



# International Endometrial Tumor Analysis (IETA) terminology in women with postmenopausal bleeding and sonographic endometrial thickness $\geq 4.5$ mm: agreement and reliability study

P. SLADKEVICIUS<sup>1</sup>, A. INSTALLÉ<sup>2,3</sup>, T. VAN DEN BOSCH<sup>4</sup>, D. TIMMERMAN<sup>4,5</sup>,  
B. BENACERRAF<sup>6</sup>, L. JOKUBKIENE<sup>1</sup>, A. DI LEGGE<sup>7</sup>, A. VOTINO<sup>8</sup>, L. ZANNONI<sup>9</sup>,  
B. DE MOOR<sup>2,3</sup>, B. DE COCK<sup>5</sup>, B. VAN CALSTER<sup>5</sup> and L. VALENTIN<sup>1</sup>

<sup>1</sup>Department of Obstetrics and Gynecology, Skåne University Hospital Malmö, Lund University, Sweden; <sup>2</sup>KU Leuven, Department of Electrical Engineering (ESAT) - STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium; <sup>3</sup>KU Leuven, iMinds Future Health Department, Leuven, Belgium; <sup>4</sup>Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium; <sup>5</sup>KU Leuven, Department of Development and Regeneration, Leuven, Belgium; <sup>6</sup>Harvard Medical School and Brigham & Women's Hospital, Boston, USA; <sup>7</sup>Department of Obstetrics and Gynecology, Catholic University of the Sacred Heart, Rome, Italy; <sup>8</sup>Department of Obstetrics and Gynecology, University Hospital Brugmann, Brussels, Belgium; <sup>9</sup>Department of Obstetrics and Gynecology, S.Orsola Malpighi Hospital, University of Bologna, Bologna, Italy

**KEYWORDS:** Doppler ultrasonography; endometrium; observer variation; reproducibility of results; ultrasonography

## ABSTRACT

**Objective** To estimate intra- and interrater agreement and reliability with regard to describing ultrasound images of the endometrium using the International Endometrial Tumor Analysis (IETA) terminology.

**Methods** Four expert and four non-expert raters assessed videoclips of transvaginal ultrasound examinations of the endometrium obtained from 99 women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm but without fluid in the uterine cavity. The following features were rated: endometrial echogenicity, endometrial midline, bright edge, endometrial–myometrial junction, color score, vascular pattern, irregularly branching vessels and color splashes. The color content of the endometrial scan was estimated using a visual analog scale graded from 0 to 100. To estimate intrarater agreement and reliability, the same videoclips were assessed twice with a minimum of 2 months' interval. The raters were blinded to their own results and to those of the other raters.

**Results** Interrater differences in the described prevalence of most IETA variables were substantial, and some variable categories were observed rarely. Specific agreement was poor for variables with many categories. For binary variables, specific agreement was better for

absence than for presence of a category. For variables with more than two outcome categories, specific agreement for expert and non-expert raters was best for not-defined endometrial midline (93% and 96%), regular endometrial–myometrial junction (72% and 70%) and three-layer endometrial pattern (67% and 56%). The grayscale ultrasound variable with the best reliability was uniform vs non-uniform echogenicity (multirater kappa ( $\kappa$ ), 0.55 for expert and 0.52 for non-expert raters), and the variables with the lowest reliability were appearance of the endometrial–myometrial junction ( $\kappa$ , 0.25 and 0.16) and the nine-category endometrial echogenicity variable ( $\kappa$ , 0.29 and 0.28). The most reliable color Doppler variable was color score (mean weighted  $\kappa$ , 0.77 and 0.69). Intra- and interrater agreement and reliability were similar for experts and non-experts.

**Conclusions** Inter- and intrarater agreement and reliability when using IETA terminology were limited. This may have implications when assessing the association between a particular ultrasound feature and a specific histological diagnosis, because lack of reproducibility reduces the reliability of the association between a feature and the outcome. Future studies should investigate whether using fewer categories of variable or offering practical training could improve agreement and reliability. Copyright © 2017 ISUOG. Published by John Wiley & Sons Ltd.

Correspondence to: Prof. L. Valentin, Department of Obstetrics and Gynecology, Skåne University Hospital Malmö, 20502 Malmö, Sweden (e-mail: lil.valentin@med.lu.se)

Accepted: 5 July 2017

## INTRODUCTION

Different endometrial pathologies may manifest diverse ultrasound features<sup>1–6</sup>. Mathematical models that include grayscale and color Doppler ultrasound variables (e.g. endometrial echogenicity, endometrial vessel morphology or color content of the endometrial scan as estimated subjectively using a visual analog scale (VAS)) have been designed to calculate the risk of malignancy in women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm but without fluid in the uterine cavity<sup>7–12</sup>. It is not a clinical priority to develop a model to predict endometrial cancer in women with postmenopausal bleeding and endometrial thickness  $\leq 4.4$  mm, because in such women endometrial cancer is very rare<sup>7–16</sup>. Moreover, in the endometrium of such women, it is usually impossible to detect color Doppler signals and there is little variation in grayscale echogenicity. To facilitate comparison between studies, make it meaningful to combine studies in meta-analyses and conduct multicenter studies, and to create a uniform clinical reporting system, the International Endometrial Tumor Analysis (IETA) group suggested a standardized terminology for describing grayscale and color/power Doppler ultrasound images of the endometrium<sup>17</sup>. Before introducing the IETA terminology into clinical practice or scientific studies, it is important to assess agreement in the use of the IETA terminology among ultrasound examiners with different levels of expertise and to what extent the IETA terms can be used to differentiate between patients. Intra- and interrater agreement and reliability when describing endometrial echogenicity and vascularity using rating categories other than the IETA categories<sup>8,18</sup> and when estimating the color content of the endometrial scan using a VAS<sup>9</sup> have been evaluated previously; however, studies on intra- and interrater agreement in the use of the IETA terminology have not been published so far.

The primary aim of this study was to estimate intra- and interrater agreement in the use of the IETA terminology by expert and non-expert ultrasound examiners in women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm but without fluid in the uterine cavity. Secondary aims were to assess intra- and interrater agreement when estimating the color content of the endometrial scan using a VAS and to report on reliability (how well ultrasound features can differentiate between patients) using kappa indices and intraclass correlation coefficients (ICC). This study was focused on agreement in the use of IETA terminology when describing ultrasound images rather than on diagnosis.

