An integrated clinical-MR radiomics model to estimate survival time in patients with endometrial cancer

Journal:	*Journal of Magnetic Resonance Imaging*
Manuscript ID	JMRI-22-1126.R2
Wiley - Manuscript type:	Research Article
Classification:	Artificial Intelligence/Machine Learning, Mathematical models of imaging processes < Imaging technology and safety < Basic Science, Pelvis (female) < Body imaging < Clinical Science, Pelvis (female) < Body imaging < Clinical Science, Mathematical models of imaging processes < Imaging technology and safety < Basic Science
Manuscript Keywords:	Endometrial cancer, survival analysis, radiomics, feature selection, Cox Proportional Hazards Model, T2-weighted MRI

Dear Dr. Schweitzer,

Thank you very much for giving us the opportunity to revise our paper. The boldfaced words below are our point-to-point responses to the questions raised by the reviewers.

Title Page

1. Title should not end in a period.

We have deleted the period in the title.

Introduction

2. P1L21 – To call it "imaging signatures" here is misleading because clinical features were also used. It should be made clear that the authors are investigating the value of T2-weighted MRI *in addition to* clinical features.

To identify clinical features and imaging signatures on T2-weighted MRI that can be used in an integrated model to estimate survival time for endometrial cancer subjects.

For Peer Review

3. P1L23 – Please verify with JMRI editors that the standard nomenclature is "T2-weighted".

Yes, it is T2-weighted MRI.

4. P2L38 – I would consider this work in the category of Stage 3 Technical Efficacy according to the definition outlined at https://onlinelibrary.wiley.com/doi/full/10.1002/jmri.25417.

Ok.

5. P3L41 – An example of several radiometric features here would be beneficial.

There have been studies evaluating the application of radiomics, usually based on multi-sequence MRI features, for example Kurtosis from contrast-enhanced T1-weighted MRI to predict survival time in endometrial cancer.

6. P3L55 – It should be "clinical prognostic variables".

It has been corrected to "these studies did not include or combine clinical prognostic variables in the CPH model for the survival time prediction…"

Materials and Methods

7. P4L36 – The meaning of "initial" was not clear the until the inclusion criteria were outlined below.

"initial" was changed. The sentence is : "and 270 of the initially considered 611 subjects were obtained from a previous study".

8. P4L41 – Clarify that this is different *imaging* acquisition parameters and protocols. I would

like to see an explanation of why only T2-weighted images were considered, i.e. if it is the most common contrast used in endometrial cancer evaluations. I would be curious to see a Table with the various sequence parameters, including TE, slice width, resolution, and field strength; perhaps as a subfield under the "MRI manufacturer" heading in Table 1.
We added Table 1 to explain the sequence parameters for training dataset, including TE,TR, slice thickness, space between slice, and field strength.

9. P5L8 – Please revise the wording of the sentence "That information was updated..", it is awkward.
This has been deleted.

10. P5L13 – Details of the motion artifact determination should be described, i.e. if they were subjectively judged by a technologist.
Yes, it is judged by radiologist. Now the sentence has been changed to "no severe motion artifacts in T2-weighted images that obscured the tumour mass, which was determined by radiologists subjectively"

11. P5L18 – Please clarify if this is one slice before or after the images are resampled to have the same voxel size. I am concerned that this criteria will add bias towards including image sets with smaller slice thickness.
It is before image resampling. The sentence has been changed to : "sufficient size of the tumor on images (i.e., the tumor could be identified on more than one MRI slice before image resampling),"

You are right, image resampling step could introduce bias for the image with smaller slice thickness. Our software for radiomics can only deal with same voxel size image, because it will be easier for us to compare the quantitative measures (e.g., volume and shape features etc.) from different subjects.

12. P5L18 – "Able to pass the image pre-processing steps". Please link to Figure 3, S2, or provide a brief description of the image pre-processing here.
Ok, the link is provided as "the T2-weighted sequence passes the image pre-processing steps (see step two in Figure 2)."

13. P5L46 – The concept of "right censoring" is unfamiliar to me and bears some additional description, or at minimum some references, especially describing how right censored data impacts the survival estimate.

Explanation is now added as "74 of those 82 cases were right censoring (i.e., 74 patients have not (yet) experienced the relevant outcome/death, by the time of the close of the study, in our case the close of the study was Dec 1st,2021) and 8 cases having died before

For Peer Review

the end of the study (Dec 1st, 2021) .The right censored survival times underestimate the true (but unknown) time to event (6).”

14. P6L18 – should be the "mean survival time".
This has been moved to “Results” section as following:

For the survival time (Figure 3B and Figure 3E), no significant differences were revealed (training dataset: 870.6±592.1 days, testing dataset: 637.1±314.2, p=0.09).

15. P6L37 – Readers will likely be more familiar with the acronyms, DICOM and PACS, so they should be included.
These have been changed. In “Radiomics study pipeline” section:

Specifically, Digital Imaging and Communications in Medicine (DICOM) file formats were downloaded from the picture archiving and communication systems (PACS), de-identified and converted to the simpler Neuroimaging Informatics Technology Initiative (NIFTI) format.

16. P6L47 – ITK-snap should be cited with the following reference: "Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 2006 Jul 1;31(3):1116-28"
This paper has been cited as reference 24.

17. P6L51 – The acronym, NIfTI should be included. Additionally, this sentence needs clarification. Was this the T2-weighted DICOM image that was converted to NIfTI? What precisely is the "relevant" image set?

This sentence has been rewritten as: After loading the T2-weighted MRI, the paintbrush tool was used to label all voxels containing visible tumor on each sagittal slice.

18. P7L20 – Please clarify whether images were resampled before or after tumour labelling. The phrase "was adopted" is unclear, were images/masked resliced in the sagittal plane?
We removed the “sagittal scan was adopted.”. The images and its masks were resampled after tumour labelling. The T2-weighted MRI was scanned in sagittal plane. We clarified the order of image resampling as:

“Following bias correction, T2 MRI and its masks were resampled to median voxel resolution (Figure 2)”

19. P7L25 – A single number for image size is incomplete. It would be clearer to specify the

For Peer Review

range of the values of in-plane resolution. It should be "median resolution" instead of "value".
We replace the "value" with "resolution" as "The median resolution (image voxel size) of all T2-weighted images..."

20. P7L32 – To avoid confusion in this sentence, I suggest: "Third, following bias correction, T2 MRI and its masks..".
We changed the sentence as: Following bias correction, T2 MRI and its masks were resampled to median voxel resolution (Figure 2)

21. P7L41 – In the Discussion, the possibility of other approaches for image intensity normalization could be discussed, such as normalization to a reference tissue outside the tumour-affected region.
We added a new paragraph as follows:

Different from quantitative MRI such as apparent diffusion coefficient (ADC) from diffusion-weighted imaging, T2-weighted MRI signal depends on the acquisition protocol, the coil profile, the scanner type, etc. and there is not standard method to normalize the image intensity for cross-subjects comparison. We adopted z-score like method to normalize the image intensity, other methods such as min/max normalization or scaling the image intensity to common max value can also be used. An alternative method to reduce the image intensity difference of T2-weighted images acquired from different centers and scanners is to normalize to a reference tissue outside the tumor-affected region such as cerebrospinal fluid (CSF) in brain or bladder where baseline water signal can be obtained. Although the image intensity features such as mean intensity value within the tumor mask will be affected by different image normalization steps, the tumor shape, volume, and image complexity radiomic features will not be affected by the image normalization step.

For Peer Review

22. P8L44 – It might be more accurate to say that the "CPH method was employed to generated the model to estimate the survival time".
The sentence is modified as :

The selected CPH model was then applied to calculate the survival time.

23. P9L3 – The sentence "The CPH was implemented.." can be omitted as it is a repetition of the previous sentence.
It has been removed and the sentence is: "The nomogram was applied to visualize the prediction survival probability."

24. P9L9 – "estimation of the two models was constructed and compared for the prediction" can be replaced with "two models were constructed and compared".

To study the influence of the radiomic features on the survival probability, two models were constructed and compared for the estimation.

Results

25. P10L30 – " Thus, there was no indeterminate index test or standard reference results." Please elaborate, with references, for users not familiar with this methodology,
This has been deleted in the manuscript.

26. P10L40 – "MRI features including age" can be replaced with "features". Similarly, on P10L47, don't call it "MRI features" if it includes the clinical features.
These have been deleted. "all 959 features (958 MRI features + age)"

27. P10L54 – Please define the lambda value / tuning parameter in the Methods.
Added in the method section as: "To avoid model overfitting, a 10-fold cross validation for penalized Cox regression models with grouped covariates was adopted to determine the optimal regularization parameter lambda (λ)."

28. P12L11 – Please revise the wording of "split with survival objects (time/death)" or define "survival object" in the Methods.
Added in the section above "Testing data": with balance the survival object (i.e., the combination of time and death information) distributions within the splits.

29. P12L18 – Some elaboration is required here. Why is the AUC a good example if the methods have randomness?
Because the bootstrap method and stratified sampling method have randomness, and as a typical example, the AUC was calculated and displayed because AUC is widely used criterion for measure discrimination.

30. P12L46 – Do not use the term "significantly" unless there is evidence of significance from a statistical test.
It has been removed "suggesting that the integrated clinical-radiomic model is superior to the clinical model"

31. P12L54 – This sentence seems unnecessarily wordy: "Similar to validation and using the same trained model (obtained from the training dataset), testing was carried out and the results are presented in Figures 5B". It could be changed to "Similarly, AUC curves were computed using the trained model on the testing dataset.."
It has been changed to : "Similarly, AUC curves were computed using the trained model on the testing dataset and the results are presented in Figure 5B."

For Peer Review

32. P13L31 – I recommend removal of "based on all training, validation and testing datasets". It infers that authors planned to combine all the datasets, but instead different combinations of the datasets were considered.
It has been deleted: "A likelihood-ratio test showed a significant difference between the integrated model and clinical model based on both training and testing datasets."

33. P13L47 – I would remove "play an important role", as it is subjective, and revise the sentence to "radiomics features improve survival prediction..".
This has been removed.

34. P13L52 – Be consistent, the "model with the age and clinical cancer grades" can be replaced with "clinical model" as it was already defined.
The sentence has been changed to: "The CI was 0.797 for the clinical model and 0.818 for the integrated model."

35. P14L18 – Please define the "threshold probability" and "net benefit" in the context of this application in the Methods.
These definitions were added:

Net benefit is calculated for each possible threshold probability which puts benefits and harms on the same scale. Threshold probability is the expected benefit of treatment is equal to the expected benefit of avoiding treatment (30). By varying the threshold probability, DCA allows us to examine whether one model is superior to another at a certain range of threshold probability with respect to the net benefit.

36. P14L34 – Should this be "when the threshold is *below* 0.5"?
Yes, it should be below 0.5: The radiomic model had a larger net benefit than the clinical model when the threshold was below 0.5 (Figures 7B and 7C).

37. P15L10 – Further description is required to justify the principal component analysis used here. Why is it not simply a comparison between the distribution of the radiomic features that were used in the integrated model?
Explanation for this has been added in the following/(main text):

It is not obvious to inspect the features difference from two dimension, i.e., an image with 413 rows and 985 feature columns. Dimension reduction method was applied to obtain the major components of the features from each type of scanner. Specifically, principal component analysis (PCA) was applied to study the effect of feature difference from different scanners.

Discussion

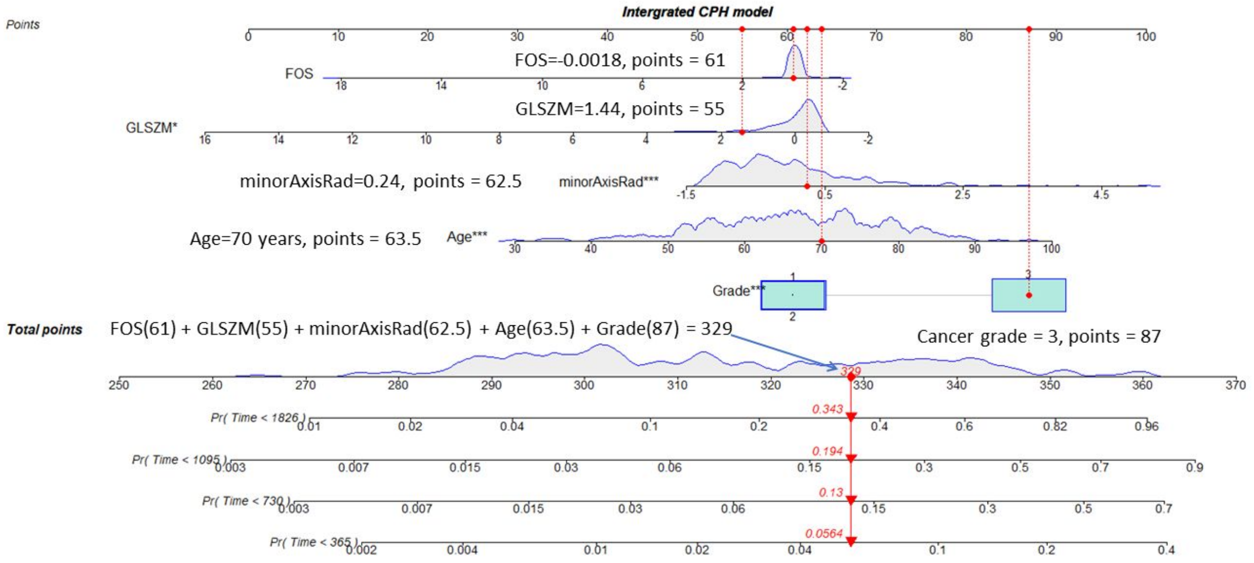
For Peer Review

38. P16L9 – Easy and straightforward are synonyms, please omit or replace one term.
Replaced with graphical, the sentence reads as:

Furthermore, the CPH model with a nomogram for visualization provided a graphical, straightforward, and non-invasive method of predicted survival…

39. P16L31 – Please include the acronyms for the quality score and transparent reporting. They may only be used once, but the acronym gives the unfamiliar reader a clue that these phrases describe established methods.
Ok, TRIPOD was added as follows: transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines

40. Figure 6 – Please use different colours for the different features (right now all are red and there is some overlap which is confusing), and indicate in the text the amount of points obtained from each. Include expansion of acronyms "FOS" and "coxph", so that labels are consistent with how these are referred to within the body of the manuscript.
It is not easy to set color for each feature. To avoid the overlay problem, we selected another subject who has cancer grade of 3 and age of 70 (see Figure below).



We also added the method to calculate total score in the figure and changed the corresponding text in the manuscript.

41. P17L6 – Is there a reason why the authors did not include T1-weighted images in their analysis? Is it not generally part of the standard imaging protocol for endometrial cancer?

T2-weighted imaging (T2WI) is the mainstay of pelvic MRI. They are best performed without fat suppression (FS) due to the inherent contrast between the signal intensity (SI) of the uterus and the surrounding fat. Thin sections (3–4 mm) and a FOV of 20–24 cm are recommended. For T2WI, image acquisition must be optimized and angled perpendicularly to the endometrium or cervix.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6375059/

We add "we obtained the features from T2-weighted image which is the mainstay of pelvic MRI" in the manuscript.

P17L40 – Can the authors comment on why gLASSO was better than composite minimax concave penalty method?
The composite minimax concave penalty is another objective function to achieve gLASSO.

https://cran.r-project.org/web/packages/grpreg/grpreg.pdf

In the study, the penalty of GEL (Group exponential lasso (Breheny, 2015)) was applied. We tried CMP (A hierarchical penalty which places an outer MCP penalty on a sum of inner MCP penalties for each group (Breheny & Huang, 2009)). We found both methods select same features.

We changed the manuscript as:

Secondly, although we had also tested the composite minimax concave penalty method for the model selection in the CPH model (which produced the same model selection results as the group exponential LASSO method), other methods such as the regular elastic net and ridge models method, which may produce better results.

P18L20 – The adjective should be "additional" not "alternative", since the method would benefit from any number of additional MRI contrasts.

It has been changed to additional as:

future work could evaluate the use of additional MRI sequences or quantitative MRI such as diffusion–weighted images with ADC maps

P18L23 – This term is typically "dynamic, contrast-enhanced MRI (DCE-MRI)".
It has been changed to "as well as dynamic contrast-enhanced MRI (DCE-MRI) (38)."

P18L25 – "Also, a possible avenue to explore in the future would be the boosting method (35) or the deep survival model (36) for the survival study." Please elaborate on the rationale for pursuing these methods.
which do not based on the assumption of proportional hazard.

These methods do not require proportional hazard assumption. This has been added as:

For Peer Review

the boosting method or the deep survival model for the study of survival because these methods do not require the proportional hazards assumption

Conclusions

P18L36 – Please remove "enhanced" since the integrated model was introduced herein. The authors should, however, highlight the beneficial addition of the T2-weighted features.
The ”enhanced” has been removed from the text.

P18L40 – Clarify that the phone application would be intended for clinicians. The statement "enable true personalized medicine" is overreaching and should be removed; I suggest removal of this entire sentence.
The sentence has been removed.

P19L3 – Please revise this sentence to clarify which AUCs the authors are referring to.
AUC denotes area under the receiver operating characteristic (ROC) curve (AUC). We added this when we first mentioned AUC as : “thereafter AUC is specified for AUC of ROC.”

P19L6 – suggest that the sentences in the conclusions are rearranged to end on a stronger note. The sentence beginning with "Also, .." could be moved to become the second sentence of the Conclusion.
As suggested by editor, “Also” has been removed, and the sentence becomes:

Furthermore, based on the testing dataset, we found that the integrated model is robust against the variability of the independent external testing dataset, as the AUC value showed only a marginal decrease, while for the clinical model the AUC decreased markedly.

For Peer Review

Minor grammatical points

42. P3L32 – "While" should be removed.
Removed.

43. P5L27 – Should be "diagnosis and surgery date".
It has been corrected to: 2) availability of age at diagnosis and surgery date,

44. P5L27 – Should be "type of cancer coexisting with..".
Changed to: 3) no other type of cancer coexisting with.

45. P8L21 – There is a typo, it should be "performed".
Replaced preformed with performed

46. P9L47 – decision curve analysis *was* applied.
replaced “were” with “was”

47. P13L23 – Can be referred to as the "external dataset" or "independent dataset".
This sentence has been removed.

48. P13L28 – "5-year survival probability".
The nomogram display the 1, 2, 3, and 5-year survival probabilities is shown in Figure 6.

49. P16L24 – should be "the CPH model".
This sentence has been removed.

50. P18L10 – Should this be "multi-center study"?
This has been removed.

51. P18L30 – This should be "as other methods", not "with".
Replaced “with” with “as”

Reviewer: 2

Comments to the Author
The authors describe an interesting study and approach to survival prediction in patients with endometrial cancer using radiomics. Albeit retrospective, the study is well-conducted with good methodological detail and sound conclusions. It could be further improved by providing greater detail on:
- MRI sequence parameters for each scanner manufacturer and field strength;

We added Table 1 to provide the scanner sequences and other parameters for training dataset.

Three types of scanners (Table 1) were used. Most T2-weighted images were collected with 1.5T scanner.

- the process of segmentation "checking' by the two experienced radiology consultants - this would have been an opportunity for inter-observer metrics to be collected / reported or consensus measures; please provide additional information;

For Peer Review

There is disagreement on the segmentation. Figure R1 below plots the difference of the number of voxels from two radiologists.

