

Title

Radiomic machine learning for pretreatment assessment of prognostic risk factors for endometrial cancer and its effects on radiologists' decisions of deep myometrial invasion

Abstract

Purpose

To evaluate radiomic machine learning (ML) classifiers based on multiparametric magnetic resonance images (MRI) in pretreatment assessment of endometrial cancer (EC) risk factors and to examine effects on radiologists' interpretation of deep myometrial invasion (dMI).

Methods

This retrospective study examined 200 consecutive patients with EC during January 2004 –March 2017, divided randomly to Discovery ($n=150$) and Test ($n=50$) datasets. Radiomic features of tumors were extracted from T2-weighted images, apparent diffusion coefficient map, and contrast enhanced T1-weighted images. Using the Discovery dataset, feature selection and hyperparameter tuning for XGBoost were performed. Ten classifiers were built to predict dMI, histological grade, lymphovascular invasion (LVI), and pelvic/paraortic lymph node metastasis (PLNM/PALNM), respectively. Using the Test dataset, the diagnostic performances of ten classifiers were assessed by the area under the receiver operator characteristic curve (AUC). Next, four radiologists assessed dMI independently using MRI with a Likert scale before and after referring to inference of the ML classifier for the Test dataset. Then, AUCs obtained before and after reference were compared.

Results

In the Test dataset, mean AUC of ML classifiers for dMI, histological grade, LVI, PLNM, and PALNM were 0.83, 0.77, 0.81, 0.72, and 0.82. AUCs of all radiologists for

dMI (0.83–0.88) were better than or equal to mean AUC of the ML classifier, which showed no statistically significant difference before and after the reference.

Conclusion

Radiomic classifiers showed promise for pretreatment assessment of EC risk factors. Radiologists' inferences outperformed the ML classifier for dMI and showed no improvement by review.

Keywords:

Radiomic machine learning; endometrial cancer.

1. Introduction

Endometrial cancer (EC) is the most frequent malignancy in gynecology in industrialized countries, with increasing incidence worldwide [1, 2]. Although prognoses of EC are predominantly favorable, 15–20% of cases show recurrence or metastasis into an aggressive course [1, 3, 4]. Multiple prognostic factors have been reported including the International Federation of Gynecology and Obstetrics (FIGO) stage, deep myometrial invasion (dMI; invasion to one-half or more of myometrium), the FIGO histological grading system, lymphovascular invasion (LVI), and pelvic/para-aortic lymph node metastasis (PLNM/PALNM) [3, 5-12]. Unfortunately, these factors can only be assessed after invasive procedures and pathological examinations. Dilation and curettage (D&C) and endometrial biopsy can yield some information about the histopathological grade, but that information might differ from the final results because of small specimen amounts and intratumoral heterogeneity [13, 14].

Radiomics is a noninvasive method to extract quantitative features from medical images to build a predictive model with machine learning (ML) techniques [15-17]. To date, many studies have examined cancer of various types. Radiomics is expected to provide relevant information such as prognostic factors and clinical outcomes [18-21]. For EC, radiomic approaches with MRI have been explored for

various risk factors: dMI, LVI, histological grade, risk categories (high-risk vs low-risk EC), PLNM, and prognosis [22-29]. These radiomics research papers have described promising results with excellent diagnostic performance. As cautioned these days in radiomic and ML studies [30, 31], ML models with a large number of explanatory variables easily overfit to the data. Thus, rigorous validation with a test dataset is critical for radiomic ML models. Additionally, clinical features including age and tumor markers play important roles in the risk assessments of EC [3, 5]. The combination of radiomic and clinical features might harness the potential.

For assessment of dMI, MRI is regarded as the most accurate imaging modality. Reportedly, the qualitative assessment of dMI can be subjective, with widely varying accuracy: 62%–90% [32]. Ueno et al. reported diagnostic accuracy using a random forest model based on imaging features of MRI was equivalent to that of three board-certified radiologists specializing in gynecologic imaging [22]. Subsequent radiomic studies also reported very good diagnostic performance for dMI [25, 27]. Still, there was limited evidence showing improvement of radiologists' interpretation of dMI with the assistance of radiomic ML classifiers [24].

This study was conducted to evaluate the performance of radiomic ML classifiers for pretreatment assessment of risk factors comprehensively in combination

with clinical features: dMI, histological grade, LVI, PLNM/PALNM based on pretreatment multiparametric MRI with a held-out test dataset to avoid overfitting. We also tested whether radiologists' interpretations of dMI using MRI would be improved with the assistance of a ML classifier.

2. Material and Methods

2.1. Patient cohort

The institutional review board of our hospital approved this retrospective study (#R1356). Informed consent of participants was waived. From our institutional database for January 2004 to March 2017, we extracted 334 patients with pathological diagnosis of EC and its surgical staging who underwent pretreatment MRI including T2-weighted image (T2WI), diffusion-weighted image (DWI), and dynamic contrast enhanced (DCE) MRI at our institution. Among them, 134 patients were excluded for one of the following reasons: 1) tumor affected by motion or metal artifact (consensus achieved by two radiologists) (n= 22); 2) tumor not recognized on MRI (n=43) or too small for segmentation (n=15); 3) omission of lymphadenectomy and no confirmation of lymph node status available (patients performed palliative surgery (n=17) or early stage patients without follow-up computed tomography (CT) imaging over two years (n=14));

4) any earlier treatment for EC before MRI (D&C (n=2), hormone therapy (n=7), or neoadjuvant chemotherapy (n=14)). The final cohort comprised 200 patients (Fig. 1).