## SUBJECTS AND METHODS

The study was carried out at six university hospitals. No formal sample size calculation was performed. We aimed to include four expert and four non-expert raters (level-III *vs* level-II examiners according to the European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB)<sup>19</sup>) based on feasibility and availability.

Guidelines suggest that the inclusion of more than three raters has limited effect on the width of confidence intervals<sup>20</sup>. In addition, we aimed to include around 100 patients based on feasibility while keeping in mind existing sample-size guidelines<sup>20–22</sup>. Raters were invited to participate in the study by one of the authors (L.V.) on the basis of their level of ultrasound experience. The raters were asked to review the same electronic videoclips of transvaginal grayscale and color Doppler ultrasound examinations of the endometrium twice, at least 2 months apart, and to describe the images using the IETA terminology. The raters received no specific practical training but were instructed to study thoroughly the IETA consensus statement<sup>17</sup> before reviewing the videoclips. The ultrasound features to be evaluated and the terminology to use when describing the images are shown in Table 1. All the features are explained and illustrated in the IETA consensus statement<sup>17</sup>.

Electronic videoclips of transvaginal grayscale and color Doppler ultrasound examinations of the endometrium were collected for the purposes of this study between April 2010 and March 2012 by the last-named author (L.V.), an expert with more than 20 years' experience in gynecological ultrasound. The videoclips were obtained at the postmenopausal bleeding clinic of Skåne University Hospital, Malmö, Sweden, in consecutive women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm. All women were examined transvaginally in the lithotomy position with an empty bladder. Exclusion criteria were refusal to provide informed consent, presence of fluid in the uterine cavity and videoclips of very poor quality, for example owing to upright position of the uterus. Women with endometrial thickness  $\leq 4.4$  mm were also excluded. Only cases without fluid in the uterine cavity were included in this study, as a different set of ultrasound features would need to be assessed in women with fluid in the uterine cavity.

The duration of the videoclips was 8 s for grayscale and 6 s for color Doppler ultrasound. A Sequoia 512 (Acuson Inc., Mountain View, CA, USA) ultrasound system with a 4–7.5-MHz transvaginal transducer was used. The color Doppler settings were optimized to detect small blood vessels with low blood-flow velocities in the endometrium while avoiding color Doppler artifacts. The videoclips showed sagittal sections through the uterine corpus with the endometrium zoomed in and centralized in the image. Examples of videoclips are available as supplementary material (Videoclips S1–S6).

The videoclips were incorporated into the web-based electronic data capture software Clinical Data Miner (CDM), developed by one of the authors (A.I.)<sup>23</sup> and modified to simplify data collection in the context of interrater agreement. When shown on the website, the videoclips were running continuously in a loop. Using a slider, the raters could scroll back and forth in the videoclips. The videoclips were assessed independently by each of the raters after they had logged on to the CDM website with their user credentials. The grayscale

**Table 1** Prevalence of ultrasound features in first round of assessment of grayscale and color Doppler ultrasound videoclips obtained from 99 women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm but without fluid in uterine cavity

Ultrasound feature	Prevalence (% (range))*	
	Expert raters (n = 4)	Non-expert raters (n = 4)
Grayscale ultrasound		
Uniform echogenicity of endometrium	32 (23–39)	29 (18–40)
Echogenicity of endometrium		
Non-uniform		
Heterogeneous with irregular cysts	8 (1–15)	7 (2–15)
Heterogeneous with regular cysts	9 (4–11)	8 (3–15)
Heterogeneous without cysts	17 (10–29)	13 (3–20)
Homogeneous with regular cysts	27 (14–48)	31 (19–42)
Homogeneous with irregular cysts	9 (2–13)	13 (0–29)
Uniform		
Homogeneous hyperechoic	17 (13–23)	20 (12–33)
Homogeneous hypoechoic	1 (0–2)	0 (0–1)
Homogeneous isoechoic	9 (5–12)	6 (2–12)
Three-layer pattern	5 (3–9)	3 (1–6)
Endometrial midline appearance		
Irregular	2 (0–4)	3 (0–8)
Linear	8 (3–11)	6 (5–6)
Non-linear	3 (1–6)	2 (0–3)
Not defined	88 (81–96)	90 (84–94)
Bright edge	27 (7–37)	17 (11–28)
Endometrial–myometrial junction		
Interrupted	18 (10–25)	9 (4–19)
Irregular	11 (2–22)	16 (3–36)
Regular	54 (38–68)	62 (31–83)
Not defined	17 (5–24)	13 (6–26)
Synechiae	2 (0–7)	2 (0–4)
Color Doppler ultrasound		
Color score		
1 (no color)	9 (6–14)	8 (6–12)
2 (minimal color)	26 (21–29)	20 (6–30)
3 (moderate amount of color)	49 (38–56)	57 (41–86)
4 (abundant color)	16 (9–18)	15 (1–22)
Vascular pattern		
Single dominant vessel without branching	11 (7–16)	11 (6–20)
Single dominant vessel with branching	15 (6–27)	17 (9–31)
Multiple dominant vessels, focal origin	21 (11–30)	16 (6–21)
Multiple dominant vessels, multifocal origin	33 (20–49)	35 (13–53)
Scattered vessels	11 (7–14)	12 (7–14)
Circular flow	0 (0–0)	1 (1–1)
No detectable color Doppler signal	9 (6–14)	8 (6–12)
Irregularly branching vessels	17 (10–24)	32 (12–56)
Color splashes	7 (5–11)	22 (10–49)