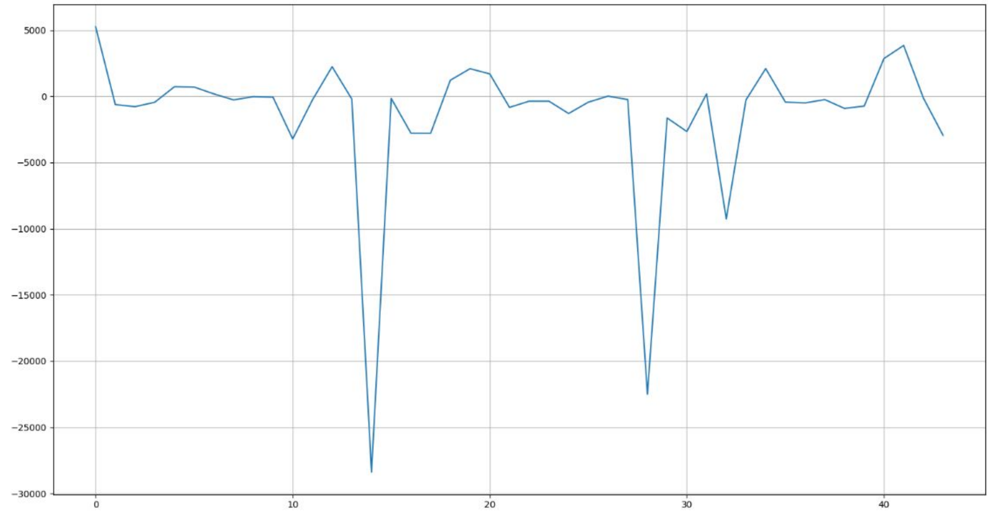


Figure R1. Difference between two radiologists for the tumor segmentation from random select 44 subjects. The X-axis is the number of subjects; the Y-axis is the number of voxel difference from two masks segmented by two radiologists.

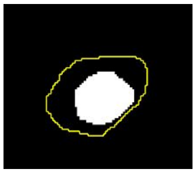


Figure R2. An example of segmentation difference from one subject (Batch_1_42.nii.gz). The white region was delineated by the first radiologist and the yellow curve region was checked and defined by the second radiologist.

From Figure R1, we can see that there was difference between two radiologists for tumor manual segmentation, but most of segmentation has smaller difference, i.e., less than 2000 voxel (Figure R1). Figure R2 show one example slice for the difference.

- more information should be provided on the other predictors that were not selected in the radiomics analysis (even if thematically grouped together given the large number analysed) and how these decisions were made;

The other predictors were from the first order statistics, image intensity-based features, and other shape-based features such as fractal dimension which measure the complexity of the shape. As a predictor in the regression model, each feature plays a different role. Within the framework of the Cox regression model, Figure R3 (below) shows the importance for each feature in terms of p value from univariate Cox regression. However, the decision was made based on statistic model selection (i.e., gLASSO in our study), which determines the features not on the individual importance of each feature, but the combination of a set of features . The model selection such as LASSO (https://en.wikipedia.org/wiki/Lasso_(statistics)) seeks to select useful features which simplifies the model but keeps a reasonable accuracy in the meantime. In this study, we adopted the bi-level method (https://onlinelibrary.wiley.com/doi/10.1111/biom.12300) to select these features automatically, because we have category variables (cancer grade and risk score) in the model.

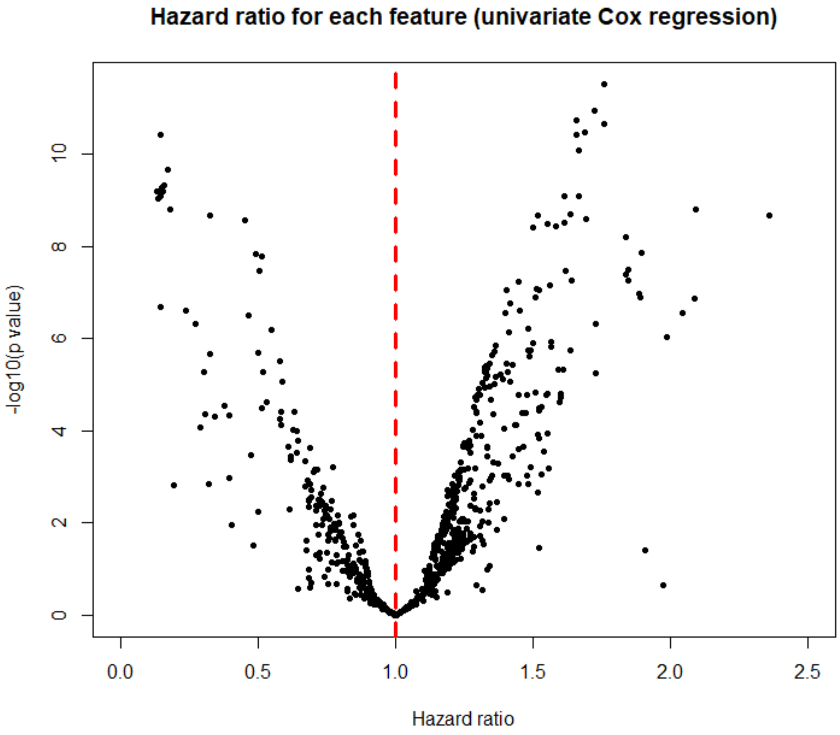


Figure R3. Volcano plot for each image features for survival analysis using the Cox regression model. The scatterplot shows statistical significance (-log(p value), Y axis) versus Hazard ratio obtained from Cox regression model (x-axis).

- a better discussion of sources of error and their magnitude in the pipeline / model build and deployment is recommended - this will be useful for future studies in same / other malignancies;

Based on Figure 2 (pipeline for the study), the sources of error can come from the first 3 steps, i.e., image segmentation, image processing, and feature extraction. In the image segmentation step, the error is generated if the tumor mask is not parcellated properly. For the image processing step, the image interpolation method is another source of error. In the feature extraction step, bias can be produced if only a fraction of image feature were extracted from the image.

This has been added in the manuscript.

Additional comments:
What does "and the image size was between 256 and 640" mean?

It means that the image reconstruction matrix is between 256 and 864 in the sagittal direction (Table 1).

A more general proffered is recommended - for example, DICOM and NIfTI acronyms should be used.

These have been changed in the manuscript.

Yours faithfully,

Xingfeng Li

For Peer Review

Title page

Title: An integrated clinical-MR radiomics model to estimate survival time in patients with endometrial cancer

Author Names and Degrees:

Xingfeng Li, PhD[1], Diana Marcus MBBS, BSc, PhD, MRCOG[2], James Russell, MBChB, BSc, MRCS, FRCR[3], Eric O. Aboagye, PhD[1], Laura Burney Ellis, MBChB, BSc[3], Alexander Sheeka, BMBS, BSc[3], Won-Ho Park, MBBS, MRCP, FRCR, MSc[3], Nishat Bharwani, BSc, MBBS, MRCP, FRCR[3], Sadaf Ghaem-Maghami, MBBS, MRCOG, PhD[3], Andrea G Rockall, MRCP, FRCR[1,3]

Author Affiliations:

From the [1] Department of Surgery and Cancer, Imperial College, London, UK; [2]Chelsea and Westminster Hospital NHS Foundation Trust, London, UK; the [3]Imaging Department, Imperial College Healthcare NHS Trust, London, UK.

Corresponding Author Information:

Address correspondence to A.G.R. (email: a.rockall@imperial.ac.uk). Tel: +44 7866 585 476, 1st Floor, ICTEM building, Hammersmith Campus, Du Cane Road, London, UK. W12 0NN.

Acknowledgments/Grant Support:

This study is funded by Imperial National Institute for Health Research (NIHR) Biomedical Research Centre, Imperial CRUK Centre and Experimental Cancer Medicines Centre.

For Peer Review

Running Title: Integrated model for endometrial cancer study

For Peer Review

An integrated clinical-MR radiomics model to estimate survival time in patients with endometrial cancer

Abstract

Background: Determination of survival time in women with endometrial cancer using clinical features remains imprecise. Features from MRI may improve the survival estimation allowing improved treatment planning.

Purpose: To identify clinical features and imaging signatures on T2-weighted MRI that can be used in an integrated model to estimate survival time for endometrial cancer subjects.

Study Type: Retrospective.

Population: 413 patients with endometrial cancer as training (n=330, 66.41 ±11.42 years) and validation (n=83, 67.60±11.89 years) data and an independent set of 82 subjects as testing data (63.26±12.38 years).

Field Strength/Sequence: 1.5-T and 3-T scanners with sagittal T2-weighted spin echo sequence.

Assessment: Tumor regions were manually segmented on T2-weighted images. Features were extracted from segmented masks, and clinical variables including age, cancer histologic grade and stage were included in a Cox proportional hazards (CPH) model. A group least absolute shrinkage and selection operator (gLASSO) method was implemented to determine the model from the training and validation datasets.

Statistical Tests: A likelihood-ratio test and decision curve analysis (DCA) were applied to compare the models. Concordance index (CI) and area under the receiver operating characteristic curves (AUCs) of were calculated to assess the model.

For Peer Review

Results: Three radiomic features (two image intensity and volume features) and two clinical variables (age and cancer grade) were selected as predictors in the integrated model. The CI was 0.797 for the clinical model (includes clinical variables only) and 0.818 for the integrated model using training and validation datasets, the associated mean AUC value was 0.805 and 0.853. Using the testing dataset, the CI was 0.792 and 0.882, significantly different and the mean AUC was 0.624 and 0.727 for the clinical model and integrated model, respectively.

Data Conclusion: The proposed CPH model with radiomic signatures may serve as a tool to improve estimated survival time in women with endometrial cancer.

Keywords: Endometrial cancer, survival analysis, radiomics, feature selection, Cox proportional hazards model, T2-weighted MRI

For Peer Review

INTRODUCTION

Endometrial cancer is the most common gynecological cancer, with 417,000 new cases diagnosed globally in 2020 (1,2). The 5-year overall survival rate of endometrial cancer patients ranges from 74% to 91% (3-5). To study the survival time of endometrial cancer patients, survival analysis methods have been extensively applied (4-6). Currently utilized examples include the non-parametric Kaplan-Meier method and the semi-parametric Cox's proportional hazards (CPH) method (4,5,7-11). Based on the CPH model, it is possible to evaluate a single covariate's and/or the combination joint covariates effects on the survival time estimation. For instance, a combination of age, cancer histologic grade, socioeconomic factors, and other clinical prognostic factors have been investigated in endometrial cancer survival studies within the framework of the CPH model (12-15).

For Peer Review

Until the advent of radiomics, image biomarkers have been insufficiently studied as potential survival predictors for endometrial cancer. Radiomics is a rapidly expanding field of research in oncology (16,17). There have been studies evaluating the application of radiomics, usually based on multi-sequence MRI features, for example Kurtosis from contrast-enhanced T1-weighted MRI to predict survival time in endometrial cancer, but these studies had several limitations (17-19). Specifically, these studies employed a dataset with less than 200 cases, thus with a small number of sample sizes and radiomic features (less than 100 features), which could lead to a large bias for the model estimation (21,22). Moreover, as these studies did not conduct a validation using independent external testing data, there was no model validation for survival time prediction (17-19). Finally, these studies did not include or combine clinical prognostic variables in the CPH model for the survival time prediction (17-19). As a result, these previous studies have most likely not evaluated the true potential of radiomic features for survival time prediction in endometrial cancer (17-19).

To overcome these limitations, this retrospective study aimed to identify a radiomic signature using pelvic MRI data that could estimate survival time in endometrial cancer. Furthermore, we sought to develop and validate an integrated clinical-radiomic model that might be used to tailor adjuvant management for women based on their personalized risk features.

For Peer Review

MATERIALS AND METHODS

This retrospective study protocol was approved by the Institutional Review Board (IRB), and the Research Ethics Committee reference number for this study is 17/LO/0173. The requirement for written informed consent was waived due to the retrospective design of this study. This retrospective study will develop and test a model which will be further validated as part of a larger prospective study (ClinicalTrials.gov NCT03543215, https://clinicaltrials.gov/).

Training and Validation Datasets

For Peer Review

Images were acquired between Feb 2007 and Aug 2017 (Figure 1), and 270 of the initially considered 611 subjects were obtained from a previous study (22). The training and validation datasets were obtained from 15 UK hospitals and centers with different parameters and protocols (Table 1). Table 1 shows the scan parameters for collecting 411 subjects of training/validation dataset which excluded two subjects because the scan parameters information was not available. The sagittal T2-weighted image was chosen for radiomic analysis as this was part of the standard protocol from all referral centers whereas availability of other sequences was more variable. As T2-weighted images were included from different centers in the study, image pre-processing and image normalization was required to minimize the difference between different scanners and sequences.

Clinical data, including the patient age at diagnosis, date of surgery, type and grade of tumor, the international federation of obstetricians and gynecologists (FIGO) stage, presence of lymphovascular space invasion, and any adjuvant or neoadjuvant treatment of these subjects were obtained from an online medical records system (23). Survival time was defined as the time from the date of surgery until the date of death, with final censor date on August 3rd, 2020.

The inclusion criteria regarding MRI were as follows: 1) no severe motion artifacts in T2-weighted images that obscured the tumour mass, as determined by radiologists subjectively, 2) sufficient size of the tumour on images (i.e., the tumour could be identified on more than one MRI slice before image resampling), and 3) the T2-weighted sequence passed the image pre-processing steps (see step two in Figure 2). The inclusion criteria regarding clinical data were: 1) availability of censoring or noncensoring survival information, information on lymphovascular space invasion, histological risk, and histological type, 2) availability of age at diagnosis and surgery date, 3) no other type of co-existing cancer. After exclusion of patients based on image and clinical criteria, 413 cases were used in this study (Table 2). The ratio for splitting the training and validation was 80:20 (n=330 for the training data; n=83 for the validation data) with balance the survival object (i.e., the combination of time and death information) distributions within the splits.

For Peer Review

Testing Dataset

Overall, 82 additional patients from three hospitals in the UK with endometrial cancer were included in the testing dataset, the scans being acquired between May 2017 and July 2019 (Table 2). For the testing dataset, the beginning time was the surgery date also, of which the earliest was in May 2017, and the ending time was in July 2019, and the close of the study was on December 1st 2021. 74 of the 82 cases were right censoring; at the close of the study on December 1st, 2021, 8 patients had died and 74 patients had survived. The right censored survival times underestimate the true (but unknown) time to event/death (6). The distribution of the training and testing datasets are displayed in Figure 3.

Radiomics Study Pipeline

Figure 2 shows the radiomics study pipeline for the survival analysis. There were five steps in this pipeline. The first and second steps were designed to analyze images, including manual image segmentation (prior to image re-sampling), MRI non-uniformity correction, image resampling, and image normalization. Specifically, Digital Imaging and Communications in Medicine (DICOM) file formats were downloaded from the picture archiving and communication systems (PACS), de-identified and converted to the simpler Neuroimaging Informatics Technology Initiative (NIFTI) format.

An interactive tool (ITK-snap, version 3.6.0, http://www.itksnap.org) for semi-automatic segmentation of sagittal orientation T2-weighted MRI was employed for manual slice-by-slice tumor segmentation by two radiologists in-training (JR, 5 years, with assistance from AS, 3 years) (24). After loading the T2-weighted MRI, the paintbrush tool was used to label all voxels containing visible tumor on each sagittal slice. Once all slices containing tumor had been labelled, the segmentation mask was saved as NIFTI format for pre-processing steps. This process was repeated for T2-weighted MRI in every image set. This was then checked by two radiology consultants (AR, 19 years' experience and NB 15 years' experience), who corrected the segmented tumor masks, without further went through all cases together again. The radiologists were blinded to the outcome measures. One example of the image segmentation is displayed in step 1 of Figure 2.

The T2-weighted images were pre-processed according to step two as shown in Figure 2. First, all image voxel sizes were obtained from NIFTI files with T2-weighted MRI header

For Peer Review

files, and the median voxel size of all data was calculated. The image reconstruction matrix size in sagittal orientation was between 256 and 864 (Table 1). The median resolution (image voxel size) of all T2-weighted images (including both training/validation and testing datasets) was 0.625 mm x 0.625 mm x 5 mm. Then, T2-weighted images were processed using an N4 toolbox for MRI non-uniformity bias correction, and to remove artifacts due to the inhomogeneity of magnetic fields (https://github.com/ANTsX/ANTs/wiki/N4BiasFieldCorrection) (25). Following bias correction, T2-weighted MRI and its masks were resampled to median voxel resolution (Figure 2). For T2-weighted MRI resampling, the cubic spline interpolation method was adopted. For segmented tumor masks (binary image), a nearest neighbor interpolation method was used for image resampling. Next, the intensity of resampled T2-weighted images was normalized using the following equation:

$$I = \frac{I - \bar{I}}{std(I)} 100 \qquad \text{(Eq. 1.)}$$

where I is image intensity, $\bar{I}$ is the mean value of the image intensity within the volume, and *std* is the standard deviation of the image volume. Finally, the TexLAB tool (version 2.0) on MATLAB (version R2019a; The MathWorks Inc., Natick, MA, USA; http://www.mathworks.com/), PyRadiomics (version 3.0.1, https://github.com/AIM-Harvard/pyradiomics), and Scikit-image (version 0.19.2, https://scikit-image.org/), both implemented in Python (Python Software Foundation, version, Python3.8, https://www.python.org/) were used to perform feature extraction as shown in Figure 2 (26,27). After elimination of identical features by a correlation method, in total 958 radiomics features were extracted from T2-weighted MRI and its associated segmentation masks. T2-weighted MRI was included because image intensity-based

For Peer Review

features were derived from T2-weighted MRI images. Endometrial cancer tumor region was the only region of interest in this study.

Feature Selection

The fourth step was to select features for survival analysis. Radiomic and clinical feature selections were performed within the framework of statistical model selection, and the CPH model was used to study the relationship between predictor variables and survival time. In the CPH model, the time and event/death were treated as dependent variables (survival object); 958 radiomics features, cancer risk score (which includes FIGO stage), cancer grade, and age were included as predictors (independent variables) for model selection (9). Cancer risk score and grade were defined according to FIGO (23,28). Before applying the model selection method, all 959 features (958 MRI features + age) were normalized using a Z-score method (similar to Eq. 1, except multiply 100). To avoid model overfitting, a 10-fold cross validation for penalized Cox regression models with grouped covariates was adopted to determine the optimal regularization parameter lambda (λ). Specifically, a group exponential least absolute shrinkage and selection operator (gLASSO) was used to select statistical models (29). The maximum iteration of the 10-fold cross validation was set to be 1 million times in the model fitting. The final selected CPH model was then applied to calculate the survival time.

For Peer Review

Statistical Analysis

The R software (version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org) was used for statistical analysis. Model selection package "grpgrep" (version 3.4.0, https://cran.r-project.org/web/packages/grpreg/index.html) was applied to determine the optimal CPH model. The criteria for the optimal model were model simplicity and accuracy (i.e., minimize the combination of the L1 and L2 norm) (29,35). The "Survival" package (version 3.4.0, https://cran.r-project.org/web/packages/survival/index.html) was used to implement CPH model. A bootstrap resampling method was developed to assess the predictive performance of the CPH model using a Score() function from a "riskRegression" R library (version 2021.10.10, https://cran.r-project.org/web/packages/riskRegression/index.html). Nomograms were generated using a "regplot" R package (version 1.1, https://cran.r-project.org/web/packages/regplot/index.html).

Survival analysis was implemented based on the selected integrated model as shown in step five of Figure 2. The gLASSO method produced model selection results with randomness. The most common output by the gLASSO method was adopted. Once the survival time prediction according to the CPH model was established with the gLASSO method, a nomogram was created as a graphical representation of the integrated model. The nomogram was applied to visualize the prediction survival probability. To study the influence of the radiomic features on the survival probability, two models were constructed and compared for the estimation. The first model was based on clinical information only; the predictors of the model included only age and cancer grade. The risk score was not included in the final model as it had not been selected by the gLASSO method, which may because the cancer grade and risk score are correlated. The second model used both clinical information (age and cancer grade) and three radiomic features selected by the gLASSO method.

Additional analyses were performed to validate the model based on the prediction using the "riskRegression" library. The time-dependent area under the receiver operating

For Peer Review

characteristic (ROC) curve (AUC) was calculated from the validation (AUC is specified for AUC of ROC in this study). For the model validation, 80% of the 413 cases were used to generate the CPH model, while the rest of the datasets were employed to validate the predictive performance. Using a stratified sampling method, the training and validation datasets were split with survival objects (time/death). To validate the CPH model, the bootstrap resampling method with a sample size of 10 at each time point was adopted. Because the bootstrap method and stratified sampling method have randomness, and as a typical example, the AUC was calculated and displayed because AUC is widely used criterion for measure discrimination. To reduce the effect of the randomness in the evaluation study by using bootstrap method, the concordance index (CI), which measures the prediction accuracy, was calculated with 10 repetitions (with different training validation datasets splits). The time point started at 100 days and terminated at 1,825 days with a 5-day interval. Similarly, additional external testing cases were used to test the CPH model prediction performance.

