Reviewer Response Document

We thank the reviewers for providing revisions to our manuscript. We agree that the major revisions suggested are reasonable and that the comments provide productive edits to improve the quality of the manuscript. We have addressed each comment and made changes in response to every suggestion, including further software development and analyses for completeness. Each reviewer comment has a corresponding author response and listed changes in the manuscript. Each reviewer response section is prefixed with the letter of that reviewer (E = editor, A = reviewer A, B = reviewer B).

We hope that the reviewers find the changes in accordance with their comments and look forward to continued supportive feedback.

Editor-in-chief:

Comment E1:
Please format the manuscript according to the journal guidelines (eg. no conclusion section foreseen).

Author Response:
(The Guide for Authors document we are referring to for this edit is the one available on the phiRO website https://www.elsevier.com/wps/find/journaldescription.cws_home/738793?generatepdf=true)

If there are any formatting issues remaining, we can promptly update the manuscript as there are some formatting specifics we have interpreted that may be incorrect (e.g. we misinterpreted that supplemental information should be labeled “appendix” following the instructions “in a subsequent appendix, Eq. (B.1) and so on. Similarly for tables and figures: Table A.1; Fig. A.1, etc”)

We have reviewed the guide for authors document and have made changes listed below for formatting:

- Conclusion title removed, now merged with discussion
- Changed appendix to “Supplementary Material” and named tables and figures consecutively
- Reformatted references to follow journal format
- Shading has been removed from tables
- Removed the “abbreviations” listed under to in-text Table 1 and listed the terms in full in the table. Note that there is still an abbreviations list under the supplementary variable distribution tables (Supplementary Tables S4 and S5) for concision. If this is the incorrect format we can certainly change it promptly.

Changes to Manuscript:
L338 (conclusion):
Removed:
“5. Conclusion”

L338 (discussion):
Added:
“In this work”

L208 (table 1):
Removed:
“Abbreviations: RT = radiotherapy, PTV = planning target volume, HAIP = hepatic arterial infusion pump, CEA = carcinoembryonic antigen, KRAS = Kirsten rat sarcoma virus, TARE = transarterial radioembolization”
Changed:

Category	**Variables**
Imaging Clinical Data	Number of lesions at RT Other sites at RT Lesion dimension 1 Lesion dimension 2 PTV (cm^3)
Treatment Clinical Data	Biologically effective dose (Gy) Minimum dose for PTV (cGy) Maximum dose (cGy) Dose for 95% of target volume (% of intended prescribed dose) Systemic treatment before RT Lines of chemotherapy HAIP before RT Reirradiation Surgery before RT Ablation before RT TARE before RT Embolization before radiotherapy
Other Clinical Data	Primary tumor subsite Metastasis at diagnosis Number of liver lesions at diagnosis Other sites at diagnosis Liver location CEA KRAS mutation

To:

Category	**Variables**
Imaging Clinical Data	Number of lesions at radiotherapy Other sites at radiotherapy Lesion dimension 1 Lesion dimension 2 PTV (cm^3)
Treatment Clinical Data	Biologically effective dose (Gy) Minimum dose for planning target volume (cGy) Maximum dose (cGy) Dose for 95% of target volume (% of intended prescribed dose) Systemic treatment before radiotherapy Lines of chemotherapy Hepatic arterial infusion pump before radiotherapy Reirradiation Surgery before radiotherapy Ablation before radiotherapy Yttrium-90 embolization before radiotherapy Arterial embolization before radiotherapy

Other Clinical Data	Primary tumor subsite Metastasis at diagnosis Number of liver lesions at diagnosis Other sites at diagnosis Liver location Carcinoembryonic antigen Kirsten rat sarcoma virus mutation

Changes to supplementary material and references are described in comment E2 and E3

Comment E2:
References should be formatted exactly according to the journal guidelines.

Author Response:
We have updated the reference formatting to follow the examples in the Guide for Authors. Namely, the following changes have been made:

1. Removed quotations for article titles
2. Listed 6 authors before “et al.” instead of 3
3. Changed date and volume format to yyyy;vol(num):page-page
4. Changed last page number to shortened form
5. Removed italics of journal names
6. Added DOIs when available and formatted at https://doi.org/[doi] instead of just doi/[doi]
7. Periods added to end of DOIs
8. Website references (the PySurvival and Hmisc code libraries) have been updated to following the example in the Guide for Authors.

Changes to Manuscript:
L354 (references):
New reference list is now formatted such as the following example for reference [1]:

[1] Abdalla EK, Vauthey JN, Ellis LM, Ellis V, Pollock R, Broglio K, et al. Recurrence and outcomes following hepatic resection, radiofrequency ablation, and combined resection/ablation for colorectal liver metastases. Ann Surg. 2004 ;239(6):818-27. https://doi.org/10.1097/01.sla.0000128305.90650.71.

Comment E3:
Appendix should be renamed "Supplementary Material". Tables and figures should be numbered consecutively "Supplementary Table S1, ...".

Author Response:
Appendices and callouts to appendices have been changed to “Supplementary Table”.

Changes to Manuscript:
L150 (methods)
Changed:
“A full list of radiomic features is available in Appendix C. ”
To:
“A full list of radiomic features is available in Supplementary Table S1. ”

L168 (methods):
Changed:
“The full algorithm is listed in Appendix B”
To:
“The full algorithm is listed in Supplementary Equation S3”

L221-222 (results):
Changed:
“The baseline distribution of clinical variables is summarized in Appendix A”
To:
“The baseline distribution of clinical variables is summarized in Supplementary Table S4”

L494 (Supplementary Material)
Changed: all “Appendix ___”
To: “Supplementary ____” (Table/Figure/Equation)

Comment E4:
Please add a word count for Abstract and Main text to the title page.

Author Response:
Word count for the abstract and main text have been added to the title page. Word count was from beginning of Introduction to end of Discussion, with section headers include and excluding tables and table captions.

Changes to Manuscript:
L5:
Added:
“Word Count: Abstract: 250, Main Text: 2996 (From Introduction to end of Discussion, includes headers)”

Comment E5:
Abstract. The authors are encouraged to make the link to radiation oncology in the Abstract already - or alternatively in the title.

Author Response:

We can certainly make an explicit link to radiation oncology. We have updated the title to include radiotherapy and updated the abstract to explicitly mention radiation oncology.

