Congenital Uterine Malformation by Experts (CUME): better criteria for distinguishing between normal/arcuate and septate uterus

Running Title: Septate and normal/arcuate uterus

Key words: Mullerian ducts, Congenital Uterine Anomalies, Uterine septum, Agreement, Uterus, Ultrasound, 3D Ultrasound, Coronal plane

Authors: Artur Ludwin[1,2], Wellington P. Martins*[3,4], Carolina O. Nastri[4], Inga Ludwin[1,2], Marcela A. Coelho Neto[3], Valeria M. Leitão[3], Maribel Acién [5], Juan L. Alcazar[6], Beryl Benacerraf[7], George Condous[8], Rudy-Leon De Wilde[9], Mark Hans Emanuel[10], William Gibbons[11], Stefano Guerriero[12], William W. Hurd[13], Deborah Levine[14], Steven Lindheim[15], Antonio Pellicer[16], Felice Petraglia[17], Ertan Saridogan[18].

Affiliations: 1) Department of Gynecology and Oncology, Jagiellonian University, Krakow, Poland; 2) Ludwin & Ludwin Gynecology, Private Medical Center, Krakow, Poland; 3) SEMEAR Fertilidade, Reproductive Medicine, Ribeirao Preto, Brazil. 4) Department of Obstetrics and Gynaecology, Faculty of Medicine of Ribeirão Preto, University of Sao Paulo (DGO-FRMP-USP), Ribeirao Preto, Brazil. 5) San Juan University Hospital/Miguel Hernández University, Alicante, Spain. 6) Department of Obstetrics and Gynecology, University of Navarra, Pamplona, Spain. 7) Harvard Medical School, Brookline, MA, United States. 8) Obstetrics and Gynaecology, Acute Gynaecology, Early Pregnancy and Advanced Endosurgery Unit, Nepean Hospital, Sydney Medical School Nepean, University of Sydney, Sydney, NSW, Australia. 9) Carl-von-Ossietzky-University Oldenburg, Oldenburg, Germany. 10) Professor at the University Medical Center Utrecht (NL) and University Hospital Ghent (B). 11) Baylor College of Medicine, Houston, TX, United States. 12) Department of Obstetrics and Gynecology, University of Cagliari, Cagliari, Italy. 13) Division of Reproductive Endocrinology and Infertility Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, NC, United States. 14) Department of Radiology Beth Israel Deaconess Medical Center, Boston, MA, United States. 15) Department of Obstetrics & Gynecology, Wright State University, Boonshoft School of Medicine, Dayton, OH, United States. 16) Instituto Valenciano de Infertilidad, Valencia, Spain. 17) University of Florence, Florence, Italy, 18) University College London Hospital, London, United Kingdom.

*Corresponding author:

Wellington P Martins, Department of Obstetrics and Gynecology, Ribeirao Preto Medical School, University of Sao Paulo (DGO-FMRP-USP), Address: Av. Bandeirantes, 3900 – 8 andar - HCRP - Campus Universitario; City: Ribeirao Preto; State: Sao Paulo; Country: Brazil; Postal code: 14048–900. Phone: +55(16)3602-2583; Fax: +55(16)3633-0946; E-mail: wpmartins@gmail.com.

ABSTRACT

Objectives: To assess whether level of agreement among experts in distinguishing between septate and normal/arcuate uterus using subjective judgments from review of coronal view from three-dimensional ultrasound. We also aim to determine the inter-observer reliability and diagnostic test accuracy of three measurements suggested by recent guidelines, using the most voted option by experts (CUME - Congenital Uterine Malformation by Experts) as a reference standard.

Methods: Images of the coronal plane of the uterus from 100 women with suspected fundal internal indentation were anonymized and submitted to 15 experts (5 clinicians, 5 surgeons and 5 sonologists). They were instructed to vote between normal/arcuate (normal uterine morphology or degree of distortion caused by the internal indentation is not clinically relevant) or septate uterus (the degree of distortion caused by the internal indentation is clinically relevant). Two other raters independently measured indentation depth, indentation angle and indentation to wall thickness (I:WT) ratio. The agreement among experts was assessed by kappa, the inter-rater reliability was assessed by concordance correlation coefficient (CCC), the diagnostic test accuracy was assessed by the area under ROC curve (AUROC) and the best cut-off value was assessed using Youden's index, considering the most voted option (CUME) as the reference standard.

Results: There was a good agreement among the impression of all experts (kappa = 0.62). There were 18 septate and 82 normal/arcuate uteri by CUME; ESHRE-ESGE criteria (I:WT ratio > 50%) resulted in 80 septate and 20 normal/arcuate, while ASRM criteria resulted in 5 septate (depth > 15 mm and angle < 90°), 82 normal/arcuate (depth < 10 mm and angle > 90°) and 13 uterus would not be classified (gray-zone). The agreement between ESHRE-ESGE and CUME was 38% (kappa=0.10); the agreement between ASRM criteria for septate and CUME was 87% (kappa=0.39), and considering both septate and gray-zone as septate, the agreement was 98% (kappa=0.93). Among the three measurements, the inter-rater reproducibility of indentation depth (CCC=0.99, 95%CI=0.98-0.99) was better than both indentation angle (CCC=0.96, 95%CI=0.94-0.97) and I:WT ratio (CCC=0.92, 95%CI=0.90-0.94). The diagnostic test accuracy of these three measurements using CUME as reference standard was very good: AUROC between 0.96 and 1.00. The best cut-off values for these measurements were: indentation depth ≥ 10 mm, indentation angle < 140°, and I:WT ratio > 110% .

Conclusions: The suggested cut-off value by ESHRE-ESGE overestimates the prevalence of septate uterus while those by ASRM underestimate this prevalence, leaving in the gray zone most of the uteri considered as being septate by experts. We recommend considering indentation depth ≥ 10 mm as septate, since it is simple, reliable and in agreement with the opinion of experts.