To study the effect of the radiomic features and clinical variables on survival time estimation, decision curve analysis (DCA) was applied to evaluate the clinical, radiomic, and integrated models for net benefit (30). Net benefit is calculated for each possible threshold probability which puts benefits and harms on the same scale. Threshold probability is the expected benefit of treatment is equal to the expected benefit of avoiding treatment (30). By varying the threshold probability, DCA allows us to examine whether one model is superior to another at a certain range of threshold probability with respect to the net benefit.

For Peer Review

A likelihood-ratio test method was applied to study the importance of the radiomic and clinical features for survival time prediction. Furthermore, training and test datasets were compared using Chi-squared tests for categorical data and two-sample t-tests for continuous data (Table 2). A P-value <0.05 was considered statistically significant.

A diagnostic analysis was carried out to study the feature variation obtained from different types of scanners using 413 training/validation cases. It is not obvious to inspect the features difference from two dimensions, for example, in an image with 413 rows and 985 feature columns. Therefore, dimension reduction method was applied to obtain the major components of the features from each type of scanner. Specifically, principal component analysis (PCA) was applied to study the effect of feature difference from different scanners. All features were normalized using the Z-score method, and then a PCA was employed to split the feature dataset into different components. Four principal components were used to compare the feature variations from different scanners. Three different manufacturers GE: 108 cases, Philips: 163 cases, and Siemens: 142 cases were used to acquire sagittal T2-weighted MRI (Table 1). Feature matrix from these three different scanners were decomposed into four components. Visual comparison was carried out to evaluate the distribution of the feature components from different scanners.

For Peer Review

RESULTS

Training and Testing Dataset Demographics

Clinical-pathological characteristics of the patients are shown in Table 2. In addition, Figure 3 plots the histograms of the testing dataset and a two-sample t-test that was applied to compare the training (including validation) and external testing datasets. Except for the survival time (Figure 3B and Figure 3E), all comparisons between training and testing datasets were significant. For the survival time (Figure 3B and Figure 3E), no significant differences were revealed (training dataset: 870.6±592.1 days, testing dataset: 637.1±314.2, p=0.09). Table 2 also includes the demographic information from the testing dataset. The age at diagnosis of the testing dataset was significantly different from the training dataset (66.64±11.51 years versus 63.26±12.38 years, Table 2).

Feature Selection Results

Figure 4 shows the gLASSO coefficient profiles selected from 961 features. Specifically, Figure 4A plots the 10-fold cross-validated error rates, and Figure 4B shows the amplified version of the gLASSO selection plot. Five features were selected from 961 predictors and were included in the integrated CPH model. They were tumor mask minor axis radius (minorAxisRad), grey level size zone matrix (GLSZM), first order statistics (FOS), patient age at diagnosis (Age), and cancer grade (Grade). Tumor minor axis radius reflects the size of the tumor indirectly; the FOS here is the coefficient of variation, which is defined as the ratio of the standard deviation to the mean, and these values were computed within the tumor mask. This was computed after the normalized T2-weighted image were filtered with low, low, and high wavelet filters in x, y, and z direction of the 3D image subsequently. The GLSZM was calculated after the normalized MRI image was converted into 25

For Peer Review

Hounsfield unit grey level, then the large zone low gray level emphasis was computed within the tumor mask. The FOS and GLSZM represent image statistical property and intensity character. These five selected features were refit into a CPH model without the normalization of the age covariate, for the purpose of displaying in the nomogram. The survival prediction was then estimated based on the refit CPH model. The final integrated model was:

$$\mathrm{Surv}(\mathrm{Time}, \mathrm{Death}) = 0.0548{*}\mathrm{Age} + 0.0025{*}\mathrm{Grade2} + 1.684{*}\mathrm{Grade3} + 0.495{*}\mathrm{minorAxisRad} - 0.263{*}\mathrm{GLSZM} - 0.179{*}\mathrm{FOS}.$$

The corresponding clinical model (excluding radiomic features) was:

$$\mathrm{Surv}(\mathrm{Time}, \mathrm{Death}) = 0.0455{*}\mathrm{Age} + 1.881{*}\mathrm{Grade2} + 2.107{*}\mathrm{Grade3},$$

where Surv is the survival object, defined as a response variable in the CPH model and age was not normalized. Time is the time (number of days) from the date of surgery to the end of the study for the right censoring data. If the subject has died before the end of the study, Time is the number of days from the surgery date to the death date. Death is a status binary variable, with 1 to represent death of the subject, and 0 to denote the survival of the subject at the close of the study. Grade2 and Grade3 are the numerical variables to represent cancer grade 2 and grade 3 which were converted from cancer grade categorical variable. minorAxisRad is the tumor minor axis radius which was calculated from the tumor mask image.

Model Training and Validation

Figure 5A plots the time-dependent AUC based on training and validation datasets. For the clinical model, the AUC accuracy was below 80% for the time points after 1,250 days, suggesting that this model is less accurate for long-time estimation. In Figure 5A, the

For Peer Review

integrated model had a larger AUC than the clinical model for all time points, suggesting that the integrated clinical-radiomic model is superior to the clinical model for the prediction based on the external testing dataset for all time points (integrated model AUC: 0.853±0.06, clinical model AUC 0.805±0.058). The results also showed that the CI value was significantly higher using the integrated model based on these 413 cases (integrated model CI: 0.825±0.010, clinical model CI: 0.806±0.011).

Similarly, AUC curves were computed using the trained model on the testing dataset and the results are presented in Figure 5B. Comparing Figure 5A with Figure 5B, the AUC obtained from the testing dataset is smaller than the AUC computed from the training dataset (integrated model AUC: 0.727±0.085, clinical model AUC 0.624±0.070). This is because the survival time and age from the testing data were significantly different from the training and validation datasets (training and validation data: 1583.4 ±669.6 days, testing data: 1318.7±306.4 days). A likelihood-ratio test showed a significant difference between the integrated model and clinical model based on both training and testing datasets. The CI was 0.797 for the clinical model and 0.818 for the integrated model. Based on the selected model from training data, the nomogram display the 1, 2, 3, and 5-year survival probabilities is shown in Figure 6.

The difference between the clinical model and the integrated model is small in terms of CI using the training and validation datasets, however, for the independent testing data, the CI was 0.792 for the model with age and clinical cancer grades, and the index was 0.882 for the integrated model, thus showing a significant difference (likelihood-ratio x^2=12.677). This suggests that the integrated model is robust to the different distributions

For Peer Review

of the data because age (Table 2) is statistically significant difference between the internal (training/validation) data and external dataset.

Decision Curve Analysis

To study the contribution of radiomic features to survival time estimation within the CHP model, a DCA was applied to compare radiomic (includes 3 radiomic features only), clinical, and integrated models at 500, 1,000, 1,500, and 2,000 days (Figure 7). The integrated model was almost consistently on the top of other curves in Figure 7, suggesting that the model has more net benefit than the other models for survival time prediction. The radiomic model had a larger net benefit than the clinical model when the threshold was below 0.5 (Figures 7B and 7C). For a larger threshold probability (>0.45), radiomic, clinical, and integrated models had similar net benefit for a short time range estimation (Figures 7A and 7B). However, for the long-time range (2,000 days or more) survival time estimation, the integrated model had a larger net benefit; comparing Figure 7A to Figure 7D, the gap between the curves is larger in Figure 7D, suggesting a larger net benefit for estimations longer than 2,000 days.

For Peer Review

Features From Different Scanners

From PCA analysis, features components from different scanners were overlaid onto each other in Figure 8. Similarities in the feature distribution was observed, although as shown in Figure 8A, radiomic features from the Siemens scanners had a larger variation. For the 3rd and 4th principal components (PC3/4) (Figure 8B), the distribution of the radiomic features obtained from different scanners were smaller, suggesting good agreement for the features from different scanners.

DISCUSSION

We have developed the CPH model using features from sagittal T2-weighted MRI and clinical variables for survival time estimation based on gLASSO method. We studied the effect of the radiomic features within the model and found radiomic features from MRI are useful biomarkers to predict survival time in patients with endometrial cancer.

We identified a set of radiomic signatures using pelvic MRI that could potentially aid in accurately estimating survival time for patients with endometrial cancer. In combination with clinical features, our integrated radiomics model outperformed the clinical model in predicting survival time. We validated and compared the integrated and clinical models using both internal (training and validation) and independent external (testing) datasets. Furthermore, the CPH model with a nomogram for visualization provided a graphical, straightforward, and non-invasive method of predicted survival, which could be used in clinical settings and therefore has potential to facilitate personalized medicine. The multiple centers and scan machines used in this study presented challenges for model building, but this setup also implies that the findings are likely to be generalizable. Multiple modelling techniques were evaluated, and feature selection was utilized to avoid overfitting of the model. The radiomics quality score which determines the validity and completeness of radiomics studies, and transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines were adhered to ensure quality of both scientific methods and reporting (31).

In contrast to quantitative MRI such as apparent diffusion coefficient (ADC) from diffusion-weighted imaging, T2-weighted MRI signal depends on many variable factors including the acquisition protocol, the coil profile, the scanner type, and therefore it is not the

For Peer Review

standard method to normalize the image intensity for cross-subjects comparison. We adopted Z-score-like method to normalize the image intensity, other methods such as min/max normalization or scaling the image intensity to common max value can also be used. An alternative method to reduce the image intensity difference of T2-weighted images acquired from different centers and scanners is to normalize to a reference tissue outside the tumor-affected region such as cerebrospinal fluid (CSF) in brain or bladder where baseline water signal can be obtained. Although the image intensity features such as mean intensity value within the tumor mask will be affected by different image normalization steps, the tumor shape, volume, and image complexity radiomic features will not be affected by the image normalization step.

Most of the published studies have focused on addressing the classification problem in endometrial cancer using radiomics (17-19). Studies have applied this radiomic technology to endometrial cancer survival prediction models (32,33). Fasmer et al. developed an MRI-based whole-volume tumor radiomic signature for the prediction of high-risk features (19). Radiomic features were studied to predict poor progression-free survival (19). Meanwhile, Ytre-Hauge et al. applied radiomics to study survival in women with endometrial cancer (17). They reported that high kurtosis in contrast-enhanced T1-weighted MRI predicted reduced recurrence and progression-free survival (hazard ratio 1.5), but their study used only 180 patients without model validation (17), compared to 495 cases in this study. Furthermore, they used contrast-enhanced T1-weighted MRI images (17). However, we obtained the features from T2-weighted image, which is a sequence that clearly delineates most endometrial cancers without the use of gadolinium and the sagittal T2-weighted sequence is the mainstay of MRI protocols for staging

For Peer Review

endometrial cancer, enabling the development of a generalizable tool. We found that tumor size reflected by minor axis radius was an important biomarker for survival time estimation. The minor axis radius describes the radius of the minor axis of the ellipse that reflects the tumor region indirectly (27). For the features of GLSZM and FOS, both features are related to the distribution of image intensities. The GLSZM was based on the image converted from Hounsfield unit, this could be due to the relationship between MRI intensity and Hounsfield unit values (34).

In addition to using CI, we adopted multiple criteria to evaluate the models. We have applied likelihood-ratio test, AUC, and DCA methods to compare different models. By considering the clinical utility of the specific model, DCA overcomes the limitations of traditional metrics such as AUC which only measures the diagnostic accuracy of the model.

The pipeline of this study (Figure 2) can be extended to other malignancies for survival analysis based on integrated features. In this method, the sources of error can come from the first 3 steps: image segmentation, image processing, and feature extraction. For example, in the image segmentation step, error can be generated if the tumor mask is not parcellated properly. For the image processing step, the image interpolation method can introduce numerical error. In the feature extraction step, bias can be produced if only a fraction of image features is extracted from the image.

Lastly, the clinical application of the nomograms could be in patient management or prioritization; as the survival time of the patient is known from the model estimation, so treatment for patients could be arranged in a more efficient way. The integrated radiomics

For Peer Review

model may also enable better stratification of patients enrolling into clinical trials, as it has higher AUC and CI value than the model with only clinical variables.

Limitations

This was a retrospective study and therefore, there was a risk of bias and missing data. The study also only included patients who had undergone surgery and had an MRI with paired clinical data available. Whilst the model was assessed based on the external testing dataset, there was slight variation in demographics when directly comparing the training and validation datasets. In the testing dataset, there were less women in the older 59-to-70-year group and more women with endometrioid low grade cases, namely more patients with low-risk scores. Debatably, this would infer that the testing dataset group would be more likely to have better survival. As radiomics models perform better with more homogenous datasets such as that generated by the low-risk cases, this may explain the slightly better performance with the testing dataset. Secondly, although we had also tested the composite minimax concave penalty method for the model selection in the CPH model (which produced the same model selection results as the group exponential LASSO method), other methods such as the regular elastic net and ridge models method, which may produce better results, have not been investigated in this study (35,36). Thirdly, survival outcomes do not only represent the effect of the disease itself, but also of patient factors (such as age or co-morbidities) and treatment factors (such as whether the patient underwent neoadjuvant radiotherapy or chemotherapy). Study has shown that adjuvant treatment predictably improves survival for high-risk patients (37). Regional and national differences in patient demographics along with treatment options offered and delivered can also impact survival disparities. This study

For Peer Review

did not consider co-morbidities, which are likely to have a relevant impact on survival. Finally, although we normalized the images to minimize the T2-weighted image differences obtained from different protocols, more work is needed to study effects on features for radiomics studies.

This study employed only T2-weighted MRI data; future work could evaluate the use of additional MRI sequences or quantitative MRI such as diffusion–weighted images with ADC maps, as well as dynamic contrast-enhanced MRI (DCE-MRI) (38). Also, a possible method to explore in the future would be the boosting method or the deep survival model for the study of survival because these methods do not require the proportional hazards assumption (39,40).

For Peer Review

Conclusion

The integrated radiomic model and the nomogram may enable us to estimate of survival with a high degree of accuracy. Furthermore, we found that the integrated model is robust; it retained a high level of accuracy despite the variability of the independent external testing dataset, as the AUC value showed only a marginal decrease when applied to the testing dataset, in comparison to the clinical model, in which the AUC decreased markedly.

For Peer Review

References

1. International Agency for Research on Cancer G. Corpus uteri factsheet, estimated cancer incidence, mortality and prevalence worldwide 2018. World Health Organisation; 2018.

2. Crosbie EJ, Kitson SJ, McAlpine JN, Mukhopadhyay A, Powell ME, Singh N. Endometrial cancer. Lancet 2022;399(10333):1412-1428.

3. Morice P, Leary A, Creutzberg C, Abu-Rustum N, Darai E. Endometrial cancer. The Lancet 2016;387(10023):1094-1108.

4. Coronado PJ, Rychlik A, Baquedano L, et al. Survival Analysis in Endometrial Carcinomas by Type of Surgical Approach: A Matched-Pair Study. Cancers (Basel) 2022;14(4).

5. Odagiri T, Watari H, Hosaka M, et al. Multivariate survival analysis of the patients with recurrent endometrial cancer. jgo 2011;22(1):3-8.

6. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. British journal of cancer 2003;89(2):232-238.

7. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American statistical association 1958;53(282):457-481.

8. Hoivik EA, Hodneland E, Dybvik JA, et al. A radiogenomics application for prognostic profiling of endometrial cancer. Communications Biology 2021;4(1):1363.

9. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological) 1972;34(2):187-220.

For Peer Review

10. de Boer SM, Powell ME, Mileshkin L, et al. Adjuvant chemoradiotherapy versus radiotherapy alone for women with high-risk endometrial cancer (PORTEC-3): final results of an international, open-label, multicentre, randomised, phase 3 trial. The Lancet Oncology 2018;19(3):295-309.

11. Uharcek P. Prognostic factors in endometrial carcinoma. J Obstet Gynaecol Res 2008;34(5):776-783.

12. Njoku K, Barr CE, Crosbie EJ. Current and Emerging Prognostic Biomarkers in Endometrial Cancer. Frontiers in oncology 2022;12.

13. Madison T, Schottenfeld D, James SA, Schwartz AG, Gruber SB. Endometrial cancer: socioeconomic status and racial/ethnic differences in stage at diagnosis, treatment, and survival. Am J Public Health 2004;94(12):2104-2111.

14. Morielli AR, Kokts-Porietis RL, Benham JL, et al. Associations of insulin resistance and inflammatory biomarkers with endometrial cancer survival: The Alberta endometrial cancer cohort study. Cancer Medicine 2022;11(7):1701-1711.

15. Caruso D, Polici M, Zerunian M, et al. Radiomics in Oncology, Part 2: Thoracic, Genito-Urinary, Breast, Neurological, Hematologic and Musculoskeletal Applications. Cancers 2021;13(11):2681.

16. Michalet M, Azria D, Tardieu M, Tibermacine H, Nougaret S. Radiomics in radiation oncology for gynecological malignancies: a review of literature. Br J Radiol 2021;94(1125):20210032.

For Peer Review

17. Ytre-Hauge S, Dybvik JA, Lundervold A, et al. Preoperative tumor texture analysis on MRI predicts high-risk disease and reduced survival in endometrial cancer. Journal of magnetic resonance imaging : JMRI 2018;48(6):1637-1647.

18. Jacob H, Dybvik JA, Ytre-Hauge S, et al. An MRI-Based Radiomic Prognostic Index Predicts Poor Outcome and Specific Genetic Alterations in Endometrial Cancer. Journal of clinical medicine 2021;10(3).

19. Fasmer KE, Hodneland E, Dybvik JA, et al. Whole-Volume Tumor MRI Radiomics for Prognostic Modeling in Endometrial Cancer. Journal of magnetic resonance imaging : JMRI 2021;53(3):928-937.

20. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLOS ONE 2019;14(11):e0224365.

21. Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage 2018;180:68-77.

22. Soneji ND, Bharwani N, Ferri A, Stewart V, Rockall A. Pre-operative MRI staging of endometrial cancer in a multicentre cancer network: can we match single centre study results? European radiology 2018;28(11):4725-4734.

23. FIGO Committee on Gynecologic Oncology. FIGO staging for carcinoma of the vulva, cervix, and corpus uteri. Int J Gynaecol Obstet 2014;125(2):97-8.

24. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 2006;31(3):1116-1128.

25. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310-1320.

For Peer Review

26. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research 2017;77(21):e104-e107.

27. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. PeerJ (San Francisco, CA) 2014;2:e453-e453.

28. Kasius JC, Pijnenborg JMA, Lindemann K, et al. Risk Stratification of Endometrial Cancer Patients: FIGO Stage, Biomarkers and Molecular Classification. Cancers 2021;13(22).

29. Breheny P. The group exponential lasso for bi-level variable selection. Biometrics 2015;71(3):731-740.

30. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research 2019;3(1):18.

31. Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. European radiology 2020;30(1):523-536.

32. Zhang X, Zhao J, Zhang Q, et al. MRI-based radiomics value for predicting the survival of patients with locally advanced cervical squamous cell cancer treated with concurrent chemoradiotherapy. Cancer Imaging 2022;22(1):35.

33. Liu D, Yang L, Du D, et al. Multi-Parameter MR Radiomics Based Model to Predict 5-Year Progression-Free Survival in Endometrial Cancer. Frontiers in oncology 2022;12.

For Peer Review

34. Kapanen M, Tenhunen M. T1/T2*-weighted MRI provides clinically relevant pseudo-CT density data for the pelvic bones in MRI-only based radiotherapy treatment planning. Acta Oncologica 2013;52(3):612-618.

35. Breheny P, Huang J. Penalized methods for bi-level variable selection. Statistics and its interface 2009;2(3):369.

36. Van De Wiel MA, Lien TG, Verlaat W, van Wieringen WN, Wilting SM. Better prediction by use of co-data: adaptive group-regularized ridge regression. Statistics in Medicine 2016;35(3):368-381.

37. Son J, Chambers LM, Carr C, et al. Adjuvant treatment improves overall survival in women with high-intermediate risk early-stage endometrial cancer with lymphovascular space invasion. International Journal of Gynecologic Cancer 2020;30(11):1738-1747.

