Answer to reviewers

Reviewer 1:

Comments to the Author

This is a well-written paper on an evaluation study of two mathematical models. However, there are several issues with the current version as follow:

1 The authors stated, in the abstract, that the objective was to evaluate the trade-off between the time investment in the algorithm, efficiency of the numerical implementation, and accuracy required in predicting the final endpoint of the therapy. However, the results and discussion were focus on the validation of the models. It was not obvious what is the trade-off between the models? What is accuracy required in predicting the final endpoint of the therapy?

Thank you for returning our focus on the important trade-off between accuracy and required time. The accuracy of the models is reported, but the time of implementation is not as you stated. The models implemented in the manuscript are 1) a homogenous, time-dependent FEM model and 2) a homogeneous steady state model. The FEM model is more accurate than the steady state model but takes considerably longer time to run. It is difficult to succinctly express the difference in computation times for at least two reasons. During optimization for a given patient, the FEM model and steady state model may have a different number of optimization realizations. Secondly, the FEM model is transient and has multiple time steps within one realization whereas the steady state model naturally has a single time step. Therefore, we consider the simplest means to summarize computation time is to report the timing of optimization and LOOCV. Optimization of the same dataset may involve differing number of realizations for the two models; LOOCV only has one forward prediction for each dataset. In optimization, the FEM model takes [Some estimate of time]; during LOOCV, it took [Some estimate of time]. The steady state model’s LOOCV takes 246.6 seconds or 11.2 seconds on average per dataset. This indicates that both models could be used before and during surgery. The Results and Discussion have been amended to describe this information; the new text is red.

Regarding the topic of what accuracy is required to consider a model useful for prediction, it is a multifaceted concept. Firstly, we consider a model prediction as useful in a particular case if the model’s ablation region and the ablation as indicated via MR temperature imaging has a Dice similarity coefficient (DSC) greater than 0.7. Secondly, the model’s overall accuracy in a cohort is measured by the ratio of useful predictions versus inaccurate predictions (*i.e.*, number of DSC > 0.7 versus DSC < 0.7). Clearly, this overall accuracy is desired to be high, but it is beyond the scope of our present investigation to name a success rate that would indicate the model would be useful in clinical practice.

2 It is not clear how this method will improve pretreatment planning?

With MR temperature images from an unprecedented number of retrospective MRgLITT treatments, the authors decided a new thermal treatment planning method was becoming available and worth investigating. With a relatively large number datasets, a model can be optimized for each dataset to train the model for predicting future datasets. The success or failure of this paradigm is thought to be related to several factors. 1) The model used must be able to describe clinically relevant ablations. 2) There must be a sufficient quantity of retrospective datasets for the training to converge. 3) The cohort of retrospective datasets must have sufficient similarity within the group and to the prediction scenario. *E.g.*, a model trained on datasets from liver ablations is unlikely to be predictive for brain. However, a training cohort of brain metastases may or may not have enough homogeneity for a trained model to be predictive. There is new text in red and is in the beginning of the Discussion section.

3 The introduction is too long. It is not clear how the second part that describes the hardware relates to the paper.

[DF, what do you want to do about this?]

4 There is no information on the model size and the actual computations time.

The time required for computation is described in your 1st critique and the Results and Discussion include that information now.

5 It is not clear which hardware was used with which particular model.

The methods are now revised to include the hardware information.

[DF, can you add this?]

Comments and suggestions to improve the manuscript:

-Please expand on any heating induced changes in model parameters. Was blood perfusion held constant for the entire treatment duration? Blood perfusion is well known to change during ablative exposures (as the authors note, this is sometimes exploited to confirm post-ablation heating with contrast-enhanced imaging). Have the authors explored any techniques for incorporating temperature/thermal damage dependencies for blood perfusion? Suggest adjusting nominal perfusion level based on time-temperature history (see, for example, Schutt, Med Phys, 2008).

The model parameters do not change with heating. We obviously must make this fact abundantly clear in the manuscript and directly address the reasons why it is appropriate. Thank you for bringing to our attention the lack of clarity regarding the use of model parameters. Accordingly, the manuscript has been modified in the following places: [list areas that were modified].

As the hyperthermia/ablation literature states in concert, tissue properties are dependent on temperature and thermal damage state and these dependencies are important to consider in order to accurately predict outcomes. However, this time-dependent phenomenon is mitigated in our investigation via shorter ablation times relative to other thermal modalities. *E.g.*, laser ablations typically last 1.5 minutes to 2 minutes whereas the RFA procedures modeled in Schutt, 2008, were 12 and 15 minutes in duration. Furthermore, and perhaps more importantly, we do not consider the parameters from model optimization (or ‘model training’) to be truly physical. In this investigation, the modeling is driven by model training, not model complexity/physical fidelity.

- Did the authors investigate the impact of temperature dependent optical properties? During the model calibration procedure, it appears a single optical attenuation coefficient was determined for all spatial location/time points. Since optical absorption properties likely change during the course of heating, one approach may be to perform the model calibration over only the first ~30-60 seconds, to estimate the nominal optical absorption coefficient. This nominal value could then modified during heating based on previously published reports of the relationship between optical attenuation and temperature. Suggest considering approach where

You are correct that the model calibration arrives at a constant effective optical attenuation coefficient, *µeff*, for one training dataset. Also, we certainly agree that patient-specific optimization via test pulse or initial 30 seconds of ablation is an excellent idea for providing a calibrated parameter value. In this manuscript we did not include patient-specific optimization because we did not think it would be clinically practical. The optimization used in calibration requires [such and such] for the FEM model and [such and such] for the steady state model. That optimization process could only begin after the low power pulse the neurosurgeon uses for position confirmation. If the neurosurgeon is satisfied with the localization, ablation would normally follow immediately. If a patient-specific optimization was used, the clinician would have to wait for the optimization and then check the predicted ablation zone.

Instead, our investigation is presented such that the present surgical technique would be mostly unmodified. The surgical planning would be performed beforehand. During this planning, one prediction would take only [such and such] for the FEM model and [such and such] for the steady state model. Even in the case that the laser fiber was placed differently than the designed surgical plan, an updated prediction could be quick generated. In summary, using parameter values calibrated by a prior cohort allows the surgically planning to be performed within a tolerably short period of time. Furthermore, even in the future case that computation was fast enough for patient-specific optimization, the surgeon would likely want a pre-surgical ‘best guess’ to inform the surgical plan. This ‘best guess’ is what is presented in this manuscript.

Our rationale for the use of constant parameters is explained in the answer to the previous critique you raised.

Finally, the last sentence in this critique seems to have been cut short. Please reiterate it, as necessity dictates.

- Please state the test performed to confirm the discretization (mesh size, time steps) did not impact the solution.

[DF, what did Rice CAAM do for this?]

- Figure 1, consider blowing up the sketch of applicator (bottom half of fig 1a) so details can be distinguished more easily.

The figure has been modified accordingly at [page and line number].

-p6, L24 – State units of frequency factor and activation energy

The units have been added at [page and line number].

- p6, L30 – can the low-power pulse (used to confirm applicator positioning) be used to estimate optical attenuation in a patient-specific manner? Did you consider evaluating this dataset and comparing to the average value obtained from the calibration process for the other datasets?

We consider it a very good idea to use the low-power pulse as a patient-specific

- Please modify references so they appear in IJH style.

The references are updated to match the *IJH* style.

- Figure 3, suggesting use same bin width for both histograms so the results for the two models can be more readily compared.

The figure has been amended.