INTRODUCTION

Distinguishing between normal uterus and congenital uterine anomalies and naming of specific malformations were described two centuries ago [1-3]. For years the term 'uterus septate/uterine septum' meant the uterus with single uterine fundus and with a divided uterine cavity into two parts without measurable criteria of deformity degree [4-6]. An intermediate benign form of anomaly between septate and normally developed uterus was called the arcuate uterus [5]. Despite the lack of robust evidence, hysteroscopic metroplasty is considered in women with septate uterus associated with infertility or miscarriages, and even in women without reproductive failures aiming to improve reproductive outcomes [6-9]; however, there are no justifications for surgical incision of internal fundal indentation in normal/arcuate uterus [6, 10]. Difficulties in differentiation between normal/arcuate and septate uterus, inconsistent definitions, and liberal indications for surgery are associated with risk of unnecessary iatrogenic treatment. Indeed, misdiagnosis and defining criteria for distinguishing of 'congenital anomaly' without consideration their relevance; especially the septate uterus can be iatrogenic, with psychological consequences.

The coronal plane of the uterus, obtained by three-dimensional ultrasound, provided an excellent tool to evaluate the level of distortion of the uterine fundus, and, aiming to improve the inter-observer variability of distinguishing between septate and normal/arcuate uterus, some objective criteria have been proposed [11, 12]. More recently, the most important societies on the field published their recommendations on how to distinguish between these two uterine morphologies using the coronal plane of the uterus: ESHRE-ESGE recommended to use an indentation to wall thickness (I:WT) ratio > 50% to diagnose septate uterus [13, 14], while ASRM recommended considering as septate when there is both an indentation depth > 15 mm and an indentation angle < 90°; while a normal/arcuate uterus should have both an indentation depth < 10 mm and an indentation angle > 90°, and uterus that does not fit those criteria would be left on a gray zone [6].

The lack of universally accepted criteria is likely to cause confusion for patients, care providers and scientific community. Additionally, the suggested cut-off values were not based on diagnostic test accuracy, probably because of it is very hard to provide a reference standard for distinguishing between these two uterine morphologies. Thus, we assumed that the most voted option among several experts blinded to other expert's opinion using images of uterus from optimal diagnostic test would be the best possible reference standard. Indeed, the reference standard is pivotal to estimate the diagnostic test accuracy of measurements and to estimate the best cut-off values for such measurements.

Our primary objective was to assess the level of agreement among experts in distinguishing between septate and normal/arcuate uterus using subjective judgments from review of coronal view from three-dimensional ultrasound, and to compare these results to two recently published guidelines from ESHRE-ESGE and ASRM. The secondary aims were: (i) to compare the preference regarding image quality provided by four different 3D ultrasound techniques when assessing the coronal plane of the uterus; (ii) to evaluate the inter-observer reliability/agreement of currently used measurements to distinguish between normal/arcuate and septate uterus; (iii) to assess the diagnostic test accuracy and best cut-off values for the most used measurements in distinguishing between normal/arcuate and septate uterus using experts' opinion as reference standard.

METHODS

Study design

This was a reliability/agreement and diagnostic test accuracy study, performed as a part of an ongoing prospective observational project on two- and three- dimensional ultrasound in screening, diagnostic and classification of female genital tract congenital anomalies (KBET/236/B/2013). The local ethics committee approved the entire project and the study. The design and report were based on the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [15] and STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) statement [16]. Written informed consent was obtained from all patients.

For this study, we planned to include 3D data-sets from 100 different uteri consecutively evaluated between Jun-2016 and Jul-2016 with suspected uterine anomaly in private medical center (Ludwin & Ludwin Gynecology, Krakow, Poland), using a single 3D data-set of each uterus acquired by an experienced observer. Fifteen invited experts were included to perform subjective judgements, 2 observers prepared the images for the experts, and 2 observers performed the measurements using the 3D data-sets.

3D data-sets of uteri were obtained from non-pregnant women in reproductive age (>18 and < 45 years). The exclusion criteria were unknown pregnancy, menopause, malignant neoplasms, specific benign lesions, and uterine surgeries [17, 18]. Moreover obvious uterine anomalies that were completely impossible to classify as either normal/arcuate or partially septate were not included in this study [5] (namely uterine agenesia, unicornuate, bicornuate, didelphys and complete septate uteri). On the contrary, T-shape and asymmetrical uterus with internal indentation were not excluded.

Ultrasound scanning

The 3D data-sets were acquired using an ultrasound system (Voluson E8 Expert BT13, GE Healthcare Ultrasound, Milwaukee, WI, USA) with volumetric intravaginal probes (GE RIC 5–9 MHz). Ultrasound scans were performed in a standardized manner by an experienced examiner (I.L) in the luteal phase (between days 17 and 25) of the menstrual cycle in women with regular cycles. Women with irregular cycles, amenorrhea and on hormonal contraception were examined regardless of the day of the cycle outside the period of menstruation. The patients were asked to hold their breath and refrain from moving during 3D volume acquisition. A maximum sweep angle of 120° after obtaining a sagittal view of the uterus was adopted and the approximate angle between the ultrasound beam and the uterine axis was 90°.

Subject selection and preparation

We included 3D datasets of consecutive women with suspicion of uterine internal indentation. A single 3D volume of the uterus of each woman was recorded, anonymized, numbered, stored and sent to two observers from another institution (MACN and VML, both with 3 years of experience with 3D-ultrasound) who prepared the images of coronal view for experts, using four different ultrasound techniques: standard coronal plane from multiplanar view (MP) [19], a rendered view of the coronal plane (volume contrast imaging, VCI), a curved rendered mode (OmniView with VCI), and a curved rendered mode using HD-Live. Each one of these observers prepared half of the images and all the images were reviewed by the two observers. The four images for each uterus were combined into a single image and submitted to the 15 experts.

