Classification of uterine anomalies by three-dimensional ultrasonography using ESHRE/ESGE criteria: inter-observer variability.

Running title: Uterine anomalies diagnosis by ultrasonography

Authors: C. Bermejo[1]*, P. Martínez-Ten[1], L. Ruíz-López[2], M. Estévez[1], M.M. Gil[1,3,4]

1. Delta Ecografía, Centro de Diagnóstico por la Imagen en Obstetricia, Ginecología y Mama, 28006 Madrid, Spain.
2. Gerencia de Atención Primaria de Toledo, SESCAM, 45005 Toledo, Spain.
3. Obstetrics and Gynecology Department, Hospital Universitario de Torrejón, 28850 Torrejón de Ardoz, Madrid, Spain.
4. Universidad Francisco de Victoria, Carretera de Pozuelo a Majadahonda Km 1,800, Pozuelo de Alarcón, 28223 Madrid, Spain.

Name of the institution where the study was performed: Delta Ecografía, Centro de Diagnóstico por la Imagen en Obstetricia, Ginecología y Mama, 28006 Madrid, Spain.

Correspondence
Dr Carmina Bermejo
Delta Ecografía
Calle del Conde de Peñalver, 14, Esc derecha, 1º izquierda
28006 Madrid. SPAIN
Tel: 00 34 91 432 08 91
Fax: 00 34 91 435 44 43
Mail: carminabermejo@gmail.com

Email address from all other co-authors

P. Martínez-Ten: pilarmten@gmail.com
L. Ruíz-López: lruizl@sescam.jccm.es
M. Estévez: m.estevezmortes@gmail.com
M.M. Gil: mmar1984@gmail.com

ABSTRACT

Objective: To evaluate the interobserver repeatability of the coronal view measurements and classification of the uterine malformations (UM) according to the ESHRE/ESGE consensus by transvaginal three-dimensional ultrasound (3D US).

Methods: 89 transvaginal 3D volumes acquired during the last two years at Delta Ecografía in Madrid, Spain, were selected from our archive by convenience sampling. Two expert operators blinded from each other, performed post-hoc analysis using render mode and multiplanar-VCI navigation. Uterine wall thickness at the fundus, indentation of the cavity and indentation of the fundus were measured, classified and sub-classified following the recommendations of the ESHRE/ESGE consensus. The reproducibility of interobserver measurements and classification was examined by calculating intraclass correlation coefficients (ICC) and their 95% confidence intervals (CI) and kappa statistic (k).

Results:
Repeatability in the measurements: uterine wall thickness: ICC = 0.93 (95% CI, 0.90-0.96), p<0.0001; indentation of the cavity: ICC = 0.93 (95% CI, 0.86-0.96), p<0.0001; indentation of the fundus: ICC = 0.93 (95% CI, 0.90-0.96), p<0.0001.
Level of agreement in the classification of classes and sub-classes: overall (U0, U1, U2, U3, U4, U5): k = 0.73 (95% CI, 0.61-0.84), p>0.0001; U2 (U2a, U2b): k = 0.56 (95% CI, 0.31-0.80), p<0.001 (0.78 observed agreement compared to 0.49 expected); U3 (U3a, U3b, U3c): k = 0.69 (95% CI, 0.16-1.00), p<0.05 (0.85 observed agreement compared to 0.53 expected); U4 (U4a, U4b): k = 1.00 (95% CI, 1.00-1.00), p<0.0001.

Conclusions: Transvaginal 3D US examination of the uterus allows assessment and classification of the UM by the ESHRE/ESGE criteria showing a good interobserver repeatability and reproducibility in most cases.