Changes to Manuscript:
L1-3(title)
Changed:
“Prediction of Local Control for Colorectal Liver Metastases using a Radiomic Artificial Intelligence Model”
To:
“Prediction of Local Control for Colorectal Liver Metastases Treated with Radiotherapy using a Radiomic Artificial Intelligence Model”

(abstract)
L38 Changed:
“Prognostic assessment of local therapies for colorectal liver metastases (CLM) is essential for guiding management”
To:
“Prognostic assessment of local therapies for colorectal liver metastases (CLM) is essential for guiding management in radiation oncology”

(conclusion)
Changed:
“As a proof of concept, this study provides support that radiomic AI methods may be further developed to aid in prognostic decision making.“
To:
“As a proof of concept, this study provides support that radiomic AI methods may be further developed to aid in prognostic decision making in radiation oncology. “

Comment E6:
Results. L209. Of the 129 lesions, ...

Author Response:
The typo has been fixed. Thank you.

Changes to Manuscript:
L220 (results):
Changed:
“f the 129 lesions”
To:
“Of the 129 lesions”

Comment E7:

Discussion.
- Please make sure to discuss and cite current literature in this field. Only very few studies are referenced here.

Author Response:
We acknowledge that we mainly focused on survival prediction models of colorectal liver metastases and that there are other papers on radiomics in liver metastases and in machine learning survival models. We have cut some redundant sections in lieu of discussing relevant literature and have added the following:

(To discuss current applications of radiomics for liver metastases, gives discussion on limitations especially due to limited validation datasets)
[15] Fiz F, Viganò L, Gennaro N, Costa G, La Bella L, Boichuk A, et al. Radiomics of Liver Metastases: A Systematic Review. Cancers. 2020;12(10):2881. https://doi.org/10.3390/cancers12102881

(To discuss other AI methods of deep learning in analysis of liver metastases images)
[36] Wei J, Cheng J, Gu D, Chai F, Hong N, Wang Y, et al. Deep learning-based radiomics predicts response to chemotherapy in colorectal liver metastases. Med Phys. 2021;48(1):513-22. https://doi.org/10.1002/mp.14563.

(To discuss usage of random survival forest vs. cox models in recent radiomicss studies)
[33] Leger S, Zwanenburg A, Pilz K, Lohaus F, Linge A, Zöphel K, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Sci Rep. 2017;7(1):13206. https://doi.org/10.1038/s41598-017-13448-3.

[34] Chang E, Joel MZ, Chang HY, Du J, Khanna O, Omuro A, et al. Comparison of radiomic feature aggregation methods for patients with multiple tumors. Sci Rep. 2021;11(1):9758. https://doi.org/10.1038/s41598-021-89114-6.

Changes to Manuscript:
L100-101(introduction):
Added:
“Specific to liver metastases, Fiz et al. report 32 different studies up to June 2020 evaluating the association of radiomics to overall survival, tumor size, or response evaluation criteria [15]”

L269 (discussion):
Added:
“Recent studies modelling survival with radiomics show no significant difference between CPH and RSF models [33, 34]. “

L319-320 (discussion):
Added:

“It has been a reported challenge of radiomics that these is no standardized cutoff or clinical interpretation of features [15]. “

L314-317 (discussion):
Added:
“With sufficient data, deep learning with convolutional neural networks is another potential method to predict survival and has been used previously to predict response to chemotherapy for CLM [36]. However, deep learning models are more difficult to interpret than radiomic features due to the multiple layers of matrix convolutions.”

Reviewer A

Comment A1: The authors should state more clearly in the abstract and in the methods what volume of interest were used to calculate Radiomic models. Only "total liver" or also the metastases GTV?

Author Response:
We agree this should be clearly listed through the manuscript. We have added clarifications of the volumes of interest studies for the radiomics model, which is 1. Liver parenchyma only 2. GTV only 3. Liver parenchyma + GTV.

Changes to Manuscript:
L46 (abstract) :
Changed :
“A time-dependent survival model was built by extracting 108 radiomic features from CT scans”
To:
“A time-dependent survival model was built by extracting 108 radiomic features from liver and tumor volumes in CT scans “

L127-128 (methods):
Changed:
“Liver and tumor volumes were segmented by radiation oncologists at MSK, as part of standard of care.”
To:
“Liver and gross tumor volumes (GTV) were segmented by radiation oncologists at MSK, as part of standard of care. This created volume subsets of liver volumes only, GTV only, and liver and GTV volume for radiomic analysis. “

Comment A2: Did the authors adhere to the IBSI guideline of radiomic features calucation? If not, please list the deviations. (https://pyradiomics.readthedocs.io/en/latest/faq.html)

Author Response:
The radiomic features resulted in several minor deviations to ISBI guidelines as we implemented pyradiomics without changing the application programming interface. The deviations are in rounding (where PyRadiomics does now round to a lower resolution when resampling) and in indexing conventions from binning and resampling. A sentence has been added commenting on the deviations to ISBI guidelines and deviations are listed in a new supplemental table.

Changes to Manuscript:
L150-152 (methods):
Added:

The majority of radiomic features follow the Image Biomarker Standardisation Initiative (IBSI) guidelines. Deviations from IBSI are listed in Supplemental Table S2.

(supplemental information):
Added:

Computation	**PyRadiomics Implementation**	**IBSI Guidelines**
Binning	Discretizes gray values with fixed bins with edges equally spaced from 0.	Discretizes using fixed bin width equally spaced from minimum of resegmentation range
Resampling	Aligns to the corner of the original voxel	Aligns to the center of the image
Gray value rounding	Does not resample to similar resolution if original intensity is lower precision	Resamples to similar resolution if original intensity is lower precision
Mask resampling	Resamples to nearest neighbor	Allows selection of different interpolators for resampling

Table S2: A list of deviations from the feature extraction guidelines by the Image Biomarker Standardisation Initiative (IBSI).

Comment A3: Could the authors provide a broad description with which RT concepts these metastases were treated? Dosage and fractionation? Prescription isodose?

Author Response:
We have added a sentence on the broad description of RT treatment from the data available (total dose and number of fractions). We have also added a supplemental table on the dose fractions and frequencies.