Sample size

We planned to include data from 100 subjects, because this is considered as the minimum sufficient sample size to obtain sufficiently precise estimation of reliability coefficients based on available guidelines [20-22]; additionally, some authors argue that the increase in precision gained from sample > 50 subjects is rarely worth the effort [23, 24]. The sample size of raters (experts) was arbitrarily chosen; although, more than three raters is rarely worth of attempts to reach more precise reliability coefficients [25], we assumed that 15 raters (5 raters for field) should provide a better representation of medical community in the study subject.

Principles for the selection of experts

We intended to include 15 independent experts around the world not involved in previous consensuses on the measurable criteria for congenital uterine malformations without known personal conflict of interest: 5 clinicians, 5 surgeons and 5 sonologists and/or radiologist specialized in gynecological imaging. The selection process and inclusion criteria were following: (i) Editors in Chief/Deputy Editors/Members of Editorial Board of journals with the highest impact in the fields: Gynecology and Reproductive Medicine/Imaging/Gynecological Surgery; (ii) Presidents or Members of Executive Committee of the targeted societies in the field; (iii) globally well-known experts in these fields due to their publications about uterine anomalies; and (iv) invited experts should have at least 50 publications in the field of Obs/Gyn/Surg/Imaging; and (v) consent to participate.

An initial list of 15 experts and supplemental list (2 experts per field) were created (A.L., W.P.M), and the invitations were consecutively sent to experts. If somebody from the initial list did not agreed on the participation, we consecutively invited another expert from supplemental list, who was representative for the same field. To avoid bias the experts were blinded to the results of measurements and the opinion of the other experts.

Rating process

An online form was created and the link was submitted to the participating experts (https://goo.gl/forms/XU51vdDe79Fw2RDE2). The experts were asked about: (i) uterine morphology using dichotomous responses; and (ii) imaging quality using a multipoint scale.

Clinical definitions for experts

Experts were asked to distinguish between normal/arcuate and septate uterus using the following definitions: (i) normal/arcuate = normal uterine morphology or degree of distortion caused by the internal indentation is not clinically relevant; and (ii) septate uterus = the degree of distortion caused by the internal indentation is clinically relevant. These definitions were constructed as potentially important for management of patients, unbiased relative to available definitions, without any hint relative to other measurable cut-off values [6, 11, 14, 26].

Imaging quality

The image quality of the four imaging techniques: (i) standard coronal view, (ii) volume contrast imaging (VCI), (iii) OmniView with VCI, and (iv) HD-Live render mode were rated by experts providing only one vote for each technique in the end of the questionnaire, using an 11-point numeric scale (0-10). We provided fifteen examples randomly chosen (using an online random numbers generator; https://www.randomizer.org/ to generate 1 set of 15 unique numbers between 1-100) among the 100 datasets for image quality voting (Supplemental Figure 1).

Measurements

Two observers using the same initial data-set, independently manipulated the uterus to obtain the coronal plane aiming to identify the visible intramural parts of both Fallopian tubes (mid-coronal plane) and performed the following measurements blinded to each other results using VCI mode: indentation depth, indentation angle, and uterine fundal wall thickness (Supplemental Figure 2); the latter was used to calculated the indentation to wall thickness (I:WT) ratio. Among the imaging methods we preferred using VCI because it is easier to use than both Omniview and HDlive and it is the suggested imaging technique for myometrial assessment [27].

The indentation depth was measured as the distance between the the internal intercornual line (line connecting the highest point of the endometrial cavity in each side of the uterus) and the lowest point of the internal indentation/partition in the lower part of the uterus [6, 14] (Supplemental Figure 2). Although some authors suggest using the interostial line [14], the position of the tubal ostia is frequently not so clear, particularly using the standard MP view (Supplemental Figure 1); using this line as reference would result in several non-measurable cases. Additionally, the interostial line is sometimes placed below the internal intercornual line, and using this line as reference would not reflect the total indentation in these cases (Supplemental Figure 3). The indentation angle was measured tracing two lines close to the indentation apex [6, 11]. The uterine wall thickness was defined and measured as the distance between the internal intercornual line and the external uterine contour [14].

These measurements were used to calculate inter-observer reliability; and the average values between the two observers were used to classify of uterine morphology as suggested by previous guidelines (ESHRE-ESGE and ASRM), to assess the diagnostic test accuracy of the measurements and to determine the best cut-off values using the most voted option of the experts as reference standard.

Statistical analysis

Analyses were carried out using GraphPad Prism version 6 (GraphPad Software Inc., San Diego, CA, USA), IBM SPSS Statistics 22 (IBM Corp.,Armonk, NY, USA), and Stata version 13.0 (StataCorp LP, College Station, TX, USA). Continuous variables were analyzed for normal distribution using the D'Agostino & Pearson omnibus normality test. Variables with normal distribution were presented as mean ± standard deviation. The other continuous variables were presented as median values with lower and upper quartiles. Categorical variables were presented as numbers of subjects and percentages.