Pathological reports of a final surgical specimen rendered by board-certified pathologists served as the reference standard for MI, LVI, and, if inferred from results of surgery, PLNM and PALNM. If PLNM or PALNM was not surgically evaluated, then the status was confirmed by a patient's clinical records and follow-up CT within two years.

2.2. Imaging techniques

All MR studies were performed at our institution using 3.0-tesla (Trio or Skyra; Siemens Healthineers, Erlangen, Germany) or 1.5-tesla (Symphony or Avanto; Siemens Healthineers, Erlangen, Germany) MRI systems with a phased-array body coil. Before examinations, 20 mg of butyl scopolamine (Buscopan®; Nippon Boehringer Ingelheim Co. Ltd., Tokyo, Japan) was administered to reduce bowel motion, unless contraindicated. The MR protocol for EC included three planes of T2-weighted images (T2WI; axial, sagittal, and oblique axial for uterine corpus), axial and sagittal T1-weighted images (T1WI) with/without fat suppression, axial or sagittal diffusion weighted images (DWI), and axial or sagittal DCE T1WI. The parameters were the

following: for T2WI — 3730–7760/81–120 ms repetition time (TR)/echo time (TE),
 260 mm × 260 mm field of view (FOV), 4–5 mm slice thickness, 448–512 × 204–512
 matrix size, 150° flip angle (FA), and 140–370 Hz/pixel bandwidth, one or two
 averages, acquisition times of 50–262 seconds; for T1WI — 400–655/11–30 ms TR/TE,
 260 mm × 260 mm FOV, 4–6 mm slice thickness, 527 × 224–348 matrix size, 80° FA,
 and 100–140 Hz/pixel bandwidth, one or two averages, acquisition times of 95–381
 seconds; for DWI — TR/TE 4800–5300/59–75 ms at 3.0-tesla and 2300–3000/70–79 ms
 at 1.5-tesla, 260 × 195–260 mm FOV, 4–5 mm slice thickness, matrix size 128 × 128 at
 3.0-tesla and 128 × 90 at 1.5-tesla, 90° FA, and, 1445–2170 Hz/pixel bandwidth, one–
 six averages, acquisition times of 94–302 seconds. The number of diffusion directions
 was three. The b-values of DWI were 0, 500, and 1000 s/mm² from January 2004 to
 March 2009 and 0, 100, 500, and 1000 s/mm² from April 2009. The ADC maps were
 generated automatically using a mono-exponential decay model with all three or four b-
 values. Fat suppression technique was spectral adiabatic inversion recovery (SPAIR) at
 3.0-tesla and chemical shift selective (CHESS) at 1.5-tesla. DCE three-dimensional T1-
 weighted images on sagittal or axial planes were obtained at 20, 40, 60, 80, 100, 120,
 and 180 s after intravenous bolus injection of 0.2 mmol/kg gadolinium contrast agent
 (Magnevist; Bayer Yakuhin, Ltd., Osaka, Japan): 4.696/2.268 TR/TE, 190–260 × 260

mm FOV, 5 mm slice thickness, 320–384× 198–230 matrix size, 10° FA, and 580 Hz/pixel bandwidth.

2.3. Tumor segmentation and radiomic feature extraction

Tumor segmentation and radiomic feature extraction were performed using open-source software (LIFEx ver. 4.90; Inserm, Orsay, France) [33]. Tumors were segmented three-dimensionally by any of three radiologists with expertise in gynecologic imaging (**, 20-year experience; **, 10-year experience; **, 7-year experience), referring to histopathological results for accurate tumor border.

In DCE T1WI, the image acquired at 180 s after injection was adopted as contrast enhanced (CE)-T1WI because the tumor border was clearly recognized at the phase. On T2WI and CE-T1WI, tumors were segmented in all slices. Considering the susceptibility of ADC map to artifacts by intestinal gas, three continuous slices of tumor without artifact were selected. Any questions arose during segmentation were resolved via discussion; consensus was obtained through agreement between at least two of the three radiologists. An example of segmentation is presented in Figure 2.

The LIFEx software parameters for extraction of radiomic features were the following: spatial resampling, $1.0 \times 1.0 \times 1.0$ mm; intensity discretization, 128 as the

number of gray levels; size of bins, 64 with the absolute minimum and maximum values (400–1400) in ADC map and relative values (mean $\pm$ 3 standard deviations) in T2WI and CE-T1WI. In all, 144 features from tumor segmentations of the three sequences (48 features from each sequence) were extracted. Details of features are given in Supplemental Table 1.

2.4. Clinical features for machine learning models

Patient age and serum level of carbohydrate antigen 125 (CA125), carcinoembryonic antigen (CEA), and carbohydrate antigen 19-9 (CA19-9) were obtained by one of the authors (**) from electronic health records. These clinical features were included to create ML classifiers for the respective risk factors.