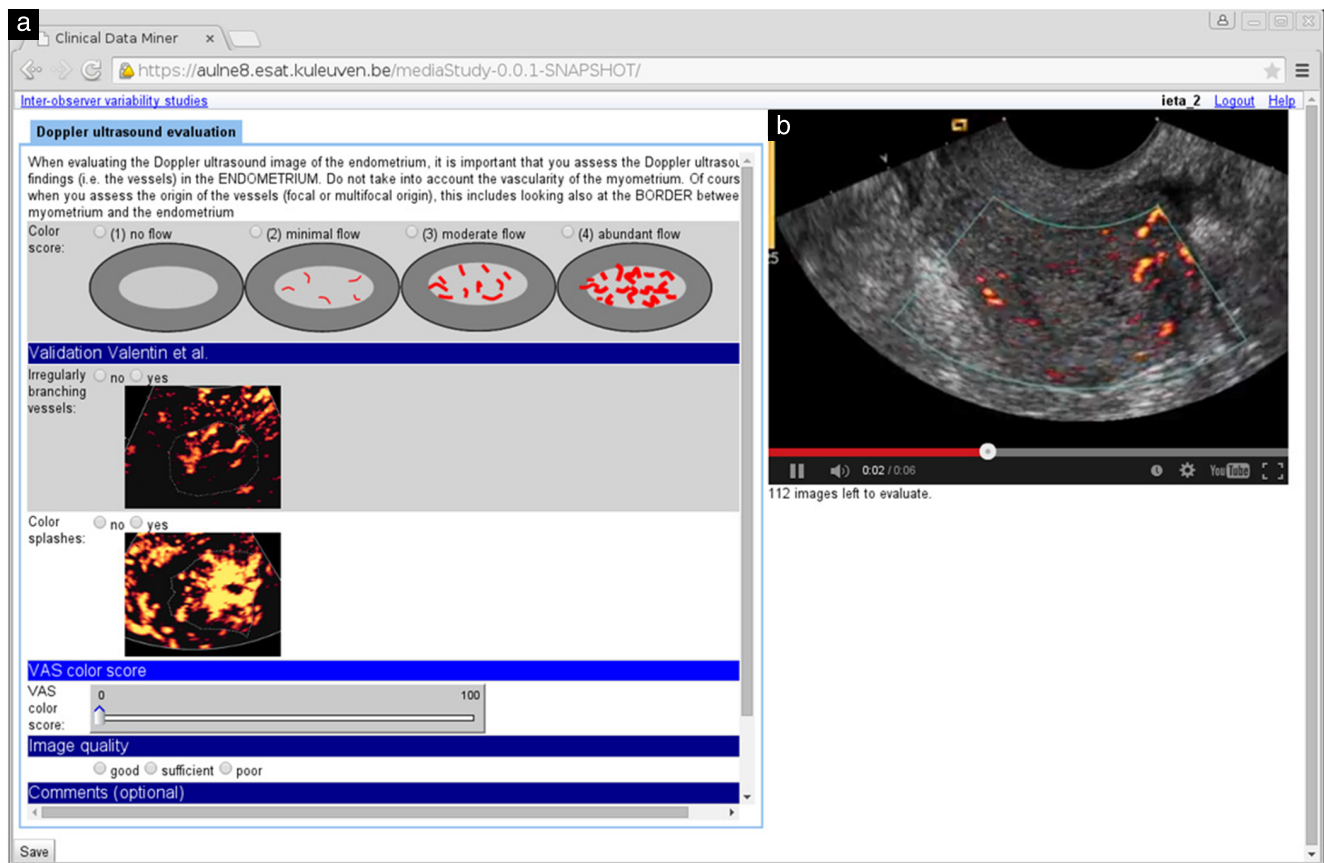
\*Overall prevalence of each category calculated as number of times that category was noted by all raters as percentage of total number of ratings (i.e. 99 patients and 4 raters, equivalent to 396 ratings).

videoclips and color Doppler videoclips were evaluated separately, as if they formed part of two separate studies. The raters were allowed to analyze the videoclips in different sessions rather than all at once.

Each ultrasound feature was rated online using the IETA terminology. The following features were rated: uniform endometrial echogenicity (yes/no), endometrial echogenicity (nine categories), endometrial midline (four categories), bright edge (yes/no), endometrial–myometrial junction (four categories), synechiae (yes/no), color score (1–4), vascular pattern (seven categories), irregularly branching vessels (yes/no) and color splashes (yes/no). To facilitate rating, and to mimic the case report forms of ongoing IETA studies, schematic drawings or ultrasound

examples were shown on the computer screen when evaluating the videoclips (Figure 1). In addition, the color content of the endometrial scans was estimated on a VAS graded from 0 to 100. No information was available to the raters other than the videoclips. To estimate intrarater agreement and reliability the same videoclips were assessed in two sessions, with a minimum of 2 months between assessments. All raters were blinded to their own results as well as to those of the other raters. In each round and for each rater, the order in which the videoclips were shown was random.

The study was approved by the Medical Ethics Committee of Lund University. Informed consent was obtained from all women who provided videoclips to



**Figure 1** Screenshot showing interface when assessing power Doppler ultrasound videoclips of endometrium in Clinical Data Miner web-based software. (a) Schematic drawings and ultrasound images illustrating ultrasound features to be assessed; (b) videoclips. When shown on website, videoclips were running continuously in a loop, and raters could navigate using a slider.

the study after the nature of the procedures had been fully explained to them. The manuscript was prepared according to the guidelines for reporting reliability and agreement studies<sup>24</sup>.

### Statistical analysis

Analysis was done separately for expert and non-expert raters with a focus on descriptive statistics of categorical variables (i.e. observed prevalence) and VAS score (i.e. mean and median scores), agreement and reliability. Agreement quantifies how often different raters agree and reliability quantifies how well the ultrasound variables can differentiate between patients. It is possible that agreement can be good when reliability is not. Reliability tends to be lower when data are highly skewed (e.g. very low or very high prevalence)<sup>25</sup>. Interrater agreement and reliability were based on the results of the first round of assessments, and intrarater agreement and reliability on the results of the first and second rounds of assessments. Statistical analysis was performed using R 3.3.2 (www.r-project.org).

### Descriptive statistics

The overall prevalence of each category of nominal and ordinal variables was computed by combining the results

of all raters. In addition, the range of prevalences given by individual raters was calculated in order to assess differences between raters with respect to which categories they considered to be present.

### Agreement

For nominal and ordinal variables, interrater agreement was investigated using the proportion of overall agreement and that of specific agreement adapted for multirater settings<sup>26</sup>. The proportion of specific agreement was calculated for every category of a variable. This estimates the probability that a specific category used by a randomly selected first rater is also used by a randomly selected second rater. For the overall and specific intrarater agreement, we computed the agreement for each of the raters, and the mean was then taken as a measure of the overall and specific intrarater agreement.