Results of the imaging quality rating summarized as median and interquartile range (IQR), and comparison across groups performed with Friedman test. Agreement across experts (assessing all experts, and then assessing each group of experts with respect to area of expertise) was assessed by kappa statistics and proportion of agreement (p_o). The κ-value was interpreted with regard to reporting the reliability/strength of agreement as follows: poor, <0.20; fair, 0.21–0.40; moderate, 0.41–0.60; substantial/good, 0.61–0.80; and almost perfect/very good, 0.81–1.00. Inter-observer reliability and agreement of indentation depth, angle and I:WT ratio were assessed by concordance correlation coefficient (CCC) and limits of agreement (LoA). CCC values were interpreted as following: very poor, < 0.70; poor = 0.70-0.90, moderate, 0.90-0.95; good, 0.95-0.99; and very good, >0.99 [22]; limits of agreement were assessed to estimate the margins of variability/error. The observed agreement, kappa, and proportion of false positive (FP) and false negative (FN) were used to express agreement between ESGE-ESHRE, ASRM and CUME (the most voted option by experts: **C**ongenital **U**terine **M**alformation by **E**xperts). Diagnostic test accuracy was assessed by the area under ROC curve (AUROC) and the best cut-off value was assessed using Youden's index. The relative risk (RR) with 95% CI, and P value were calculated to estimate the probability of diagnosis the septate uterus using ESHRE-ESGE relative to ASRM, and ESHRE-ESGE and ASRM relative to CUME.

RESULTS

The actual number of experts for subjective assessment (N=15), for measurements (N=2), and subjects/uteri (N=100) were exactly the same as planned. Details of the included fifteen experts (5 clinicians, 5 imaging, and 5 surgeons) are presented in Supplemental Table 1 (field, years of experience, country of residence).

3D datasets were selected from 143 women who were potentially eligible with suspected anomaly and all of them were examined for eligibility. Four women declined for participate and 39 women were excluded because of the following reasons: unknown pregnancy = 1, myomas = 8, surgeries = 11, Asherman syndrome = 2, Mayer-Rokitansky-Küster-Hauser syndrome = 2, unicornuate uterus = 3, bicornuate or uterus with external cleft = 6, dydelphys uterus = 1, complete septate uterus = 5. Finally, datasets of 100 women were included in the study and analyzed as planned. The results for individual measurements (indentation depth, indentation angle, and uterine wall thickness) for each observer are presented on Supplemental Table 2. There were 18 septate and 82 normal/arcuate uteri using the most voted option of all the 15 experts as reference standard. ESHRE-ESGE criteria resulted in 80 septate and only 20 normal/arcuate, while ASRM criteria resulted in 5 septate, 82 normal/arcuate and 13 uteri would not be classified (gray-zone) (Table 1).

Agreement among experts (Table 2)

Considering all the experts, 357/1500 (24%) of the votes were for septate uterus with a good overall agreement (kappa = 0.62, 95%CI = 0.48-0.73; p_o = 86.3%). Considering the three groups separately, the sonologists were more likely to consider the uteri as being normal and had a better inter-observer agreement (septate uterus = 92/500 = 18%; kappa = 0.74, 95%CI = 0.64-0.82; p_o = 92.2%) than both clinicians (septate uterus = 138/500 = 28%; kappa = 0.53, 95%CI = 0.37-0.66; p_o = 81.2%) and surgeons (septate uterus = 127/500 = 25%; kappa = 0.56, 95%CI = 0.41-0.68; p_o = 83.2%).

Agreement between ESHRE/ESGE and ASRM criteria and with the experts' opinion (Table 3)

The agreement between ESHRE-ESGE and the opinion of experts was only 38% (FP=62%, FN=0%, kappa = 0.10, 95%CI = -0.10 to 0.29); the agreement between ASRM criteria for septate and the opinion of experts was

87% (FP=0%, FN=13%, kappa = 0.39, 95%CI = 0.21 to 0.55), and considering both septate and gray zone as septate, the agreement was 98% (FP=1%, FN=1%, kappa = 0.93, 95%CI = 0.90 to 0.95).

Difference in the proportion of septate uteri using different criteria

The proportion of septate uteri if using the ESHRE-ESGE criteria would be much higher than if using the ASRM criteria (RR 13.9, CI 5.9-32.7, P < 0.01). In comparison with the most voted option by experts, the proportion of septate uteri would be significantly higher if using the ESHRE-ESGE criteria (RR 4.5, CI 2.9-6.8, P < 0.01), and significantly lower if using the ASRM criteria (RR 0.3, CI 0.1-0.8, P = 0.01).

Inter-observer reliability/agreement of currently used measurements (Table 4, Figure 1)

The inter-rater reliability of indentation depth (CCC=0.99, 95%CI=0.98-0.99, very good) was significantly better than both indentation angle (CCC=0.96, 95%CI=0.94-0.97, good) and I:WT ratio (CCC=0.92, 95%CI=0.90-0.94, moderate). The LoA were: indentation depth = -1.7mm to +2.1mm; indentation angle = -17° to +16°; and I:WT ratio = -75% to + 96%.

Diagnostic test accuracy and best cut-off values for currently used measurements (Table 5)

The diagnostic test accuracy of the three measurements was very good: AUROC=1.00/0.96/0.99, 95%CI=0.96-1.00/0.90-0.99/0.94-1.00; indentation depth, indentation angle and I:WT ratio respectively. The best cut-off values determined by Youden's Index were: indentation depth ≥ 10 mm, indentation angle ≤ 136°, and I:WT ratio > 111%.

Septate uterus by best cut-off values for currently used measurements and their agreement with the experts' opinion are presented on Supplemental Table 3.

3D techniques and image quality

Considering the opinion of all fifteen experts, the standard coronal plane from multiplanar view (MP) was considered as providing worse imaging quality: MP = 7 (6-9), VCI = 8 (8-10), Omniview = 8 (8-9), and HDlive = 9 (8-10); *P* = 0.002. Considering only the five clinicians the results were: MP=7 (5-9.5), VCI = 8 (7.5-9.5), Omniview = 8 (7.5-9.5), and HDlive = 8 (8-9.5); *P* = 0.39. Considering only the five surgeons, the results were: MP = 7 (5-

9.5), VCI = 8 (7.5-9.5), Omniview = 8 (7.5-9.5), and HDlive = 8 (8-9.5); P=0.14. Considering only the five sonologists, the results were: MP = 7 (6-8.5), VCI = 8 (7.5-10), Omniview = 9 (8.5-9.5), and HDlive = 10 (9-10); P=0.07.