2.5. Machine learning classifiers based on radiomic and clinical features

The ML classifiers were constructed using the following steps for the respective five outcome variables (prognostic risk factors): dMI, histological grade (grade 1/2 vs. grade 3), PLNM/PALNM, and LVI. Patients were allocated in a 3:1 ratio to the Discovery dataset of 150 patients and the Test dataset of 50 patients with an equivalent proportion of dMI using random split method. Feature selections and

creations of ML classifiers were performed using the Discovery dataset. For building ML classifiers, python (ver. 3.6.5, <https://www.python.org/>) was used as the programming language. Feature selections for respective five outcome variables were done by application of the Boruta package (ver. 0.3, https://github.com/scikit-learn-contrib/boruta_py) on the Discovery dataset. Boruta, a feature selection method, is originally based on random forest and statistical testing [34]. It can provide unbiased and stable selection of important features by statistical comparison of the importance of attributes with that of shadow attributes generated by random fluctuations.

After feature selection, ML classifiers for the respective five outcomes were built using the XGBoost package (ver. 0.82, <https://xgboost.readthedocs.io/en/latest/>) [35]. XGBoost implements gradient tree boosting, which yields state-of-the-art results on many standard classification benchmarks. The diagnostic accuracy of XGBoost can be varied by changing the hyperparameters of XGBoost [36]. The optimal hyperparameters of XGBoost were found automatically using Bayesian optimization with the Optuna package (ver. 0.8, <https://optuna.org/>). The following ranges of hyperparameters were searched: eta 0.1–0.9, max_depth 3–12, min_child_weight 1–10, subsample 0.1–0.9, colsample_bytree 0.1–0.9, gamma 0.1–8, reg_alpha 0–2, and reg_lambda 0–2. Details of these hyperparameters are presented in the XGBoost

documentation. For the respective five outcomes, Bayesian optimization (5000 times) was used for hyperparameter optimization using ten-fold internal cross validation on the Discovery dataset. After the process, the optimal hyperparameters were obtained.

Construction of 10 ML classifiers for the respective five outcomes was performed using the optimal hyperparameters and ten-fold internal cross validation on the Discovery dataset. To avoid overestimation of their diagnostic performance, the diagnostic performance of the 10 ML classifiers was evaluated respectively, using the Test dataset. Here, model ensemble was not used in the evaluation of the Test dataset. Then, the mean AUC of 10 ML classifiers was calculated based on classification results of the Test dataset. The testing process above was performed just once.

2.6. Evaluation of the added value of the machine learning classifier for diagnosis of deep myometrial invasion

To evaluate the added value of the ML classifier on radiologists' interpretations of dMI, we used a sequential-read protocol with four board-certified radiologists' interpretations (**, **, **, and **) in the Test dataset (n=50). The study design was selected to mimic clinical settings. All readers were board-certified radiologists with more than ten years of experience in radiology. Readers 1 and 2 were experts of

gynecologic imaging, while Readers 3 and 4 were not. Among the earlier described ten results of the ML classifier for dMI using the Test dataset, one with the nearest AUC to the mean AUC of the ten results was chosen as the inference of the ML classifier to refer. Readers were informed the mean AUC of ten ML classifiers before their interpretations. First, each radiologist independently interpreted all MRI sequences, blinded to clinical information except that patients had EC, and recorded its interpretation of dMI using a Likert scale with scores of 1–5: 1, definitely absent; 2, probably absent; 3, equivocal; 4, probably present; 5, definitely present. Just after the interpretation of respective cases, referring the inference of the ML classifier (probability of dMI (%)), each radiologist revised the Likert score when necessary. Then, the AUCs of radiologists' assessments before and after reference were obtained.

2.7. Statistical analysis

Statistical analyses were conducted using software (JMP ver. 14; SAS Institute Inc., Cary, North Carolina, USA). Chi-square testing was used to compare histologic subtype, FIGO stage, dMI, histological grade (grade 1 or 2 vs. 3), LVI and PLNM/PALNM of tumors in data from the Discovery and Test datasets. Receiver operating characteristic (ROC) curve analyses were performed for the four radiologists'

Likert scores before and after referring to the ML inference. Area under the curve (AUC) with the standard error was computed using DeLong method: AUC 0.9–1.0 was considered excellent; 0.8–0.9, very good; 0.7–0.8, good; 0.6–0.7, moderate; 0.5–0.6, poor; less than 0.50, test not useful. ROC curves of the respective readers' interpretations before and after reference to the ML inference were compared using the DeLong method. A p value of less than 0.05 was inferred as significant. To obtain sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy, Likert scores were dichotomized as follows: scores of 4 and 5 reflected positive dMI, whereas scores of 1–3 were regarded as negative. Readers were instructed the cutoff of Likert scores before the interpretations.