Interrater agreement for the VAS score was assessed using Bland–Altman plots modified for the multirater setting<sup>27</sup>. The mean VAS score for each patient was plotted on the *x*-axis and the difference between each rater's score and the mean was plotted on the *y*-axis. Limits of agreement were derived as approximate 2.5 and 97.5 percentile curves based on the method of Royston and Wright<sup>28</sup>. This was necessary because the mean VAS score is, by definition, related to the difference in scores,



because the scale is bounded by 0 and 100: differences are smaller for mean scores close to 0 and 100. This approach assumes that the difference between an observer's VAS score and the mean VAS score has a normal distribution. This distribution has a mean of 0 by definition, and a SD that varies with the mean VAS score. For each mean VAS score, the 2.5 and 97.5 percentiles of the differences were calculated as  $0 \pm 1.96 \times \text{SD}$ , and a smooth curve was then fitted using local regression analysis (LOESS).

### Reliability

Interrater reliability for nominal variables was estimated using Fleiss's multirater kappa ( $\kappa$ )<sup>29</sup>. For the ordinal variable color score, ICC2, which assumes that the raters are a random sample from a larger population of raters, was calculated. It is asymptotically equivalent to Cohen's  $\kappa$  with squared weights, but it can accommodate more than two raters<sup>30,31</sup>. For the continuous VAS measurement of the color content of the endometrial scan, interrater reliability was quantified using ICC2<sup>32</sup>. Intrarater reliability for each rater was computed using Cohen's  $\kappa$  for nominal variables<sup>33</sup>, Cohen's  $\kappa$  with squared weights for the ordinal variables<sup>34</sup> and ICC2 for the VAS measurement of color content<sup>32</sup>; mean values over all raters for the nominal and ordinal variables are reported. 95% CIs were computed using the jackknife method (Fleiss's multirater  $\kappa$ ), bias-corrected and accelerated bootstrap method (Cohen's  $\kappa$ , ICC2 for interrater reliability of the ordinal variables) or Shrout & Fleiss's method (ICC2 for VAS score)<sup>32,35,36</sup>.

For the interpretation of  $\kappa$  (nominal and ordinal variables) and ICC2 (continuous variables), the categories recommended in a review of reproducibility studies in obstetrics and gynecology were used<sup>37</sup>. According to this review,  $\kappa < 0.20$  corresponds to very poor reliability, 0.21–0.40 to poor reliability, 0.41–0.60 to moderate reliability, 0.61–0.80 to good reliability and  $> 0.80$  to very good reliability, while  $\text{ICC} < 0.70$  reflects very poor reliability, 0.70–0.90 poor reliability, 0.90–0.95 moderate reliability, 0.95–0.99 good reliability and  $> 0.99$  very good reliability.

## RESULTS

The first round of videoclip assessments was carried out between July and September 2012 and the second round between November 2012 and January 2013.

Eight raters with different levels of experience in gynecological ultrasound were involved. Four raters were experts, with 21 to 33 years' experience in gynecological ultrasound (P.S., D.T., T.Vd.B., B.B.). They all fulfilled the EFSUMB criteria of a level-III examiner<sup>19</sup>. Four raters were moderately experienced and fulfilled the EFSUMB criteria of a level-II examiner (non-expert raters)<sup>19</sup>. All the non-expert raters had received at least 1 year of training in gynecological ultrasound at one of the following ultrasound centers: Skåne University Hospital, Malmö, Lund University, Sweden (L.J.); Catholic University of the

**Table 2** Demographics and histological diagnoses of 99 women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm but without fluid in uterine cavity whose videoclips of transvaginal ultrasound examinations were included in study

Characteristic	Value
Age (years)	62 (45–89)
Years postmenopausal	11 (1–40)
Body mass index (kg/m <sup>2</sup> )	28.4 (20.0–46.9)
Parity	2 (0–9)
Hormone replacement therapy	24 (24)
Diabetes	16 (16)
Hypertension	46 (46)
Anticoagulants	24 (24)
Histological diagnosis	
Endometrial carcinoma	16 (16)
Endometrial polyp	44 (44)
Polypoid endometrium	1 (1)
Polyps or polypoid endometrium plus endometrial hyperplasia	6 (6)
Hyperplasia without atypia	11 (11)
Hyperplasia with atypia	3 (3)
Other benign lesion	16 (16)
Not available*	2 (2)

Data are given as median (range) or  $n$  (%). \*One woman was not operable; one was operated on in another hospital and we could not retrieve her histological report.

Sacred Heart, Rome, Italy (A.dL.); University Hospital Brugmann, Brussels, Belgium (A.V.); and St Orsola Malpighi Hospital, University of Bologna, Italy (L.Z.).

Grayscale and the corresponding color Doppler ultrasound videoclips obtained from 99 women with postmenopausal bleeding were included; data from the women contributing the videoclips have been included in other studies<sup>10,11</sup>. Demographic data and histological diagnoses of the women are shown in Table 2.

The prevalence of the different ultrasound features in the first round of assessments for expert and non-expert raters is shown in Table 1; the prevalence of the features according to each individual rater is shown in Appendix S1. There were substantial interrater differences in the prevalence of most categories of most IETA variables. The smallest differences in prevalence were observed in the four categories of the appearance of the endometrial midline and in the three-layer endometrial pattern. The presence of synechiae was rarely recorded, thus meaningful assessment of intra- and interrater reliability for this variable was not feasible.

### Agreement

Overall interrater agreement for expert and non-expert raters was best for endometrial midline (87% and 90%), uniform echogenicity of the endometrium (81% and 80%) and bright edge (74% and 80%). Differences in interrater agreement were observed between experts and non-experts for irregularly branching vessels (87% and 70%) and color splashes (96% and 74%) (Table 3).