DISCUSSION

The most important findings of this study, using the most voted option among independent expert's judgments on clinical relevance of uterine deformity as the reference standard for the classification of uterine anomalies caused by internal fundal indentation were:

- The level of agreement among experts using coronal view of uterus from 3D ultrasound is good;

- The agreement between ESHRE-ESGE and ASRM criteria with the experts' opinion is poor: ESHRE-ESGE definition result in a much higher proportion of septate uteri, and the ASRM criteria result in a much lower proportion of septate uteri, leaving a large proportion of septate uteri by experts' opinion in the gray zone (neither normal/arcuate nor septate);

- The three measurements suggested by those criteria (indentation depth, indentation angle and I:WT ratio) have a good diagnostic test accuracy when using the experts' opinion as the reference standard, however the suggested cut-off values by this study for indentation angle and I:WT ratio are substantially different from the values suggested by ESHRE-ESGE and ASRM definitions (Table 6).

- The best cut-off values for defining septate uterus are: indentation depth ≥ 10 mm, indentation angle ≤ 136° (or < 140° for rounding), and I:WT ratio > 111% (or > 110% for rounding) (Figure 2);

- From these three measurements, the indentation depth has the best inter-observer reliability (CCC=0.99; very good level of agreement) and it is the simplest to be performed.

- The experts had preferred rendered modes (VCI, Omniview with VCI or HDlive) in visualization of uterine coronal view.

The most important limitation of this study is that our reference standard might not be good for assessing clinical relevance: appropriateness of our reference standard criteria for patient management should be confirmed by studies assessing whether using such cut-off values are appropriate for distinguishing uterus at low and high risk of infertility and miscarriage. Additionally, all the estimates of reliability/agreement and diagnostic accuracy might be somewhat overestimated since it was performed by highly trained raters using a single 3D dataset for measurements. On the other hand, the raters had to independently manipulate the 3D data-set before measurements, and we believe that such manipulation is probably one of the most important sources of variability [17] as the same uterus in the same data-set might provide different images depending on the angle that the coronal plane is obtained (Supplemental Figure 4).

ESHRE/ESGE consensus wrote that endoscopy should be used in debatable cases of anomalies [14], but endoscopy is an expensive and invasive tool when used in the diagnostic setting [6], moreover hysteroscopy without true measurements is unreliable [28, 29], and reliability of laparoscopy have not been tested yet. International multi-rater agreement study without any criteria for septate, arcuate and normal uterus [28], and other study [29], with and without criteria and hysteroscopic videos as subjects, showed poor to moderate agreement among raters. Although limitations of these estimations were shown [30, 31], these two studies may indicate that hysteroscopy with subjective judgments should not be used as the reference standard for distinguishing between normal/arcuate and septate uterus in borderline cases: all the most important definitions of uterine anomalies, particularly the most used one [5], use pattern recognition of the coronal plane of the uterus for distinguishing anomalies and hysteroscopy/laparoscopy does not provide such an image.

Reliability of currently used measurements may be compared with four available studies [11, 17, 32, 33]. Revised interpretation of these studies results by new proposed cut-offs for assessment of the level of reliability in ultrasound confirm that the reliability of measurements of internal indentation depth is significantly better than the angle, and uterine wall thickness [22]. Moreover, the problem with the paradox of use the index based on uterine wall (a variable parameter that is independent from deformity) regarding uterine cavity shape was earlier highlighted [34]: a small internal indentation may be recognized as septate uterus, and larger internal indentation as normal uterus depending on the uterine wall thickness [18, 33, 35]. Suggested cut-offs for uterine wall thickness by the study results (septate when I:WT > 110%) showed that I:WT can be accurate in diagnosis, but wide margin of error for this benchmark (approximately ± 80% just by repeating the measurement by another observer) indicate problems of its use in clinical practice.

Currently, there is no evidence to support any surgical procedure for women with septate uterus [9]. There are two ongoing trials and none of them contain a reliable/accurate definition of septate uterus in the registered study protocol [36, 37]. Our study should not be used as a support for metroplasty in women with internal indentation depth of 1 cm, but our findings could be helpful for future studies to define the eligibility criteria.

CONCLUSIONS

Experts showed a preference for rendered imaging techniques when assessing the coronal plane of the uterus. Experts, particularly sonologists, have a good agreement in distinguishing between normal/arcuate and septate uterus by just assessing the coronal plane of the uterus obtained by 3D ultrasound. Using expert opinion as a reference standard, the suggested cut-off by ESHRE-ESGE greatly overestimates the prevalence of septate uterus, while the definition by ASRM underestimates the prevalence of septate uteri, leaving most of them in the gray zone. Indentation depth, indentation angle and I:WT ratio have good diagnostic test accuracy for distinguishing between normal/arcuate and septate uterus; however, the suggested cut-off values for indentation angle (<90° by ASRM) and for I:WT ratio (>50% by ESHRE-ESGE) should be revised, as they are not in agreement with the best cut-off values considering the experts' opinion as the reference standard: indentation depth ≥ 10 mm, indentation angle <140°, and I:WT ratio >110%. We suggest using the internal indentation depth ≥ 10 mm to distinguish between normal/arcuate and septate, as it is the simplest and the most reliable measurement of these three.

ACKNOWLEDGMENTS

We want to thank Professor T. Falcone (Obstetrics, Gynecology, and Women's Health Institute, Cleveland Clinic, Cleveland, Ohio, USA) for his participation in this study as an expert; additionally we want to thank both Marzena Stryjak (Ludwin & Ludwin Gynecology, Krakow, Poland) and Doctor Fabio Cruz (Instituto Valenciano de Infertilidad, Valencia, Spain) for their help in data management.