Changes to Manuscript:
(results):
Added:
“Dosage and fractions are summarized in Supplementary Table S6. Liver metastases were treated with total dose range of 24Gy-80Gy (mean: 52.6Gy) and number of fractions from 3-50 (mean: 8.6).”

(supplementary information):
Added:

Total Dose (Gy)	Fractions	BED_{10} (Gy)	Patient Count
24	1	82	9
24	3	43	1
27	3	51	1
30	3	60	1
30	5	48	8
30	10	39	1
35	3	60	2
36	6	68	1
38	15	48	1
40	5	72	4
45	3	113	5
45	5	86	1
50	5	100	12
50	10	75	1
60	3	180	2
60	5	132	4
60	6	120	2
60	10	96	5
60	15	84	1
67.5	15	98	15
70	10	119	9
75	3	263	1

75	5	188	2
75	15	113	1
75	25	98	5
75	50	86	1
80	10	144	1

Supplementary Table S6: A list of doses to liver metastases, fractions, biologically effective dose (BED), and number of patients treated with the combination.

Comment A4: Radiomic models appear to be show a predicitive benefit, however, similar to "Other Clinical Data" and inferior to "Treatment Clinical data". This should be stated more clearly.

Author Response:
We agree with this comment that the similar of accuracies should be more explicitly stated. We note that a major implication of the study is that radiomic models and clinical models are not significantly different in predictive accuracy if machine learning is used and agree this should be emphasized.

Changes to Manuscript:
L54-55 (abstract):
Added:
“The clinical data only model achieved a similar C-index of 0.62 (CI: 0.56 - 0.69), suggesting that predictive signals exist in radiomics and clinical data.”

L59-61 (abstract):
Changed:
“The AI radiomics model achieves high prediction accuracy for progression-free survival of CLM, providing support that radiomics combined with machine learning may aid in clinical decision making that requires prognostic assessment.”
To:
“The AI model achieves good prediction accuracy for progression-free survival of CLM, providing support that radiomics or clinical data combined with machine learning may aid prognostic assessment and management.”

L295-296(discussion):
Added:
“Utilizing only clinical data did not result in a statistically significant decrease in accuracy than with radiomics alone.”

Comment A5: Treatment data is generally dependent on the treatment strategy of the physician and partially dictated by clinical circumstances (e.g. size, location of metastases). If a patient stratification would be possible prior to treatment planning e.g. based on a radiomic model - the treatment itself could be adjusted. A "pre-treatment" model would this have more clinical relevance.

a) Thus it would be interesting to determine the best "Pre-therapy model" -> Is there a benefit in combining "other Clinical Data" and "imagin Clinical Data" with radiomics?

b) Does the best pre-therapy model allow for significant patient stratification (e.g. cross validatd Kaplan Meier analysis)

Author Response:
This is indeed an excellent hypothesis to explore - if pre-treatment data provides some significant stratification, then we would have earlier markers than imaging or treatment data. We agree that investigating this in combination with radiomics would add to the completeness of the analysis. We have re-ran the algorithm combining the pre treatment ("other clinical data" and "imaging clinical data" with radiomics to observe differences in accuracies. Informally, the accuracies were similar and pre-treatment model does allow for patient stratification that is better than random chance and similar to the models with treatment data. We have commented this as a strength in using radiomic and pre-treatment variables in the discussion.

We now present a clinical implication that with the pre-treatment information, the good prediction accuracies can support usage of radiomics and clinical data for prognostic decision making, rather than relying on treatment data and physician judgment (which still performs similar). We have also decreased the strength of some of our statements so that we present the performance of each model accurately.

Changes to Manuscript:
L188-203 (methods):
Updated the list of feature sets, changed:
1. Non-imaging and non-treatment clinical data: baseline patient variables not related to treatment information or tumor geometry from CT imaging.
2. Treatment clinical data: variables related to treatment parameters, including dosimetric variables.
3. Imaging clinical data: variables related to tumor geometry measure in CT imaging.
4. All clinical data: The union feature sets 1-3.
5. Radiomics: tumor volume: radiomic features computed from the tumor volume only.
6. Radiomics: liver parenchyma: radiomic features computed from the liver parenchyma only.

7. Radiomics: liver parenchyma + tumor: radiomic features computed from the union of the tumor volume and liver parenchyma.
8. Treatment clinical data and radiomics from liver parenchyma + tumor: the union of feature sets 2. And 7.
9. All clinical data and radiomics from liver paraenchyma + tumor: the union of feature sets 4. And 7.

To:
1. Non-imaging and non-treatment clinical data: baseline patient variables not related to treatment information or tumor geometry from CT imaging.
2. Treatment clinical data: variables related to treatment parameters, including dosimetric variables.
3. Imaging clinical data: variables related to tumor geometry measure in CT imaging.
4. All clinical data: The union feature sets 1-3.
5. Radiomics: tumor volume: radiomic features computed from the tumor volume only.
6. Radiomics: liver parenchyma: radiomic features computed from the liver parenchyma only.
7. Radiomics: liver parenchyma + tumor: radiomic features computed from the union of the tumor volume and liver parenchyma.
8. Treatment clinical data and radiomics from liver parenchyma + tumor: the union of feature sets 2. and 7.
9. All clinical data and radiomics from liver paraenchyma + tumor: the union of feature sets 4. And 7.
To:
1. Non-imaging and non-treatment clinical data: baseline patient variables not related to treatment information or tumor geometry from CT imaging.
2. Treatment clinical data: variables related to treatment parameters, including dosimetric variables.
3. Imaging clinical data: variables related to tumor geometry measured in CT imaging.
4. All pre-treatment clinical data: All clinical data except treatment clinical data, which do not utilize variables based on physician judgment for treatment planning.
5. All clinical data: The union feature sets 1-3.
6. Radiomics: tumor volume: radiomic features computed from the tumor volume only.
7. Radiomics: liver parenchyma: radiomic features computed from the liver parenchyma only.
8. Radiomics: liver parenchyma + tumor: radiomic features computed from the union of the tumor volume and liver parenchyma.
9. Treatment clinical data and radiomics from liver parenchyma + tumor: the union of feature sets 2. and 8.
10. Non treatment clinical data and radiomics from liver parenchyma + tumor: the union of feature sets 4. and 8.
11. All clinical data and radiomics from liver paraenchymaparenchyma + tumor: the union of feature sets 5. and 8.