**Table 3** Overall interrater agreement and reliability for expert and non-expert raters assessing grayscale and color Doppler ultrasound videoclips obtained from 99 women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm but without fluid in uterine cavity

Ultrasound feature	Expert raters (n = 4)		Non-expert raters (n = 4)	
	Agreement (%)	Multirater $\kappa$ (95% CI)	Agreement (%)	Multirater $\kappa$ (95% CI)
Grayscale ultrasound				
Uniform echogenicity of endometrium (yes/no)	81	0.55 (0.43–0.67)	80	0.52 (0.40–0.64)
Echogenicity of endometrium (9 categories)	40	0.29 (0.22–0.36)	41	0.28 (0.21–0.35)
Endometrial midline (4 categories)	87	0.40 (0.21–0.59)	90	0.47 (0.25–0.69)
Bright edge (yes/no)	74	0.35 (0.25–0.45)	80	0.30 (0.18–0.42)
Endometrial–myometrial junction (4 categories)	53	0.25 (0.18–0.32)	52	0.16 (0.07–0.25)
Color Doppler ultrasound				
Color score (4 ordinal categories)*	70	0.77 (0.70–0.84)	64	0.69 (0.60–0.77)
Vascular pattern (7 categories)	48	0.35 (0.27–0.43)	47	0.32 (0.24–0.40)
Irregularly branching vessels (yes/no)	87	0.56 (0.42–0.70)	70	0.30 (0.19–0.41)
Color splashes (yes/no)	96	0.69 (0.46–0.92)	74	0.25 (0.11–0.39)

\*Reliability was calculated using intraclass correlation coefficient, ICC2. In this context, ICC2 can be interpreted in same manner as weighted Cohen's  $\kappa$ <sup>30,31,34</sup>.

Specific interrater agreement for expert and non-expert raters was better for absence than for presence of color splashes, irregularly branching vessels and bright edge, and for the presence of non-uniform rather than uniform endometrium (Table 4). For variables with more than two outcome categories, specific interrater agreement for expert and non-expert raters was best for not-defined endometrial midline (93% and 96%), regular endometrial–myometrial junction (72% and 70%), three-layer endometrial pattern (67% and 56%), homogeneously hyperechoic endometrium (53% and 55%), homogeneous endometrium with regular cysts (48% and 55%) and multiple dominant vessels with multifocal origin (57% and 55%).

Overall and specific intrarater agreement data are shown in Tables S1 and S2. Intrarater percentage agreement was higher than was interrater percentage agreement. The best intrarater agreement was observed for the same variables as those with the best interrater agreement.

The recorded VAS scores for the color content of the endometrial scan ranged from 0 to 100 (Table 5). The interrater differences in VAS values for the same subject were substantial and were larger for expert than for non-expert raters (Figure 2).

### Reliability

Interrater reliability for the grayscale ultrasound variables was very poor to moderate, with multirater  $\kappa$  ranging from 0.25 to 0.55 for the expert and from 0.16 to 0.52 for the non-expert raters, with no substantial differences between the two groups (Table 3). Reliability for expert and non-expert raters was best for endometrial echogenicity 'uniform vs non-uniform' (multirater  $\kappa$ , 0.55 and 0.52) and poorest for the appearance of the endometrial–myometrial junction (multirater  $\kappa$ , 0.25 and 0.16). Interrater reliability for color score was good for both expert and non-expert raters (multirater  $\kappa$ , 0.77 and 0.69), while it was poor for the 7-category vascular

pattern variable (multirater  $\kappa$ , 0.35 and 0.32). Interrater reliability was better for expert than non-expert raters for irregularly branching vessels (multirater  $\kappa$ , 0.56 vs 0.30) and color splashes (multirater  $\kappa$ , 0.69 vs 0.25).

Intrarater reliability was moderate to good and substantially better than interrater reliability (Table S1).

The ICC2 value indicated poor interrater and intrarater reliability for the VAS score. Interrater ICC2 was 0.71 (95% CI, 0.58–0.80) for the expert raters and 0.77 (95% CI, 0.69–0.83) for the non-expert raters. Mean intrarater ICC2 was 0.78 (95% CI, 0.74–0.82) for the expert raters and 0.82 (95% CI, 0.78–0.85) for the non-expert raters.

### DISCUSSION

We found substantial interrater differences in the prevalence of most categories of most IETA variables, and some categories were rarely observed. Agreement was better for absence than for presence of the binary variables (bright edge, irregularly branching vessels, color splashes), and for the presence of non-uniform vs uniform endometrium. For the variables with more than two outcome categories, specific agreement was best for not-defined endometrial midline, regular endometrial–myometrial junction, the three-layer endometrial pattern, homogeneously hyperechoic endometrium, homogeneous endometrium with regular cysts and multiple dominant vessels with multifocal origin. Interrater reliability was poor for most variables, moderate for some and good only for the color score. The grayscale ultrasound variable with the best reliability was echogenicity uniform vs non-uniform and those with the lowest reliability were the appearance of the endometrial–myometrial junction and the nine-category endometrial echogenicity variable. The most reliable of the color Doppler variables was color score, while the seven-category vascular pattern variable and the VAS score were the least reliable. For almost all variables, agreement and reliability were similar for experts and non-experts.

**Table 4** Interrater specific agreement per category for expert and non-expert raters assessing grayscale and color Doppler ultrasound videoclips obtained from 99 women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm but without fluid in uterine cavity

