analyzed." (N4)

Physicians' assessment of heatmap plausibility directly influences their trust in the specific AI decision, and their overall trust in the AI model in general.

"I don't think that it's going to be a binary thing (that heatmaps can help to judge whether the AI is right or wrong), but I think it (explanation) assists in judging whether or not you can clinically trust the conclusion of the (AI) algorithm." (N5)

"I don't know whether it's correct to use these explanations to kind of triangulate accuracy of the model. So far a lot of the explanations have been factually incorrect, and so the more that I see that there's a discordance between what I think is important and what actually is the tumor, the more I trust my own initial decision of this being a high grade versus a low grade (as predicted by AI).

Here you can see the only areas that are red are outside of the tumor. Overall this these explanations have reinforced my initial skepticism about the determination of the model, because even if it used its information and made correct predictions, let's say it was correct for this patient, and I was incorrect. I wouldn't trust it. Because every time every one of these explanations suggests that it's used wrong information to make that decision. It's like prioritizing ventricles over anywhere where the tumor is, even if you (AI) are predicting it correctly, I won't take it that's right.

If a system made its prediction based upon these areas (outside the tumor), I would definitely not trust that system, but I would be very reassured that the system is telling me that. ...You show me this (heatmap), and I see that it's focusing a lot of its decision-making on areas that have nothing to do with the tumor. That decreases the trust in that specific model (AI decision on this specific case), but it increases the overall trust in AI. So I'm less likely to use this model, but I'm more likely to use a model that does a better job than this, because I am reassured that when I see that better model, that I will be able to have access to that back-end explanation. So you're not asking me to believe in the system without evidence." (N1)

"There were areas which were clearly had nothing to do with the decision making which were clearly outliers, like it would often pick up interhemispheric frontal and show me that as part of the heatmap, or temporal poles and stuff like that where you might get artifact, and that has nothing to do with the decision making. So they're kind of areas which seem like red herrings kind of irrelevant, and I don't know why it would be using some of that in its decision-making process. So the only thing is that makes me feel a little less confident in this machine learning ability, because it's showing areas that don't seem relevant to making the decision presented to me." (N4)

Explanation was also regarded by clinicians as a "debugging" tool for physicians to verify AI decisions and cross-check whether AI is malfunctioning. This is an important clinical utility of explanation to *"ensure safety"* (N1) in AI-assisted *"life-and-death clinical decisions"* (N3).

"I think it's gonna be a big leap one day, if you just press a button, and the AI says it's GBM. Well, clinicians are gonna want an explanation. We're not like, because what if there's a malfunction in the computer? We like to cross-check. And if the AI said, 'based on the contrast enhancement, based on the T2 in this case showed that edema, then that is why the AI thinks it's a high likelihood.' Then I could look at the pictures, I see what the AI was thinking, 'Okay, good, I agree.' So that's the ideal expect." (N3)

"If the model keeps making the same mistake for a subset of patients, it's then important for you to be able to tell me how it (AI) made that mistake, and that's a utility of explanation." (N1)

U2.4 Making medical discoveries

One physician mentioned that the explanation can be used to make new medical discoveries by identifying new features that are different from humans' pattern recognition, providing the AI has a good performance, and the explanation truthfully expresses the AI model's decision process.

"The other fascinating thing is that, maybe there is a lot of information encoded in a different dimension than what we look at. Because as humans we use our own pattern recognition, and we've gotten very good at differentiating contrast enhancement and necrosis. But maybe there's more information to be gathered from the AI perspective. ... So helping you to identify patterns and features that I would have previously not thought are important. So for example, if you tell me an AI machine that every time it picks up a lot of weird things on Flair (modality), and then has the best accuracy, and it makes you think maybe there is something important on the Flair that we haven't been paying attention to. So that's the added value of explanation." (N1)

U3 Clinical requirements of explainable AI

U3.1 Limitations of existing heatmap explanation

We analyze the limitations of the existing heatmap explanation by corresponding the information provided by the form of heatmap to the clinical image interpretation process. In general, doctors' image interpretation consists of two steps: 1) First, they perform pattern recognition and extract key features. This includes recognizing or localizing key features, and identifying their pathology. And then 2) they perform medical reasoning and construct multiple diagnostic hypotheses (differential diagnosis) based on the image feature evidence.

"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 GBM. And I have the same expectations of AI (explanation)." (N3)

A complete explanation may correspond to the above clinical image interpretation process. A clinically relevant explanation at least corresponds to the aforementioned step 1), i.e., identifying important features and describing their pathology. The existing form of heatmap explanation, however, only localizes important features, but lacks the description of the pathology of important features. This fatal drawback of heatmap explanation made all physician participants confused, and they requested an explanation for the meaning of the highlighted regions in the heatmap.