3. Results

3.1. Patient characteristics

Patient characteristics are shown in Table 1. The interval between pretreatment MRI and initial treatment was 27.1 ± 17.4 days (mean $\pm$ standard deviation). All patients underwent hysterectomy with bilateral salpingo-oophorectomy. Pelvic lymphadenectomy was performed in 97% (195/200). In 123 patients (62%), para-aortic lymphadenectomy was also performed. No significant difference was found between the

Discovery and Test datasets with respect to histologic subtype, FIGO stage, dMI, histological grade, LVI, PLNM, or PALNM.

3.2. Diagnostic performance of radiomic ML classifiers

Lists of the most frequently selected features in the ten-time tests for the respective outcomes are shown in Supplemental Table 2. Tumor markers were selected ten times for PLNM (CA125 and CEA) and PALNM (CA125 and CA 19-9).

Radiomic ML classifiers for the respective five outcomes showed good to very good diagnostic performance with the Test dataset. Mean AUCs of ten evaluations with Test dataset were the following: 0.83 for dMI, 0.77 for histological grade 3, 0.81 for LVI, 0.72 for PLNM, and 0.82 for PALNM (Table 2). All AUCs for the Test dataset were within the range of the 95% confidence interval for the Discovery dataset.

3.3. Added value of the ML classifier for diagnosis of deep myometrial invasion

The AUC, sensitivity, specificity, PPV, NPV, and accuracy of radiologists' interpretations for dMI before and after referring the ML classifier's inference are in Table 3. The AUCs of the four readers for dMI were 0.83–0.88, superior or equal to the mean AUC of the ML classifiers (AUC 0.83). AUC was improved slightly in two of

four radiologists (one expert and one non-expert radiologist), slightly deteriorated in one expert radiologist after referring to the ML classifier inference, but these changes were not statistically significant. Numbers of cases revised after the reference were the following: Reader 1, three; Reader 2, six; Reader 3, five; Reader 4, five. In all cases of all readers, the change of the Likert scale before and after the reference was limited to one score increment.

4. Discussion

This retrospective study rigorously evaluated the diagnostic performance of ML classifiers using radiomic and clinical features for the comprehensive pretreatment assessment of risk factors in EC with repeated ten-time tests in the Test dataset. The ML classifiers for respective risk factors showed good to very good diagnostic performance (AUC 0.72–0.83, Table 2). Regarding dMI, all four radiologists were superior or equal to the mean AUC of the ML classifiers. Reference to the ML classifier's inference revealed no significant improvement in radiologists' interpretations. Nonetheless, radiomic approaches could also infer risk factors including histological grade, LVI and LN metastases. These risk factors are difficult to obtain accurately at pretreatment workup. This information would contribute to give another option for pretreatment risk

stratification and the optimization of treatment strategy.

MRI-radiomic approaches have shown promising results and potential clinical applications in pretreatment risk stratification of EC. There are several earlier reports about radiomic models for dMI [22, 25, 27]. Ueno Y et al. reported AUC of 0.84 using first-order texture features, and Han Y et al reported AUC of 0.85 using radiomic features from whole-uterus segmentation. Their ML classifiers for dMI were equivalent to the performance of radiologists, whereas AUC of ML classifier in our study (0.83 for Test dataset) was not superior to radiologists' performance. About histological grade, Yamada et al. showed much higher AUC (0.97) than ours (0.77 for Test dataset) with radiomic features extracted from ADC map using the same software we used [23]. In addition to the differences of the cohorts and methods, one of the reasons for the difference in the performance of ML classifiers might be the methods to evaluate robustness of ML classifiers. In the studies above, they did not perform validations with test datasets and mentioned it as a limitation [22, 23]. Concerns about the issue of overfitting are now highlighted in radiomic and ML studies [30, 31]. To avoid overfitting and to ensure repeatability and reproducibility [30], we performed validation with the Test dataset ten times and calculated the mean value of ten AUCs.

Exploration of optimal selections of MRI sequences and radiomic/clinical

features for respective risk factors would be challenges for the future. As shown in Supplemental table 2, tumor volume and features extracted from T2WI were most frequently selected for dMI in this study, while tumor markers were in PLNM and PALNM. In the qualitative assessment of dMI, tumor volume is an important factor, and T2WI is a key sequence. Thus, ML classifiers based on tumor volume and T2WI features might be easy to accept for clinicians. On the other hand, Fasmer KE et al. reported comparable AUC of ML classifiers for dMI (0.84 in a training dataset and 0.76 in a test dataset) based on the radiomic features extracted from CE-T1WI [27]. Unlike our results, tumor volume was not selected in their ML classifiers. Using radiomic features from whole-uterus segmentation, Han Y et al reported that the combination of T2WI and DWI improved the diagnostic performance for dMI. Further studies are needed to find the optimal selections and to maximize the potential of MRI radiomics.