L226 (results):
Updated the results table (in red are new data points, note that CPH model results have been updated after re-optimization following suggestions from reviewer 2) (table 2):

Input Features	Concordance Index (95% CI)	Integrated Brier Score (95% CI)
(No Feature Selection, Local Progression as Outcome)		
Other Clinical Data	0.64 [0.54, 0.75]	0.18 [0.15, 0.22]
Imaging Clinical Data	0.66 [0.61, 0.71]	0.17 [0.14, 0.20]
Treatment Clinical Data	0.69 [0.62, 0.77]	0.17 [0.14, 0.20]
All Pre-treatment Clinical Data	0.63 [0.55, 0.71]	0.22 [0.19, 0.25]
All Clinical Data	0.67 [0.58, 0.75]	0.16 [0.15, 0.18]
Radiomics: Tumor Volume	0.64 [0.52, 0.76]	0.18 [0.17, 0.18]
Radiomics: Liver Parenchyma	0.61 [0.53, 0.69]	0.21 [0.19, 0.23]
Radiomics: Liver Parenchyma + Tumor	0.66 [0.58, 0.74]	0.20 [0.17, 0.22]
Treatment Clinical Data + Radiomics from Liver Parenchyma and Tumor	0.66 [0.59, 0.73]	0.19 [0.18, 0.21]
All Pre-treatment Clinical Data + Radiomics from Liver Parenchyma and Tumor	0.66 [0.55, 0.77]	0.21 [0.17, 0.25]
All Clinical Data and Radiomics from Liver Parenchyma + Tumor	0.64 [0.60, 0.68]	0.19 [0.16, 0.22]
(With Feature Selection, Local Progression as Outcome)		
Other Clinical Data	0.66 [0.56, 0.76]	0.19 [0.16, 0.22]
Imaging Clinical Data	0.61 [0.56, 0.66]	0.17 [0.14, 0.19]
Treatment Clinical Data	0.72 [0.64, 0.79]	0.18 [0.15, 0.21]
All Pre-treatment Clinical Data	0.65 [0.58, 0.72]	0.21 [0.18, 0.24]

All Clinical Data	0.62 [0.56, 0.69]	0.19 [0.16, 0.22]
Radiomics: Tumor Volume	0.58 [0.51, 0.84]	0.19 [0.16, 0.24]
Radiomics: Liver Parenchyma	0.66 [0.60, 0.72]	0.20 [0.18, 0.22]
Radiomics: Liver Parenchyma + Tumor	0.68 [0.62, 0.74]	0.20 [0.16, 0.25]
Treatment Clinical Data + Radiomics from Liver Parenchyma and Tumor	0.73 [0.64, 0.82]	0.18 [0.15, 0.20]
All Pre-treatment Clinical Data + Radiomics from Liver Parenchyma and Tumor	0.66 [0.57, 0.75]	0.20 [0.17, 0.23]
All Clinical Data and Radiomics from Liver Parenchyma + Tumor	0.69 [0.65, 0.74]	0.23 [0.21, 0.26]
With Cox Proportional Hazards Model		
Other Clinical Data	0.53 [0.50, 0.56]	0.20 [0.18, 0.22]
Imaging Clinical Data	0.56 [0.45, 0.67]	0.25 [0.22, 0.28]
Treatment Clinical Data	0.50 [0.48, 0.52]	0.24 [0.20, 0.28]
All Pre-treatment Clinical Data	0.54 [0.48, 0.60]	0.19 [0.15, 0.23]
All Clinical Data	0.57 [0.48, 0.66]	0.21 [0.16, 0.26]
Radiomics: Tumor Volume	0.47 [0.42, 0.52]	0.22 [0.17, 0.27]
Radiomics: Liver Parenchyma	0.49 [0.42, 0.56]	0.24 [0.22, 0.26]
Radiomics: Liver Parenchyma + Tumor	0.43 [0.40, 0.46]	0.25 [0.21, 0.29]
Treatment Clinical Data + Radiomics from Liver Parenchyma and Tumor	0.53 [0.45, 0.61]	0.19 [0.15, 0.23]
All Pre-treatment Clinical Data + Radiomics from Liver Parenchyma and Tumor	0.55 [0.49, 0.61]	0.20 [0.17, 0.23]
All Clinical Data and Radiomics from Liver Parenchyma + Tumor	0.58 [0.47, 0.67]	0.22 [0.19, 0.25]

Reviewer B

.

Comment B1: The manuscript lacks technical details regarding the radiomics process. I would recommend the authors peruse the IBSI reporting guidelines [1], and report missing elements in the manuscript or as supplementary materials. Aside from the usual, e.g. interpolation and discretisation, please report on the following:

- The use of contrast agents for pre-treatment CT imaging.
- Whether the version of pyradiomics used for the study is IBSI-compliant. This is not necessarily the case, despite claims on the pyradiomics website [2,3].

Author Response:

a) IV contrast was for CT imaging. This has been added to the description of the dataset.
b) We agree that addition of explicit details of the radiomics process is important for transparency and reproducibility. We had added a sentence on the IBSI deviations using Pyradiomics. The radiomic features resulted in several minor deviations to ISBI guidelines as we implemented pyradiomics without changing the application programming interface. The deviations are in rounding (where PyRadiomics does now round to a lower resolution when resampling) and in indexing conventions from binning and resampling. A sentence has been added commenting on the deviations to ISBI guidelines and deviations are listed in a new supplemental table.

Changes to Manuscript:

L219 (methods):
Changed:
“The query resulted in obtaining imaging and chart data for N=97 patients”
To:
“The query resulted in obtaining CT imaging with intravenous contrast and chart data for N=97 patients”

L150-152 (methods):
Added:
The majority of radiomic features follow the Image Biomarker Standardisation Initiative (IBSI) guidelines. Deviations from IBSI are listed in Supplemental Table S2.

L488-491 (supplementary material):
Added:

Computation	**PyRadiomics Implementation**	**IBSI Guidelines**

Binning	Discretizes gray values with fixed bins with edges equally spaced from 0.	Discretizes using fixed bin width equally spaced from minimum of resegmentation range
Resampling	Aligns to the corner of the original voxel	Aligns to the center of the image
Gray value rounding	Does not resample to similar resolution if original intensity is lower precision	Resamples to similar resolution if original intensity is lower precision
Mask resampling	Resamples to nearest neighbor	Allows selection of different interpolators for resampling

Table S2: A list of deviations from the feature extraction guidelines by the Image Biomarker Standardisation Initiative (IBSI).

