Inter-observer agreement of transvaginal ultrasound and magnetic resonance imaging in local staging of cervical cancer

K. Pálsdóttir* a, b and S. Fridsten* c, d, L. Blomqvist c, d, Z. Alagic d, e, D. Fischerova f, A. Gaurilcikas g, K. Hasselrot h, i, F. Jäderling c, d, j, A. Testa k, l, A. Sundin m, E. Epstein n

*Both authors contributed equally

a Department of Women’s and Children’s Health, Karolinska Institutet, 17177 Stockholm, Sweden
b Division of pelvic cancer, Theme Cancer, Karolinska University Hospital, 17176 Stockholm, Sweden
c Department of Molecular Medicine and Surgery, Karolinska Institutet, 17177 Stockholm, Sweden
d Department of Diagnostic Radiology, Karolinska University Hospital, 17176 Stockholm, Sweden
e Department of Clinical Science, Intervention and Technology, Karolinska Institutet, 17177, Stockholm, Sweden
f Gynecologic Oncology Center, Department of Obstetrics and Gynecology, First Faculty of Medicine, Charles University and General University Hospital in Prague
g Obstetrics and Gynecology, Lithuanian University of Health Sciences, Kaunas 50009, Lithuania
h Department of Obstetrics and Gynecology, Danderyd Hospital, 18288 Stockholm, Sweden
i Department of Clinical Sciences at Danderyd Hospital, Division of Obstetrics and Gynecology, Karolinska Institutet, 17176 Stockholm, Sweden
j Department of Radiology, Capio S:t Göran Hospital, 112 19 Stockholm, Sweden
k Dipartimento Scienze della Salute della Donna e del Bambino, Fondazione Policlinico Universitario A. Gemelli, IRCCS, Rome, Italy
l Dipartimento Scienze della Vita e Sanità pubblica. Università Cattolica del Sacro Cuore, Roma, Italia
m Department of Surgical Sciences, Section for Radiology, Uppsala University, Uppsala University Hospital, 75185 Uppsala, Sweden
n Department of Clinical Science and Education, Karolinska Institutet and Department of Obstetrics and Gynecology Södersjukhuset, 11883 Stockholm, Sweden

Corresponding Author: Kolbrún Pálsdóttir
Department of Women’s and Children’s Health, Karolinska Institutet
17177 Stockholm Sweden E-mail: kolbrun.palsdottir@sll.se

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/uog.23662

This article is protected by copyright. All rights reserved.
Running Head: Ultrasound and MRI in staging of cervical cancer

Keywords: Uterine Cervical Neoplasm, Ultrasonography, Magnetic Resonance Imaging, Observer Variation, Neoplasm Staging

What are the novel findings of this work?
This is the first study we are aware of, that approaches the inter-observer agreement between observers with different experience for US and MRI in women with cervical cancer. We have found comparable agreement among less experienced and more experienced observers for both imaging modalities after small educational investments.

What are the clinical implications of this work?
We conclude that with minimal efforts and training inter-observer agreement seems to be valuable. We further conclude that parametrial assessment demands experience in ultrasound. We found better agreement among MRI observers in general when compared to US observers.
Abstract

Objectives: To evaluate the inter-observer agreement in relation to observer experience for the assessment of local tumor extension in women with cervical cancer, using transvaginal ultrasound (US) and magnetic resonance imaging (MRI).

Methods: The observers comprised six US specialists with, and seven without previous experience of cervical cancer assessment, five experienced radiologists in pelvic MRI, and four less experienced radiology residents without previous MRI experience. The less experienced US observers, and all MRI observers underwent comprehensive training on assessment of cervical tumor extension while experienced US observers received written directives. All observers were assigned the same images from cervical cancer patients of all stages (n=60) for off-line evaluation on tumor detection, cervical stromal- (>1/3), and parametrial invasion. Inter-observer agreement was measured using Fleiss kappa, with 95% CI.