38. Ueno Y, Forghani B, Forghani R, et al. Endometrial Carcinoma: MR Imaging-based Texture Model for Preoperative Risk Stratification-A Preliminary Analysis. Radiology 2017;284(3):748-757.

39. De Bin R. Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. Computational Statistics 2016;31(2):513-531.

40. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology 2018;18(1):24.

For Peer Review

TABLES

TABLE 1. Scan information for training/validation datasets

MRI manufacturer (413 cases)	**GE**	**Philips**	**Siemens**
	108	163	142
Scanner parameters (411 cases)	**Mean**	**Median**	**Std**
Echo Time (ms)	101.81	100	16.64
Repetition Time (ms)	4653.02	4100	2100.78
Slice Thickness (mm)	4.31	4	0.66
Spacing Between Slices (mm)	4.95	5	0.55
Reconstruct Matrix Size	441.47	512	113.71
Magnetic Field Strength (T)	1.51	1.5	0.15

Std: standard deviation ‘ ms: millisecond; mm: millimeter; T: tesla . GE: General Electric Company, NY, USA.

review

TABLE 2. Training (including validation) and testing patient demographics

Clinical information	N(training)	%(training)	N(testing)	%(testing)	P value
Age: mean (SD)	**66.64±11.5**		**63.25±12.4**		**0.024**
under 50	29	7	11	13.4	0.79
50-59	78	18.9	24	29.3	0.55

60-69	133	32.2	15	18.3	0.18
70 and older	173	41.9	32	39.0	0.18
Histological type					**0.0011**
Endometrioid	304	73.6	69	84.1	
Carcinosarcoma	44	10.7	1	1.2	
Serous	39	9.4	4	4.9	
Clear cell	18	4.4	2	2.4	
Mixed high grade	7	1.7	2	1.7	
Undifferentiated	1	0.2	3	3.7	
NET small cell			1	1.2	
Grade					**3.72e-04**
1 (low grade)	124	30.0	43	52.4	
2 (intermediate grade)	130	31.5	20	24.4	
3 (high grade)	159	38.5	19	23.2	
Overall FIGO stage					**0.1738**
Stage I	292	70.7	59	72	
➢ IA	199	48.2	45	54.9	
➢ IB	93	22.5	14	17.1	
Stage II	31	7.5	5	6.1	
Stage III	64	15.5	7	8.5	

For Peer Review

➢ IIIA	18	4.4	4	4.9	
➢ IIIB	6	1.4	0	0	
➢ IIIC	40	9.7	3	3.7	
Stage IV	25	6.1	1	1.2	
➢ IVA	18	4.4	0	0	
➢ IVB	7	1.7	1	1.2	
Other (missing)	1	0.2	1	1.2	
Clinical risk score					**0.0388**
Low	150	36.3	41	50	
Intermediate	78	18.9	15	18.3	
High	96	23.2	9	11.0	
Advanced	89	21.5	16	19.5	
Unknown			1	1.2	
Censored					**0.0096**
Censoring	317	76.8	74	90.2	
Death	96	23.2	8	9.8	

Note. N = 413 (training/validation), *N* = 82 (testing).SD: standard deviation.NET: neuroendocrine tumor. FIGO: The International Federation of Gynecology and Obstetrics. Staging version

For Peer Review

Figure Legends

FIGURE 1: Flow chart of patient selection. After exclusion, 413 cases were included and used to generate the final model. Eighty-two cases were used as external testing dataset.

FIGURE 2: Pipeline for the study. Five steps were included as shown in the column. LR: likelihood-ratio test; gLASSO: group Least Absolute Shrinkage and Selection Operator; CPH: Cox proportional hazards model; DCA: decision curve analysis; AUC: area under the receiver operating characteristic (ROC) curve.

For Peer Review

FIGURE 3: Histograms of the survival data for all training/validation (A) and testing (D) datasets and histograms of the censoring and the noncensoring datasets. The noncensoring data (B) and the right censoring data (C) distributions from the training dataset. The noncensoring data (E) and the right censoring data (F) distributions from the testing dataset.

FIGURE 4: Group least absolute shrinkage and selection operator (LASSO) or feature selection. A: 10-fold cross-validated error rates for the model selection. B: amplified version of Figure 4A at the optimal lambda value. The vertical dotted lines indicate the minimum error, and the top of the plot gives the size of each model. Each red dot represents a λ value along the path. In the group LASSO method, the cross-validation method was applied to select the tuning parameter (λ). Dotted vertical lines were drawn at the optimal λ values by using the minimum criteria (i.e., cross-validation error). A

Lambda value of 0.067 (log(λ)=-2.7) according to the 10-fold cross-validation method was computed.

FIGURE 5: A: Time-dependent AUC summary at evaluation time points from the validation dataset; the AUC values are within the range of 0.5 and 0.9. The AUC for the integrated model (red curve) is consistently larger than the AUC obtained from using clinical model (green curve). B: AUC obtained from the testing data; AUC values are within the range between 0.5 and 0.85.

For Peer Review

FIGURE 6: Nomogram visualization for the survival time prediction. At the top of the nomogram, a point scale was included. Beneath the scale, 3 radiomic features, age, and the clinical cancer grade were displayed. The refitted CPH model was adopted to predict the survival probability for 1 (365 days), 2 (730 days), 3 (1095 days) and 5 (1826 days) year periods as shown at the bottom of the Figure 6. The dotted red vertical line in the figure indicates one example of observation with an age of 70, cancer grade of 3, minor axis radius of 0.24, GLSZM of 1.44, and FOS of -0.0018. The aggregate score for this case is 329 as indicated by the red arrow vertical line at the bottom of the figure. The corresponding probability to the survival for the 5-year, 3-year, 2-year, and 1-year periods is 0.657 (1-0.343), 0.806 (1-0.194), 0.87(1-0.13), and 0.944(1-0.056), respectively. GLSZM: grey level size zone. FOS: first order statistics. CPH: Cox proportional hazards.

FIGURE 7: Decision curve analysis at 500(A), 1,000(B), 1,500(C), and 2,000(D) days. The net benefit is plotted against the threshold probability. If the curve is closer to the

right top corner, then the corresponding model is better as it has larger net benefit. The “all” curve shows the net benefit by treating all patients, while the “none” curve denotes net benefit for treating no patients.

FIGURE 8: Scanner difference study. Principal component analysis for radiomics features from a different scanner. A: first principal component (PC1) vs second principal component (PC2); a relatively larger variation was observed using the Siemens scanner. B: Third principal component (PC3) and fourth principal component (PC4) explain smaller percentages of the total variation, and the three different scanners show good agreement. The dots in the figure represent samples; the colors represent groups (scanner types); and the legends have three groups at the top. The ellipse represents the core area added by the default confidence interval of 68%, which facilitates the separation between the observation groups. No clear separation of the sample based on the three MRI vendors was observed. var.: variance. PC: principal component.

For Peer Review

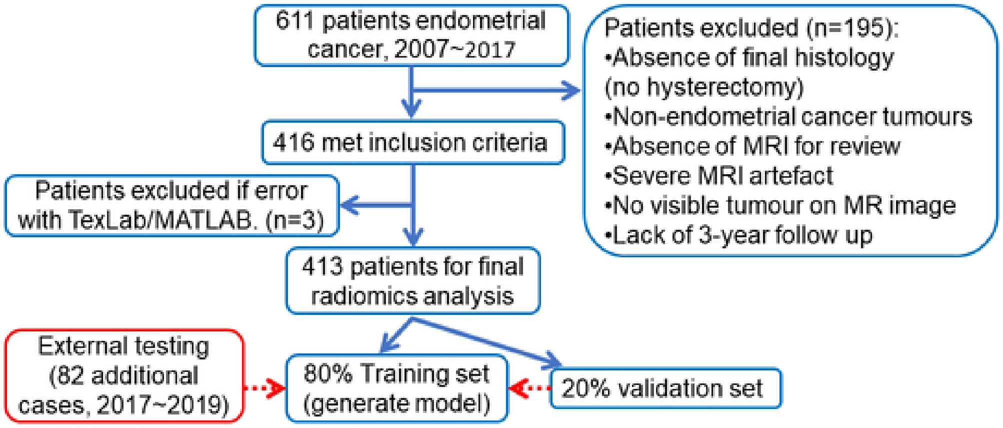

611 patients endometrial cancer, 2007~2017
Patients excluded (n=195):
•Absence of final histology (no hysterectomy)
•Non-endometrial cancer tumours
•Absence of MRI for review
•Severe MRI artefact
•No visible tumour on MR image
•Lack of 3-year follow up
416 met inclusion criteria
Patients excluded if error with TexLab/MATLAB. (n=3)
413 patients for final radiomics analysis
External testing (82 additional cases, 2017~2019)
80% Training set (generate model)
20% validation set

FIGURE 1: Flow chart of patient selection. After exclusion, 413 cases were included and used to generate the final model. Eighty-two cases were used as external testing dataset.

45x19mm (300 x 300 DPI)

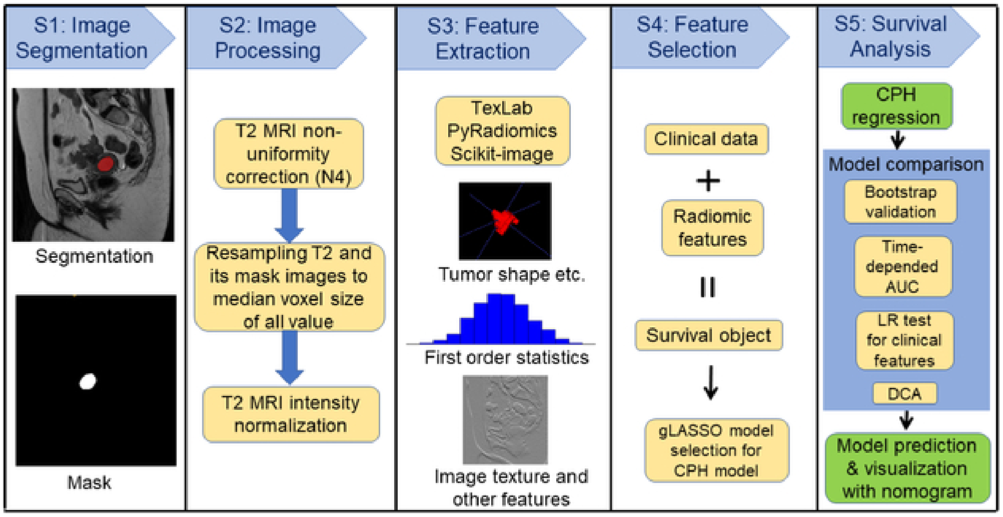


FIGURE 2: Pipeline for the study. Five steps were included as shown in the column. LR: likelihood-ratio test; gLASSO: group Least Absolute Shrinkage and Selection Operator; CPH: Cox proportional hazards model; DCA: decision curve analysis; AUC: area under the receiver operating characteristic (ROC) curve.

53x27mm (300 x 300 DPI)

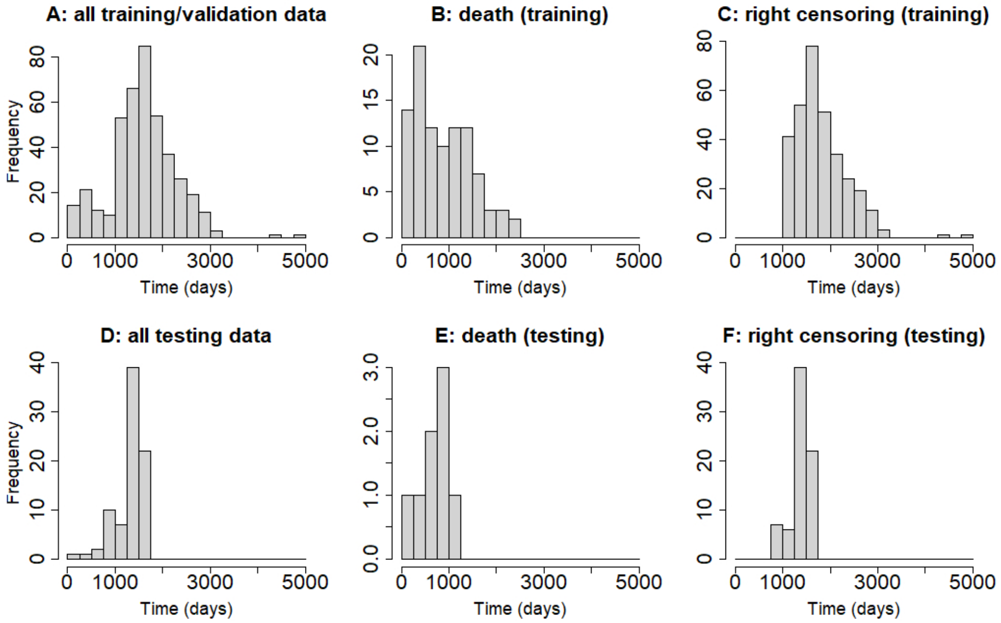


FIGURE 3: Histograms of the survival data for all training/validation (A) and testing (D) datasets and histograms of the censoring and the noncensoring datasets. The noncensoring data (B) and the right censoring data (C) distributions from the training dataset. The noncensoring data (E) and the right censoring data (F) distributions from the testing dataset.

86x54mm (300 x 300 DPI)

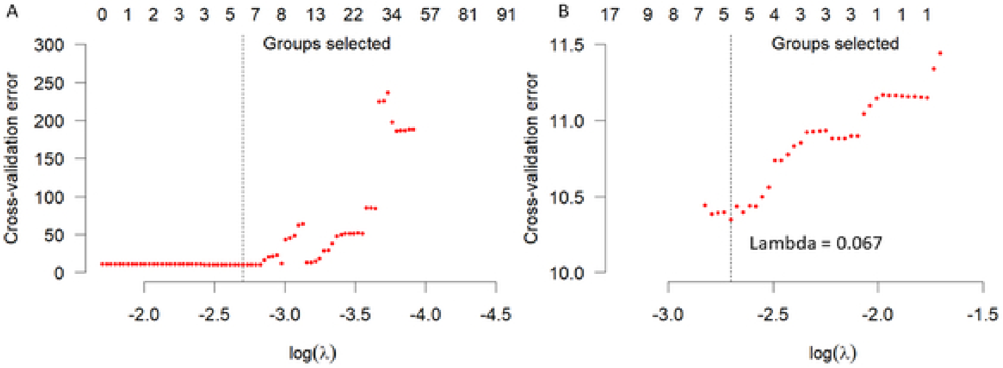


FIGURE 4: Group least absolute shrinkage and selection operator (LASSO) or feature selection. A: 10-fold cross-validated error rates for the model selection. B: amplified version of Figure 4A at the optimal lambda value. The vertical dotted lines indicate the minimum error, and the top of the plot gives the size of each model. Each red dot represents a λ value along the path. In the group LASSO method, the cross-validation method was applied to select the tuning parameter (λ). Dotted vertical lines were drawn at the optimal λ values by using the minimum criteria (i.e., cross-validation error). A Lambda value of 0.067 (log(λ)=-2.7) according to the 10-fold cross-validation method was computed.

49x18mm (300 x 300 DPI)

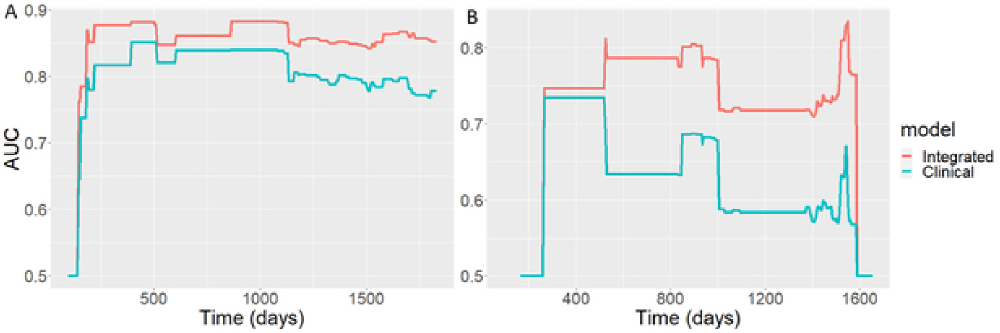


FIGURE 5: A: Time-dependent AUC summary at evaluation time points from the validation dataset; the AUC values are within the range of 0.5 and 0.9. The AUC for the integrated model (red curve) is consistently larger than the AUC obtained from using clinical model (green curve). B: AUC obtained from the testing data; AUC values are within the range between 0.5 and 0.85.

52x17mm (300 x 300 DPI)

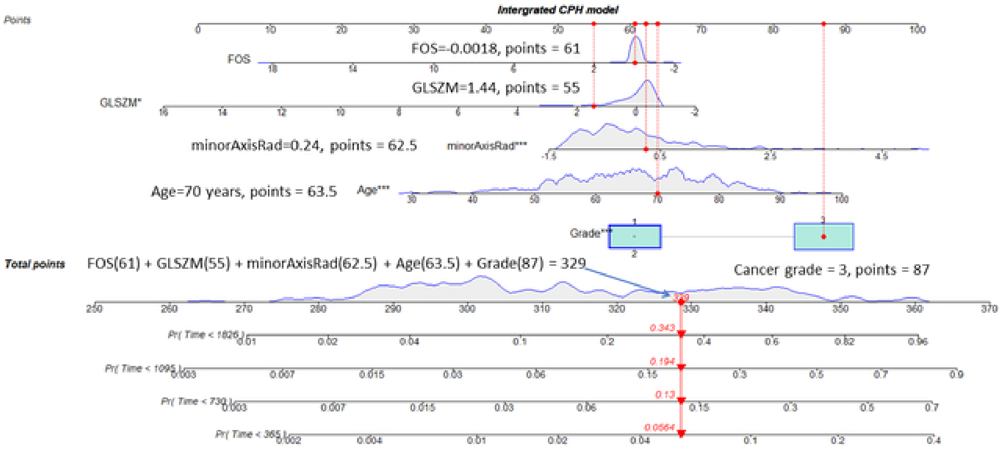


FIGURE 6: Nomogram visualization for the survival time prediction. At the top of the nomogram, a point scale was included. Beneath the scale, 3 radiomic features, age, and the clinical cancer grade were displayed. The refitted CPH model was adopted to predict the survival probability for 1 (365 days), 2 (730 days), 3 (1095 days) and 5 (1826 days) year periods as shown at the bottom of the Figure 6. The dotted red vertical line in the figure indicates one example of observation with an age of 70, cancer grade of 3, minor axis radius of 0.24, GLSZM of 1.44, and FOS of -0.0018. The aggregate score for this case is 329 as indicated by the red arrow vertical line at the bottom of the figure. The corresponding probability to the survival for the 5-year, 3-year, 2-year, and 1-year periods is 0.657 (1-0.343), 0.806 (1-0.194), 0.87(1-0.13), and 0.944(1-0.056), respectively. GLSZM: grey level size zone. FOS: first order statistics. CPH: Cox proportional hazards.

54x24mm (300 x 300 DPI)

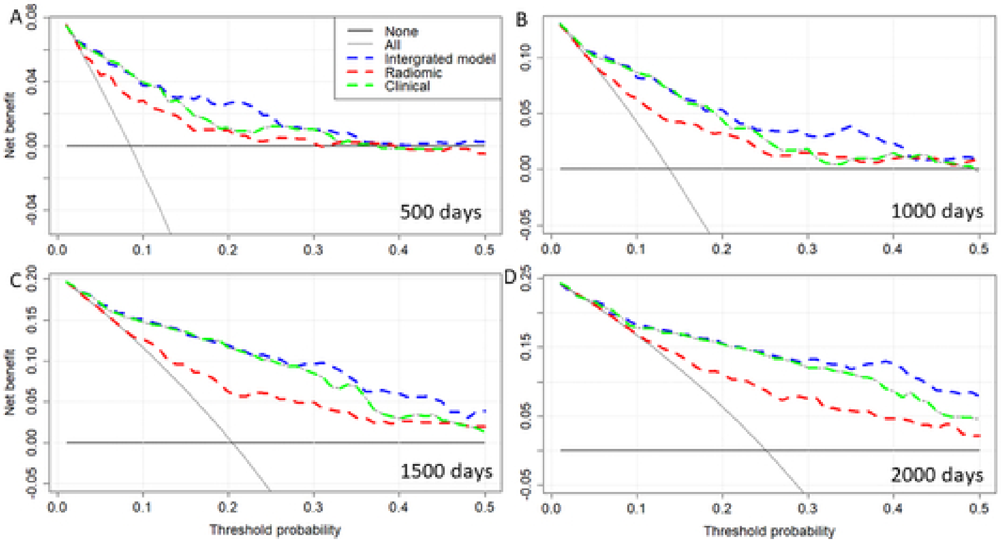


FIGURE 7: Decision curve analysis at 500(A), 1,000(B), 1,500(C), and 2,000(D) days. The net benefit is plotted against the threshold probability. If the curve is closer to the right top corner, then the corresponding model is better as it has larger net benefit. The "all" curve shows the net benefit by treating all patients, while the "none" curve denotes net benefit for treating no patients.