Ultrasound feature	Agreement (%) (95% CI)	
	Expert raters (n = 4)	Non-expert raters (n = 4)
Grayscale ultrasound		
Echogenicity of endometrium		
Uniform	69 (59–77)	66 (54–75)
Non-uniform	86 (81–90)	86 (82–90)
Echogenicity of endometrium		
Heterogeneous with irregular cysts	20 (12–26)	26 (6–49)
Heterogeneous with regular cysts	25 (11–38)	24 (9–37)
Heterogeneous without cysts	40 (29–50)	34 (23–43)
Homogeneous with regular cysts	48 (37–59)	55 (44–64)
Homogeneous with irregular cysts	19 (8–28)	17 (9–24)
Homogeneous hyperechoic	53 (37–67)	55 (38–67)
Homogeneous hypoechoic	0 (NA)	0 (NA)
Homogeneous isoechoic	31 (18–42)	20 (6–33)
Three-layer pattern	67 (22–86)	56 (0–83)
Endometrial midline		
Irregular	7 (NA)	0 (0–0)
Linear	56 (25–73)	64 (19–86)
Non-linear	20 (0–43)	38 (NA)
Not defined	93 (90–96)	96 (93–98)
Bright edge		
No	82 (77–87)	88 (84–92)
Yes	52 (44–60)	41 (31–51)
Endometrial–myometrial junction		
Interrupted	26 (17–35)	29 (7–47)
Irregular	21 (9–31)	11 (6–16)
Regular	72 (65–78)	70 (64–76)
Not defined	40 (29–51)	34 (20–46)
Color Doppler ultrasound		
Color score		
1 (no color)	74 (49–89)	69 (44–86)
2 (minimal color)	63 (50–73)	49 (38–58)
3 (moderate amount of color)	74 (67–80)	73 (65–79)
4 (abundant color)	67 (50–79)	48 (36–58)
Vascular pattern		
Single dominant vessel without branching	52 (33–67)	43 (25–57)
Single dominant vessel with branching	38 (22–50)	44 (29–58)
Multiple dominant vessels, focal origin	38 (26–47)	31 (19–43)
Multiple dominant vessels, multifocal origin	57 (48–65)	55 (46–62)
Scattered vessels	33 (16–49)	31 (16–44)
Circular flow	0 (NA)	100 (NA)
No detectable color Doppler signal	74 (48–90)	69 (45–84)
Irregularly branching vessels		
No	92 (89–95)	78 (72–83)
Yes	64 (51–74)	52 (44–60)
Color splashes		
No	98 (96–99)	83 (79–87)
Yes	71 (43–88)	42 (27–53)

NA, not applicable as too few observations.

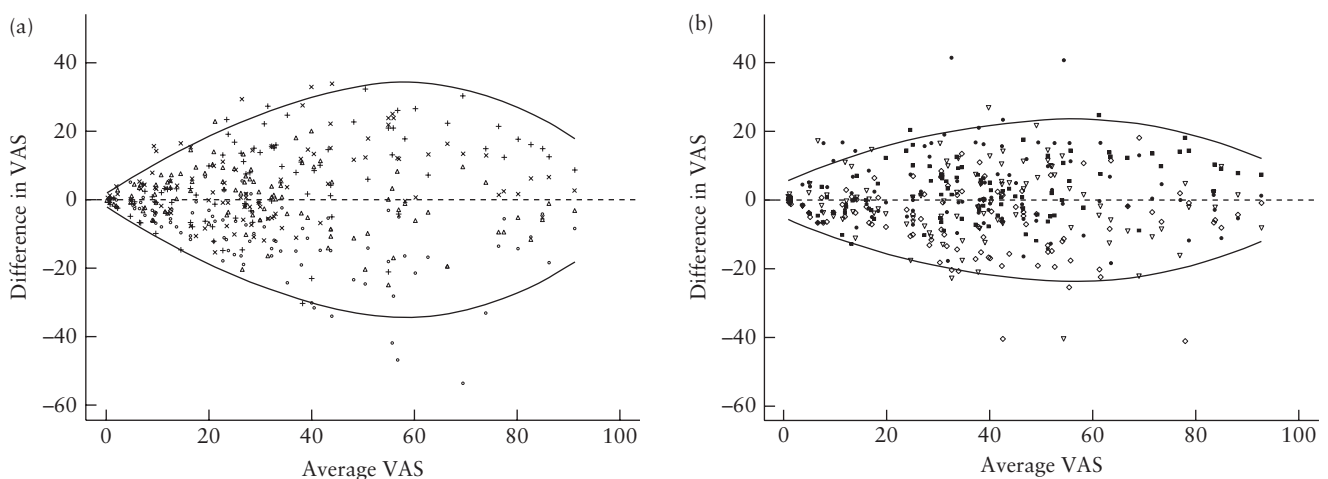
To the best of our knowledge there are no published results on intra- or interrater agreement and reliability when using IETA terminology to describe ultrasound images of the endometrium. Strengths of our study are the inclusion of as many as four raters for each level of ultrasound expertise, and the fact that the raters came from different centers, which should make our results generalizable.

A limitation of our study is the use of videoclips to estimate intra- and interrater agreement and reliability,

thus our results are not necessarily generalizable to live scanning. However, it is difficult to involve more than two ultrasound examiners in an interrater agreement study based on live vaginal ultrasound examinations, because it is unlikely that women would accept being scanned vaginally by more than two sonologists in the same session. We found that it was good to use several raters with different levels of ultrasound experience and from different ultrasound centers to obtain generalizable results, and feel that assessment

**Table 5** Descriptive statistics and intraclass correlation coefficient (ICC)2 values for color content of endometrial scan assessed using visual analog scale (VAS)

Rater	VAS score			
	First assessment round		Second assessment round	
	Mean (SD)	Median (range)	Mean (SD)	Median (range)
Expert				
1	21 (19)	16 (0–83)	22 (16)	18 (0–78)
2	31 (22)	31 (0–88)	33 (22)	30 (0–91)
3	36 (31)	32 (0–100)	20 (20)	17 (0–78)
4	36 (28)	27 (1–94)	38 (28)	29 (1–97)
Interrater ICC2 (95% CI)	0.71 (0.58–0.80)			
Intrater ICC2 (95% CI)	0.78 (0.74–0.82)			
Non-expert				
1	31 (23)	30 (0–92)	31 (26)	26 (0–93)
2	37 (24)	36 (2–95)	30 (22)	26 (1–91)
3	40 (27)	39 (0–100)	39 (27)	34 (0–96)
4	41 (25)	39 (0–96)	37 (25)	32 (1–98)
Interrater ICC2 (95% CI)	0.77 (0.69–0.83)			
Intrater ICC2 (95% CI)	0.82 (0.78–0.85)			



**Figure 2** Modified Bland–Altman plot for expert raters ( $n = 4$ ) (a) and non-expert raters ( $n = 4$ ) (b) showing relationship between mean color content of endometrial scan as estimated using a visual analog scale (VAS) of all four expert or non-expert raters per subject (average VAS) and difference between a rater's VAS value and mean VAS value for all four expert or non-expert raters per subject (difference in VAS). Solid lines indicate normal range of differences (modified limits of agreement), i.e. 95% of differences fall within these lines. Differences that lie outside solid lines are considered to be extreme.  $\circ$ , expert 1;  $\Delta$ , expert 2;  $+$ , expert 3;  $\times$ , expert 4;  $\diamond$ , non-expert 1;  $\nabla$ , non-expert 2;  $\blacksquare$ , non-expert 3;  $\bullet$ , non-expert 4.

of digital videoclips is an acceptable alternative to live scanning.