"What does that (heatmap 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 the accelerator had a little bit of red on it, this button had some red, that button had some red, but it's not an explanation. A picture is never an explanation. Like more red indicates the region is more important, what about that region? Like 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." (N3)

"Though the heatmap 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." (N2)

In addition, since the heatmap cannot explain the reasons for localizing clinically irrelevant features, physicians preferred to see an explanation that aligns with their prior knowledge, simply to avoid confusion during the interpretation of clinically irrelevant features.

"My priority would be to have the AI show and explain the features inside the tumor that are important. Because I'm not sure what it's capturing outside the tumor that is important to the (AI) system actually." (N2)

"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)

"Although this appears to do a reasonable job at predicting disease grade, it does not appear to diagnose the presence of a lesion versus normal, nor to have the reliability required to forego a biopsy." (O2)

U3.2 Desirable explanation

Physicians described the ideal explanation could be some simple linear or rule models, with clinically relevant features as the variables.

"I know it's an AI model, so it's not like a regression model. But it would be helpful to discuss what variables the model is using. ... So maybe an explanation beginning with factors: the enhancement pattern, the midline shift, the amount of edema, more like radiologists language, that's what clinicians would use to guess if it's GBM or not." (N3)
— Use clinically relevant features as variables for the explanation model

"If this was like a logistic regression model, you could say, well based on the fact that it's enhancement: yes/no? yes, multicentricity: yes/no? yes, and then mass effect: yes/no, yes, edema: yes. You got four yeses, so your chances of having this be GBM is 94%. So like you use old-school statistics to explain." (N3)
— Use a linear model as explanation

"I want to be able to work with the (AI) model to go down like a flow chart to, I understand that's not how necessarily machine learning works, but if you're trying to explain, or show the relevant areas if I had some degree of insight into how the model was trained, or what the relative importance of each of these heatmaps was in making its decision, then it might help me build confidence with it." (N4)
— Use rule-based explanation to dissect the decision

In addition, compared with a global explanation that explains the AI model's behavior in general, doctors have more demand for a local explanation for a specific case.

"I expected the explanations to be different for every patient, every scan (the explanation) should be different." (N3)

U3.3 Making AI transparent by providing information on performance, training dataset, and decision confidence

Besides the heatmap explanation, some physicians mentioned that other information is required to make the AI model decision process transparent, and help build trust in the AI model and its suggestions. The requested information includes model performance, information on training and validation dataset (such as "*distribution of the patient's demographics, distribution of lesions diagnosed*" (N5)), and decision uncertainty overall and at the modality level for a given case.

> *"The trust in the AI model comes from performance ultimately. Heatmaps are secondary to the performance. So if you tell me system A has a 99% accuracy, and system B has a 22% accuracy. I don't care how it figured that out as long as it (has high accuracy). Like I said, clinicians are numbers people. If you tell me over a thousand patients this did really well, and this one didn't do really well, I'm okay with that black box designation. If you tell me that for this study or this system is 85 (percent accuracy), and you're asking me how likely I am to trust it, then I think about, 'okay, let's see how it made that decision.' So explain to me how you did that, and that's the value that you're adding with the explainable AI."* (N1)

> *"I'm not sure what the relative importance of each one of those heatmaps is in making its decision. So if it had like, as it could express to me it's confidence for each of those images of it being a high-grade glioma, and then its overall confidence based upon those individual heatmaps. Because all of this is based on a threshold model of confidence. So if I have an idea about what that threshold is, saying there's a 92% chance this is a high-grade glioma, and these are the important areas, or these are the relevant areas that make that confidence. Because for me, I don't know how much that heatmap relates to the confidence overall in its prediction. ... I think as someone who reads (MRI) scans and gives diagnosis, we are very happy to deal with probabilities. I imagine there's just a threshold that has been reached that you're using that as a confidence for the model. So if I knew what that threshold was, and how close it (the prediction) was to that threshold, then it would help me at least understand a little bit more about the model, and having that transparency in the models. Because if we were to do this without AI, and if you're sitting down with a few radiologists in a room, they would all tell you what they think it is, and explain why it is that, and then you could question them to say, 'Okay, what is your likelihood? Or what's in your differential diagnosis? What are your relative probabilities of each of these things in your differential diagnosis?' "* (N4)

> *"I know that a data set that's not trained properly does not come with good answers, so to calibrate my trust, I need to know that the data set that was used is relevant to the data that I'm looking at to making sure that I'm inputting the same, that my scan is no different than (the training dataset). In a clinical trial, you need to know your patient that is in front of you is of similar demographic and condition as the people who are in a clinical trial to determine how applicable it is. It's the same thing with any AI model is that, you need to know that the AI model is representative of what you're looking at at the same time. Because without that, it's not relevant in you (task), and you can't trust it."* (N4)

U4 Multi-modal medical image interpretation and clinical requirements for its explanation

U4.1 Clinical interpretation of multi-modal medical images

In their multi-modal medical image interpretation process, physicians seek modality-specific features, and rely on the cross-modality comparison for clinical reasoning.