Results: Experienced and less experienced US observers had moderate agreement regarding tumor detection Fleiss κ [95% CI] (0.46 [0.40-0.53], and 0.46 [0.41-0.52]), stromal invasion (0.45 [0.38-0.51] and 0.53 [0.40-0.58]) and parametrial invasion (0.57 [0.51-0.64], and 0.44 [0.39-0.50]), respectively. Experienced and less experienced MRI observers had good κ [95% CI] (0.70 [0.62-0.78]) and moderate agreement (0.51 [0.41-0.62]), regarding tumor detection, good agreement regarding stromal invasion (0.80 [0.72-0.88] and 0.71 [0.61-0.81]) and parametrial invasion (0.69 [0.61-0.77] and 0.71[0.61-0.81]), respectively.
Conclusion: The inter-observer agreement was moderate for US, and moderate - good for MRI regarding the assessment of local tumor extension. The level of inter-observer agreement was only associated with experience among US observers regarding parametrial invasion.
Introduction

According to the 2018 staging criteria for cervical cancer from the International Federation of Gynecology and Obstetrics (FIGO), imaging is currently included in the staging to achieve a more accurate evaluation of local and systemic spread. Accepted imaging modalities are magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography/CT (PET/CT) and transvaginal/transrectal ultrasound (US), the method of choice depending on local availability and expertise. This is an important milestone, since the therapeutic choice is based on more accurate staging compared to the FIGO 2009 criteria where discrepancy between clinical staging and surgical results was apparent, mainly in terms of clinical overestimation of tumor size, failure to detect parametrial invasion and to assess lymph node status. Further, the inter-observer agreement for clinical staging is at best moderate.

MRI is widely accepted for cervical cancer imaging providing high accuracy to determine the tumor size (especially for lesions > 10 mm) and high specificity (> 90%) to detect parametrial invasion. MRI also allows for evaluation of tumor extension in relation to adjacent organs, such as the urinary bladder and rectum, and evaluation of pelvic lymph node status. Transvaginal and transrectal US both have high accuracy (> 90%) for tumor detection, and a high specificity to identify parametrial invasion. US is especially valuable for small tumors ≤ 1 cm. An important aspect of US is that image acquisition and interpretation is usually performed during the examination and dependent on the operator’s technical skill and expertise. By contrast, MRI acquisition is performed according to routine protocols in standardized anatomical planes while evaluation is performed after the examination. The operator dependency has led to doubts regarding the reproducibility of US for evaluation of cervical cancer. There are a
few studies published on inter-observer agreement between experienced MRI observers, 
however no study has previously addressed how observer experience affects the 
evaluation of patients with cervical cancer, neither for US nor MRI. 
Therefore, the aim of this study was to assess the inter-observer agreement for US and 
MRI, in relation to experience, in a cohort of cervical cancer patients.
Material and methods

The study was performed at Karolinska University Hospital in Stockholm, Sweden and was approved by the Regional Ethical committee (Dnr 2011/1925-31/3). Written informed consent was obtained from all patients.

Study participants

During the study period between July 2011 and August 2015, all women with biopsy verified invasive cervical cancer in the region of Stockholm/Gotland, referred to the tertiary hospital Karolinska University Hospital, Stockholm, Sweden were eligible for inclusion in this study, irrespective of disease stage. Before referral, some of the women had undergone a diagnostic cone biopsy, not all of them had residual disease according to surgery. Pregnancy and other histological diagnosis verified after surgery were exclusion criteria as well as insufficient imaging material. Insufficient US was defined as cases including only still images, absence of Doppler images/cine-clips and cine-clips not including the entire parametria. Insufficient MRI was defined as lacking pulse sequences, such as contrast-enhanced images, presence of artifacts from metal implants, presence of severe motion-related artifacts or severe obesity impairing image quality.

All women underwent a routine work-up including MRI and clinical examination under anesthesia for clinical staging according to FIGO’s revised staging criteria from 2009 \(^\text{15}\).

Women with early-stage disease (≤ IB1) underwent additional radical surgical treatment while those with advanced stages received radiochemotherapy. US was performed as part of the study protocol. Baseline demographic data on age, histology type, prior diagnostic cone biopsy and clinical stage was collected prospectively. Information on examination dates for MRI, US as well as date of surgery for women with early-stage disease were also collected.
All US examinations were performed by two US experts, one with 17 years of experience as a consultant sonographer (EE) and the second with 3 years of experience (KP). A Voluson E8 US system (GE Medical Systems, Zipf, Austria) equipped with a 5-9 MHz three-dimensional vaginal transducer or an IU22 US system (Philips Healthcare, Best, The Netherlands) applying a 3–9 MHz vaginal transducer, was used. The examination was performed transvaginally, applying the standardized protocol according to Fisherova’s systematic method \(^\text{16}\), with the woman lying in a lithotomy position having emptied her bladder prior to the examination. Still images and cine-clips of the conventional grayscale and Power Doppler US examinations were recorded in the sagittal and transverse planes with the image optimized focusing on the uterine cervix. The image set recorded for each patient included between 3 and 10 cine-clips with and without Doppler technique. When Doppler cine-clips were missing, Doppler still images of the cervix were instead provided in the image set. All US and MR-images were free from any annotations indicating image findings. The patients’ images were coded and randomly assigned a number 1-60. The image material was prepared for the study by one of the US examiners (KP), who did not participate in the evaluations.