FUNDING

This study was supported by Jagiellonian University (grant no. K/ZDS/005568).

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

References

1. Acien P, Acien MI. The history of female genital tract malformation classifications and proposal of an updated system. *Hum Reprod Update* 2011; **17**: 693-705.
2. Cruveilhier J. *Anatomie pathologique du corps humain*: Bailliere; 1842.
3. Doerr W. [Jean Cruveilhier, Carl v. Rokitansky, Rudolf Virchow. Fundaments of pathology, thoughts on the 100th anniversary of Rokitansky's death]. *Virchows Arch A Pathol Anat Histol* 1978; **378**: 1-16.
4. Buttram VC, Jr., Gibbons WE. Mullerian anomalies: a proposed classification. (An analysis of 144 cases). *Fertil Steril* 1979; **32**: 40-46.
5. Buttram VCJ, Gomel V, Siegler A, DeCherney A, Gibbons W, March C. The American Fertility Society classifications of adnexal adhesions, distal tubal occlusion, tubal occlusion secondary to tubal ligation, tubal pregnancies, mullerian anomalies and intrauterine adhesions. *Fertil Steril* 1988; **49**: 944-955.
6. ASRM. Uterine septum: a guideline. *Fertil Steril* 2016; **106**: 530-540.
7. Valle RF, Ekpo GE. Hysteroscopic metroplasty for the septate uterus: review and meta-analysis. *J Minim Invasive Gynecol* 2013; **20**: 22-42.
8. Kowalik CR, Goddijn M, Emanuel MH, Bongers MY, Spinder T, de Kruif JH, Mol BW, Heineman MJ. Metroplasty versus expectant management for women with recurrent miscarriage and a septate uterus. *Cochrane Database Syst Rev* 2011; CD008576.
9. Rikken JF, Kowalik CR, Emanuel MH, Mol BW, Van der Veen F, van Wely M, Goddijn M. Septum resection for women of reproductive age with a septate uterus. *Cochrane Database Syst Rev* 2017; **1**: CD008576.
10. Venetis CA, Papadopoulos SP, Campo R, Gordts S, Tarlatzis BC, Grimbizis GF. Clinical implications of congenital uterine anomalies: a meta-analysis of comparative studies. *Reprod Biomed Online* 2014; **29**: 665-683.
11. Salim R, Woelfer B, Backos M, Regan L, Jurkovic D. Reproducibility of three-dimensional ultrasound diagnosis of congenital uterine anomalies. *Ultrasound Obstet Gynecol* 2003; **21**: 578-582.
12. Ludwin A, Ludwin I, Banas T, Knafel A, Miedzyblocki M, Basta A. Diagnostic accuracy of sonohysterography, hysterosalpingography and diagnostic hysteroscopy in diagnosis of arcuate, septate and bicornuate uterus. *J Obstet Gynaecol Res* 2011; **37**: 178-186.
13. Grimbizis GF, Gordts S, Di Spiezio Sardo A, Brucker S, De Angelis C, Gergolet M, Li TC, Tanos V, Brolmann H, Gianaroli L, Campo R. The ESHRE/ESGE consensus on the classification of female genital tract congenital anomalies. *Hum Reprod* 2013; **28**: 2032-2044.
14. Grimbizis GF, Di Spiezio Sardo A, Saravelos SH, Gordts S, Exacoustos C, Van Schoubroeck D, Bermejo C, Amso NN, Nargund G, Timmerman D, Athanasiadis A, Brucker S, De Angelis C, Gergolet M, Li TC, Tanos V, Tarlatzis B, Farquharson R, Gianaroli L, Campo R. The Thessaloniki ESHRE/ESGE consensus on diagnosis of female genital anomalies. *Hum Reprod* 2016; **31**: 2-7.
15. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011; **64**: 96-106.
16. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, Initiative S. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Int J Surg* 2014; **12**: 1495-1499.
17. Ludwin A, Ludwin I, Kudla M, Kottner J. Reliability of the European Society of Human Reproduction and Embryology/European Society for Gynaecological Endoscopy and American Society for Reproductive Medicine classification systems for congenital uterine anomalies detected using three-dimensional ultrasonography. *Fertil Steril* 2015; **104**: 688-697 e688.
18. Ludwin A, Ludwin I. Comparison of the ESHRE-ESGE and ASRM classifications of Mullerian duct anomalies in everyday practice. *Hum Reprod* 2015; **30**: 569-580.
19. Martins WP, Raine-Fenning NJ, Leite SP, Ferriani RA, Nastri CO. A standardized measurement technique may improve the reliability of measurements of endometrial thickness and volume. *Ultrasound Obstet Gynecol* 2011; **38**: 107-115.