> *"For my interpretation, the things that people would typically look for are the pre-T1 signal to see if there are any blood products or hemorrhage within it. Then the T1 post-contrast (modality)*

to compare that to pre-contrast to highlight the areas which are enhancing. Although the T1 is helpful, there's not a lot of value in my mind to differentiate between high grade and low grade when you look at the T1 pre-contrast, other than ruling other things and ensuring that it's true contrast enhancement. And then the Flair and T2 (modality), they show very similar things, the only difference I would see in between those two would be, for the high-grade gliomas looking for vasogenic, and even for the low grade we almost see Flair, where there's which would really show more of the T2 signal or more difference than the T2 signal may have in the cortical thickening of the low-grade glioma." (N4)

U4.2 Role and prioritization of multiple modalities

Additional modalities give a full range of differential diagnosis and help to rule out similar diagnosis.

> *"For me (ADC modality and the diffusion modality) it's differentiating between cystic areas and either necrosis or degree of cellularity. It's helpful for ensuring that you're not dealing with something else. Because we're trying to differentiate high grade and low grade, but a lot of them are pretty slam dunk, there are other diagnostic possibilities. So the problem is you're not always starting with a high-grade glioma, you could be starting with a lymphoma or other enhancing lesion like an abscess, and putting in diffusion (modality) restriction, you can automatically weed out the abscess or the hypercellularity of the lymphoma in comparison which might have some similar appearance. So it depends on your start. You have to be fairly confident that it's a glioma to put it down this data set pathway to just differentiate between high grade and low grade. But it (the current four MRI modalities) doesn't rule out the entirety of other mimickers."* (N4)

The modality difference reflects how the model uses the modality information, and such information should be reflected in the explanation.

> *"I like the heatmap being displayed on everyone (MRI modality). It's just I didn't know what the difference between the heatmaps, I didn't understand what the difference between each map was in making its decision. Because we're trying to get out is, does the model evaluate each sequence (modality) independently, and then make a decision upon the individual sequences? Or does it evaluate them in total, like all of them at the same time? If it's evaluating all the sequences at the same time, the heat maps should be the same. If it's evaluating each sequence independently, then the heatmaps would be different."* (N4)

Since doctors' reasoning relies on specific modalities, they regard prioritizing important modalities as a useful explainability feature. If such modality prioritization explainability feature is made available, doctors should be aware of it.

> *"The fact that it's using the T1 with contrast (modality) which is the most important sequence to make the differentiation between high grade and low grade, that's reassured that means it's doing a good job."* (N1)

> *"If you look at that (heatmap) picture, that you're showing the red blotches look the same in all of them. Maybe more in the second one (T1CE modality), but not that different."* (N3)

> *"It would be more helpful to display a different map for each of the sequences (modalities), and have each map 'point out' distinctive features of each sequence: edema seen on Flair (modality), hypodensity seen on T1, enhancement on the T1 enhanced study, etc. However, for this to be useful, I would think the best way to go is to tell your participants (doctors) that there are differences between color maps created by each sequence, so they can look for the differences."* (N2)

> *"The only one (modality) I really looked at was MRI the T1 with contrast (modality). 90% of my time is on the T1 with contrast, and then I will spend 2% on each of the other ones."* (N1)

The modality prioritization may or may not necessarily need to be in concordance with physicians' prior knowledge on modality prioritization.

> *"This one (Feature Ablation heatmap) is not bad on the Flair (modality), it (the tumor) is very well detected. I wouldn't give it a perfect mark, because I would like it to prioritize the T1C (modality) instead. But I'll give it (a score of) 75 (out of 100)."* (N1)

> *"I don't mind if it's like showing a discrepancy between the two sequences, I'm willing to consider that as an important prediction requirement, but I'm not willing to accept the machine using wrong areas to make this decision."* (N1)

> *"Maybe I can learn from the AI, maybe the lesson here is to stop relying on T1 contrast enhancement (modality), and rely more on T2 or what AI is trying to teach us. It's like, 'oh, what is AI going to see that I've never seen before. Oh Okay.' "* (N3)

In addition, such modality prioritization is task-specific.

> *"When I'm looking to differentiate stroke from tumor or from infection, certainly I look at different sequences (modalities). But when I'm looking at tumors only, the single most important sequence is the T1 of contrast. And then everything (modality) else is you look at them but far less time."* (N1)

U4.3 Modality-specific features

Doctors require the heatmaps to capture the modality-specific features. Such features may or may not totally align with doctors' prior knowledge, but should at least be a subset and not deviate too much.