MRI

MRI was performed on four different MRI scanners (1.5T Siemens Magnetom Aera, 1.5T Siemens Magnetom Avanto, 3T Siemens Magnetom Verio Medical Solutions, Erlangen, Germany and 1.5 T Philips Intera, Medical Systems, Best, The Netherlands) using a phased array body coil. The patients fasted for at least four hours prior to imaging and to minimize bowel motion-related artefacts, they received an intramuscular injection of antiperistaltic agent, either 1 mg glucagon (Glucagen®, Novo Nordisk,
Bagsværd, Denmark) or 20 mg butyl-scopolamine (Buscopan®, Boehringer Ingelheim GmbH, Ingelheim, Germany). They were also instructed to use a small enema to empty the rectum before arriving at the radiology department. Acquisition included high-resolution T2-weighted axial and sagittal images, oblique coronal images (i.e. along the longitudinal axis of the tumor/cervical canal) and oblique transaxial images (i.e. perpendicular to the tumor/cervical canal) as well as T1-weighted transaxial/transverse images of the pelvis before and after intravenous administration of a gadolinium-chelate based contrast agent (Gadopentetic acid, Magnevist® 469 mg/mL, 0.2 mL/kg bodyweight, Bayer AB, Solna, Sweden or Gadoteric acid, Dotarem® 279.3 mg/mL, 0.2 mL/kg bodyweight, Gothia Medical AB, Billdal, Sweden). Diffusion weighted images were not available for all patients and were hence not included for review. (The patients received the same study number for MRI as for US). The radiologist (SF) who prepared the MR images for the observers and held the workshop, was not participating in the evaluations.

Image review and observer performance

For both US and MRI there were two groups of observers: experts on cervical cancer imaging and those without such previous experience (Supplementary Table 1). The US experts comprised European Gynecologists and consultant US specialists (n=6) with 3 to 18 years of experience in cervical cancer imaging, onwards named “experienced”. The US specialists with less experience (n=7) comprised Swedish Gynecologists and consultant US specialists with 1 to 2 years of experience, although not in cervical cancer imaging, onwards named “less experienced”. The MRI observers comprised a group of specialists in radiology (n=5) with 5 to 25 years of experience in abdominal radiology including 1 to 25 years of experience in
MRI of the pelvis, onwards named “experienced”, and a group of radiology residents (n=4) with 3 to 5 years of experience in general radiology, but without previous experience in MRI of the pelvis, onwards named “less experienced”.

US images and cine-clips were stored on flash drives. All observers received a written manual on how to review the US examinations. The less experienced observers additionally attended a one-hour lecture/workshop by one of the authors (KP) on US imaging of cervical cancer with emphasis on tumor assessment and detection of stromal- and parametrial invasion, according to Fischerovas method 16. The observers then individually performed an off-line evaluation of the coded US images sets on a personal computer with a high-resolution screen, blinded to all imaging results and clinical information, except for the patients’ cervical cancer diagnosis.

All MRI observers were introduced to MRI of cervical cancer during a two-hour workshop by one of the authors (SF). The following two days (weekend) they reviewed the coded MRI examinations on the Karolinska University Hospital PACS system (SECTRA PACS, IDS7, version 19.3.6.3510, Linköping, Sweden).

All observers submitted their imaging findings by using a personalized link to an online survey (Survey Monkey®) (Please see US Survey link and MRI Survey link). A complete imaging assessment was submitted consecutively for each patient before continuing to the next subject, with a possibility to edit the imaging evaluation until the survey was completed.

For each case, the observers provided replies to three questions: 1) Is there a visible primary tumor? Yes/No. 2) Does the tumor infiltrate > 1/3 of the cervical stroma (deep stromal invasion)? Yes/No and 3) Is there parametrial invasion? Yes/No. For each question, the observers additionally rated their confidence regarding the respective
image finding on a visual analogue scale (VAS) ranging between 0 and 100, where 0 = very uncertain and 100 = absolutely confident. Further, the image quality for each examination was rated on a VAS 0-100, where 0 = very unsatisfactory and 100 = perfect.