45x24mm (300 x 300 DPI)

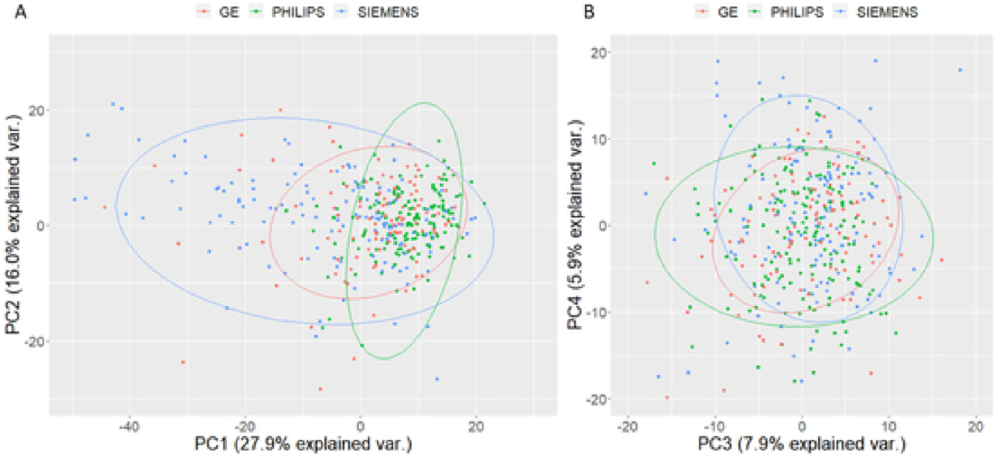


FIGURE 8: Scanner difference study. Principal component analysis for radiomics features from a different scanner. A: first principal component (PC1) vs second principal component (PC2); a relatively larger variation was observed using the Siemens scanner. B: Third principal component (PC3) and fourth principal component (PC4) explain smaller percentages of the total variation, and the three different scanners show good agreement. The dots in the figure represent samples; the colors represent groups (scanner types); and the legends have three groups at the top. The ellipse represents the core area added by the default confidence interval of 68%, which facilitates the separation between the observation groups. No clear separation of the sample based on the three MRI vendors was observed. var.: variance. PC: principal component.

46x21mm (300 x 300 DPI)

~~Title of the manuscript:~~

An integrated MR radiomics model to estimate ~~predicts~~ survival time in patients with endometrial cancer~~.~~

Abstract

Background: ~~Prediction~~ Determination of survival time in women with endometrial cancer using standard clinical features remains imprecise. Identification of radiomic signatures may improve these estimates.

Purpose: ~~To identify imaging signatures in T2-weighted sequences that predict survival time in women with endometrial cancer, using T2 weighted MRI and to develop a clinically useful nomogram to provide afor more personalized and accurate estimation of survival time.~~ To identify clinical features and imaging signatures on T2-weighted MRI that can be used in an integrated model to estimate survival time for endometrial cancer subjects.

Study Type: Retrospective.

Population: ~~A total of 413 patients with endometrial cancer were included as training and validation cases data and an independent set of 82 subjects were used as testing datasets~~. 413 patients with endometrial cancer as training (n=330, 66.41 ±11.42 years) and validation (n=83, 67.60±11.89 years) data and an independent set of 82 subjects as testing data (n=63.26±12.38 years).

Field Strength/Sequence: ~~1.5-T and 3-T, scanners with standard sagittal T2 -weighted sequence covering the pelvis and abdomen were used.~~ 1.5-T and 3-T scanners with sagittal T2-weighted spin echo sequence.

Assessment: ~~Tumor regions in MRI were manually segmented. Imaging features were extracted from the segmented masks and T2- weighted images. A group lLeast aAbsolute~~

Commented [ME1]: Retrospective studies show associations.
Remove the words “predict” / “prediction” throughout the title, abstract and main manuscript text.

Commented [LX2R2]: Changed as : to estimate

Commented [MS3]: Retropectice studies show associations
Do not use the word predict or prediction anywhere in this paper

Commented [LX4R4]: Replaced "Prediction" in the text where it is necessary.

Commented [ME5]: Background of the abstract should lay foundation for the paper – end with a clear transition statement towards the purpose.

Commented [LX6R6]: Changed.

Commented [ME7]: Include information on age.

Commented [MS8]: State numbers in each

Commented [ME9]: Report the exact numbers of subjects used for the groups of training / validation / test.

Commented [LX10R10]: Included now.

Commented [ME11]: Unclear what “standard” refers to regarding the sequence – please remove.

Commented [LX12R12]: Spin echo sequence was used.

Commented [ME13]: Include base sequence type (e.g., spin/gradient echo).

Commented [ME14]: Add info on the clinical variables that are then reported under the results section. Where do they come from and what did they represent?

Commented [ME15]: Which type of imaging features (e.g., second-order etc.)? Please specify.

Commented [LX16R16]: Features include tumour size/shape, first order statistics etc. Abstract only allows 300 words, we cannot include these information in abstract.

Commented [ME17]: The segmentation masks are on the T2-weighted images. Were there any additional areas used for feature extraction besides the segmentation mask enclosing the tumor?

For Peer Review

~~sShrinkage and sSelection oOperator (gLASSO) logistic regression model with a stratified 10-fold cross validation method was implemented to determine the optimal model from the training dataset.~~ Tumor regions were manually segmented on T2-weighted images. Features were extracted from segmented masks, and clinical variables including age, cancer histologic grade and stage were included in a Cox proportional hazards (CPH) model. A group least absolute shrinkage and selection operator (gLASSO) method was implemented to determine the model from the training and validation datasets.

Statistical Tests: ~~A likelihood test and dDecision Ccurve aAnalysis (DCA) were applied to compare the models from the gLASSO method. Time-dependent aArea Uunder the cCurves (AUCs) for the models were calculated using a bootstrap method. A P-value <0.05 was considered statistically significant.~~ A likelihood-ratio test and decision curve analysis (DCA) were applied to compare the models. Concordance index (CI) and area under the receiver operating characteristic curves (AUCs) of were calculated to assess the model..

Results: ~~Three radiomic features and two clinical variables were selected as predictors in the Cox proportional hazards (CPH) model. The concordance index (CI) was 0.797 for the clinical model and 0.818 for the integrated clinical-radiomic model using training dataset. Based on the testing dataset, the CI was 0.792 and 0.882 for the clinical model and integrated model, respectively. A likelihood ratio test showed a significant difference between the integrated model and clinical model based on both training (p=7.097e-06) and testing datasets (p=0.0054).~~ Three radiomic features (e.g., two image intensity and volume features) and two clinical variables (age and cancer grade) were selected as predictors in the integrated model. The CI was 0.797 for the clinical model (includes

Commented [ME18]: Optimal model in what sense? Please specify.

Commented [LX19R19]: We removed the "optimal". It means the best model under L1 and L2 constraints, i.e., the best model when considering model accuracy and parsimony.

Commented [ME20]: For which parameter? Please specify.

Commented [LX21R21]: It is AUC of ROC, included now.

Commented [ME22]: Report threshold for statistical significance at the end of the statistical analysis section.

Commented [LX23R23]: Due to 300 words constraints, we moved this sentence into method section.

Commented [ME24]: Remains unclear what the "clinical model" and "integrated clinical-radiomic" model are – define their composition in the assessment section.

Commented [LX25R25]: Clinical model includes age and cancer grade only, integrated model includes both clinical and radiomic features.

Commented [ME26]: See comment above – mention in the assessment section which kind of features were used.

Commented [LX27R27]: Image intensity within the tumour mask and tumour mask volume., mentioned now.

Commented [ME28]: Clinical variables are not mentioned in the assessment section. Add there.

Commented [LX29R29]: Indicated by (age and cancer grade)

Commented [ME30]: Cox proportional hazard model needs to be mentioned first under the statistical analyses. Please add there so that related findings can then be reported here under results.

Commented [LX31R31]: Mentioned CPH model in

Commented [MS32]: You have to discuss outcomes in

Commented [MS33R33]:

Commented [LX34R33]: Moved to assessment.

Commented [MS35]: These. Need to be listed in your

Commented [LX36R36]: Moved to statistics section,

Commented [ME37]: Concordance index needs to be

Commented [LX38R38]: Mentioned in statistical

Commented [MS39]: You need to state clearly what

Commented [LX40R40]: Specified with (with age and

Commented [ME41]: Do not report p-values for

Commented [LX42R42]: Removed these.

Commented [ME43]: How about the AUC values? Please

Commented [LX44R44]: Added.

For Peer Review

clinical variables only) and 0.818 for the integrated model using training and validation datasets, the associated mean AUC value was 0.805 and 0.853. Using testing dataset, the CI was 0.792 and 0.882, significantly different and the mean AUC was 0.624 and 0.727 for the clinical model and integrated model, respectively..

Data Conclusion: ~~The newly developed radiomic signature was a~~may serve as a ~~powerful predictor of survival time in women with endometrial cancer, outperforming current models using either clinical information or radiomic features only~~. The proposed CPH model with radiomic signatures may serve as a tool to improve estimated survival time in women with endometrial cancer.

~~**Level of Evidence:** 4~~

~~**Technical Efficacy Stage:** 2~~

Keywords: Endometrial cancer, ~~MRI,~~ survival analysis, radiomics, feature selection, Cox p~~P~~roportional h~~H~~azards m~~M~~odel

Commented [ME45]: Retrospective studies show associations.
Remove the words “predict” / “prediction” throughout when related to the findings of your study.

Commented [LX46R46]: Replace predict with estimate.

Commented [ME47]: Retrospective studies show associations.
Remove the words “predict” / “prediction”.

Commented [LX48R48]: Replace with estimate/estimation.

Commented [MS49]: This is not mentioned above
Nor is any numerical data provided

Commented [LX50R50]: CI/AUC was provided in Results section now.

Commented [ME51]: No comparison made to “current models”. Only report conclusions that are directly linked to your reported findings.

Commented [LX52R52]: Removed.

For Peer Review

For Peer Review

Introduction

Endometrial cancer is the most common gynecological cancer, ~~in the western world with over 120,000 new cases diagnosed each year in Europe (1), and~~ with 417,000 new cases diagnosed globally in 2020 (2). ~~It has been reported that t~~The 5-year overall survival rate of endometrial cancer patients ranges from 74% to 91% ~~worldwide~~ (3) (4) (5).

To study the survival time of endometrial cancer patients, survival analysis methods have been extensively applied (6) (7)~~ have been extensively applied~~. Currently utilized examples include the non-parametric Kaplan-Meier method (8) (5) (9) and the semi-parametric Cox's proportional hazards (CPH) method (10) (4) (11). Based on the CPH model, it is possible to evaluate ~~the~~ a single covariate's and/or the combination joint covariates effects on the survival time estimation. For instance, a combination of age, cancer histologic grade (12), socioeconomic factors (13) (14), and other clinical prognostic factors have been investigated in endometrial cancer survival studies within the framework of the CPH model (15).

Until the advent of radiomics,~~While~~ image biomarkers ~~were typically~~have been ~~less well-studied~~insufficiently studied as potential survival predictors for endometrial cancer~~, until the advent of radiomics~~. Radiomics is a rapidly expanding field of research in oncology (16) (17). There have been~~are~~ studies evaluating the application of radiomics, usually based on multi-sequence MRI features, for example Kurtosis from contrast-enhanced T1-weighted MRI~~,~~ to predict survival time in endometrial cancer, but these studies had several limitations (18-20)~~, but these studies had several limitations~~. ~~Firstly~~Specifically, these studies (18-20) employed a dataset with less than 200 cases, thus with a small number of sample sizes and radiomic features (less than 100 features), which could lead

Commented [ME53]: As per JMRI reference style, all in-text reference numbers need to be between parentheses () and combined within () – e.g. (3-5) here instead of (3) (4) (5). Please modify reference style throughout.

Commented [LX54R54]: Changed in the clean version.

Commented [ME55]: As per JMRI reference style, all in-text reference numbers need to be between parentheses () and combined within () – e.g. (3-5) here instead of (3) (4) (5). Please modify reference style throughout.

Commented [LX56R56]: All the references have been changed in the clean version.

Commented [ME57]: Move references to the end of this sentence.

Commented [LX58R58]: Changed.

to a large bias for the model estimation (21) (22). ~~and some were missing important radiomic features in their analysis. Secondly~~Moreover, as these studies did not conduct a validation using independent external testing data, there was no model validation for survival time prediction(17-19). Finally, these studies did not include or combine clinical~~ly~~ prognostic variables in the CPH model for the survival time prediction(17-19). As a result, ~~they~~ those previous studies ~~did~~ have most likely not evaluated the true potential of radiomic features for survival time prediction in endometrial cancer(17-19).

To overcome these limitations, this retrospective study aimed~~s~~ to identify a radiomic signature using pelvic MRI data ~~images~~ that ~~can~~ could predict survival time in endometrial cancer. Furthermore, we aimed to ~~and~~ develop and validate an integrated clinical-radiomic model that ~~can~~ might be used to tailor adjuvant management for women based on their personalized risk features.

Commented [ME59]: Add references to those studies to this sentence as well.

Commented [LX60R60]: added

Commented [ME61]: Add references to those studies to this sentence as well.

Commented [LX62R62]: Added (17-19)

Commented [ME63]: Add references to those studies to this sentence as well.

Commented [LX64R64]: Added (17-19)

Commented [ME65]: Retrospective studies show associations.
Remove the words "predict" / "prediction" throughout.

Commented [LX66R66]: Replace with estimate in the clean version,

For Peer Review

Materials and methods

~~This was a retrospective study and was approved by Bloomsbury ethics. The requirement for written informed consent was waived due to the retrospective design of this study.~~ ~~; the ClinicalTrials.gov Identifier was NCT03543215 on clinicaltrials.org website. The Institutional Review Board (IRB) approved the study protocol, and the Research Ethics Committee reference number was for this study is 17/LO/0173.~~ ~~The ClinicalTrials.gov identification number is NCT03543215 (https://clinicaltrials.gov/).~~ ~~The requirement for written informed consent was waived due to the retrospective design of this study.~~ This retrospective study protocol was approved by the Institutional Review Board (IRB), and the Research Ethics Committee reference number for this study is 17/LO/0173. The requirement for written informed consent was waived due to the retrospective design of this study. This retrospective study will develop and test a model which will be further validated as part of a larger prospective study (ClinicalTrials.gov NCT03543215, https://clinicaltrials.gov/).

Commented [ME67]: First line of the methods section should be ethics approval and written informed consent or waiver.

Commented [LX68R68]: Ok, included.

Commented [MS69]: What does this mean?

Commented [LX70R70]: Removed.

Commented [ME71]: Include information on informed consent – please check if this is correct and a waiver has been granted?

Commented [LX72R72]: Yes.

Commented [MS73]: How is there a clinical trials number for a retrospective study?

Commented [LX74R74]: This is a retrospective study in a big prospective study.

Commented [ME75]: Include information on informed consent – please check if this is correct and a waiver has been granted?

Commented [LX76R76]: Yes, included.

Training and validation datasets

~~This was a retrospective study and approved by Bloomsbury ethics; the ClinicalTrials.gov Identifier was NCT03543215 on clinicaltrials.org website. The Institutional Review Board (IRB) approved the study protocol, and the Research Ethics Committee reference number was 17/LO/0173.~~ The time range of image acquisition ~~for the~~ ~~datasets~~ was between October 2011 and September 2017 (Figure 1), and 270 of the initially considered 611 subjects were obtained from a previous study (23). The training and validation datasets

Commented [ME77]: Please use a unique style/format for all headings/subheadings – subheadings have been changed to italic style.

Commented [LX78R78]: Ok, we fixed the style.

Commented [ME79]: Make clear whether the reported range refers to the time point of image acquisitions.

Commented [LX80R80]: It is image acquisition time.

For Peer Review

were obtained from 15 UK hospitals and centers with different parameters and protocols (Table 1). Table 1 shows the scan parameters for collecting 411 subjects of training/validation dataset which excluded two subjects because the scan parameters information was not available. As T2-weighted images were included from different centers in the study, we rely on image pre-processing and image normalization to minimize the difference between different scanners and sequences.

Clinical data, including the patient age at diagnosis, date of surgery, type and grade of tumor, the international federation of obstetricians and gynecologists (FIGO) stage~~stage~~, presence of lymphovascular space invasion, and any adjuvant or neoadjuvant treatment of these subjects were obtained from an online medical records system (23). Survival time was defined as the time from the date of surgery until the date of death. ~~That~~ This information was last updated on August 3rd, 2020~~, therefore the ending time for the training and validation datasets was on August 3rd, 2020~~.

~~T2-weighted MR images~~The inclusion criteria regarding MRI were as follows: 1) no severe motion arti~~e~~facts in T2-weighted images, 2) sufficient size of the tumor ~~in MRI~~on images~~,~~ (i.e., the tumor ~~can~~ could be ~~found~~ identified o~~i~~n more than one MRI~~ image slice~~), and 3) the T2-weighted sequence ~~able to~~ passes the image pre-processing steps. The ~~clinical data were obtained from an online medical record system, and the~~ inclusion criteria regarding clinical data were: 1) availability of censoring or noncensoring survival information, information on ~~L~~lymphovascular space invasion, histological risk, and histological type~~ information availabilities~~, 2) availability of ~~the~~ age at diagnos~~e~~is and surgery date, 3) no other type of cancer coexist~~s~~ing with endometrial cancer. After exclusion ~~from~~ of patients based on image and clinical criteria, 413 cases were ~~eventually~~

Commented [ME81]: Section on imaging parameters for T2-weigthed images entirely missing – needs to be added with info on base sequence types and most common parameters of acquisition.

Commented [LX82R82]: Add as Table 1.

Commented [ME83]: Type, grade and stage according to which scheme? Please specify.

Commented [LX84R84]: FIGO scheme.

Commented [ME85]: Use American English spellings throughout the manuscript.

Commented [LX86R86]: Changed tumor to tumour etc.

For Peer Review

used in this study (Table 2). The ratio for splitting the training and validation was 80:20 (n=330 for the training data; n=83 for the validation data) with balance the survival object (i.e., the combination of time and death information) distributions within the splits. ~~1). The distribution of the training dataset is displayed in Figure 2A,2B and 2C.~~

Commented [ME87]: Report exact numbers for subjects used for the groups of training / validation / test.

Commented [LX88R88]: added

Testing dataset

~~An additional 102 patients with endometrial cancer were collected from 3 hospitals in the UK. After excluding any patients with additional other types of cancers such as breast and ovary cancers~~Overall, 82 additional patients from three hospitals in the UK with endometrial cancer were included in the testing dataset (Table ~~1~~2). Specifically, 74 of those 82 cases were right censoring and 8 cases died before the end of the study (Dec 1st, 2021). For this testing dataset, the beginning time was the surgery date (the earliest date was in 2017), and the ending time was ~~on~~ Dec 1st, 2021. Table 1 shows the demographic information from the testing dataset. The age at diagnosis of the testing dataset was significantly different from the training dataset (~~the age of the training dataset was~~ 66.64±11.5 years~~, the age of the testing dataset was~~ versus 63.26±12.38 years, ~~p =0.024, see~~ Table 1).

Commented [ME89]: Do not report p-values for statistically significant findings. Only report exact p-values for non-significant findings.