Key words: uterine anomalies; classification; inter-observer variability; three-dimensional imaging; ultrasonography

INTRODUCTION

According to the European Society of Human Reproduction and Embryology (ESHRE) and the European Society for Gynaecological Endoscopy (ESGE), transvaginal three-dimensional ultrasound (3D US) assessment of the uterus is the recommended technique for diagnosis and classification of female genital tract anomalies (FGTA) (Grimbizis et al, 2016). The coronal view or ‘c’ plane was the main contribution of the 3D US in this field because it allows the assessment of the relationship between the fundus of the uterus and the uterine cavity. Additionally, the coronal view is the view where, according to the ESHRE/ESGE consensus, the measurements of the uterine wall thickness at the level of the fundus, the indentation of the uterine cavity and the indentation of the fundus should be taken (Grimbizis et al, 2016). Both, 3D US and magnetic resonance (MR) can be used for this purpose, with a good correlation between them (Bermejo et al, 2010). Using these measurements, it is possible to classify the anomalies in classes and sub-classes and just in very few cases other views will be needed to perform the measurements (Grimbizis et al, 2016).

Our main objectives were firstly, to evaluate the variability of the measurements of the uterine wall thickness at the level of the fundus, the indentation of the uterine cavity and the indentation of the fundus, and secondly, to analyse the level of agreement in the classification of the FGTA of the corpus of the uterus following ESHRE/ESGE criteria between two expert operators using 3D US.

MATERIAL AND METHODS

89 three-dimensional volumes acquired during 2015 and 2016 at Delta Ecografía in Madrid, Spain, were selected by convenience sampling from our archive and two expert operators (C.B. and P.M) performed a post-hoc analysis using render mode and multiplanar-VCI navigation. C.B and P.M. have over twenty years of experience in the field, and many patients are referred to their clinic yearly as proven by the number of cases examined during the study period and their previous publications (Bermejo et al, 2010 and 2014). Two machines (Voluson Expert 730 an Voluson E8, GE Medical Systems, Zipf, Austria) with 4D vaginal probes (RIC5-9H 5–9 MHz 4D) were used for the acquisition and 4D-View 10.0 software was used for the post-hoc analysis. Using the coronal view, the uterine wall thickness at the level of the fundus, the indentation of the uterine cavity and the indentation of the fundus were independently measured by the two operators following the ESHRE/ESGE recommendations (Figures 1-3) (Grimbizis et al, 2016). The uterine wall thickness at the level of the fundus was calculated by measuring the distance in millimetres between the line joining tubal ostia (interostial line) and a parallel line on the top of the fundus. The indentation of the cavity or internal midline indentation is the distance between the interostial line and a parallel line on the top of the internal midline indentation. Finally, the indentation of the fundus or external midline indentation is also the distance between two parallel lines, one on the top of the fundus and the second on the top of the external midline indentation. Once all the measurements had been independently recorded by the two operators, each of them separately proceeded to classify and sub-classify the anomalies according to the ESHRE/ESGE classification (Grimbizis et al, 2013):

> **Class U0** or *normal uterus*: any uterus having either straight or curved fundal and interostial lines but with the internal or external indentation at the fundal midline not exceeding 50% of the uterine wall thickness (Figures 1a and 2a).
> **Class U1** or *dysmorphic uterus*: includes all cases with normal uterine outline but with an abnormal shape of the uterine cavity excluding septa. This class is further subdivided into three categories: **Class U1a** or *T-shaped uterus*, characterized by a narrow uterine cavity due to thickened lateral walls with a

correlation 2/3 uterine corpus and 1/3 cervix (Figure 4a); **Class U1b** or *uterus infantilis*, also characterized by a narrow uterine cavity without lateral wall thickening and an inverse correlation of 1/3 uterine body and 2/3 cervix (Figure 4b); **Class U1c** or *others*, which includes all minor deformities of the uterine cavity (Figure 4c and 4d).

Class U2 or *septate uterus*: includes all cases with a normal outline and an internal indentation at the fundal midline exceeding 50% of the uterine wall thickness. This septate uterus can be partial (**U2a)** (Figure 2b and 2c) or complete (**U2b)** (Figure 2d) depending on whether the septum reaches the internal os (IO).