Statistical analysis

The chi-square test was used for all categorical data, the independent two-sample t-test for normally distributed continuous data, and the Mann-Whitney U-test for non-normally distributed data. Fleiss’ kappa was calculated for groups of multiple observers. Agreement was interpreted as; poor for $\kappa = 0\text{.}2$, fair for $\kappa 0.21\text{--}0.40$, moderate for $\kappa 0.41\text{--}0.60$, good for $\kappa 0.61\text{--}0.80$ and very good for $\kappa 0.81\text{--}1$ \textsuperscript{17}. Observer confidence and image quality (VAS) was tested with the Wilcoxon rank sum test and Spearman correlation was performed to correlate observer confidence with image quality. For women undergoing surgery, histology was used as gold standard to calculate sensitivity and specificity. The statistical analysis was performed in SPSS (version 25, Statistical Package for Social Science, IBM Corporation, Armonk, NY). The level of significance was set at $p < 0.05$. Data were given as means/median, depending on the distribution, with 95\% confidence intervals (CI).
Results

Patients

During the study period 483 patients were diagnosed with cervical cancer, 89 accepted participation and were eligible for inclusion in the present study, 60 of those were included after exclusion due to cancer in the cervical stump (n=1), cancer in situ (n=1), lack of MRI (n=10), insufficient US and MRI (n=3), insufficient MRI (n=3) or insufficient US (n=11). The demographic data of the patients are shown in Table 1. The time interval between US and MRI was median 5 days (range 0-45), between MRI and surgery median 42 days (range 4-71) and between US and surgery median 36 days (range 9-69).

Inter-observer agreement

In Table 2 the inter-observer agreement (Fleiss Kappa values) within the observer groups is presented. Experienced and less experienced US observers showed moderate inter-rater agreement ranging from 0.44 to 0.57 for tumor detection, stromal and parametric invasion. Regarding parametrial invasion, the experienced observers showed significantly better agreement than the less experienced ones, with Fleiss kappa 0.57 (95% CI; 0.51-0.64) versus 0.44 (95% CI; 0.39-0.50). Both experienced and less experienced MRI observers achieved good inter-rater agreement 0.69-0.80 regarding deep stromal invasion and parametrial invasion, but for tumor detection, the agreement was moderate 0.51 (95% CI; 0.41-0.62) for the less experienced observers and good for the experienced 0.70 (95% CI; 0.62-0.78). In a sub-analysis of 41 patients, excluding the 19 patients who had undergone cone biopsy prior to imaging, the Fleiss kappa was unchanged for all US observers (numbers not presented here), but dropped for the
combined group of MRI observers regarding tumor detection, from Fleiss’ 0.64 (95% CI; 0.60-0.68) to Fleiss’ 0.56 (95% CI; 0.51-0.61), respectively.

**Observer confidence**

Table 3, shows the VAS for observer confidence and assessment of image quality for the observer groups. Experienced US observers were more confident than the less experienced US observers for all three components of the assessment. A positive correlation between observer confidence and image quality was found for the experienced US observers regarding all three components of the assessment. The strongest correlation was found for tumor detection ($\rho = 0.672, p < 0.001$). The same was true for the less experienced US observers regarding tumor detection and deep stromal invasion but not for parametrial invasion ($\rho = 0.214, p = 0.1$).

For all parts of the MRI assessment, the experienced MRI observers were more confident than the less experienced observers ($p < 0.001$). Observer confidence and assessed image quality were found unrelated ($p = 0.318$) except for tumor detection among the less experienced observers ($p = 0.01$).

Experienced MRI observers rated image quality higher than did the experienced US observers, VAS 85 (84-89) and VAS 80 (73-83), respectively ($p = 0.004$). The situation was the reverse for the less experienced observers who rated US image quality VAS 76 (71-80) and MRI quality 71 (68-73) ($p < 0.001$). Supplementary Table 2 show the sensitivity and specificity in tumor detection and cervical stromal invasion for US and MRI assessment in the 31 women undergoing surgery.
**Discussion**

In this study, the inter-observer agreement was found moderate for US, and moderate - good for MRI regarding assessment of primary tumor extension in patients with cervical cancer. An unexpected finding was the similar inter-observer agreement (Fleiss kappa) for both experienced and less experienced US observers and MRI observers, respectively. Only for parametrial invasion the experienced US observers showed significantly better agreement than the less experienced. Since this study was not designed as a diagnostic accuracy study, sensitivity and specificity for the observer groups is published only as a reference in supplementary Table 2. There were no significant differences between experienced and non-experienced observers in sensitivity or specificity regarding tumor detection and deep stromal invasion for either US or MRI, fortifying our findings regarding similar agreement.