Commented [LX90R90]: ok

In addition, Figure 2~~D, 2E and 2F~~ plots the histograms of the testing dataset and a two-sample t-test was applied to compare the training (including validation) and external testing datasets. Except for the survival time (Figure 2B and Figure 2E), all comparisons between training and testing datasets were significant~~ (P<0.05)~~. For the survival time (Figure 2B and Figure 2E), ~~the t test does not shown~~no significant differences were revealed (~~the survival time for the~~ training dataset: ~~was~~ 870.6±592.1 days, ~~and the~~

Commented [ME91]: This information should be deleted here and reported under the statistical analysis paragraph. Please modify accordingly.

Commented [LX92R92]: Removed to results section.

For Peer Review

~~survival time for the ~~testing dataset: ~~was ~~637.1±314.2, p=0.09). ~~This is because there are only 8 cases which had died in the dataset (Figure 2E), and the degree of freedom for the t test is small.~~

Radiomics study pipeline

Figure ~~3~~ 2 shows the radiomics study pipeline for the survival analysis. There were five steps in this study~~ as shown in the columns of Figure 3~~. The first and second steps were designed to analyze images, including manual image segmentation, MRI non-~~-~~uniformity correction, image resampling, and image normalization. Specifically, Digital ~~i~~Imaging and C~~c~~ommunications in ~~m~~Medicine (DICOM) file formats were downloaded from the picture archiving and communication systems (PACS), de-identified and converted to the simpler N~~n~~euroimaging ~~i~~Informatics T~~t~~echnology I~~i~~nitiative (NIFTI) format.

An interactive tool (ITK-snap, version 3.6.0, http://www.itksnap.org~~-~~) for semi-automatic segmentation of sagittal orientation T2-weighted MRI ~~multi-modality biomedical images~~ was employed for manual slice-by-slice tumor segmentation by two radiologists in-training (JR, 5 years, with assistance from AS, 3 years) (24)~~by a radiologist (**) with a minimum of 5 years' of experience~~. After loading a relevant ~~neuroimaging informatics technology and an initiative format image~~NIFTI dataset, the paintbrush tool was used to label all voxels containing visible tumor on each sagittal slice. Once all slices containing tumor had been labelled, the segmentation mask was saved for pre-processing steps. This process was repeated for T2-weighted MRI in every image set~~every image set~~. This was then checked by two ~~experienced ~~radiology consultants (AR, 19 years' experience and NB 15 years' experience), who corrected the segmented tumor masks, without further

Commented [ME93]: This should all be part of the results instead, given that it reports calculated values based on tests that have to be reported in the statistical analysis paragraph of the methods. Please move accordingly.

Commented [LX94R94]: Removed to results section.

Commented [ME95]: Include version.

Commented [LX96R96]: added

Commented [ME97]: Using only the T2-weighted images? In which orientation? Please add.

Commented [LX98R98]: Yes, only T2-weighted image. sagittal orientation

Commented [ME99]: Report the exact years of experience.

Commented [LX100R100]: Included.

Commented [ME101]: T2 of each patient – or more than one sequence analysed in the patients? Please clarify.

Commented [LX102R102]: The segmentation was carried out on T2-weighted MRI only.

For Peer Review

went through all cases together again.~~(** and ***; each haveing more than 10 years’ of experience)~~, ~~who provided the ground truth~~. The radiologists were blinded to the outcome measures. One example of the image segmentation is displayed in ~~the first step/column of~~ Figure ~~3~~2.

Commented [ME103]: Report the exact years of experience for each rater.

Commented [LX104R104]: included

Commented [ME105]: Were they allowed to correct the masks – and went through all cases together again? Please specify.

Commented [LX106R106]: no

The T2-weighted images were pre-processed according to step 2 ~~in~~ as shown in Figure ~~3~~2. First, all image voxel sizes were obtained from ~~neuroimaging informatics technology initiative format~~NIFTI files with T2 image header files, and the median voxel size of all data was calculated. The sagittal scan was then adopted.~~;~~ T~~t~~he image slice thickness was between 0.5 mm and 5 mm~~;~~, and the image size was between 256 and 640. The median resolution (image voxel size)~~value~~ of all T2-weighted images (including both training/validation and testing datasets) was 0.625 mm x 0.625 mm x 5 mm. ~~Second~~Then, T2-weighted images were processed using an N4 toolbox for MRI non-uniformity bias correction, and to remove artifacts due to the inhomogeneity of magnetic fields (https://github.com/ANTsX/ANTs/wiki/N4BiasFieldCorrection) (24)~~ for MRI non-uniformity bias correction; to remove artefacts due to the inhomogeneity of MRI magnetic fields~~. ~~Third, a~~All non-uniformity corrected T2-weighted data ~~MRI~~ and ~~its~~ the segmentation masks were resampled to median voxel values. For T2-weighted images, the cubic spline interpolation method was adopted. ~~; while f~~For segmented tumor masks (binary image), a nearest neighbor interpolation method was used for image resampling. Next, the intensity of resampled T2-weighted images was normalized using the following equation:

Commented [ME107]: Unit missing for this?

Commented [LX108R108]: This is image reconstruction matrix size. There is no unit for this.

Commented [ME109]: Include version.
Also include URL (for non-commercially available software) or vendor name / city / state / country for commercially available software.

Commented [LX110R110]: Included now.

$$I = \frac{I - \bar{I}}{std(I)} 100 \qquad \text{(Eq. 1.)}$$

where I is image intensity, $\bar{I}$ is the mean value of the image intensity within the volume, and *std* is the standard deviation of the image volume. Finally, the TexLAB tool (version

For Peer Review

2.0) on MATLAB (version ~~R17a~~R19a; ~~(~~The MathWorks~~,~~ Inc., Natick, ~~Massachusetts~~MA, ~~United States.~~USA; http://www.mathworks.com/), PyRadiomics (25), and Scikit-image (26), both implemented in Python (Python Software Foundation, https://www.python.org/) were used to perform feature extraction as shown in ~~the third step in~~ Figure ~~3~~2. After eliminating the identical features by a correlation method, in total~~,~~ 958 radiomics features were extracted from T2-weighted MRI and ~~its~~ the segmentation masks~~ images~~. The reason to include T2-weighted MRI was because image intensity-based features were derived from T2-weighted MRI images. Endometrial cancer tumor region is the only region of interest included for this study.

Commented [ME111]: Include versions for those applications.

Commented [LX112R112]: added

Commented [ME113]: Any segmentations performed other than for the tumor? Why is the T2 reported here in addition to the extraction of metrics from the masks? Please clarify.

Commented [LX114R114]: This is because the image intensity-based features were derived from T2-weighted MRI. Added in the clean version.

Feature selection

The fourth step was to select features for survival analysis. Radiomic and clinical feature~~s~~ selections were preformed within the framework of statistical model selection, and the ~~a Cox proportional hazards (~~CPH~~)~~ model (27) was used to study the relationship between predictor variables and survival time. In the CPH model, 958 radiomics features, cancer risk score (which includes FIGO stage), cancer grade, and age ~~variables~~ were included as predictors for model selection. Cancer risk score and grade were defined according to FIGO (23,28). To avoid model overfitting, a 10-fold cross validation for penalized Cox regression models with grouped covariates was adopted. Specifically, a group ~~exponential~~ least absolute shrinkage and selection operator (gLASSO) was used to select statistical models (28). The maximum iteration of the 10-fold cross validation was set to be 1 million times in the model fitting. The selected model was then applied to calculate the survival time.~~;~~ ~~t~~The time and event/death was treated as a dependent variable and

Commented [ME115]: Unclear how cancer risk score and cancer grade were defined and calculated – please specify.

Commented [LX116R116]: The definition and calculation were based on references (23,28).

For Peer Review

radiomic features, cancer grade, and age were included as the independent variables.~~;~~ T~~t~~he CPH model ~~(27)~~ was employed to estimate the survival time.

~~Survival and s~~Statistical analys~~i~~es

The R ~~language~~software (version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria;~~,~~ http://www.R-project.org) was ~~employed in the~~used for statistical analysis. Model selection package “grpgrep” (version 3.4.0, https://cran.r-project.org/web/packages/grpreg/index.html) was applied to determine the optimal CPH model. The criteria for the optimal model were model simplicity and accuracy (i.e., minimize the combination of the L1 and L2 norm) (29,35). The “Survival” package (version 3.4.0, https://cran.r-project.org/web/packages/survival/index.html) was used to implement CPH model. A bootstrap resampling method was developed to assess the predictive performance of the CPH model using a Score() function from a “riskRegression” R library (version 2021.10.10, https://cran.r-project.org/web/packages/riskRegression/index.html). Nomograms were generated using a “regplot” R package (version 1.1, https://cran.r-project.org/web/packages/regplot/index.html).

~~Nomograms were generated using a ‘regplot’ R package with an R version of 4.0.2.~~

~~The last step was to implement s~~Survival analysis ~~with the~~was implemented in the selected integrated model as shown in step 5 of Figure ~~3~~2. Before applying the model selection method, all 959 features (958 MRI features + age) were normalized using a Z-score method (similar to Eq. 1, except multiply 100). The gLASSO method produced model selection results with randomness. We adopted the most common output by the

Commented [ME117]: Risk score mentioned above – why not included here?

Commented [LX118R118]: This is because gLASSO method did not select risk score in the model.

Commented [ME119]: Include version for R.

Commented [LX120R120]: Included.

Commented [ME121]: The software used for statistical analysis should be reported at the beginning of the statistical analysis section.

Commented [LX122R122]: Changed and included.

Commented [ME123]: Include version, if available for this package.

Commented [LX124R124]: Included.

For Peer Review

gLASSO method in our ~~. Once~~study. Once the survival time prediction as the CPH model was established with the LASSO method, a nomogram was created as a graphical representation of the integrated model. The CPH was implemented to predict the survival probability and a nomogram was applied to visualize the prediction. To study the influence of the radiomic features on the survival probability, estimation of the two models was constructed and compared for the prediction. The first model was based on clinical information only~~,~~ (i.e., the predictors of the model included only age and clinical cancer grade). The second model use~~s~~d both clinical information (age and clinical cancer grade) and ~~3~~ three radiomic features selected by the bi-level gLASSO method. The survival probabilities~~y~~ of 1, 2 ,3, and 5 years ~~was~~ were estimated based on the CPH model.

Additional analyses were performed to validate the model based on the prediction. A bootstrap resampling method was developed to assess the predictive performance of the CPH model using a s~~S~~core function from a “riskRegression” R library. The time-dependent area under the receiver operating characteristic (ROC) curve (AUC) was calculated from the validation (AUC is specified for AUC of ROC in this study)~~area under the curve (AUC) was calculated from the validation~~. For the model validation, 80% of the 413 cases were used to generate the CPH model, while the rest of the datasets were employed to validate the predictive performance. The size of bootstrap resampling was set to be 10 at each time point. The time point started at 100 days and terminated at 1,825 days with a 5-day interval. Similarly, additional external testing cases were used to test the CPH model prediction performance.

To study the effect of the radiomic features and clinical variables on survival time estimation, decision curve analysis (DCA) (29) ~~were~~ was applied to evaluate the clinical,

Commented [ME125]: Risk score mentioned above – why not included here?

Commented [LX126R126]: This is because after applying gLASSO, the risk score was not be selected in the model. However, cancer grade was selected with the model selection method.

Commented [ME127]: Include version and URL, if available for this package.

Commented [LX128R128]: Add and moved the beginning of this section.

Commented [ME129]: Define all abbreviations at first mention.

For Peer Review

radiomic, and integrated models for net benefit. A likelihood ratio test method was applied to study the importance of the radiomic and clinical features for survival time prediction. Furthermore, ~~The R language (R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org) was employed in the analysis. ~~t~~T~~raining and test datasets were compared using ~~c~~Chi-~~-~~squared tests for categorical data and ~~a ~~two-sample t-tests for continuous data. ~~Nomograms were generated using a 'regplot' R package with an R version of 4.0.2.~~A P-value <0.05 was considered statistically significant.

Commented [ME130]: If more than one statistical software has been used, report info for each software incl. version.

Commented [LX131R131]: Included and moved to the beginning of this section.

Commented [ME132]: Report threshold for statistical significance at the end of the statistical analysis section.

A diagnostic analysis was carried out to study the feature variation obtained from different types of scanners using 413 training/validation cases. It is not obvious to inspect the features difference from two dimension, i.e., an image with 413 rows and 985 feature columns. Therefore, dimension reduction method was applied to obtain the major components of the features from each type of scanner. Specifically, principal component analysis (PCA) was applied to study the effect of feature difference from different scanners. All features were normalized using a Z-score method, and then a PCA was employed to split the feature dataset into different components. Four principal components were used to compare the feature variations from different scanners. Three different manufacturers, i.e., GE (108 cases), Philips (163 cases), and Siemens (142 cases) were used to acquire sagittal T2-weighted MRI (Table 1). Feature matrix from these three different scanners were decomposed into four components. Visual comparison was carried out to evaluate the distribution of the feature components from different scanner.

For Peer Review

Results

Training and testing dataset demographics

Clinical-pathological characteristics of the patients are shown in Table 2. In addition, Figure 3 plots the histograms of the testing dataset and a two-sample t-test was applied to compare the training (including validation) and external testing datasets. Except for the survival time (Figure 3B and Figure 3E), all comparisons between training and testing datasets were significant. For the survival time (Figure 3B and Figure 3E), no significant differences were revealed (training dataset: 870.6±592.1 days, testing dataset: 637.1±314.2, p=0.09). Table 2 also includes the demographic information from the testing dataset. The age at diagnosis of the testing dataset was significantly different from the training dataset (66.64±11.5 years versus 63.26±12.38 years, Table 2).~~Out of the total 611 consecutive patients treated for endometrial cancer, 413 were included in the final analysis (Figure 1). The breakdown of cClinical-pathological characteristics of these the patients is are seen shown in the first 3 columns of Table 1. Any patients missing key data, such as survival, were excluded from the outset (Figure 1). Thus, there was no indeterminate index test or standard reference results.~~

Commented [ME133]: Please use a unique style/format for all headings/subheadings – subheadings have been changed to italic style.

Commented [ME134]: Do not report info from the methods section in the results section.

Commented [ME135]: Do not report info from the methods section in the results section.

Feature selection ~~results~~

~~Before applying the model selection method, all 961 features (958 MRI features + age + risk score + cancer grade) MRI features including age were normalized using a Z -score method. The group LASSO method produced model selection results with randomness. We adopted the most common output by the group gLASSO method in our study.~~ Figure 4 shows the gLASSO coefficient profiles of the 961 ~~(958+clinical variables) MRI~~ features.

Commented [ME136]: This needs to be part of the methods section and should be removed from the results section.

For Peer Review

For Peer Review

Specifically, Figure 4A plots the 10-fold cross-validated error rates, and Figure 4B ~~is~~ shows the amplified version of the g~~roup~~ LASSO selection plot~~ (Figure 4A)~~. ~~In Figure 4, each red dot represents a λ value along the path. In the group LASSO method, the cross-validation method was applied to select the tuning parameter (λ). Dotted vertical lines were drawn at the optimal λ values by using the minimum criteria (i.e., cross-validation error). A Lambda value of 0.067 (log(λ)=-2.7) according to the 10-fold cross-validation method was computed.~~

Commented [ME137]: This can be moved to the legend of Figure 4 and deleted here.

~~After applying the bi-level LASSO method (4), f~~Five features ~~(Figure 4)~~ were selected from 961 predictors and were included in the integrated CPH model. They were tumor mask minor axis ~~length~~radius~~ from Python library Sci-Image~~, grey-level size zone and first order statistics~~ from TexLab~~, patient~~s'~~ age at diagnosis, and patient~~s'~~ cancer grade.

Tumor minor axis radius reflects the size of the tumor indirectly; the FOS here is the coefficient of variation, which is defined as the ratio of the standard deviation to the mean, and these values were computed within the tumor mask. This was computed after the normalized T2-weighted image were filtered with low, low, and high filters in x, y, and z direction of the 3D image subsequently. The GLSZM was calculated after the normalized MRI image was converted into 25 Hounsfield unit grey level, then the large zone low gray level emphasis was computed within the tumor mask. The FOS and GLSZM represent image statistical property and intensity character. These five selected features were refit into a CPH model without the normalization of the age covariate, for the purpose of displaying in the nomogram. The survival prediction was then estimated based on the refit CPH model. The final integrated model was:

Surv(Time, Death) = 0.0548*Age + 0.0025*Grade2 + 1.684*Grade3 + 0.495*minorAxisRad - 0.263*GLSZM - 0.179*FOS. The corresponding clinical model (excluding radiomic features) was:

Surv(Time, Death) = 0.0455*Age + 1.881*Grade2 + 2.107*Grade3,

where Surv is the survival object, defined as a response variable in the CPH model and age was not normalized.~~The minor axis length describes the length of the minor axis of the ellipse that has the same normalized second central moments as the tumor region from the image (26) and is a statistical measure which quantifies the variation of an image pixels intensities. For the features of a grey-level size zone and first order statistics, both features are related to the distribution of intensities. In summary, theseThe two selected image-based features reflected the tumor shape, image intensity character, and statistical properties.~~

~~The patients' age at diagnosis, patients' cancer grade, and three radiomic features, e.g., the (tumor minor axis length, grey-level size zone, and first order statistics) were selected for survival time estimation. After these 5 features have been selected using the group LASSO method, and for the purpose of investigating age influence on the survival time estimation clearer, tThese features were refit into a CPH model without the normalization of the age covariate, for the purpose of displaying in the nomogram. The survival prediction was then estimated based on the refit CPH model.~~

Commented [ME138]: Avoid including any references in the results section.

Commented [LX139R139]: Removed.

Commented [ME140]: This should be part of the discussion section, not a result of your analysis but an interpretation.

Commented [LX141R141]: Removed.

Model training and validation

~~Eighty percent of the included 413 cases were used as training data to build the CPH model, and the rest of the data were used as a validation dataset. Using a stratified~~

Commented [ME142]: Already report exact numbers for subjects used for the groups of training / validation / test in the methods section.

Commented [LX143R143]: Removed.

For Peer Review

~~sampling method, the training and validation datasets were split with survival objects (time/death). The CPH model with 5 predictors was determined by a group LASSO method. To validate the CPH model, the bootstrap resampling method with a sample size of 10 at each time point was adopted. Because the bootstrap method and stratified sampling method have randomness, and as a typical example, the area under the receiver operating characteristic curve (AUC) was calculated and displayed. To reduce the effect of the randomness in the evaluation study, the concordance index (CI), which measures the prediction accuracy, was calculated with 10 repetitions. The results showed that the CI value was significantly higher using the integrated model based on these 413 cases (integrated model CI: was 0.825±0.010, clinical model CI: was 0.806±0.011, p=0.00097).~~

Commented [ME144]: Do not repeat methodological details in the results section, only report in the methods section.
This entire section should be integrated in the statistical analysis section and removed here.
Also define the CI in the stats section.

Figure 5A plots the time~~-~~-dependent AUC based on training and validation datasets. For the clinical model, the AUC accuracy ~~is~~ was below 80% for the time points after 1,250 days, suggesting that there is an AUC decrease for ~~the~~ long~~er~~-time prediction. In Figure 5A, the integrated model ~~(red curve) has~~had a larger AUC than the clinical model ~~(green curve) in~~for all time points, suggesting that the integrated clinical-radiomic model is significantly superior to the clinical model for the prediction based on the external testing dataset for all time points (integrated model AUC: 0.853±0.06, clinical model AUC 0.805±0.058~~, p=2.2e-16~~).