The substantial interrater differences in the prevalence of most categories of most IETA variables indicate that the raters used the IETA terminology differently, or that they had difficulty discriminating between the different grayscale and color Doppler ultrasound categories. The poor specific agreement for many of the rating categories, and the fact that some of the categories were rarely observed, suggest that combining some categories might improve agreement and reliability. For example, the endometrial midline could be described as not-defined or other (or as not-defined, linear or other), the endometrial–myometrial border as regular or not regular, the nine categories of endometrial echogenicity could be collapsed into five (heterogeneous with or without cysts,

homogeneously hyperechoic without cysts, homogeneous with regular cysts, three-layer appearance and other) and the seven categories of vascular morphology could be collapsed into five (single dominant vessel with or without branching, multiple dominant vessels with focal origin, multiple dominant vessels with multifocal origin, scattered vessels and other). It must be emphasized that the suggested collapse of categories applies only to women with postmenopausal bleeding, endometrial thickness  $\geq 4.5$  mm and no fluid in the uterine cavity. In women of fertile age, in whom the sonographic appearance of the endometrium may be different from that of women with postmenopausal bleeding, other categories of variable might be more appropriate. It is an interesting observation that, despite the prevalence of the three-layer endometrium being rare, the stated prevalence of this



category differed little between raters, and the specific agreement for this variable was quite good, indicating that this particular pattern is easy to recognize.

We are planning to develop risk models to see if any of the IETA ultrasound features helps in predicting endometrial malignancy in women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm. The importance of establishing reproducibility of predictors in risk models is debated. Some researchers recommend using reproducible predictors<sup>38</sup>, but others state that lack of reproducibility is a self-penalizing characteristic in multicenter datasets as it will automatically reduce the strength of a possible relationship of a feature with the outcome. The color content of the endometrial scan as estimated on a VAS has been used in mathematical models to predict endometrial malignancy in women with postmenopausal bleeding<sup>7,9,10</sup>, while we know of no such model incorporating the IETA color score. Our results suggest that the color score is a better reproducible measure of the color content of the endometrial scan than is subjective assessment using a VAS. We think that it would be a good idea to test its ability to predict endometrial malignancy together with other variables in a risk-prediction model.

The IETA statement is a consensus on terminology, definitions and measurements for describing the sonographic features of the endometrium and uterine cavity on grayscale and color Doppler ultrasound, which was developed with the aim of promoting consistent reporting of research results, thus facilitating interpretation and comparison of results between studies. It becomes clear from the current study that the intra- and interrater agreement and reliability with regard to the use of the IETA terminology are disappointing, and further work is required to improve reproducibility. The use of fewer categories for some of the proposed variables could potentially improve agreement and reliability, as a higher number of categories is associated with difficulty in achieving agreement. In addition, offering workshops in which sonologists view videoclips of the endometrium together and try to agree on which IETA terms should be used to describe the images might also improve agreement and reliability. Estimating agreement and reliability before and after such workshops and when using fewer categories of variable could be the aim of future studies. Intra- and interrater agreement and reliability when using the IETA terminology to describe the endometrium when there is fluid in the uterine cavity and when using it in women of fertile age also need to be assessed.

## ACKNOWLEDGMENTS

This study was supported by funds administered by Skåne University Hospital, two Swedish government grants (regionalt forskningsstöd and ALF-medel), KU Leuven Research Fund (grant C24/15/037) and Industrial R, Flemish Government (Research Foundation – Flanders (FWO) grants G0B4716N, G087112N, TBM-Logic Insulin (100793), TBM Rectal Cancer (100783), TBM

IETA (130256); Industrial Research Fund (IOF) fellowship 13-0260; iMinds Medical Information Technologies SBO 2015, ICON projects MSIPad and MyHealth-Data), Belgian Federal Government (FOD: Cancer plan 2012-2015 KPC-29-023 – prostate), a COST grant (Action: BM1104: Mass Spectrometry Imaging), and VLK Stichting E. van der Schueren.