20. 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.
21. Coelho Neto MA, Roncato P, Nastri 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.
22. Martins WP, Nastri CO. Interpreting reproducibility results for ultrasound measurements. *Ultrasound Obstet Gynecol* 2014; **43**: 479-480.
23. Cocchetti DV. Sample size requirements for increasing the precision of reliability estimates: problems and proposed solutions. *J Clin Exp Neuropsychol* 1999; **21**: 567-570.
24. Cicchetti DV. The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. *J Clin Exp Neuropsychol* 2001; **23**: 695-700.
25. Shoukri M, Asyali M, Donner A. Sample size requirements for the design of reliability study: review and new results. *Stat Methods Med Res* 2004; **13**: 251-271.
26. Ludwin A, Pitynski K, Ludwin I, Banas T, Knafel A. Two- and three-dimensional ultrasonography and sonohysterography versus hysteroscopy with laparoscopy in the differential diagnosis of septate, bicornuate, and arcuate uteri. *J Minim Invasive Gynecol* 2013; **20**: 90-99.
27. Van den Bosch T, Dueholm M, Leone FP, Valentin L, Rasmussen CK, Votino A, Van Schoubroeck D, Landolfo C, Installe AJ, Guerriero S, Exacoustos C, Gordts S, Benacerraf B, D'Hooghe T, De Moor B, Brolmann H, Goldstein S, Epstein E, Bourne T, Timmerman D. Terms, definitions and measurements to describe sonographic features of myometrium and uterine masses: a consensus opinion from the Morphological Uterus Sonographic Assessment (MUSA) group. *Ultrasound Obstet Gynecol* 2015; **46**: 284-298.
28. Smit JG, Kasius JC, Eijkemans MJ, Veersema S, Fatemi HM, Santbrink van EJ, Campo R, Broekmans FJ. The international agreement study on the diagnosis of the septate uterus at office hysteroscopy in infertile patients. *Fertil Steril* 2013; **99**: 2108-2113 e2102.
29. Smit JG, Overdijkink S, Mol BW, Kasius JC, Torrance HL, Eijkemans MJ, Bongers M, Emanuel MH, Vleugels M, Broekmans FJ. The impact of diagnostic criteria on the reproducibility of the hysteroscopic diagnosis of the septate uterus: a randomized controlled trial. *Hum Reprod* 2015; **30**: 1323-1330.
30. Ludwin A, Ludwin I. Reliability of hysteroscopy-based diagnosis of septate, arcuate and normal uterus: estimate or guestimate? *Hum Reprod* 2016; **31**: 1376-1377.
31. Smit JG, Torrance HL, Eijkemans MJ, Broekmans FJ. Reply: Reliability of hysteroscopy-based diagnosis of septate, arcuate and normal uterus: estimate or guestimate? *Hum Reprod* 2016; **31**: 1377-1378.
32. Saravelos SH, Li TC. Intra- and inter-observer variability of uterine measurements with three-dimensional ultrasound and implications for clinical practice. *Reprod Biomed Online* 2015; **31**: 557-564.
33. Ludwin A, Ludwin I, Pitynski K, Banas T, Jach R. Role of morphologic characteristics of the uterine septum in the prediction and prevention of abnormal healing outcomes after hysteroscopic metroplasty. *Hum Reprod* 2014; **29**: 1420-1431.
34. Ludwin A, Ludwin I, Pitynski K, Jach R, Banas T. Are the ESHRE/ESGE criteria of female genital anomalies for diagnosis of septate uterus appropriate? *Hum Reprod* 2014; **29**: 867-868.
35. Ludwin A, Ludwin I. Diagnostic rate and accuracy of the ESHRE-ESGE classification for septate uterus and other common uterine malformations: why do we not see that the Emperor is naked? *Ultrasound Obstet Gynecol* 2015; **46**: 634-636.
36. Mol BWJ, Rikken JFW. The Randomised Uterine Septum Transsection Trial (TRUST). *NTR1676* 2009; http://www.trialregister.nl/trialreg/admin/rctview.asp?TC=1676.
37. Prior M. Pilot randomised controlled trial of hysteroscopic septal resection. *ISRCTN* 2015; **ISRCTN28960271**: http://www.isrctn.com/ISRCTN28960271.

Table 1 Diagnostic rate of normal/arcuate and septate uterus according to the original ESHRE-ESGE, ASRM and other previously suggested cut-offs, and the most voted option by experts (CUME)

	Normal/Arcuate	Gray Zone	Septate
ESHRE-ESGE*	20	0	80
ASRM**	82	13	5
CUME	82	0	18

* septate = indentation to wall thickness ratio > 50%; ** normal = indentation depth < 10 mm AND indentation angle > 90°; septate = indentation depth > 15 mm AND indentation angle < 90°; Gray zone = not have the criteria to be classified as either septate or normal/arcuate.

Table 2 Agreement among experts on voting between normal/arcuate (normal uterine morphology or degree of distortion caused by the internal indentation is not clinically relevant) or septate (the degree of distortion caused by the internal indentation is clinically relevant).

	Votes for septate	kappa	95% CI	Proportion of agreement
Clinicians	138/500 (28%)	0.53	0.37-0.66	81.2%
Sonologists	92/500 (18%)	0.74	0.64-0.82	92.2%
Surgeons	127/500 (25%)	0.56	0.41-0.68	83.2%
Overall	357/1500 (24%)	0.62	0.48-0.73	86.3%

Table 3 Agreement between ESHRE-ESGE and ASRM criteria with the experts' opinion (reference standard)

	Septate	Agreement	TP	TN	FP	FN	kappa	95% CI
ESHRE-ESGE*	80	38%	18	20	62	0	0.10	-0.10 to 0.29
ASRM only septate**	5	87%	5	82	0	13	0.39	0.21 to 0.55
ASRM septate and gray-zone***	18	98%	17	81	1	1	0.93	0.90 to 0.95

* septate = indentation to wall thickness ratio > 50%; ** septate = indentation depth > 15 mm AND indentation angle < 90°; *** septate = indentation depth > 10 mm OR indentation angle < 90°; FP = false positive; FN = false negative; TP = true positive; TN = true negative.

Table 4 Inter observer agreement and reliability of the measurements.

	Difference between observers						
	Mean	SD	LoA		CCC	95% CI	
Indentation depth (mm)	0.02	0.10	-1.7	2.1	0.987	0.982	0.991
Uterine wall thickness (mm)	0.03	0.98	-1.9	2.0	0.864	0.817	0.905
Indentation angle (°)	-0.3	8.4	-16.8	16.1	0.960	0.941	0.973
I:WT ratio (%)	10.4	43.4	-74.8	95.5	0.922	0.903	0.938

SD = standard deviation; LoA = limits of agreement; CCC = concordance correlation coefficient; CI = confidence interval; I:WT = indentation to wall thickness.