Similar to validation and using the same trained model (obtained from the training dataset), testing was carried out and the results are presented in Figure~~s~~ 5B. Comparing Figure 5A with Figure 5B, the AUC obtained from the testing dataset is smaller than the AUC computed from the training dataset. This is because the survival time and age from the

For Peer Review

testing data ~~are~~ were significantly different from the training and validation datasets (training and validation data: 1583.4 ±669.6 days, testing data: 1318.7±306.4 days. A likelihood-ratio test showed a significant difference between the integrated model and clinical model based on both training and testing datasets. The CI was 0.797 for the clinical model and 0.818 for the integrated model. Based on the selected model from training data, the nomogram display the 1, 2, 3, and 5-year survival probabilities is shown in Figure 6.~~, p=5.27e-08, also see section of "Testing dataset" in the materials and methods). It suggested that the integrated model is robust to the changing of the input dataset.~~

~~*Nomogram visualization for CPH prediction*~~

~~Once the model had been tested on the dataset externally, the integrated model was used for survival probability prediction. AThe nomogram was adopted to display the 1, 2, 3, and 5-year's survival probabilityies is shown in (Figure 6). Additionally, based on all training, validation, and testing datasets, a likelihood ratio test was adopted to study the significance of the radiomic features. The model with radiomic features (age, cancer grade, and 3 radiomic features) and the model without radiomic features (age and cancer grade) were compared using the likelihood ratio test. Using training and validation datasets, the likelihood ratio test showed a significant difference between the integrated model and clinical model based on both training and testing datasets (likelihood ratio x^2=26.613, p=7.10e-06);. tThe CI was 0.797 for the clinical model and 0.818 for the integrated model, suggesting radiomics features play an important role in improving survival prediction using the CPH model. Based on the testing dataset, the CI was 0.792~~

Commented [ME145]: Do not repeat methodological details in the results paragraph.

Commented [LX146R146]: Deleted.

Commented [ME147]: Only use "significant" / "significance" in the statistical sense throughout the manuscript.

Commented [LX148R148]: Ok, changed.

Commented [ME149]: Do not repeat methodological details in the results paragraph.

Commented [LX150R150]: Deleted.

Commented [ME151]: This sentence is an interpretation and should instead be part of the discussion paragraph.

Commented [LX152R152]: Removed.

For Peer Review

~~for the model with the age and clinical cancer grades, but the index was 0.882 for the integrated model, thus showing a significant difference (likelihood ratio χ^2=12.677, p=0.0054).~~

Decision curve analysis

To study the contribution of radiomic features ~~for the~~to survival time estimation within the ~~framework~~ CHP model, a DCA was applied to compare radiomic, clinical, and integrated models at 500, 1,000, 1,500, and 2,000 days (Figure 7). ~~By considering the clinical utility of the specific model, DCA overcomes the limitations of traditional metrics such as AUC which only measures the diagnostic accuracy of the model.~~ In Figure 7, the net benefit is plotted against the threshold probability. If the curve is closer to the right top corner, then the corresponding model is better as it has larger net benefit. The “all” curve shows the net benefit by treating all patients, while the “none” curve denotes net benefit for treating no patients. The integrated model ~~is~~ was almost consistently on the top of other curves in Figure 7, suggesting that the model has more net benefit than the other models for ~~endometrial cancer~~ survival time prediction. ~~It is interesting to note that t~~The radiomic model ha~~s~~d a larger net benefit than the clinical model when the threshold ~~is~~ was 0.5 (Figures 7B and ~~Figure~~ 7C). For a larger threshold probability (>0.45), radiomic, clinical, and integrated models ha~~ve~~d similar net benefit for a short time range estimation (Figures 7A and ~~Figure~~ 7B). However, ~~the general trend appears to bethere was a trend in that~~ for the long-time range (i.e., 2,000 days) of survival time estimation, the integrated model had~~s~~ a larger net benefit (comparing Figure 7A to Figure 7D, the gap is larger in Figure 7D, suggesting a larger net benefit for 2,000 days~~' estimation~~).

Commented [ME153]: This sentence is an interpretation and should instead be part of the discussion paragraph.

Commented [LX154R154]: Moved to discussion.

Commented [ME155]: This could be condensed, most information would be better suited for the respective figure legend.

Commented [LX156R156]: Removed to figure legend.

Commented [ME157]: Statistical trend? Report the related non-significant p-value.

Commented [LX158R158]: It is not a trend.

For Peer Review

Features from different scanners

~~A diagnostic analysis was carried out to study the feature variation obtained from different types of scanners. All features were normalized using a Z-score method (similar to Eq. 1, except multiply 100), and then a principal component analysis method was employed to split the feature dataset into different components. Four components were used to compare the feature variations from different scanners.~~

~~Three different manufacturers, i.e., GE (108 cases), Philips (163 cases), and Siemens (142 cases) were employed to collect theused to acquire sagittal T2-weighted MRI image (Table 1). In Figure 8, the value of coordinate axis PC1/2 (Figure 8A) and PC3/4 (Figure 8B) are the explanatory rates of the overall difference. The dots in the figure represent samples; the colors represent groups (scanner types); and the legends have three groups at the top. The ellipse represents the core area added by the default confidence interval of 68%, which facilitates the separation between the observation groups. No clear separation of the sample based on the three scan machineMRI vendors types is seen was observed. Features from different scanners overlaid onto each other, suggesting similarities in the feature distribution, although in Figure 8A, radiomic features from the Siemens scanners hasd a relatively larger variation. For the PC3/4 (Figure 8B), the scanners from 3 three different manufacturers are were much smaller, as there is was a very good overlay between different scanners.~~

From PCA analysis, features components from different scanners were overlaid onto each other in Figure 8. Similarities in the feature distribution was observed, although in Figure 8A, radiomic features from the Siemens scanners had a larger variation. For the

Commented [ME159]: This section is not covered by the methods section, neither regarding the assessment nor the statistical analyses. It is required to include info on this evaluation early in the methods section – before related results can be reported here.

Commented [LX160R160]: Merged into method section.

Commented [ME161]: Not part of the results, needs to be integrated in the methods section.

Commented [LX162R162]: Moved to method section.

Commented [ME163]: Unclear what PC stands for. Please clarify.

Commented [LX164R164]: Principal component, added in the clean version.

Commented [ME165]: This should be removed from the text here and integrated in the respective figure legend.

Commented [LX166R166]: Moved to the legend.

Commented [ME167]: Unclear sentence – please revise.

Commented [LX168R168]: Removed.

For Peer Review

3rd and 4th principal components (PC3/4) (Figure 8B), the distribution of the radiomic features obtained from different scanners were smaller, suggesting good agreement for the features from different scanners.

For Peer Review

Discussion

We have developed the CPH model using features from sagittal T2-weighted MRI and clinical variables for survival time estimation based on gLASSO method. We studied the effect of the radiomic features within the model and found radiomic features from MRI are useful biomarkers to predict survival time in patients with endometrial cancer.

~~In this study, w~~We identified a set of radiomic signatures using pelvic MRI that ~~can~~ could potentially aid ~~in~~ to accurately predict~~ing~~ survival time for patients with endometrial cancer. In combination with clinical features, our integrated radiomics model outperformed the clinical model in predicting survival time. We validated and compared the integrated and clinical models using both internal (training and validation ~~datasets~~) and external (testing) datasets. Furthermore, the CPH model with a nomogram for visualization provide~~s~~d an easy, straightforward, and non-invasive method of predicted survival, ~~that can~~which could be used in clinical settings and therefore has potential to facilitate personalized medicine. The multiple centers and scan machines used in this study presented challenges for model building, but ~~have meant~~this setup also implies that the findings are likely to be ~~relatively~~ generalizable. ~~In addition, this is one of the largest radiomics model studies in endometrial cancer with over 400 patients, and the largest study to explore whether radiomic features in endometrial cancer can predict survival time based on CPH model.~~ Multiple modelling techniques were evaluated, and feature selection was utilized to avoid overfitting of the model. ~~Most importantly, t~~The radiomics quality score which determines the validity and completeness of radiomics studies, and transparent reporting of a multivariable prediction model for individual prognosis or diagnosis guidelines were

Commented [ME169]: It is recommended that the first paragraph of the discussion section highlights all the major findings of this study in brief. Thus, the first paragraph should not include references or further interpretations, which can be included in the subsequent paragraphs of the discussion. Please modify accordingly.

Commented [LX170R170]: Added a new paragraph now.

Commented [ME171]: Unclear what the radiomics quality score is – please define.

Commented [LX172R172]: Add "which determines the validity and completeness of radiomics studies, "

For Peer Review

adhered to ensure quality of both scientific methods and reporting (30). ~~Finally, the integrated model was validated externally based on an independent testing dataset.~~ Different from quantitative MRI such as apparent diffusion coefficient (ADC) from diffusion-weighted imaging, T2-weighted MRI signal depends on the acquisition protocol, the coil profile, the scanner type, etc. and there is not standard method to normalize the image intensity for cross-subjects comparison. We adopted z-score like method to normalize the image intensity, other methods such as min/max normalization or scaling the image intensity to common max value can also be used. An alternative method to reduce the image intensity difference of T2-weighted images acquired from different centers and scanners is to normalize to a reference tissue outside the tumor-affected region such as cerebrospinal fluid (CSF) in brain or bladder where baseline water signal can be obtained. Although the image intensity features such as mean intensity value within the tumor mask will be affected by different image normalization steps, the tumor shape, volume, and image complexity radiomic features will not be affected by the image normalization step.

Most of the ~~current~~ published studies have focused on addressing the classification problem in endometrial cancer ~~study~~ using radiomics (20). ~~Few s~~Studies have applied ~~yet to apply~~ this radiomic technology to endometrial cancer survival prediction models. Fasmer et al.~~,~~ developed an MRI-based whole-volume tumor radiomic signature for the prediction of high-risk features (20). Radiomic features were studied to predict poor progression~~-~~-free survival (20). Meanwhile, Ytre-~~h~~Hauge et al.~~,~~ applied radiomics to study survival in women with endometrial cancer (18). They reported that high kurtosis in T1-

Commented [ME173]: Sentence refers to "studies" (plural), but only one reference provided. Please add references.

Commented [LX174R174]: (17-19) were added in the clean version.

Commented [ME175]: Include references for those "few studies" here.

Commented [LX176R176]: References (32,33) were added.

For Peer Review

weighted MRI~~ images~~ predicted reduced recurrence and progression-free survival (hazard ratio 1.5~~, p< 0.001~~), but their study used only 180 patients without model validation, compared to 495 cases in this study. Furthermore, they used T1~~-~~-weighted MRI images~~ (not the current gold standard (9) for endometrial cancer study)~~ (17). However, we obtained the features from T2-weighted image, which is a sequence that clearly delineates most endometrial cancers without the use of gadolinium and the sagittal T2-weighted sequence is the mainstay of MRI protocols for staging endometrial cancer, allowing the development of a generalizable tool. We found that tumor size reflected by minor axis radius was an important biomarker for survival time estimation. The minor axis radius describes the radius of the minor axis of the ellipse that reflects the tumor region indirectly (27). For the features of GLSZM and FOS, both features are related to the distribution of image intensities. The GLSZM was based on the image converted from Hounsfield unit, this could due the relation between MRI intensity and Hounsfield unit values (34).

Commented [ME177]: Add reference to this previous study again here.

Commented [LX178R178]: Added.

The difference between clinical model and integrated model is small in terms of CI using training and validation dataset, however, for the independent testing data, the CI was 0.792 for the model with age and clinical cancer grades, and the index was 0.882 for the integrated model, thus showing a significant difference (likelihood-ratio x^2=12.677). This suggested that the integrated model is robust to the different distributions of the data because there is significant different between internal (training/validation) data and external dataset. In addition to use CI, we adopted multiple criteria to evaluate the models. We have applied likelihood-ratio test, AUC, and DCA methods to compare different models. By considering the clinical utility of the specific model, DCA overcomes the

For Peer Review

limitations of traditional metrics such as AUC which only measures the diagnostic accuracy of the model.

The pipeline of this study (Figure 2) can be extended to other malignancies for survival analysis based on integrated features. In this method, the sources of error can come from the first 3 steps, i.e., image segmentation, image processing, and feature extraction. For example, in the image segmentation step, the error can be generated if the tumor mask is not parcellated properly. For the image processing step, the image interpolation method can introduce numerical error. In the feature extraction step, bias can be produced if only a fraction of image feature were extracted from the image.

Lastly, the clinical application of the nomograms could be patient management or prioritization, as the survival time of the patient is known from the model estimation, so treatment for patients could be arranged in a more efficient way. The integrated radiomics model may also enable better stratification of patients enrolling into clinical trials, as it has higher AUC and CI value than the model with only clinical variables.

.

Commented [ME179]: Add reference to this previous study again here.

For Peer Review

Limitations

~~This study had several limitations. Firstly, it~~This was a retrospective study and therefore, there was ~~at~~a risk of bias and missing data. The study also only included patients who had received surgery and had MRI ~~imaging~~ with paired clinical data available. Whilst the model was assessed based on the external testing dataset, there was slight variation in demographics when directly comparing the training and validation datasets. In the testing dataset, there were less women in the older 59~~–~~to 70-year group and more women with endometrioid low grade cases,~~;~~ namely more ~~'low risk' score~~ patients with low risk scores ~~in this dataset~~. Debatably, this would infer that the testing dataset group would be more likely to have better survival. As radiomics models perform better with more homogenous datasets such as that generated by the low-risk cases, this may explain the slightly better performance with the testing dataset. Secondly, although we had also tested the composite minimax concave penalty method (31) for the model selection in the CPH model, other methods such as the regular elastic net and ridge models method (32), which may produce better results, have not been investigated in this study. Thirdly, survival outcomes do not only represent ~~not only~~ the effect of the disease itself, but also of patient factors (such as age or~~,~~ co-morbidities) and treatment factors (such as whether the patient underwent neoadjuvant radiotherapy or chemotherapy). Studies have show~~ed~~n that adjuvant treatment predictably improves survival for high-~~-~~risk patients (33). Regional and national differences in patient demographics along with treatment options offered and delivered ~~will~~ can also impact survival disparities. This study did not consider co-morbidities, which are likely to have a ~~significant~~ relevant impact on survival. Finally, we studied the feature differences based on MRI systems from three different manufacturers, but we did not investigate the influence ~~from~~ of different sequences~~, as this is a multi-center with different image sequences~~. Finally, a~~A~~lthough we normalized the images to minimize the T2-~~-~~weighted image differences obtained from different protocols, more work is needed to study ~~this~~ effects on features for radiomics studies.

~~Furthermore, In this study~~This study ~~only~~employed only T2-weighted MRI data ~~images~~ were used~~;~~, and future work could evaluate the use of alternative MRI sequences such as~~:~~ diffusion~~-~~-weighted images with~~,~~ apparent diffusion coefficient maps, ~~and~~ as well as

Commented [ME180]: Only use "significant" / "significance" in the statistical sense throughout the manuscript.

Commented [LX181R181]: Significant was removed.

For Peer Review

dynamic contrast~~ material-~~-enhanced images (34). Also, a~~n~~ possible ~~avenue~~ method to explore in the future would be the boosting method (35) or the deep survival model (36) for the ~~survival~~ study of survival~~. It would be interesting to see if the radiomic signature reported in our study identified could identify the same subsets and survival time with other methods~~ because these methods do not require the proportional hazards assumption (39,40).

For Peer Review

Conclusion~~s~~

~~Overall, t~~The ~~enhanced~~ integrated radiomic model and ~~in particular~~ the nomogram~~,~~ may enable~~s~~ prediction of survival with a high degree of accuracy. ~~The nomograms have potential be used in clinic or developed into a phone application to enable true personalized medicine. Another clinical application could be patient management or prioritization, as the survival time of the patient is known from the model estimation, so treatment for patients could be arranged in a more efficient way. The integrated radiomics model may also enable better stratification of patients enrolling into clinical trials, as it has higher AUC and CI value than the model with only clinical variables.~~ ~~Also~~Furthermore, based on the testing dataset, we found that the integrated model is robust against the variability of the independent external testing dataset, as the AUC value showed only a marginal ~~little~~ decrease, while for the clinical model~~,~~ the AUC decreased ~~dramatically~~markedly.

Commented [ME182]: Retrospective studies show associations.
Remove the words "predict" / "prediction" throughout when related to the findings of your study.

Commented [LX183R183]: Replaced predict with estimate.

Commented [ME184]: Provide a conclusion that is based only on the results shown. Avoid any projections or interpretations; those can be part of the discussion paragraph.

Commented [LX185R185]: Ok,

For Peer Review

References

1. International Agency for Research on Cancer G. Corpus uteri factsheet, estimated cancer incidence, mortality and prevalence worldwide 2018. World Health Organisation; 2018.
2. Crosbie EJ, Kitson SJ, McAlpine JN, Mukhopadhyay A, Powell ME, Singh N. Endometrial cancer. The Lancet 2022;399(10333):1412-1428.
3. Morice P, Leary A, Creutzberg C, Abu-Rustum N, Darai E. Endometrial cancer. The Lancet 2016;387(10023):1094-1108.
4. Coronado PJ, Rychlik A, Baquedano L, et al. Survival Analysis in Endometrial Carcinomas by Type of Surgical Approach: A Matched-Pair Study. Cancers (Basel) 2022;14(4).
5. Odagiri T, Watari H, Hosaka M, et al. Multivariate survival analysis of the patients with recurrent endometrial cancer. jgo 2011;22(1):3-8.
6. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer 2003;89(2):232-238.
7. Bewick V, Cheek L, Ball J. Statistics review 12: survival analysis. Crit Care 2004;8(5):389-394.
8. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American statistical association 1958;53(282):457-481.
9. Hoivik EA, Hodneland E, Dybvik JA, et al. A radiogenomics application for prognostic profiling of endometrial cancer. Communications Biology 2021;4(1):1363.

Commented [ME186]: As per JMRI reference style, all in-text reference numbers need to be between parentheses () and combined within () – e.g. (3-5) here instead of (3) (4) (5). Please modify reference style throughout.

Commented [LX187R187]: Changed.

For Peer Review

10. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 1972;34(2):187-202.

11. de Boer SM, Powell ME, Mileshkin L, et al. Adjuvant chemoradiotherapy versus radiotherapy alone in women with high-risk endometrial cancer (PORTEC-3): patterns of recurrence and post-hoc survival analysis of a randomised phase 3 trial. The Lancet Oncology 2019;20(9):1273-1285.

12. Uharček P. Prognostic factors in endometrial carcinoma. Journal of Obstetrics and Gynaecology Research 2008;34(5):776-783.

13. Njoku K, Barr CE, Crosbie EJ. Current and Emerging Prognostic Biomarkers in Endometrial Cancer. Frontiers in oncology. Volume 12; 2022. p. 890908.

14. Madison T, Schottenfeld D, James SA, Schwartz AG, Gruber SB. Endometrial cancer: socioeconomic status and racial/ethnic differences in stage at diagnosis, treatment, and survival. American journal of public health 2004;94(12):2104-2111.

15. Morielli AR, Kokts-Porietis RL, Benham JL, et al. Associations of insulin resistance and inflammatory biomarkers with endometrial cancer survival: The Alberta endometrial cancer cohort study. Cancer Medicine 2022;11(7):1701-1711.

16. Caruso D, Polici M, Zerunian M, et al. Radiomics in Oncology, Part 2: Thoracic, Genito-Urinary, Breast, Neurological, Hematologic and Musculoskeletal Applications. Cancers 2021;13(11):2681.

For Peer Review

17. Michalet M, Azria D, Tardieu M, Tibermacine H, Nougaret S. Radiomics in radiation oncology for gynecological malignancies: a review of literature. The British Journal of Radiology 2021;94(1125):20210032.

18. Ytre-Hauge S, Dybvik JA, Lundervold A, et al. Preoperative tumor texture analysis on MRI predicts high-risk disease and reduced survival in endometrial cancer. Journal of magnetic resonance imaging : JMRI 2018;48(6):1637-1647.

19. Jacob H, Dybvik JA, Ytre-Hauge S, et al. An MRI-Based Radiomic Prognostic Index Predicts Poor Outcome and Specific Genetic Alterations in Endometrial Cancer. Journal of clinical medicine 2021;10(3).

20. Fasmer KE, Hodneland E, Dybvik JA, et al. Whole-Volume Tumor MRI Radiomics for Prognostic Modeling in Endometrial Cancer. Journal of magnetic resonance imaging : JMRI 2021;53(3):928-937.

21. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PloS one 2019;14(11):e0224365.

22. Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 2018;180:68-77.

23. Soneji ND, Bharwani N, Ferri A, Stewart V, Rockall A. Pre-operative MRI staging of endometrial cancer in a multicentre cancer network: can we match single centre study results? European radiology 2018;28(11):4725-4734.

24. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310-1320.