## REFERENCES

- Epstein E, Van Holsbeke C, Mascilini F, Måsbäck A, Kannisto P, Ameje L, Fischerova D, Zannoni G, Vellone V, Timmerman D, Testa AC. Gray-scale and color Doppler ultrasound characteristics of endometrial cancer in relation to stage, grade and tumor size. *Ultrasound Obstet Gynecol* 2011; 38: 586–593.
- Timmerman D, Verguts J, Konstantinovic ML, Moerman P, Van Schoubroeck D, Deprest J, van Huffel S. The pedicle artery sign based on sonography with color Doppler imaging can replace second-stage tests in women with abnormal vaginal bleeding. *Ultrasound Obstet Gynecol* 2003; 22: 166–171.
- Hulka CA, Hall DA, McCarthy K, Simeone JF. Endometrial polyps, hyperplasia, and carcinoma in postmenopausal women: differentiation with endovaginal sonography. *Radiology* 1994; 191: 755–758.
- Atri M, Nazarnia S, Aldis AE, Reinhold C, Bret PM, Kintzen G. Transvaginal US appearance of endometrial abnormalities. *Radiographics* 1994; 14: 483–492.
- Baldwin MT, Dudiak KM, Gorman B, Marks CA. Focal intracavitary masses recognized with the hyperechoic line sign at endovaginal US and characterized with hysterosonography. *Radiographics* 1999; 19: 927–935.
- Davis PC, O'Neill MJ, Yoder IC, Lee SI, Mueller PR. Sonohysterographic findings of endometrial and subendometrial conditions. *Radiographics* 2002; 22: 803–816.
- Epstein E, Skoog L, Isberg PE, De Smet F, De Moor B, Olofsson PA, Gudmundsson S, Valentin L. An algorithm including results of gray-scale and power Doppler ultrasound examination to predict endometrial malignancy in women with postmenopausal bleeding. *Ultrasound Obstet Gynecol* 2002; 20: 370–376.
- Opolskiene G, Sladkevicius P, Valentin L. Ultrasound assessment of endometrial morphology and vascularity to predict endometrial malignancy in women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm. *Ultrasound Obstet Gynecol* 2007; 30: 332–340.
- Opolskiene G, Sladkevicius P, Valentin L. Prediction of endometrial malignancy in women with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm. *Ultrasound Obstet Gynecol* 2011; 37: 232–240.
- Sladkevicius P, Valentin L. Prospective validation of two mathematical models to calculate the risk of endometrial malignancy in patients with postmenopausal bleeding and sonographic endometrial thickness  $\geq 4.5$  mm. *Eur J Cancer* 2016; 59: 179–188.
- Sladkevicius P, Opolskiene G, Valentin L. Prospective temporal validation of mathematical models to calculate risk of endometrial malignancy in patients with postmenopausal bleeding. *Ultrasound Obstet Gynecol* 2017; 49: 649–656.
- Dueholm M, Møller C, Rydberg S, Hansen ES, Ørtoft G. An ultrasound algorithm for identification of endometrial cancer. *Ultrasound Obstet Gynecol* 2014; 43: 557–568.
- Smith-Bindman R, Kerlikowske K, Feldstein VA, Subak L, Scheidler J, Segal M, Brand R, Grady D. Endovaginal ultrasound to exclude endometrial cancer and other endometrial abnormalities. *JAMA* 1998; 280: 1510–1517.
- Gull B, Karlsson B, Milsom I, Granberg S. Can ultrasound replace dilation and curettage? A longitudinal evaluation of postmenopausal bleeding and transvaginal sonographic measurement of the endometrium as predictors of endometrial cancer. *Am J Obstet Gynecol* 2003; 188: 401–408.
- Gull B, Carlsson S, Karlsson B, Ylöstalo P, Milsom I, Granberg S. Transvaginal ultrasonography of the endometrium in women with postmenopausal bleeding: is it always necessary to perform an endometrial biopsy? *Am J Obstet Gynecol* 2000; 182: 509–515.
- Valentin L. Imaging techniques in the management of abnormal vaginal bleeding in non-pregnant women before and after menopause. *Best Pract Res Clin Obstet Gynaecol* 2014; 28: 637–654.
- Leone FP, Timmerman D, Bourne T, Valentin L, Epstein E, Goldstein SR, Marret H, Parsons AK, Gull B, Istre O, Sepulveda W, Ferrazzi E, Van den Bosch T. Terms, definitions and measurements to describe the sonographic features of the endometrium and intrauterine lesions: a consensus opinion from the International Endometrial Tumor Analysis (IETA) group. *Ultrasound Obstet Gynecol* 2010; 35: 103–112.
- Alcázar JL, Ajossa S, Floris S, Bargellini R, Gerada M, Guerriero S. Reproducibility of endometrial vascular patterns in endometrial disease as assessed by transvaginal power Doppler sonography in women with postmenopausal bleeding. *J Ultrasound Med* 2006; 25: 159–163.
- Education and Practical Standards Committee, European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB). Minimum training recommendations for the practice of medical ultrasound. *Ultraschall Med* 2006; 27: 79–105.
- Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. *Stat Methods Med Res* 2004; 13: 251–271.
- Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005; 85: 257–268.
- Donner A, Rotondi MA. Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. *Int J Biostat* 2010; 6: Article 31.

23. Installé AJ, Van den Bosch T, De Moor B, Timmerman D. Clinical data miner: an electronic case report form system with integrated data preprocessing and machine-learning libraries supporting clinical diagnostic model research. *JMIR Med Inform* 2014; 2: e28.
24. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011; 64: 96–106.
25. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology* 2003; 228: 303–308.
26. Uebersax JS. A design-independent method for measuring the reliability of psychiatric diagnosis. *J Psychiatr Res* 1982–1983; 17: 335–342.
27. Jones M, Dobson A, O'Brian S. A graphical method for assessing agreement with the mean between multiple observers using continuous measures. *Int J Epidemiol* 2011; 40: 1308–1313.
28. Royston P, Wright EM. How to construct 'normal ranges' for fetal variables. *Ultrasound Obstet Gynecol* 1988; 11: 30–38.
29. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; 76: 378–382.
30. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational Psychol Measurement* 1973; 33: 613–619.
31. Schuster C. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational Psychol Measurement* 2004; 64: 243–253.
32. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; 86: 420–428.
33. Cohen J. A coefficient of agreement for nominal scales. *Educational Psychol Measurement* 1960; 20: 37–46.
34. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70: 213–220.
35. Fleiss JL, Davies M. Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *Am J Epidemiol* 1982; 115: 841–845.
36. Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc* 1987; 82: 171–185.
37. Coelho Neto MA, Roncato P, Nastro CO, Martins WP. True Reproducibility of UltraSound Techniques (TRUST): systematic review of reliability studies in obstetrics and gynecology. *Ultrasound Obstet Gynecol* 2015; 46: 14–20.
38. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Ann Emerg Med* 1999; 33: 437–447.

## SUPPORTING INFORMATION ON THE INTERNET

The following supporting information may be found in the online version of this article:

 **Videoclip S1** Power Doppler ultrasound videoclip from a postmenopausal woman with histologically confirmed endometrial cancer.


**Videoclip S2** Grayscale ultrasound videoclip from a postmenopausal woman with histologically confirmed endometrial cancer (same woman as S1).

**Videoclip S3** Power Doppler ultrasound videoclip from a postmenopausal woman with histologically confirmed endometrial hyperplasia.

**Videoclip S4** Grayscale ultrasound videoclip from a postmenopausal woman with histologically confirmed endometrial hyperplasia (same woman as S3).

**Videoclip S5** Power Doppler ultrasound videoclip from a postmenopausal woman with histologically confirmed endometrial polyp.

**Videoclip S6** Grayscale ultrasound videoclip from a postmenopausal woman with histologically confirmed endometrial polyp (same woman as S5).

 **Appendix S1** Prevalence of features according to each rater.

**Table S1** Intrarater percentage agreement and reliability for expert raters ( $n = 4$ ) and non-expert raters ( $n = 4$ )

**Table S2** Intrarater specific percentage agreement per category