Diagnostic performance for dMI did not show statistically significant improvement with support of the ML classifier in any of the four readers, contrary to a previous report suggesting improvement of diagnostic performance of a radiologist (from 82% to 100%) [24]. This difference might be because of a small number of their study (total 11 patients in the test set). No statistically significant improvement in our study might be because the AUC of the ML classifier was not superior to those of

radiologists. For breast cancer, integration of ML with high AUC into MRI interpretation could produce objective and accurate decision rules for the management of suspicious breast masses [37]. To improve the ML classifier for dMI further, feature extraction from the whole uterus and tumor–myometrial interface or deep learning features might be options. Han Y et al suggested higher radiomic feature consistency of the whole uterus than that of the tumor [25]. Chen X et al reported a deep learning model with sensitivity of 67% and specificity of 88% for dMI [29]. Another reason might be skepticism by radiologists in the ML classifiers, because the ML classifiers did not disclose its internal decision process. Explainable ML decision support system might be required to overcome this skepticism and make a ML classifier for dMI to be used in clinical practice [38]. One of the approaches could be an image-based disclosure of inference, where a ML algorithm present possible location of dMI within the image. How radiologists and ML models work together should be studied further.

We also created ML classifiers for lymph node metastasis. Although PLNM and PALNM are important prognostic factors of endometrial cancer, non-invasive pretreatment assessments are quite challenging. The size criterion of 10 mm on the short axis is well known to lead to low sensitivity and high specificity. Clinical features, such as tumor markers, play important roles in the prediction. Many researchers have tried to

create predicting model for LN metastases. Their AUCs were 0.66–0.78 [39]. Our ML classifiers showed comparable AUC of 0.72/0.82 for PLNM/PALNM derived from primary cancer without assessing pelvic/para-aortic lymph nodes. Using larger patient cohorts, Yan BC et al reported radiomic models with higher diagnostic performance for PLNM (AUC 0.91 and 0.89 in validation datasets) than ours [29]. The clinical-radiomics combined ML classifiers might assist pretreatment assessment of lymph node status, providing prior probability.

Our study has some limitations. First, in this study, the reproducibility of radiomic features with respect to tumor ROI and the difference of diagnostic performance of ML because of the variation of extracted features were not examined. For this assessment, it would be necessary to conduct additional analyses specifically, as reported by Pavic et al. [40]. Nevertheless, extracted features using 10-fold CV within the Discovery dataset were to some extent reproducible. This could be because feature selection using Boruta worked effectively. Second, the number of patients in the Discovery dataset was not large, because of train-test split of the dataset. In this work we prioritized validation with a separate dataset, which is crucially important to avoid issues of overfitting and to assess the ML classifier performance more reliably. Third, we did not perform subset analyses for respective histological subtypes of EC.

Radiomic ML classifiers for respective subtypes might improve diagnostic performance, which should be studied further for precision medicine. Fourth, the diversity of the MR scanners and protocols might affect the results. The diversity potentially strengthened our study by mirroring clinical practice.

5. Conclusions

Radiomic classifiers based on multiparametric MRI showed promise for pretreatment assessment of EC risk factors. Radiologists showed superior or equal diagnostic performance for dMI to the ML classifier. They showed no statistically significant improvement by review of related inferences of the ML classifier. Risk factors that are difficult to obtain accurately at pretreatment workup might be good targets for radiomic approaches.

Disclosures

Nothing to declare.

References

- [1] Yamagami W, Aoki D. Annual report of the Committee on Gynecologic Oncology, the Japan Society of Obstetrics and Gynecology. *J Obstet Gynaecol Res* 2015;41(2):167-77.
- [2] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin* 2019;69(1):7-34.
- [3] Salvesen HB, Haldorsen IS, Trovik J. Markers for individualised therapy in endometrial carcinoma. *The Lancet Oncology* 2012;13(8):e353-61.
- [4] Colombo N, Creutzberg C, Amant F, Bosse T, Gonzalez-Martin A, Ledermann J, et al. ESMO-ESGO-ESTRO Consensus Conference on Endometrial Cancer: Diagnosis, Treatment and Follow-up. *Int J Gynecol Cancer* 2016;26(1):2-30.
- [5] Binder PS, Mutch DG. Update on prognostic markers for endometrial cancer. *Women's health (London, England)* 2014;10(3):277-88.
- [6] Murali R, Delair DF, Bean SM, Abu-Rustum NR, Soslow RA. Evolving Roles of Histologic Evaluation and Molecular/Genomic Profiling in the Management of Endometrial Cancer. *J Natl Compr Canc Netw* 2018;16(2):201-9.
- [7] AlHilli MM, Mariani A, Bakkum-Gamez JN, Dowdy SC, Weaver AL, Peethambaram PP, et al. Risk-scoring models for individualized prediction of overall survival in low-grade and high-grade endometrial cancer. *Gynecol Oncol* 2014;133(3):485-93.
- [8] Werner HM, Trovik J, Marcickiewicz J, Tingulstad S, Staff AC, Amant F, et al. Revision of FIGO surgical staging in 2009 for endometrial cancer validates to improve risk stratification. *Gynecol Oncol* 2012;125(1):103-8.
- [9] Prat J. Prognostic parameters of endometrial carcinoma. *Human pathology* 2004;35(6):649-62.
- [10] Zhang GY, Wu LY, Li B, Huang MN, Zhang R, Li XG. Retrospective analysis of prognostic variables and clinical outcomes in surgically staged intermediate risk endometrial carcinoma. *Eur J Obstet Gynecol Reprod Biol* 2013;169(2):309-16.
- [11] Bosse T, Peters EE, Creutzberg CL, Jurgensliemk-Schulz IM, Jobsen JJ, Mens JW, et al. Substantial lymph-vascular space invasion (LVSI) is a significant risk factor for recurrence in endometrial cancer--A pooled analysis of PORTEC 1 and 2 trials. *Eur J Cancer* 2015;51(13):1742-50.
- [12] Han KH, Kim HS, Lee M, Chung HH, Song YS. Prognostic factors for tumor recurrence in endometrioid endometrial cancer stages IA and IB. *Medicine (Baltimore)* 2017;96(21):e6976.
- [13] Traen K, Holund B, Mogensen O. Accuracy of preoperative tumor grade and intraoperative gross examination of myometrial invasion in patients with endometrial cancer. *Acta obstetrica et gynecologica Scandinavica* 2007;86(6):739-41.