Class U3 or *bicorporeal uterus***:** defines the uterus in which there is an external indentation at the fundal midline exceeding 50% of the uterine wall thickness. This class is further subdivided into three categories: **Class U3a** or *partial bicorporeal*, where the external fundal indentation does not reach the IO (Figure 3b); **Class U3b** or *complete bicorporeal uterus*, where the external fundal indentation is prolonged up to the IO (Figure 3c); **Class U3c** or *bicorporeal septate uterus*, which are mixed forms where, in addition to the criteria needed for U3, the width of the midline fundal indentation exceeds by 150% the uterine wall thickness (Figure 3d).

Class U4 or *hemi-uterus*: includes all cases of unilateral formed uterus. The contralateral part could be either incompletely formed or absent, being **Class U4a** or *hemi-uterus with a rudimentary* (functional) *cavity*, connected or not with the functional contralateral horn (Figure 5b); or **Class U4b** or *hemi-uterus without rudimentary* (functional) *cavity*, characterised either by the presence of a non-functional uterine horn or by aplasia of the contralateral part (Figure 5a).

Class U5 or *aplastic uterus*: includes all cases of uterine aplasia.

Class U6 is used for still unclassified cases.

For the comparison of measurements of the uterine wall thickness and the indentation of the fundus we analysed the results from U0, U1, U2 and U3 cases (n=76).
For the comparison of measurements of the indentation of the cavity we analysed U0, U2a and U3c cases (n=46).

Statistical analysis

Intra-class correlation coefficients (ICC) were calculated to analyse the inter-observer variability of the measurements and kappa statistic was calculated to analyse the level of agreement between the two operators classifying the cases into classes and sub-classes. The values for both statistics range from zero (the agreement is likely to have occurred by chance) to one (perfect agreement). The 95% confidence intervals (CI) and p values were also obtained. The kappa statistic was calculated without weighting; very good levels of agreement were considered when it is > 0.80, good 0.80-0.60, moderate 0.60-0.40, poor 0.40-0.20 and very poor < 0.20 (Lantz and Nebenzahl, 1996). ICC was calculated based on two-factors and random-effects alpha model for absolute agreement, considering it to be very good when > 0.90, good 0.90-0.80, moderate 0.80-0.60 and poor < 0.60. We also used Bland and Altman figures to graphically represent the differences between the paired measurements (y axis) and their mean (x axis) (Figures 6-8) (Bland and Altman, 1986). When there is good agreement the values tend to group close to the horizontal line, which represents complete agreement. To assess the magnitude of the difference, the mean of the differences was calculated with a 95% CI (Bland and Altman, 1986).
The statistical software packages SPSS 21.0 (SPSS Inc., Chicago, IL, USA) and EPIDAT 3.1 (Xunta de Galicia, Spain) were used for data analysis.

Ethical approval

Ethical approval was obtained from the Local Research Ethics Committee.

RESULTS
Classification of cases by each operator is shown in table 1. 73 of the 89 cases (82%) were equally classified by both operators.

Classification of U2 and U3 subclasses is represented in table 2. There was agreement between the two operators in the sub-classification of 38 cases and disagreement in 10 cases.

Variability in the measurements
Uterine wall thickness: ICC = 0.93 (95% CI, 0.90-0.96), p<0.0001 (Figure 6); indentation of the cavity: ICC = 0.93 (95% CI, 0.86-0.96), p<0.0001 (Figure 7); indentation of the fundus: ICC = 0.93 (95% CI, 0.90-0.96), p<0.0001 (Figure 8).

Level of agreement in the classification of classes and sub-classes
Overall (U0, U1, U2, U3, U4, U5): k = 0.73 (95% CI, 0.61-0.84), p>0.0001; U2 (U2a, U2b): k = 0.56 (95% CI, 0.31-0.80), p<0.001 (0.78 observed agreement compared to 0.49 expected agreement); U3 (U3a, U3b, U3c): k = 0.69 (95% CI, 0.16-1.00), p<0.05 (0.85 observed agreement compared to 0.53 expected agreement); U4 (U4a, U4b): k = 1.00 (95% CI, 1.00-1.00), p<0.0001.