To our knowledge, no results on inter-observer variability for US on women with cervical cancer have previously been published. However, in patients with endometrial cancer, Eriksson and colleagues reported similar agreement for US observers irrespective of their previous experience, regarding assessment of myometrial invasion, but not for cervical stromal invasion where experienced US observers had better agreement than the less experienced (gynecologists)\(^\text{18}\). The comprehensive education together with repeated image reading in both studies might have influenced the agreement and clinical performance of the less experienced groups. Previous publications have reported on the reproducibility of MRI and CT in cervical cancer. Our results show considerably higher agreement for the MRI observers than in a previous retrospective study, in 152 patients with early-stage disease, reporting $\kappa\ 0.32$ for tumor
detection and 0.11 for parametrial invasion \(^9\). However, the study included patients with less advanced tumors than those included in our study population. Hence, these results are not directly comparable to our data as tumor invasion is easier to identify in more advanced disease. Also, we used a dichotomized scoring system, as opposed to multiple choice questions in the previous report, allowing for larger variations in the evaluation and consequently for larger inter-observer variability. Further, MRI quality has improved since the previous study was performed.

At our institution, the experienced MRI observers do generally not include assessment of stromal invasion in the routine setting, which may explain their lesser confidence regarding deep stromal invasion compared to parametrical invasion and tumor detection. The clinical importance of deep stromal invasion is under debate, although it has since the Gynecologic Oncology Group study \(^9\) publication in 1999 been used together with tumor size and lymph-vascular space invasion (LVSI) to select patients for adjuvant treatment postoperatively \(^20\).

Before imaging, a significant proportion of our patients 32\% (19/60) had undergone diagnostic cone biopsy. This reflects the clinical situation in which approximately 30\% of all cervical cancers are found by screening, especially in younger patients \(^21,22\). Previous studies have shown that edema and hemorrhage following cone biopsy affect the MRI interpretation \(^23,24\) and can mimick small tumors, which is not apparent with Doppler US \(^10\). This may have resulted in more false positive cases leading to differences in kappa value for MRI observers when conizised patients were excluded, however not statistically significant.
The present results need to be interpreted in the light of the study set up situation, with very limited clinical information available for the observers and an evaluation proforma only allowing for a dichotomized assessment of the imaging findings. The intrinsic differences between the experimental off-line US setting and the unique dynamic US examination technique in real-life setting, may also have impacted the results. For instance, the possibility to feel the firmness of a cervical tumor during a dynamic examination is a feature that is completely lost in an off-line setting and could have affected the assessment especially regarding parametrial invasion. Further, the larger observer group for US than for MRI may have impacted their respective inter-observer variability as well as the fact that not all but only some of the US observers attended the initial teaching session. The multiple institutions for US, as opposed to a single center for MRI, may also have affected the results due to probable differences in the frame of reference for imaging evaluations at different centers, diminishing generalizability. The positive correlation found between image quality and observers’ confidence for US but not for MRI observers, is interesting and indicates that the former may have been more dependent on image quality in the off-line setting to feel confident in their assessment, especially regarding parametrial invasion. A potential weakness of the study is the change in MR protocols during the study period because only the basic common sequences could be used in the present study, leaving out diffusion-weighted imaging (DWI), as DWI may improve tumor detection and the evaluation of local tumor spread. A strength of this study is the mixed cohort of patients including all stages of cervical cancer and that the imaging assessment was performed by four different observer groups, each including multiple observers with varying experience. Another
strength is that study data was completed for all observers, because of the mandatory questions in the Survey Monkey questionnaire.

In conclusion, the inter-observer agreement was moderate for US and moderate – good for MRI, regarding assessment of local tumor extension. The level of inter-observer agreement was only associated with experience among US observers regarding parametrial invasion. Our results indicate that with small investments in education, improvements in evaluation of US and MRI in women with cervical cancer can be accomplished.
Acknowledgements:

Thanks to Alessia Di Legge, Chikako Suzuki, Dorella Franchi, Erik Rollvén, Filip Fruhauf, Guðrún Óladóttir, Hampus Josefsson, Jovana Muhleip, Leonidas Magarakis, Luca Savelli, Mirka Blommé, Nikolaos Voulgarakis, Nina Betsler, Peter Anfelter, Radek Buksowicz and Sigurveig Þórisdóttir. Your contribution made this study possible.
References

1. Bhatla N, Berek JS, Cuello Fredes M, Denny LA, Grenman S, Karunaratne K, Kehoe ST, Konishi I, Olawaiye AB, Prat J, Sankaranarayanan R, Brierley J, Mutch D, Querleu D, Cibula D, Quinn M, Botha H, Sigurd L, Rice L, Ryu HS, Ngan H, Maenpaa J, Andrijono A, Purwoto G, Maheshwari A, Bafna UD, Plante M, Natarajan J. Revised FIGO staging for carcinoma of the cervix uteri. *Int J Gynaecol Obstet* 2019; 145: 129-135.