- [14] Helpman L, Kupets R, Covens A, Saad RS, Khalifa MA, Ismiil N, et al. Assessment of endometrial sampling as a predictor of final surgical pathology in endometrial cancer. *Br J Cancer* 2014;110(3):609-15.
- [15] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016;278(2):563-77.
- [16] Sala E, Mema E, Himoto Y, Veeraraghavan H, Brenton JD, Snyder A, et al. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clin Radiol* 2017;72(1):3-10.
- [17] Pinker K, Shitano F, Sala E, Do RK, Young RJ, Wibmer AG, et al. Background, current role, and potential applications of radiogenomics. *J Magn Reson Imaging* 2018;47(3):604-20.
- [18] Torheim T, Malinen E, Kvaal K, Lyng H, Indahl UG, Andersen EK, et al. Classification of dynamic contrast enhanced MR images of cervical cancers using texture analysis and support vector machines. *IEEE Trans Med Imaging* 2014;33(8):1648-56.
- [19] Shen WC, Chen SW, Liang JA, Hsieh TC, Yen KY, Kao CH. [18]Fluorodeoxyglucose Positron Emission Tomography for the Textural Features of Cervical Cancer Associated with Lymph Node Metastasis and Histological Type. *Eur J Nucl Med Mol Imaging* 2017;44(10):1721-31.
- [20] Lucia F, Visvikis D, Desseroit MC, Miranda O, Malhaire JP, Robin P, et al. Prediction of outcome using pretreatment (18)F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy. *Eur J Nucl Med Mol Imaging* 2018;45(5):768-86.
- [21] Meier A, Veeraraghavan H, Nougaret S, Lakhman Y, Sosa R, Soslow RA, et al. Association between CT-texture-derived tumor heterogeneity, outcomes, and BRCA mutation status in patients with high-grade serous ovarian cancer. *Abdom Radiol (NY)* 2019;44(6):2040-7.
- [22] Ueno Y, Forghani B, Forghani R, Dohan A, Zeng XZ, Chamming's F, et al. Endometrial Carcinoma: MR Imaging-based Texture Model for Preoperative Risk Stratification-A Preliminary Analysis. *Radiology* 2017;284(3):748-57.
- [23] Yamada I, Miyasaka N, Kobayashi D, Wakana K, Oshima N, Wakabayashi A, et al. Endometrial Carcinoma: Texture Analysis of Apparent Diffusion Coefficient Maps and Its Correlation with Histopathologic Findings and Prognosis. *Radiology: Imaging Cancer* 2019;1(2).
- [24] Stanzione A, Cuocolo R, Del Grosso R, Nardiello A, Romeo V, Travaglino A, et al. Deep Myometrial Infiltration of Endometrial Cancer on MRI: A Radiomics-Powered Machine Learning Pilot Study. *Academic radiology* 2020.
- [25] Han Y, Xu H, Ming Y, Liu Q, Huang C, Xu J, et al. Predicting myometrial invasion in endometrial cancer based on whole-uterine magnetic resonance radiomics. *J Cancer Res Ther* 2020;16(7):1648-55.
- [26] Yan BC, Li Y, Ma FH, Feng F, Sun MH, Lin GW, et al. Preoperative Assessment for High-Risk