DISCUSSION
During the last four decades, several classifications for Müllerian anomalies have been proposed, but the one published by the American Fertility Society (AFS, now called American Society for Reproductive Medicine, ASRM) (The American Fertility Society, 1988) was probably the most accepted among authors until 2013 when the ESHRE/ESGE work team published their recommendations. This new classification by the ESHRE/ESGE is probably the most complete to date because, based on morphometric criteria and considering the morphology and characteristics of the uterus, cervix and vagina, it has been able to classify anomalies which were previously unclassified by the AFS or ASRM (Di Spiezio Sardo et al, 2015). Although this classification does not consider the embryologic development, it is easier to use in clinical practice than that more recently proposed by Acien et al (Acien and Acien, 2016) and it only requires a diagnostic imaging technique, like the MR or the 'c' plane obtained by 3D US, able to assess the coronal view of the uterus to precisely take all the measurements described above (Graupera et al, 2015). Another advantage from these two techniques is the possibility of storing the images for a post-hoc analysis by the same or different operators. It has already been mentioned that there is a good correlation between both, MR as 3D US, when assessing the relationship between the fundus and the cavity as well as the anomalies of the cervix and the vagina (Bermejo et al, 2014). The only differences essentially relate to the cost, accessibility and patient acceptance and that is why 3D US is the recommended technique for the diagnosis of FGTA by the ESHRE/ESGE consensus leaving the MR just for selected cases (Grimbizis et al, 2016). It was out of the scope of this study to assess the accuracy of US in the diagnosis of the uterine anomalies but to ensure reproducible results, and therefore no other imaging or surgical procedures, like MR or laparoscopy, were analysed.

Main findings of the study
In this study, we have demonstrated that the inter-observer variability of the measurements taken by expert operators using 3D US volumes is very low and, consequently, there is a good level of inter-observer agreement when the same operators classify the uterine malformations. There was especially good correlation in the measurements of the uterine wall thickness, the indentation of the cavity and the indentation of the fundus. Conversely, most of the disagreements were related to the

classification of normal uterus (U0), partial septate uterus (U2a) and partial bicorporeal uterus (U3a), when the indentations of the cavity or the uterine wall were around 50% of the uterine wall thickness. Additionally, the quality of the 3D US volumes in many of these cases was poor, and an accurate measurement of the structures was not possible. Notably, the highest level of agreement (k = 1) was obtained in the classification of the hemi-uterus (U4), which is only based on morphological rather than quantitative criteria. In the sub-classification of septate uterus in U2a or U2b (partial or complete), the k statistic was only 0.56 but, due to the symmetrical distribution of the cases (represented in table 2), the observed level of agreement was 0.78 in comparison to that expected of 0.49. There was agreement in the sub-classification of 32 cases and disagreement in only 9 cases. The level of agreement was slightly better in the sub-classification of the U3 group (represented in table 2), 0.85 observed *vs* 0.53 expected, with 6 cases equally classified by both operators and one only case of disagreement. Although the sub-classification between partial septate and complete septate does not have major clinical significance from a therapeutic point of view, it is important diagnostically. A proposal for improvement could be to precisely locate the internal OS through multiplanar-VCI navigation and, once located, to use the coronal view to draw a line at that level, parallel to the interostial one. Then, it would be possible to assess the distance from that line to the septus.

When specifically comparing the measurement of the uterine wall thickness, there is no significant increase in the difference with increasing or decreasing thickness, which is rather slight (mean = 0.53 mm). In contrast, the mean difference of the measurement of the internal midline indentation was 1.58 mm, increasing with increasing values. When comparing the measurements of the internal midline indentation (septal length) we excluded the complete septate uterus because in these cases it is not possible to visualise the top of the septum, which is actually the deepest part, and their classification mainly relies on morphologic criteria based on the relation between the septum and the IO (Figure 2d). When analysing the reproducibility of the measurement of the external midline indentation, in addition to a high level of agreement between operators, it is important to emphasize that most of the cases (62 out of 76) presented a difference of zero.