Comment B2: The analysis is potentially biased.

- From the manuscript I could not establish if training and validation folds were strictly separated at all stages of the modelling process. Therefore, I have to assume that feature selection, hyperparameter optimisation, training and assessment of model performance were done consecutively. The authors performed training and subsequent assessment as part of a 4-fold cross-validation procedure. To avoid positively biasing results, feature selection and hyperparameter optimisation should be nested in this cross-validation procedure [4,5].
- Since we are dealing with metastatic colorectal cancer, I would expect that part of the patients succumb to the primary disease before lesion progress can be observed. How did the authors account for such competing risks? The random forest implementation cited by the authors supports modelling of competing risks (at least in the R implementation).
- It is not clear to me if cross-validation took place at the lesion or at the patient level. A patient may have multiple lesions. If cross-validation takes place at the level of the lesion, data specific to a patient may leak between training and va

Author Response:

a) We agree that the proper separation of training/validation sets is to perform feature selection and optimisation within the training subset for the cross validation loop. This was done but not explicitly stated. We have added a sentence to clarify this.

b) For this study, we did not query for patients who have succumbed to disease after radiotherapy and only queried for patients who are either living or have data of local progression. Unfortunately we do not have data on patient death to model and assess on the

effect of competing risks. We acknowledge this as a limitation in the study and have added a statement in the discussion section to highlight this.
c) Cross-validation was done at the patient level. That is, each “sample” is one lesion but multiple lesions pertaining to the same patient were grouped so that during cross validation splitting they are not placed in both the training and validation set. We have added a sentence to clarify this.

Changes to Manuscript:
L175-177 (methods)
Changed:
“The dataset is shuffled and partitioned into 4 equal sized subsets. Each is then used as a test data set for a model trained on the union of the remaining 3 subsets, giving 4 separate models trained and evaluated. “
To:
“The data was partitioned into 4 subsets of equal size and proportion of recurrences. The survival model was built by performing feature selection, training the RSF model, and hyperparameter optimization on 3 of the subsets and then evaluated with the remaining subset. This was repeated 4 times with a different testing subset. “

L312-314 (discussion):
Added:
“.Another exclusion is of patients who are deceased, and we were unable to evaluate the effect of death on the recurrence prediction model, which may require reparameterization with competing risks.”

L152-154 (methods):
Added:
“A set of radiomic features was computed for each lesion. Lesions were grouped together so that when they are shuffled into validation sets that no patient will have lesions both in the training and validation subset.“

Comment B3: In the introduction (lines 86-89), the authors refer to findings in the study by Wang et al. The reader may be helped by clarifying the issue with results whose confidence intervals contain the 0.50 threshold.

Author Response:
We have added a sentence to explain why <0.50 is an issue (i.e. random prediction). We have also refined the sentence to quantify how many studies resulted in a C-index range under 0.5 to reduce the subjective language.

Changes to Manuscript:
L89 (introduction):

Changed:
“Wang et al. [14] evaluated the accuracy of 9 different survival prediction scoring systems with the majority of these scoring systems resulting in a concordance index (C-index) crossing below the 0.50 threshold in its 95% confidence interval.”
To:
“Wang et al. [14] evaluated the accuracy of 9 different survival prediction scoring systems with f6 of these scoring systems resulting in a concordance index (C-index) crossing below the 0.50 threshold in its 95% confidence interval. A C-index crossing 0.50 would indicate prediction no better than random chance.

Comment B4: In the introduction (lines 108-112) the authors state that an advantage of AI is an automated and iterative approach where model parameters are repeatedly optimised. That is true for truly end-to-end machine learning software. However, the work described by the authors does not qualify as such due to the point described in comment 2a.

Author Response:
We agree that the line holds true for a training process that is truly separated from testing data. Similar to the response to comment 2, the process was indeed repeating features selection and hyperparameter optimization within each k-fold so that optimized parameters do not leak to another k-fold. Upon revision, we recognize that the statement in the manuscript can also be misleading and we have simplified it following the below changes. The idea that we attempted to state was that AI models can handle higher dimension parameterization compared to CPH for instance.

Changes to Manuscript:
(introduction):
L111 Changed:
“A strength of AI, specifically machine learning, is an automated and iterative approach, where model parameters are repeatedly optimized.“
To:
“AI, specifically machine learning, initializes models with parameters that can be optimized as more training data is available.”

Comment B5: In the introduction (lines 114-116), the authors claim to have developed software for the automated extraction of radiomics features from CT liver scans, for which the presented evidence is insufficient:

- The publicly available pyradiomics package is used by the authors to extract features. Implementation of this package does not constitute a major, original work.
- Liver and tumor volumes were segmented by radiation oncologists - not automatically.

Author Response:

a)This is correct that we implemented the package without modifying the API to extract radiomic features and that segmentations were extracted manually. We have reduced the strength of the statement in the introduction to remove the package implementation as original work.
b) We have also changed the automatic extraction of segmentations - this was unclear writing to express that we programmed an interface to map the manual outlines (which is standard of care) to convert the DICOM volume to a Nearly Raw Raster Data format. For clarity, we have changed this entire section to specify which portions we programmed (e.g. programmed a pipeline to utilize existing radiomics packages, random survival forest libraries, and manual segmentations to train the machine learning model).

This is certainly an important point to raise as transparency and writing an appropriate level of strength of claims is important to us and we do not want to overstate the contributions.

Changes to Manuscript:
L116-120 (introduction):
Changed:
“To do this, we developed software for the automated extraction of radiomic features from CT liver scans, a machine learning prediction model to utilize radiomic features to predict survival, and the validation of the performance of the model when enhancing radiomic features with clinical data.”
To:
“To do this, we programmed a machine learning pipeline by implementing existing radiomic libraries to extract features from liver volumes that were manually segmented by radiation oncologists. We then implemented known machine learning models to be trained with the radiomic data and validated it with known patient recurrence outcomes.”

Comment B6: Generally, the authors should make clearer that their results only pertain to that part of the patient population with CLM that were treated with primary or adjuvant radiotherapy. Title, highlights, abstract and other parts of the manuscript should be updated to indicate this.