2. Lagasse LD, Creasman WT, Shingleton HM, Ford JH, Blessing JA. Results and complications of operative staging in cervical cancer: experience of the Gynecologic Oncology Group. *Gynecol Oncol* 1980; 9: 90-98.

3. Qin Y, Peng Z, Lou J, Liu H, Deng F, Zheng Y. Discrepancies between clinical staging and pathological findings of operable cervical carcinoma with stage IB-IIB: a retrospective analysis of 818 patients. *Aust N Z J Obstet Gynaecol* 2009; 49: 542-544.

4. Tummers WP, Gerestein WK, Mens WJ, Verstraelen WH, Van Doorn WH. Interobserver Variability of the International Federation of Gynecology and Obstetrics Staging in Cervical Cancer. *International Journal of Gynecological Cancer* 2013; 23: 890-894.

5. Thomeer MG, Gerestein C, Spronk S, van Doorn HC, van der Ham E, Hunink MG. Clinical examination versus magnetic resonance imaging in the pretreatment staging of cervical carcinoma: systematic review and meta-analysis. *European radiology* 2013; 23: 2005-2018.
6. Choi SH, Kim SH, Choi HJ, Park BK, Lee HJ. Preoperative magnetic resonance imaging staging of uterine cervical carcinoma: results of prospective study. J Comput Assist Tomogr 2004; 28: 620-627.

7. Sheu MH, Chang CY, Wang JH, Yen MS. Preoperative staging of cervical carcinoma with MR imaging: a reappraisal of diagnostic accuracy and pitfalls. European radiology 2001; 11: 1828-1833.

8. Sahdev A, Sohaib SA, Wenaden AE, Shepherd JH, Reznek RH. The performance of magnetic resonance imaging in early cervical carcinoma: a long-term experience. Int J Gynecol Cancer 2007; 17: 629-636.

9. Fischerova D, Cibula D, Stenhova H, Vondrichova H, Calda P, Zikan M, Freitag P, Slama J, Dundr P, Belacek J. Transrectal ultrasound and magnetic resonance imaging in staging of early cervical cancer. Int J Gynecol Cancer 2008; 18: 766-772.

10. Epstein E, Testa A, Gaurilcikas A, Di Legge A, Ameye L, Atstupenaite V, Valentini AL, Gui B, Wallengren NO, Pudaric S, Cizauskas A, Masback A, Zannoni GF, Kannisto P, Zikan M, Pinkavova I, Burgetova A, Dundr P, Nemejcova K, Cibula D, Fischerova D. Early-stage cervical cancer: tumor delineation by magnetic resonance imaging and ultrasound - a European multicenter trial. Gynecol Oncol 2013; 128: 449-453.

11. Moloney F, Ryan D, Twomey M, Hewitt M, Barry J. Comparison of MRI and high-resolution transvaginal sonography for the local staging of cervical cancer. J Clin Ultrasound 2016; 44: 78-84.
12. Testa AC, Ludovisi M, Manfredi R, Zannoni G, Gui B, Basso D, Di Legge A, Licameli A, Di Bidino R, Scambia G, Ferrandina G. Transvaginal ultrasonography and magnetic resonance imaging for assessment of presence, size and extent of invasive cervical cancer. *Ultrasound Obstet Gynecol* 2009; **34**: 335-344.

13. Qu JR, Qin L, Li X, Luo JP, Li J, Zhang HK, Wang L, Shao NN, Zhang SN, Li YL, Liu CC, Li HL. Predicting Parametrial Invasion in Cervical Carcinoma (Stages IB1, IB2, and IIA): Diagnostic Accuracy of T2-Weighted Imaging Combined With DWI at 3 T. *AJR Am J Roentgenol* 2018; **210**: 677-684.

14. Lin Y, Chen Z, Kuang F, Li H, Zhong Q, Ma M. Evaluation of international federation of gynecology and obstetrics stage IB cervical cancer: comparison of diffusion-weighted and dynamic contrast-enhanced magnetic resonance imaging at 3.0 T. *J Comput Assist Tomogr* 2013; **37**: 989-994.