> *"It's (the heatmap) looking at the tumor which is red that's fine. I wish it also picked up on the necrotic area, but it's not bad."* (N1)

> *"This is a very good model (heatmap). It almost perfectly matched on the T1 contrast with areas that the tumor was. So all of your previous models (heatmaps) were making errors here, were not picking up the tumor at all. But this one is picking up some of the tumor in that area, and then definitely picking it up strongly here, which is important."* (N1)

U5 Clinical assessment of explainable AI

In this section, we summarize factors that are related to physicians' quantitative assessment of heatmaps. The findings together with the clinical requirements for explanation (Section U3) can be used to abstract the clinical requirements and set up automatic quantitative assessments.

U5.1 Heatmap plausibility assessment

All participants assessed the heatmap plausibility based on the agreement of heatmap localization with their prior knowledge on the clinical task. Physicians gave lower ratings to heatmaps that do not explicitly highlight clinically meaningful features.

> *"I had ranked it (the heatmap) lower, because it would show red dots outside of where the tumor was. I felt like it was pulling information away from where the tumor was, and like from a normal brain parenchyma. And I wasn't exactly sure why it was taking information from there. That's why I thought to decrease my rating of the explainability."* (N2)
> — False positives

"I put like one or two or like negative two all the time (on evaluating heatmap quality and trust in AI), because it wasn't explained to me, I didn't trust it." (N3)

"I find that it doesn't account for mass effect very well. So it doesn't seem to notice expansion of a region or compression on adjacent regions. The other thing that it does is when there's large areas of enhancement and Flair signal and edema, it samples pieces of that, or it seems to based on the heatmap without necessarily considering the entire area. I'm not sure if that's a matter of filtering, how much heatmap to show, so as not to make the scan just look red entirely. But I've noticed that for example in the scan that you're showing me now, you see kind of those two darker denser focuses (on the heatmap). But when you look at the lesion itself, it's clearly not organized in that way. It's larger than that, and it has medial and anterior components that don't have much of a heatmap signal.

The reason I think that's important is, you can see the part that appears to be crossing midline, which would be a concerning feature and an impact to surgical planning to some extent. But the heatmap doesn't indicate something like that." (N5)
— False negatives

U5.2 Bias and limitation of physicians' quantitative rating

One physician mentioned that the absolute rating of the heatmap quality may be biased towards physicians' own judgment and anchors used in comparison. In comparison with the quantitative rating, physicians' qualitative feedback and comments are more helpful.

"I think it's important to say the clinician is influenced, because that's what happened in real life. Every one of these (heatmap) that I'm giving it a score relative to the one I thought has been good. I'm not quantifying this in a vacuum. I've already seen the previous ones and so I'm biased by that. I don't think the exact reliance of a clinician on these heatmaps could be the focus. Like whether I give it a 75 or a 95 is far less important than whether I think this is helpful or not. Because these numbers I'm giving are arbitrary. The overall perspective that you get is more important than the actual numbers that are given." (N1)

U6 User Study Method

U6.1 Study design

The user study consists of a survey and an optional within-/post-survey interview. The interview sessions were one-on-one, remote, open-ended, and semi-structured. The explainable AI (XAI) prototype was embedded in the survey, and it was designed to mimic the real-world usage scenarios in clinical decision-support settings. The doctors were first introduced to the AI model and got to know its performance on the test set. The doctor then uses the AI model on a new patient MRI, and inspect AI's prediction and its explanation. The MRI and its heatmaps are 3D images presented in axial view slice-by-slice in video format.

We recruited participants by directly contacting the researchers' clinical collaborators and snowball sampling participants nationally in Canada. The inclusion criteria were: the participant must hold a Doctor of Medicine degree and work in neurosurgery, radiology, or neuro-radiology specialty. The participants were thanked with $50 CAD for their time and effort in the study. The user study is approved by the Research Ethics Board of Simon Fraser University (Ethics No.: H20-03588).

U6.2 Qualitative data analysis

We analyzed the qualitative data including the interview transcript and open-ended questions in the survey using an inductive thematic analysis approach [1]. A total of 180 minutes of interviews were recorded and transcribed. We performed open coding on the qualitative data. A total of 85 codes were generated. We then performed an axial coding and affinity diagram process to discover the hierarchical structure among the

emerging concepts. The first author who has expertise in clinical medicine and human-computer interaction research conducted the interview and qualitative data analysis process.

References

[1] Virginia Braun and Victoria Clarke. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, APA handbooks in psychology®., pages 57–71. American Psychological Association, Washington, DC, US, 2012.