- Endometrial Cancer by Developing an MRI- and Clinical-Based Radiomics Nomogram: A Multicenter Study. *J Magn Reson Imaging* 2020;52(6):1872-82.
- [27] Fasmer KE, Hodneland E, Dybvik JA, Wagner-Larsen K, Trovik J, Salvesen O, et al. Whole-Volume Tumor MRI Radiomics for Prognostic Modeling in Endometrial Cancer. *J Magn Reson Imaging* 2021;53(3):928-37.
- [28] Chen J, Gu H, Fan W, Wang Y, Chen S, Chen X, et al. MRI-Based Radiomic Model for Preoperative Risk stratification in Stage I Endometrial Cancer. *J Cancer* 2021;12(3):726-34.
- [29] Yan BC, Li Y, Ma FH, Zhang GF, Feng F, Sun MH, et al. Radiologists with MRI-based radiomics aids to predict the pelvic lymph node metastasis in endometrial cancer: a multicenter study. *Eur Radiol* 2021;31(1):411-22.
- [30] Chalkidou A, O'Doherty MJ, Marsden PK. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. *PloS one* 2015;10(5):e0124165.
- [31] Mongan J, Moy L, Kahn Jr CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. Radiological Society of North America; 2020.
- [32] Beddy P, O'Neill AC, Yamamoto AK, Addley HC, Reinhold C, Sala E. FIGO staging system for endometrial cancer: added benefits of MR imaging. *Radiographics : a review publication of the Radiological Society of North America, Inc* 2012;32(1):241-54.
- [33] Nioche C, Orlhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, et al. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. *Cancer research* 2018;78(16):4786-9.
- [34] Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010;36(11):1-13.
- [35] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016:785-94.
- [36] Nishio M, Nishizawa M, Sugiyama O, Kojima R, Yakami M, Kuroda T, et al. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PloS one* 2018;13(4):e0195875.
- [37] Ellmann S, Wenkel E, Dietzel M, Bielowski C, Vesal S, Maier A, et al. Implementation of machine learning into clinical breast MRI: Potential for objective and accurate decision-making in suspicious breast masses. *PloS one* 2020;15(1):e0228446.
- [38] Kawagishi M, Kubo T, Sakamoto R, Yakami M, Fujimoto K, Aoyama G, et al. Automatic inference model construction for computer-aided diagnosis of lung nodule: Explanation adequacy, inference accuracy, and expert's knowledge. *PloS one* 2018;13(11).
- [39] Koskas M, Fournier M, Vanderstraeten A, Walker F, Timmerman D, Vergote I, et al. Evaluation of models to predict lymph node metastasis in endometrial cancer: A multicentre study. *Eur J Cancer* 2016;61:52-60.

- [40] Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta oncologica* (Stockholm, Sweden) 2018;57(8):1070-4.

Table legends

Table 1. Patient characteristics

FIGO = International Federation of Gynecology and Obstetrics, dMI = deep myometrial invasion, LVI = lymph vascular invasion, PLNM/PALNM = pelvic/paraaortic lymph node metastasis.

* Data are mean $\pm$ standard deviation, with the range in parentheses.

Table 2. Diagnostic performances of radiomic machine learning classifiers

AUC = area under the curve, dMI = deep myometrial invasion, LVI = lymph vascular invasion, PLNM = pelvic lymph node metastasis, PALNM = paraaortic lymph node metastasis.

* Data in parentheses are 95% confidence intervals

Table 3. Diagnostic performances of four radiologists for deep myometrial invasion

before and after referring machine learning classifier's inference.

PPV = positive predictive value, NPV = negative predictive value

* Reader1,2 were experts and 3,4 were non-experts of gynecologic imaging

† Data in parentheses are 95% confidence intervals

Figure legends

Figure 1. Flow chart of patient selection.

EC, endometrial cancer; MRI, magnetic resonance imaging

Figure 2. Tumor segmentation.

Sagittal T2-weighted image (T2WI) (A), apparent diffusion coefficient (ADC) map (B) and contrast enhanced T1-weighted image (CE-T1WI) (C) show endometrial cancer dilating endometrial cavity. On T2WI and CE-T1WI, tumor was segmented on all slices. On the ADC map, considering susceptibility to artifacts, tumors were segmented on three continuous slices with no artifact.

Table1

	Training set	Test set	<i>p</i> value
Number of patients	150	50	
Age			
mean*	60±10.1	55.6±9.54	
range	31-85	36-80	
Histologic subtype			0.79
Endometrioid carcinoma	123	39	
Serous adenocarcinoma	11	5	
Others	16	6	
FIGO Stage			0.09
I	106	31	
II	9	7	
III	34	10	
IV	1	2	
dMI			0.93
positive	55	18	
negative	95	32	
Histological grade			0.73
Low grade(grade1 or 2)	98	34	
High grade(grade3)	52	16	
LVI			0.79
positive	45	16	
negative	105	34	
PLNM			0.81
positive	23	7	
negative	127	43	
PALNM			0.53
positive	20	5	
negative	130	45	

Table2

	Discovery dataset		Test dataset	
	AUC*	Accuracy*	AUC*	Accuracy*
dMI	0.76 (0.68-0.84)	0.65 (0.57-0.73)	0.83 (0.82-0.85)	0.76 (0.74-0.78)
Histological				
grade	0.70 (0.59-0.81)	0.65 (0.56-0.74)	0.77 (0.72-0.83)	0.74 (0.69-0.78)
LVI	0.71 (0.61-0.82)	0.70 (0.60-0.80)	0.81 (0.78-0.83)	0.75 (0.73-0.78)
PLNM	0.80 (0.69-0.93)	0.83 (0.72-0.93)	0.72 (0.70-0.74)	0.76 (0.74-0.78)
PALNM	0.76 (0.67-0.86)	0.70 (0.59-0.81)	0.82 (0.80-0.83)	0.73 (0.70-0.77)