15. Pecorelli S, Zigliani L, Odicino F. Revised FIGO staging for carcinoma of the cervix. *Int J Gynaecol Obstet* 2009; **105**: 107-108.

16. Fischerova D. Ultrasound scanning of the pelvis and abdomen for staging of gynecological tumors: a review. *Ultrasound Obstet Gynecol* 2011; **38**: 246-266.

17. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992; **304**: 1491-1494.

18. Eriksson LS, Lindqvist PG, Floter Radestad A, Dueholm M, Fischerova D, Franchi D, Jokubkiene L, Leone FP, Savelli L, Sladkevicius P, Testa AC, Van den Bosch T, Ameye L, Epstein E. Transvaginal ultrasound assessment of myometrial and cervical stromal invasion in women with endometrial cancer: interobserver reproducibility among ultrasound experts and gynecologists. *Ultrasound Obstet Gynecol* 2015; **45**: 476-482.
19. Hricak H, Gatsonis C, Coakley FV, Snyder B, Reinhold C, Schwartz LH, Woodward PJ, Pannu HK, Amendola M, Mitchell DG. Early invasive cervical cancer: CT and MR imaging in preoperative evaluation - ACRIN/GOG comparative study of diagnostic performance and interobserver variability. *Radiology* 2007; **245**: 491-498.

20. Sedlis A, Bundy BN, Rotman MZ, Lentz SS, Muderspach LI, Zaino RJ. A randomized trial of pelvic radiation therapy versus no further therapy in selected patients with stage IB carcinoma of the cervix after radical hysterectomy and pelvic lymphadenectomy: A Gynecologic Oncology Group Study. *Gynecol Oncol* 1999; **73**: 177-183.

21. Castanon A, Leung VM, Landy R, Lim AW, Sasieni P. Characteristics and screening history of women diagnosed with cervical cancer aged 20-29 years. *Br J Cancer* 2013; **109**: 35-41.

22. Andrae B, Andersson TM, Lambert PC, Kemetli L, Silfverdal L, Strander B, Ryd W, Dillner J, Tornberg S, Sparen P. Screening and cervical cancer cure: population based cohort study. *BMJ* 2012; **344**: e900.

23. Charles-Edwards E, Morgan V, Attygalle AD, Giles SL, Ind TE, Davis M, Shepherd J, McWhinney N, deSouza NM. Endovaginal magnetic resonance imaging of stage 1A/1B cervical cancer with A T2- and diffusion-weighted magnetic resonance technique: effect of lesion size and previous cone biopsy on tumor detectability. *Gynecol Oncol* 2011; **120**: 368-373.

24. Otero-Garcia MM, Mesa-Alvarez A, Nikolic O, Blanco-Lobato P, Basta-Nikolic M, de Llano-Ortega RM, Paredes-Velazquez L, Nikolic N, Szewczyk-Bieda M. Role of MRI in staging and follow-up of endometrial and cervical cancer: pitfalls and mimickers. *Insights Imaging* 2019; **10**: 19.
25. Exner M, Kuhn A, Stumpp P, Hockel M, Horn LC, Kahn T, Brandmaier P. Value of diffusion-weighted MRI in diagnosis of uterine cervical cancer: a prospective study evaluating the benefits of DWI compared to conventional MR sequences in a 3T environment. *Acta Radiol* 2016; 57: 869-877.

26. Bakai OA, Golovko TS, Ganich AV, Gavriluk ON, Ashykhmin AV. Radiation research to determine local tumor invasion in patients with cervical cancer. *Exp Oncol* 2019; 41: 57-60.
Supplementary Legends

Supplementary Table 1. Observers experience

Supplementary Table 2. The sensitivity and specificity of all rater groups for patients with early-stage disease (n = 31)
Table 1. Demographic data for study participants (n=60)

| Demographics          | Number of patients (%) |
|-----------------------|------------------------|
| **Age**               | 46 (24–85) *           |
| **Tumor stage (FIGO)**|                        |
| IA1                   | 1 (1.7)                |
| IA2                   | 1 (1.7)                |
| IB1                   | 34 (56.7)              |
| IB2                   | 1 (1.7)                |
| IIA                   | 7 (11.7)               |
| IIB                   | 13 (21.7)              |
| IIIB                  | 2 (3.3)                |
| IV                    | 1 (1.7)                |
| **Histology**         |                        |
| Squamous cell carcinoma | 40 (66.7)           |
| Adenocarcinoma        | 18 (30.0)              |
| Other                 | 2 (3.3)                |
| **Treatment**         |                        |
| Diagnostic cone biopsy | 19 (31.6)            |
| Surgery               | 31 (51.7)              |
| Radiochemotherapy     | 29 (48.3)              |