Table 5 Diagnostic test accuracy of the measurements considering the most voted option by experts as the reference standard.

	AUROC	95%CI	Best cut-off	Sensitivity	Specificity
Indentation depth	1.00	0.96-1.00	≥ 10 mm	100%	99%
Indentation angle	0.96	0.90-0.99	≤ 136°	94%	91%
I:WT ratio	0.99	0.94-1.00	> 111%	100%	96%

AUROC = area under ROC curve; CI = confidence interval.

Table 6 ESHRE-ESGE, ASRM and CUME definitions of internal indentation for normal/arcuate and septate uterus and suggested cut-off values according this study results.

	I:WT	Indentation Angle	Indentation Depth	CUME/ASRM 1988 definition
Suggested by	*ESHRE-ESGE*	*ASRM*	*ASRM*	
Normal/arcuate	I:WT < 50 %	Angle > 90°	Depth < 10 mm	Not clinically relevant
Septate	I:WT > 50 %	Angle < 90°	Depth > 15 mm	Clinically relevant
Suggested	*Suggested after this study*			Best criteria
Normal/arcuate	I:WT ≤ 110%	Angle ≥ 140°	Depth < 10 mm	Depth < 10 mm
Septate	I:WT > 110	Angle < 140°	Depth ≥ 10 mm	Depth ≥ 10 mm

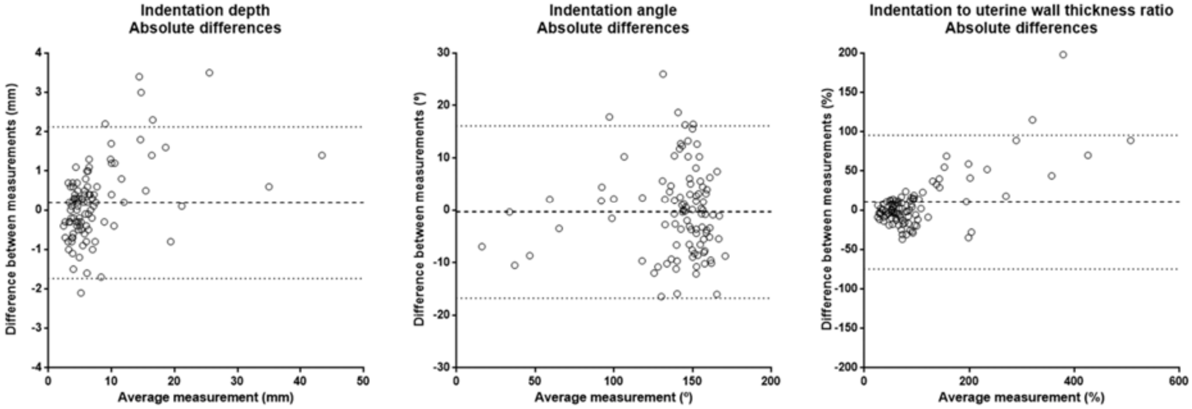

Figure 1 Bland-Altman plots for the absolute difference observed between measurements of the two observers.

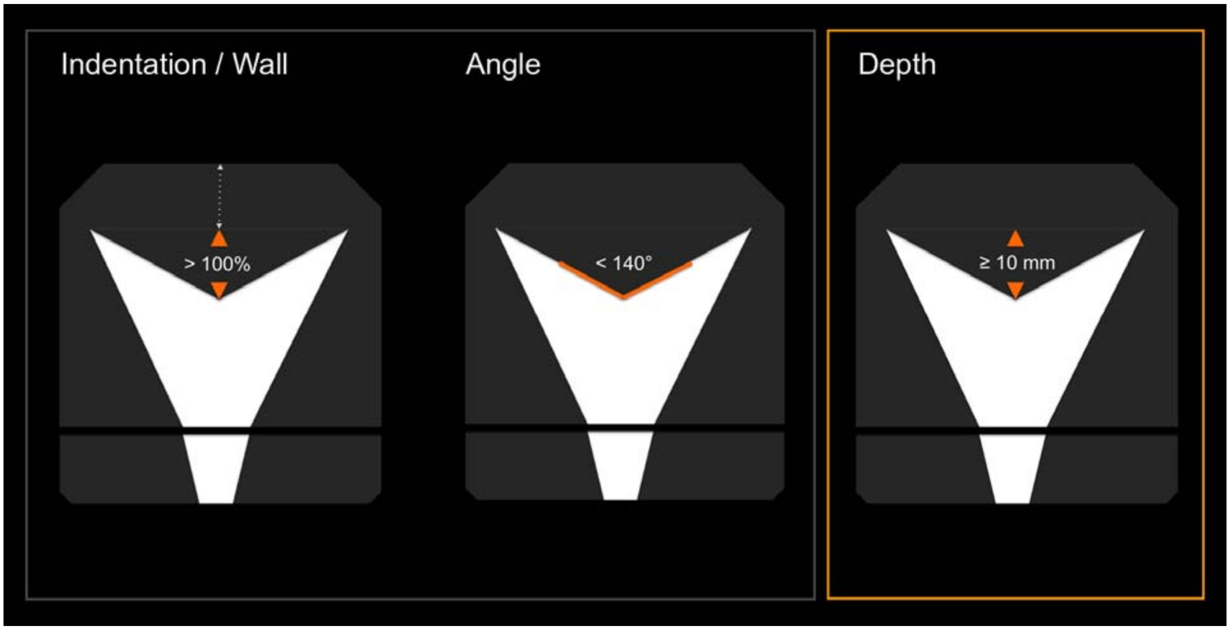

Figure 2 Criteria and the best cut-offs for distinguishing between normal/arcuate and septate uterus. Indentation depth had the highest inter-observer reliability and it is the simplest to be performed.