Table3

	Reader1*			Reader2*			Reader3*			Reader4*		
	Before	After	<i>p</i> value	Before	After	<i>p</i> value	Before	After	<i>p</i> value	Before	After	<i>p</i> value
AUC†	0.86 (0.70-0.94)	0.85 (0.70-0.94)	0.63	0.88 (0.73-0.95)	0.90 (0.76-0.96)	0.47	0.83 (0.67-0.93)	0.83 (0.67-0.92)	0.84	0.86 (0.70-0.94)	0.87 (0.72-0.94)	0.44
Sensitivity	0.61	0.61		0.77	0.72		0.72	0.72		0.66	0.66	
Specificity	0.96	0.96		0.78	0.87		0.87	0.87		0.9	0.87	
PPV	0.91	0.91		0.66	0.76		0.76	0.76		0.8	0.75	
NPV	0.81	0.81		0.86	0.84		0.84	0.84		0.82	0.82	
Accuracy	0.84	0.84		0.78	0.82		0.82	0.82		0.82	0.8	

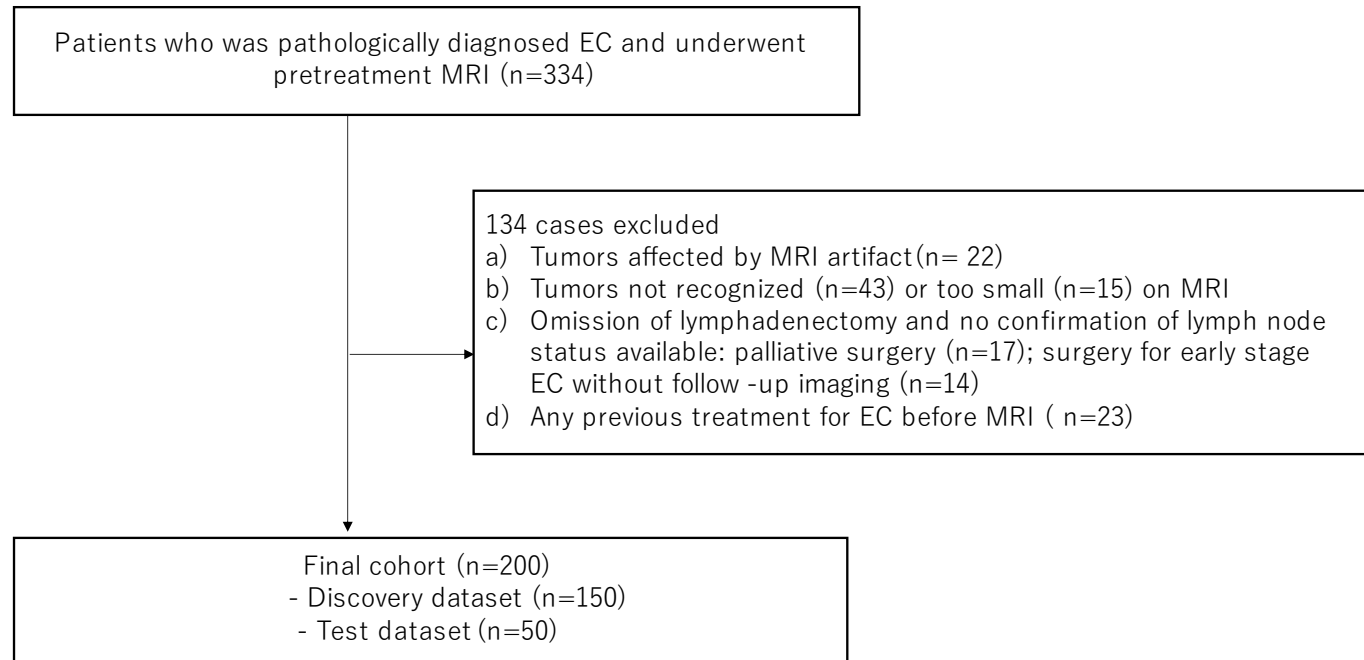
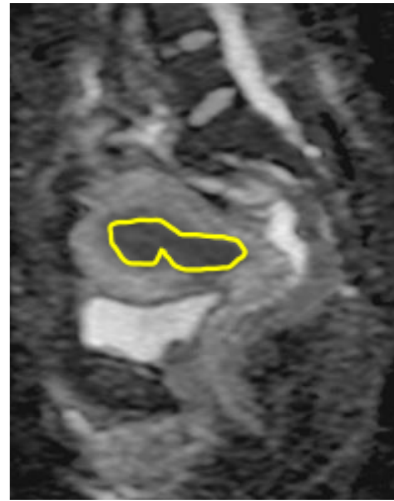


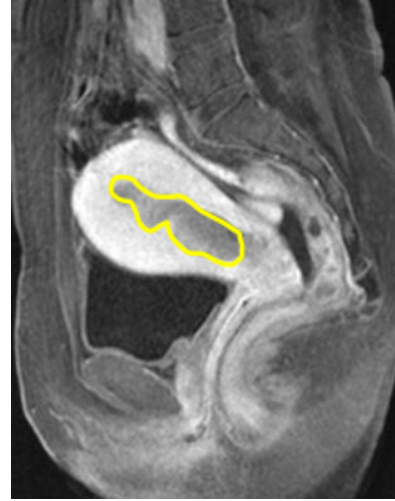
Figure1: A Flow chart of patient selection criteria



(A)



(B)



(C)

Figure2: Tumor segmentation