*median (range)
Table 2. Fleiss kappa for the group of observers.

| Group of observers (number) | Fleiss Kappa | 95% CI     | Agreement |
|-----------------------------|--------------|------------|-----------|
| Tumor seen yes/no           |              |            |           |
| US Experienced (6)          | 0.46         | 0.40-0.53  | Moderate  |
| US Less experienced (7)     | 0.46         | 0.41-0.52  | Moderate  |
| US Whole group (13)         | 0.46         | 0.43-0.49  | Moderate  |
| MRI Experienced (5)         | 0.70         | 0.62-0.78  | Good      |
| MRI Less experienced (4)    | 0.51         | 0.41-0.62  | Moderate  |
| MRI Whole group (9)         | 0.64         | 0.60-0.68  | Good      |
| Deep stromal invasion yes/no|              |            |           |
| US Experienced (6)          | 0.45         | 0.38-0.51  | Moderate  |
| US Less experienced (7)     | 0.53         | 0.40-0.58  | Moderate  |
| US Whole group (13)         | 0.50         | 0.47-0.53  | Moderate  |
| MRI Experienced (5)         | 0.80         | 0.72-0.88  | Good      |
| MRI Less experienced (4)    | 0.71         | 0.61-0.81  | Good      |
| MRI Whole group (9)         | 0.73         | 0.69-0.77  | Good      |
| Parametrial invasion yes/no |              |            |           |
| US Experienced (6)          | 0.57         | 0.51-0.64  | Moderate  |
| US Less experienced (7)     | 0.44         | 0.39-0.50  | Moderate  |
| US Whole group (13)         | 0.51         | 0.48-0.54  | Moderate  |
| MRI Experienced (5)         | 0.69         | 0.61-0.77  | Moderate  |
| MRI Less experienced (4)    | 0.71         | 0.61-0.81  | Moderate  |
| MRI Whole group (9)         | 0.68         | 0.64-0.72  | Moderate  |
Table 3. Observer confidence and image quality assessment with Spearman correlation for observer groups.

| Observers               | Confidence (VAS) | Image quality (VAS) | Spearman (ρ)² | p-value* |
|-------------------------|------------------|---------------------|---------------|----------|
| **Tumor seen**          |                  |                     |               |          |
| US Experienced          | 100 (93-100)     | 80 (73-83)          | 0.672         | <0.001   |
| US Less experienced     | 80 (71-85)       | 76 (71-80)          | 0.573         | <0.001   |
| p-value**               | < 0.001          |                     |               |          |
| MRI Experienced         | 100 (99-100)     | 85 (84-89)          | 0.153         | 0.243    |
| MRI Less experienced    | 92 (84-98)       | 71 (68-73)          | 0.318         | 0.010    |
| p-value**               | < 0.001          |                     |               |          |
| **Deep stromal invasion** |                |                     |               |          |
| US Experienced          | 90 (80-95)       | 80 (73-83)          | 0.468         | <0.001   |
| US Less experienced     | 74 (70-80)       | 76 (71-80)          | 0.463         | <0.001   |
| p-value**               | < 0.001          |                     |               |          |
| MRI Experienced         | 100 (100-100)    | 85 (84-89)          | 0.221         | 0.090    |
| MRI Less experienced    | 90 (87-95)       | 71 (68-73)          | 0.220         | 0.090    |
| p-value**               | < 0.001          |                     |               |          |
| **Parametrial invasion** |                |                     |               |          |
| US Experienced          | 85 (80-91)       | 80 (73-83)          | 0.297         | 0.020    |
| US Less experienced     | 73 (70-77)       | 76 (71-80)          | 0.214         | 0.100    |
| p-value**               | < 0.001          |                     |               |          |
| MRI Experienced         | 99 (90-100)      | 85 (84-89)          | 0.166         | 0.204    |
| MRI Less experienced    | 82 (78-90)       | 71 (68-73)          | 0.090         | 0.488    |
| p-value**               | < 0.001          |                     |               |          |

*Wilcoxon rank-sum p < 0.05 is significant for Spearman correlation from mean, **Wilcoxon rank-sum. Observer confidence and image quality presented as median (95% CI)