Author Response:
We have updated the title, highlights, abstract, discussion, and conclusion to reflect the limited scope in that it only pertains to populations treated with RT.

Changes to Manuscript:
L1-2 (title):
Changed:
“Prediction of Local Control for Colorectal Liver Metastases using a Radiomic Artificial Intelligence Model”
To:
“Prediction of Local Control for Colorectal Liver Metastases Treated with Radiotherapy using a Radiomic Artificial Intelligence Model”

L44 (abstract):
Changed:
“Liver CT scans and outcomes for N=97 CLM patients“
To:
“Liver CT scans and outcomes for N=97 CLM patients treated with radiotherapy “

L311 (discussion):
Changed:
“Future studies may include patients before and after radiotherapy “
To:
“As the samples are limited to patients treated with primary or adjuvant RT, future studies may include patients before and after radiotherapy“

L351-352 (conclusion):
Changed:
“We have developed a time-dependent tumor progression prediction model for CLM treated with RT “
To:
“We have developed a time-dependent tumor progression prediction model for CLM treated with primary or adjuvant RT “

Comment B7: The comparison with Cox models could be skewed:

- It is unclear to me if the same feature selection procedure was applied for training the Cox models. If not, the Cox model may contain many variables with non-zero coefficients - essentially overfitting the model.
- The authors should be aware that the implementation of the Cox PH model in survival is (by default) subject to l2-regularisation. It is not a "standard" Cox PH model that one would find in STATA or SPSS. This l2-regularisation parameter is a model hyperparameter that should be tuned.
- The integrated Brier score is interpreted as indicating non-random predictions for experiments using random forests - but very similar values for Cox models are ignored.

This proposes an alternative interpretation for the authors findings: the Cox model against which the authors compare their random forest model(s) is poorly optimised. This would be more in-line with results present by e.g. Leger et al. [6], who did not observe a major performance boost by using random forests, albeit for a different survival endpoint.

Author Response:
a) The same feature selection procedure was indeed applied for the Cox model to prevent overfitting. This was unclear in the manuscript and a sentence has been added to state that the same feature dimensionality cutoff applied to both the RSF and cox model.
b) We did not optimize the L2 parameter. We agree that for validity of the Cox model, we need to perform optimization.

The model used (in python with the Lifelines library https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html) defines the penalization as follows:

$$\frac{1}{2}\text{penalizer}\left((1-\text{l1-ratio})\cdot||\beta||_2^2+\text{l1-ratio}\cdot||\beta||_1\right)$$

We have re-ran the model by implementing a gridsearch on different parameter values for the penalizer coefficient (0.001 to 0.3 in 50 linearly spaced increments) and L1 ratio (0 to 1 in 0.1 increments) and selecting the best one for the Cox model for the validation set. We have updated the results table with the optimized Cox models. Informally, we observed 2 differences:

1. The mean C-index increased, though selecting the penalizer parameter combination with highest c-index resulted in a larger variance through the validation sets. This was more apparent in the radiomics models. As a result, the confidence intervals did overlap with the random survival forest model, though it also was below 0.50 at the lower bound.
2. With lower penalizer values, the Cox model did not converge for the clinical variable models (the radiomic models with continuous variables all converged).

We interpret this as inconclusive. We cannot say for certain that the random survival forest model performs better than the CPH, however with our dataset the RSF consistent predicts better than random chance.

To refine the manuscript to acknowledge that the original comparison is likely skewed, we have made some changes reflecting this:

1. We decreased the strength of the statements in the discussion of the RSF outperforming to CPH. Instead we say that the CPH performs within range of the RSF but also has high variability and further studies reproducing both previous methods (such as by Leger et al.) and our method for feature selection and optimization are needed to investigate further.
2. We have removed the section in the discussion regarding the limitations of the CPH model to avoid conjecture without evidence now that the CPH model is not consistently significantly worse than the RSF model. This also reduces the word count as we were over the word limit with the additions. The discussion now has been reduced to that theoretically an RSF has more degree of freedom and can be parameterized with more complexity; however, the results are indeterminate.
3. We have added 2 more references to support this in the discussion where we discuss other studies showing good performance with the Cox model.

   .

   Leger, S., Zwanenburg, A., Pilz, K., Lohaus, F., Linge, A., Zöphel, K., ... & Richter, C. (2017). A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Scientific reports, 7(1), 1-11.

Chang, E., Joel, M. Z., Chang, H. Y., Du, J., Khanna, O., Omuro, A., ... & Aneja, S. (2021). Comparison of radiomic feature aggregation methods for patients with multiple tumors. Scientific reports, 11(1), 1-7.

c) This is correct that the Cox values for IBS are similar and we did not discuss it in detail. We like to extend specific thanks to the reviewer for taking time to find relevant literature sources for this.

Changes to Manuscript:

(introduction):
Removed:
“that Cox-regression may not capture nonlinear relationships between clincopathological variables “

L213 (methods):
Changed:
“and with a CPH model”
To:
“and with a CPH model with gridsearch optimization of the regularization parameter”

L226 (results) Changed results Table 2:

With Cox Proportional Hazards Model		
Other Clinical Data	0.55 [0.52, 0.58]	0.21 [0.19, 0.23]
Imaging Clinical Data	0.51 [0.49, 0.53]	0.22 [0.20, 0.24]
Treatment Clinical Data	0.47 [0.41, 0.53]	0.24 [0.21, 0.27]
All Clinical Data	0.52 [0.50, 0.54]	0.19 [0.15, 0.23]
Radiomics: Tumor Volume	0.48 [0.46, 0.50]	0.20 [0.18, 0.22]
Radiomics: Liver Parenchyma	0.48 [0.46, 0.50]	0.22 [0.20, 0.24]
Radiomics: Liver Parenchyma + Tumor	0.41 [0.39, 0.43]	0.24 [0.20, 0.28]

Treatment Clinical Data + Radiomics from Liver Parenchyma and Tumor	0.51 [0.47, 0.55]	0.17 [0.14, 0.20]
All Clinical Data and Radiomics from Liver Parenchyma + Tumor	0.52 [0.50, 0.54]	0.19 [0.16, 0.22]

To:

With Cox Proportional Hazards Model		
Other Clinical Data	0.53 [0.50, 0.56]	0.20 [0.18, 0.22]
Imaging Clinical Data	0.56 [0.45, 0.67]	0.25 [0.22, 0.28]
Treatment Clinical Data	0.50 [0.48, 0.52]	0.24 [0.20, 0.28]
All Pre-treatment Clinical Data	0.54 [0.48, 0.60]	0.19 [0.15, 0.23]
All Clinical Data	0.57 [0.48, 0.66]	0.21 [0.16, 0.26]
Radiomics: Tumor Volume	0.47 [0.42, 0.52]	0.22 [0.17, 0.27]
Radiomics: Liver Parenchyma	0.49 [0.42, 0.56]	0.24 [0.22, 0.26]
Radiomics: Liver Parenchyma + Tumor	0.43 [0.40, 0.46]	0.25 [0.21, 0.29]
Treatment Clinical Data + Radiomics from Liver Parenchyma and Tumor	0.53 [0.45, 0.61]	0.19 [0.15, 0.23]
All Pre-treatment Clinical Data + Radiomics from Liver Parenchyma and Tumor	0.55 [0.49, 0.61]	0.20 [0.17, 0.23]
All Clinical Data and Radiomics from Liver Parenchyma + Tumor	0.58 [0.47, 0.67]	0.22 [0.19, 0.25]

L263-269 (discussion)
Changed:
“There are several opportunities we aimed to address to improve on existing methods. First, the Cox regression models data with an exponential hazard function with variables parameterized in a linear combination [30].“
To:

There are several opportunities we aimed to address to improve on existing methods. First, the CPH model in theory is parameterized with lower complexity than RSF and may be unable to capture nonlinear dependencies [32]. However, from our results, this is indeterminate as the although the CPH did not significantly perform better than random chance, there was a wide confidence interval to overlap with the RSF model. When assessing performance with IBS, the CPH model was not greater than the 0.25 threshold for only the combined radiomics and clinical subsets. Recent studies modelling survival with radiomics show no significant difference between CPH and RSF models [33, 34]. Comparison of our model may require a larger sample size and to evaluate the feature selection and optimization methods other studies have used. “

Comment B8: Table 1 contains both transarterial radioembolization and embolization before radiotherapy as features. What is the difference?

Author Response:
We acknowledge this is unclear from the table. The two categories should be Y90 radioembolization and arterial embolization.

Changes to Manuscript:
L206 (table 1)
Changed:
“Transarterial embolization before radiotherapy
Embolization before radiotherapy”

To:
“Arterial embolization before radiotherapy
Yttrium-90 embolization before radiotherapy”

Comment B9: From which time point were survival endpoints determined?

Author Response:
The time point from which followup was started was after radiation therapy in the institution. This has been added to the definition in the methods section.

Changes to Manuscript:
L134-135 (methods):
Changed:
“The task for the AI model was to predict the primary endpoint, defined as time until local tumor progression”
To:
“The task for the AI model was to predict the primary endpoint, defined as time from CLM radiation therapy until local tumor progression”

Comment B10: I am really confused by the use of the Gini importance score as a measure of feature importance. Gini importance is associated with binary endpoints, which is not the case here. Could the authors clarify this point?

Author Response:
This is an error and we have updated the description (the correct statement was in place in the abstract). We like to clarify that the algorithm for importance score is a perturbation error rate (termed “variables importance’ in the PySurvival API), which is not technically a Gini computation.

Changes to Manuscript:
L172 (methods):
Changed:
“feature importances were computed by the Gini importance score.”
To:
“feature importances were computed by error rates between the perturbed and unperturbed model for that feature.”

Comment B11: How were confidence intervals treated when averaging over the cross-validation steps?

Author Response:
We have rewritten the methods section on confidence interval computation to reflect how CIs were averaged. The confidence intervals were computed by taking the C-indices of the 4 k-folds as a Guassian distribution, computing the confidence interval from the mean and Gaussian standard deviation. The k-folds were stratified to mitigate bias.. We acknowledge that this has limitations in that there still is a risk of bias due to the low sample size and may not have reached central limit sufficiently. Additionally, the data was from one institute which may leak common features to different samples. An alternative we could do is bootstrapping N times (e.g. 1000) to get a distribution that is more likely to be Guassian but we did not opt for this to keep the test sets distinct.

We have added the following citations to the manuscript:

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence, 32(3), 569-575.

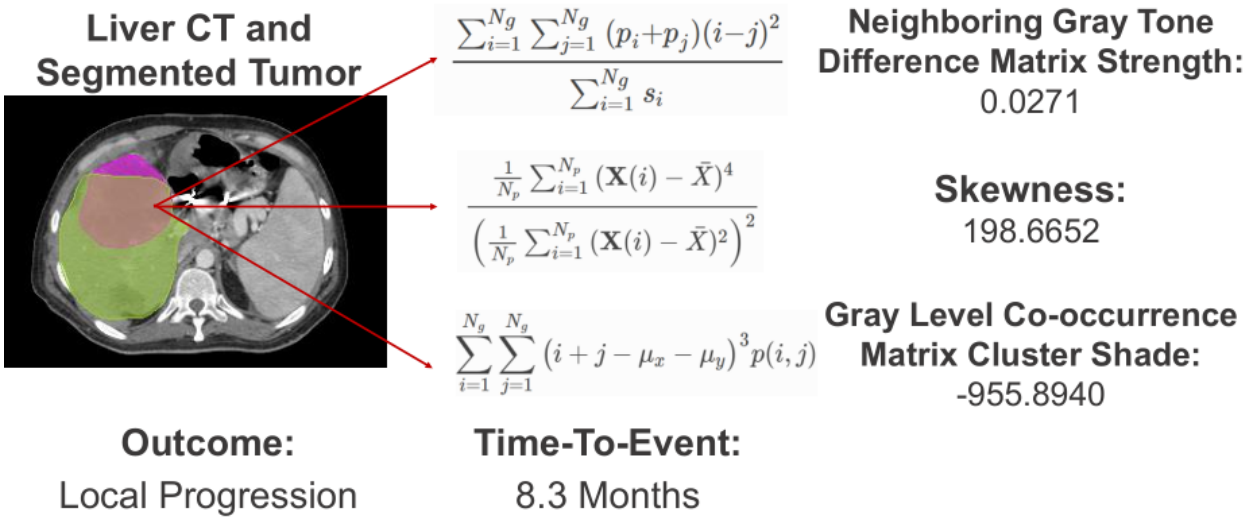
Changes to Manuscript:
L179-182 (methods):
Changed:

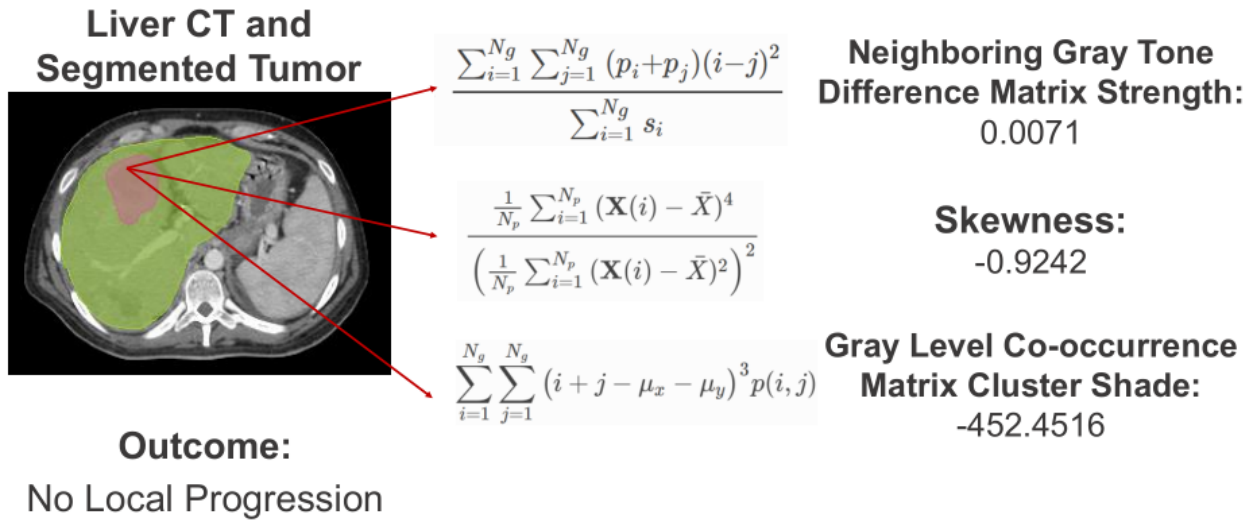
“The concordance index (C-index) integrated Brier score (IBS) were computed as accuracy metrics to evaluate the models in each k-fold. The confidence interval for the C-index was computed using Somers’ Dxy rank correlation “
To:
“The concordance index (C-index), computed by Somers’ Dxy rank correlation [29], and integrated Brier score (IBS) were averaged over 4 k-folds with confidence intervals computed by using the standard error of the distribution of C-indices. “

Comment B12: Figure 2, in my view, adds very little to the manuscript, and could be omitted or replaced. Of interest would be a figure that shows examples of lesions and associated progression-free survival as well as values of relevant radiomics features.

Author Response:
We have updated figure 2 as recommended to be of lesions with associated progression-free survival and 3 selected high-importance radiomic features. The figure is pasted below as an image for this document and the final uploaded version is a .pdf with vectorized shapes.

Changes to Manuscript:
Added Figure 2:





With the caption:
Figure 2: Two examples of lesions and selected radiomic features. A total of 108 radiomic features were computed for each patient's tumor volume. Many features have mathematical significance, such as neighboring gray tone difference strength representing coarse differences in gray level intensities, but may be difficult to interpret physiologically. The example features displayed were ones that were predictive of local progression in the final machine learning model.

Comment B13: Appendix A mixes both patient and lesion characteristics, which is confusing. It may be preferable to organise patient characteristics by patient and lesion characteristics by lesion.

Author Response:

We agree that it is confusing to mix both patient and lesion distributions. We have now made separate tables for patient and lesion characteristics. There is one separate table for averages computed from lesions and averages computed from patients. For the patient table, lesion-specific variables (essentially treatment variables for the lesion such as PTV, or lesion dimension) were excluded. The original table was all computed for lesions.

Changes to Manuscript:
L532 (supplemental information):
Added:

Baseline variable distributions by patient

Characteristics	**All Patients (n=97)**
Sex, n (%)	
Male	63 (64.9)
Female	34 (35.1)
Metastasis at time of diagnosis, n (%)	
M0	31 (32)
M1	65 (67)
Other sites at diagnosis, n (%)	
None	74 (76.3)
Lung	9 (9.3)
Non-regional LN	3 (3.1)
Lung and non-regional LN	4 (4.1)
Other	5 (5.2)
Undetermined	2 (2.1)
RT to other sites, n (%)	
No	57 (58.8)
Before liver RT	23 (23.7)
After Liver RT	13 (13.4)

Before and after liver RT	2 (2.1)
Undetermined	1 (1)
Number of liver lesions at RT, n (%)	
1	56 (57.7)
2	25 (25.8)
3	6 (6.2)
≥ 4	9 (9.3)
Undetermined	1 (1)
Other sites at RT, n (%)	
None	37 (38.1)
Lung	21 (21.6)
Non-regional LN	8 (8.2)
Lung and non-regional LN	17 (17.5)
Other	14 (14.4)
Freedom from local progression (FFLP), n (%)	
Progression	50 (51.5)
No progression	40 (41.2)
Undetermined	7 (7.2)
Mean time to FFLP (months) ± SD	10.5 (8.8)
Any hepatic progression (AHP), n (%)	
Progression	76 (78.4)
No progression	16 (16.5)
Undetermined	5 (5.2)
Mean time to AHP (months) ± SD	7.4 (6.9)

Abbreviations: LN = lymph node, RT = radiotherapy

Table S1: A table of baseline clinical variables recorded as part of standard of care, with averages computed from the set of variables per patient. Lesion-specific variables were excluded.

Comment B14: Can the authors make their models publicly available to allow for external validation?

Author Response:
Yes, the models can certainly be public. A sentence has been added to link to a public repository containing the models.

Changes to Manuscript:
L226-227 (results):
Added:
“The final models were uploaded to a public repository linked in Supplementary Table S7.”

L550 (supplementary material)
Added:
“The survival models from the different feature sets are uploaded to an open repository at: https://github.com/ricky-hu/local_control_radiomics_survival_model”