For Peer Review

25. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research 2017;77(21):e104-e107.

26. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. PeerJ (San Francisco, CA) 2014;2:e453-e453.

27. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological) 1972;34(2):187-220.

28. Breheny P. The group exponential lasso for bi-level variable selection. Biometrics 2015;71(3):731-740.

29. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research 2019;3(1):18.

30. Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. European radiology 2020;30(1):523-536.

31. Breheny P, Huang J. Penalized methods for bi-level variable selection. Statistics and its interface 2009;2(3):369.

32. Van De Wiel MA, Lien TG, Verlaat W, van Wieringen WN, Wilting SM. Better prediction by use of co-data: adaptive group-regularized ridge regression. Statistics in Medicine 2016;35(3):368-381.

33. Son J, Chambers LM, Carr C, et al. Adjuvant treatment improves overall survival in women with high-intermediate risk early-stage endometrial cancer with

For Peer Review

lymphovascular space invasion. International Journal of Gynecologic Cancer 2020;30(11):1738-1747.

34. Ueno Y, Forghani B, Forghani R, et al. Endometrial Carcinoma: MR Imaging-based Texture Model for Preoperative Risk Stratification-A Preliminary Analysis. Radiology 2017;284(3):748-757.

35. De Bin R. Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. Computational Statistics 2016;31(2):513-531.

36. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology 2018;18(1):24.

For Peer Review

TABLES

Table 1. Training and testing patient demographics

Clinical information	n(training)	%(training)	n(testing)	%(testing)	**P value**
Age: mean (SD)	**64±11.5**		**63±12.4**		**0.024**
under 50	29±5.9	7	11±4.6	13.4	0.79
50-59	78±2.5	18.9	24±2.2	29.3	0.55
60-69	134±2.7	32.4	15±2.5	18.3	0.18
70 and older	172±5.4	41.6	32±5.9	39.0	0.18
Histological type					**0.0011**
Endometrioid	304	73.6	69	84.1	
Carcinosarcoma	44	10.7	1	1.2	
Serous	39	9.4	4	4.9	
Clear cell	18	4.4	2	2.4	
Mixed high grade	7	1.7	2	1.7	
Undifferentiated	1	0.2	3	3.7	
NET small cell			1	1.2	
Grade					**3.72e-04**
1	124	30.0	43	52.4	
2	130	31.5	20	24.4	
3	159	38.5	19	23.2	

Commented [ME188]: In the table footnote, define all abbreviations used in the table (e.g., SD etc.).

Commented [LX189R189]: Included.

Commented [ME190]: Make clear in the table legend what grades and stages are reported in this table (related to which grading/staging systems?).

Commented [LX191R191]: IT is FIGO system, included now.

For Peer Review

Overall stage					**0.1738**
Stage I	292	70.7	59	72	
➢ IA	199	48.2	45	54.9	
➢ IB	93	22.5	14	17.1	
Stage II	31	7.5	5	6.1	
Stage III	64	15.5	7	8.5	
➢ IIIA	18	4.4	4	4.9	
➢ IIIB	6	1.4	0	0	
➢ IIIC	40	9.7	3	3.7	
Stage IV	25	6.1	1	1.2	
➢ IVA	18	4.4	0	0	
➢ IVB	7	1.7	1	1.2	
Other (missing)	1	0.2	1	1.2	
Clinical risk score					**0.0388**
Low	150	36.3	41	50	
Intermediate	78	18.9	15	18.3	
High	96	23.2	9	11.0	
Advanced	89	21.5	16	19.5	
Unknown			1	1.2	
Censored					**0.0096**

For Peer Review

Censoring	317	76.8	74	90.2	
Death	96	23.2	8	9.8	
MRI manufacturer					
GE	108	26.2			
Philips	163	39.5			
Siemens	142	34.2			

Note. N = 413 (training), *N* = 82 (testing).

Commented [ME192]: In the table footnote, define all abbreviations used in the table (e.g., SD etc.).

Commented [LX193R193]: Included now.

For Peer Review

Figure Legends

Figure 1: Flow chart of patient selection. After exclusion, 413 cases were included and used to generate the final model. Eighty-two cases were used as external testing dataset.

Figure 2: Histograms of the survival data for all training/validation (A) and testing (D) datasets and histograms of the censoring and the noncensoring datasets. The noncensoring data (B) and the right censoring data (C) distributions from the training dataset. The noncensoring data (E) and the right censoring data (F) distributions from the testing dataset.

Figure 3: Pipeline for the study. Five steps were included as shown in the column. LR: likelihood test; gLASSO: group Least Absolute Shrinkage and Selection Operator; CPH: Cox proportional hazards model; DCA: decision curve analysis.

Figure 4: Group least absolute shrinkage and selection operator (LASSO) or feature selection. A: 10-fold cross-validated error rates for the model selection. B: amplified version of Figure 4A at the optimal lambda value. The vertical dotted lines indicate the minimum error, and the top of the plot gives the size of each model.

Figure 5: A: Time-dependent AUC summary at evaluation time points from the validation dataset; the AUC values are within the range of 0.5 and 0.9. The AUC for the integrated model (red curve) is consistently larger than the AUC obtained from using clinical model

For Peer Review

(green curve). B: AUC obtained from the testing data; AUC values are within the range between 0.5 and 0.85.

Figure 6: Nomogram visualization for the survival time prediction. At the top of the nomogram, a point scale was included. Beneath the scale, 3 radiomic features, age, and the clinical cancer grade were displayed. The refitted CPH model was adopted to predict the survival probability for 1 (365 days), 2 (730 days), 3 (1095 days) and 5 (1826 days) year periods as shown at the bottom of the Figure 6. The dotted red vertical line in the figure indicates one example of observation with an age of 62, cancer grade of 1, minor axis length of 0.51, Grey-Level Size Zone (GLSZM) of 0.41, and First Order Statistics (FOS) of -0.1. The aggregate score for this case is 302 as indicated by the red arrow vertical line at the bottom of the figure. The corresponding probability to the survival for the 5-year, 3-year, 2-year, and 1 year periods is 0.93 (1-0.0736), 0.96, 0.975, and 0.99, respectively.

Figure 7: Decision curve analysis at 500(A), 1000(B), 1500(C), and 2000(D) days.

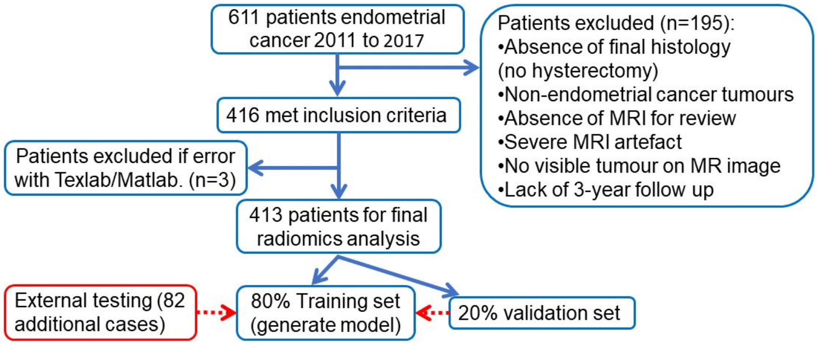
Figure 8: Scanner difference study. Principal component analysis for radiomics features from a different scanner. A: First principal component vs second principal component; a relatively larger variation was observed using the Siemens scanner. B: Third principal component and fourth principal component explain smaller percentages of the total variation, and the three different scanners show good agreement. var.: variance.

For Peer Review

~~Figures~~

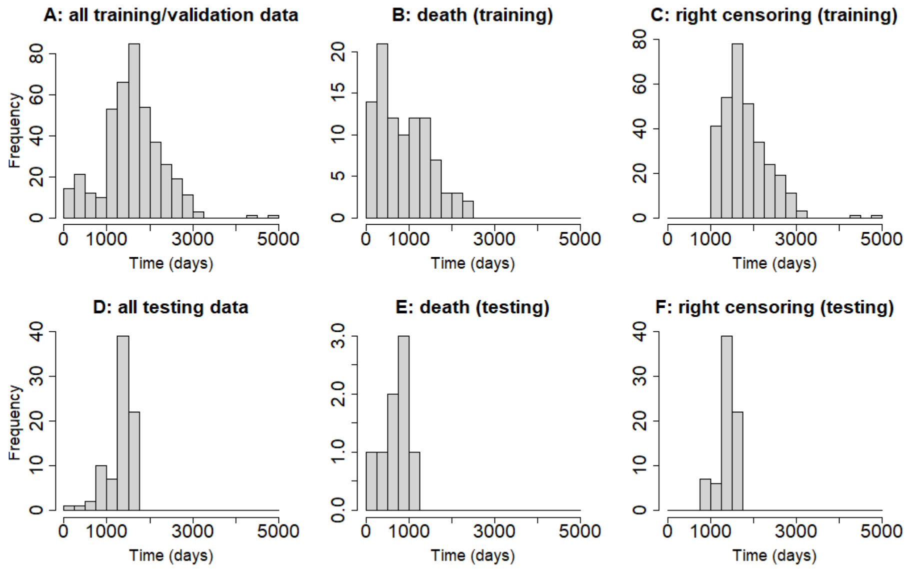
Commented [ME194]: Enough to submit the figures as separate files, no need to also include here and repeat the figure legends.

Commented [LX195R195]: ok


611 patients endometrial cancer 2011 to 2017
Patients excluded (n=195):
•Absence of final histology (no hysterectomy)
•Non-endometrial cancer tumours
•Absence of MRI for review
•Severe MRI artefact
•No visible tumour on MR image
•Lack of 3-year follow up
416 met inclusion criteria
Patients excluded if error with Texlab/Matlab. (n=3)
413 patients for final radiomics analysis
External testing (82 additional cases)
80% Training set (generate model)
20% validation set

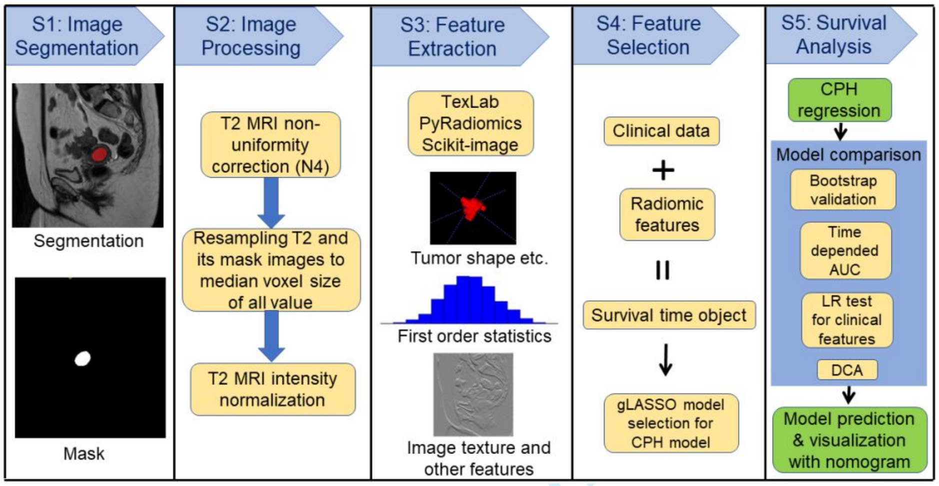
~~**Figure 1:** Flow chart of patient selection. After exclusion, 413 cases were included and used to generate the final model. Eighty-two cases were used as external testing dataset.~~

For Peer Review



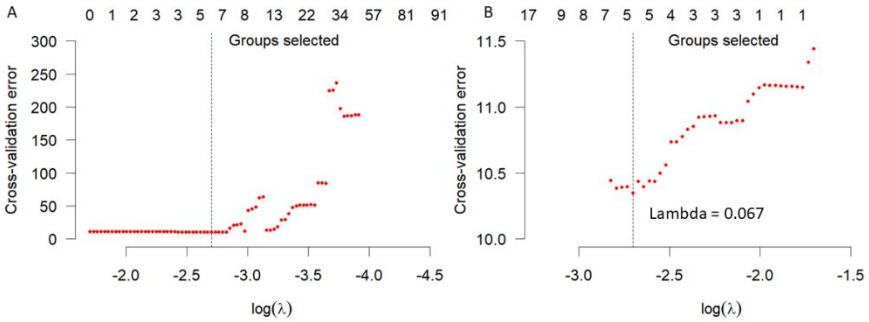
~~**Figure 2:** Histograms of the survival data for all training/validation (A) and testing (D) datasets and histograms of the censoring and the noncensoring datasets. The noncensoring data (B) and the right censoring data (C) distributions from the training dataset. The noncensoring data (E) and the right censoring data (F) distributions from the testing dataset.~~

Review


S1: Image Segmentation
Segmentation
Mask
S2: Image Processing
T2 MRI non-uniformity correction (N4)
Resampling T2 and its mask images to median voxel size of all value
T2 MRI intensity normalization
S3: Feature Extraction
TexLab PyRadiomics Scikit-image
Tumor shape etc.
First order statistics
Image texture and other features
S4: Feature Selection
Clinical data
+
Radiomic features
=
Survival time object
gLASSO model selection for CPH model
S5: Survival Analysis
CPH regression
Model comparison
Bootstrap validation
Time depended AUC
LR test for clinical features
DCA
Model prediction & visualization with nomogram

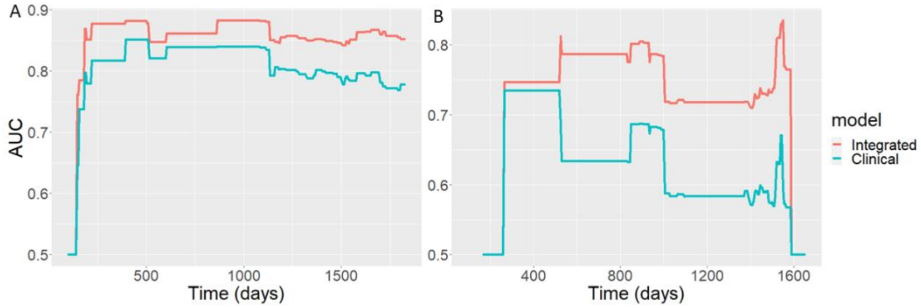
~~**Figure 3:** Pipeline for the study. Five steps were included as shown in the column. LR: likelihood test; gLASSO: group Least Absolute Shrinkage and Selection Operator; CPH: Cox proportional hazards model; DCA: decision curve analysis.~~

er Review



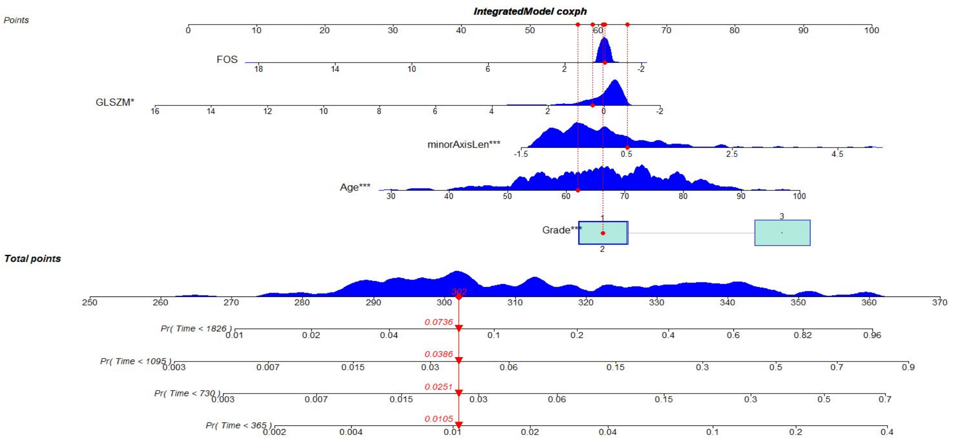
~~**Figure 4:** Group least absolute shrinkage and selection operator (LASSO) or feature selection. A: 10-fold cross-validated error rates for the model selection. B: amplified version of Figure 4A at the optimal lambda value. The vertical dotted lines indicate the minimum error, and the top of the plot gives the size of each model.~~

For Peer Review



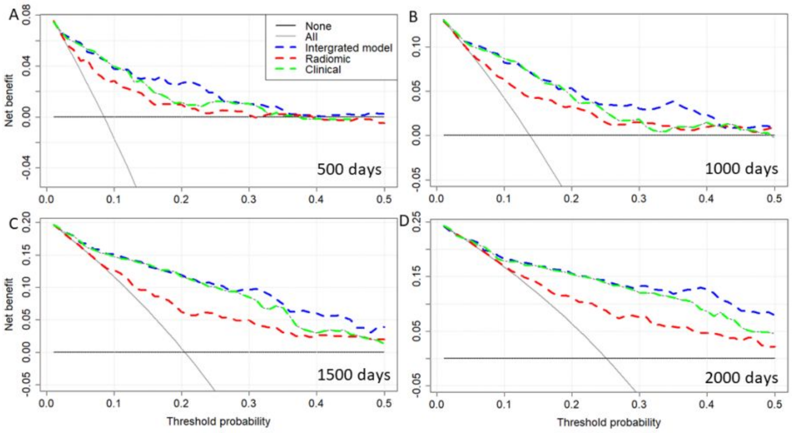
~~**Figure 5:** A: Time-dependent AUC summary at evaluation time points from the validation dataset; the AUC values are within the range of 0.5 and 0.9. The AUC for the integrated model (red curve) is consistently larger than the AUC obtained from using clinical model (green curve). B: AUC obtained from the testing data; AUC values are within the range between 0.5 and 0.85.~~

For Peer Review



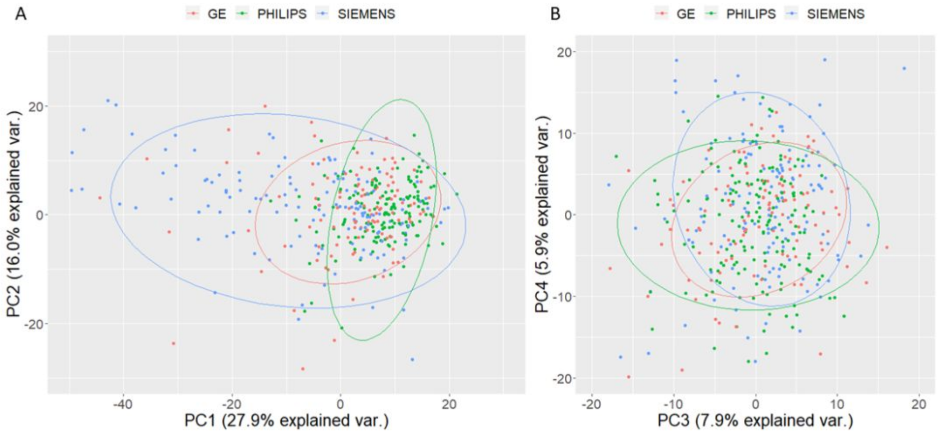
~~**Figure 6:** Nomogram visualization for the survival time prediction. At the top of the nomogram, a point scale was included. Beneath the scale, 3 radiomic features, age, and the clinical cancer grade were displayed. The refitted CPH model was adopted to predict the survival probability for 1 (365 days), 2 (730 days), 3 (1095 days) and 5 (1826 days) year periods as shown at the bottom of the Figure 6. The dotted red vertical line in the figure indicates one example of observation with an age of 62, cancer grade of 1, minor axis length of 0.51, Grey-Level Size Zone (GLSZM) of 0.41, and First Order Statistics (FOS) of -0.1. The aggregate score for this case is 302 as indicated by the red arrow vertical line at the bottom of the figure. The corresponding probability to the survival for the 5-year, 3-year, 2-year, and 1 year periods is 0.93 (1-0.0736), 0.96, 0.975, and 0.99, respectively.~~

er Review



~~**Figure7:** Decision curve analysis at 500(A), 1000(B), 1500(C), and 2000(D) days.~~

eer Review



~~**Figure 8:** Scanner difference study. Principal component analysis for radiomics features from a different scanner. A: First principal component vs second principal component; a relatively larger variation was observed using the Siemens scanner. B: Third principal component and fourth principal component explain smaller percentages of the total variation, and the three different scanners show good agreement. var.: variance. PC: Principal component.~~

er Review