Limitations of the study

The main limitation of the study is the convenience selection of cases and the fact that, our sample, enriched with borderline cases, does not correlate with the distribution of FGTA in the general population. However, previous studies have demonstrated that the disagreements are normally found in this group (Bermejo et al, 2014; Graupera et al, 2015; Ludwin and Ludwin, 2016) and therefore, our positive results may even be enhanced by this biased distribution.

In this study, all the acquisitions and measurements were performed by two expert operators in the diagnosis of FGTA by US examination. However, this is a relative limitation since the ESHRE/ESGE consensus recommends that this kind of assessment should be only performed by experts. Some of the volumes included of this study were classified as “poor quality” and accurate measurements were not possible to obtain. This is something that will always occur in cases of hypertrophic and/or adenomyotic uterus in addition to those with multiple fibroids or indifferent position, regardless of the experience of the operator. In these cases, a second volume acquired from the transverse view may help; non-the-less, a poor quality 2D image will always correlate to a poor quality 3D image.

Comparison with previous studies

In a recent study by Graupera et al, 60 women were studied using both 3D US and MR following a conventional US examination which was suggestive of a UA (Graupera et al, 2015). They applied the ESHRE/ESGE recommendations to classify the anomalies and reached a good level of agreement between the two techniques. The only two cases of aberrations where the classification was different by each technique (U2/U0 and U2/U3) were attributed to shubtle differences found when the measurements of the indentations were close to 50% of the uterine wall thickness. Similarly, in our previous study we also found the interobserver disagreements within the same group of cases (Bermejo et al, 2014).

Another study by Ludwin and Ludwin, prospectively analysed 261 women and classified the FGTA found according to both, the ASRM (with additional morphometric criteria) and the ESHRE/ESGE criteria (Ludwin and Ludwin, 2016). Using the latter, they over-diagnosed septate uterus arising from those cases classified as arcuate uterus by the ASRM plus additional criteria (1.5 mm > internal midline indentation >1.0 mm). They conclude that, since an over-diagnosis leads to an overtreatment, the ESHRE/ESGE classification should not be used until it is fully reviewed and validated. As per our study, their sample did not represent the distribution of FGTA in a general population, and most of the selected cases were found around the marginal indentations (50% normal / 50% septate) following ESHRE/ESGE classification but not according to their version of the ASRM classification, which represents a major source of bias in their study.

Clinical implications

The septate uterus is the consequence of a lack of reabsorption of the intermullerian septum after both sides are fused. The arcuate uterus represents the mildest side of the spectrum, thus corresponding to a very mild form of septate uterus[12]. In the ASRM classification it has its own category because of its clinical irrelevance. Most of the authors use additional non-standardised criteria such as the internal indentation angle or the length of the septum in order to diagnose this kind of uterus. Currently, the ASRM has also recommended using morphometric criteria, like the internal indentation angle and the measurement of the internal midline indentation, to distinguish between the arcuate uterus and the septate uterus (indentation angle < 90º and length of the internal midline indentation > 1.5 cm), because only the latter are related to a poor reproductive prognosis; but still there is no clear definition of what a septate uterus is. In contrast, the ESHRE/ESGE consensus does not contemplate the arcuate uterus as a FGTA and considers all cases with internal midline indentation < 50% of the uterine wall thickness to be normal uterus. Therefore, most of the arcuate uterus by the ASRM would be classified as U0 and some as U2a by ESHRE/ESGE. The concern here would arise only if all partial septate uterus or non-classifiable malformations were surgically managed. By this assumption, the few millimetres difference in measurement and therefore in classification showed in our study and also in Graupera's study, could be of clinical importance since they may lead to unnecessary treatment. It is important to be mentioned that it is not yet clear what length of internal midline indentation has any clinical significance and the 1.5 cm cut-off was arbitrarily chosen but not based on the existing data. Interestingly, the ASRM arcuate uterus has been related to adverse outcome, such as second trimester pregnancy losses (Venetis et al, 2014).

Actually, there is no standardised treatment for the septate uterus since there are no randomized control trials comparing surgery versus expectant management. The conclusion from a meta-analysis of 204 studies performed by the ASRM was that, with grade B evidence, 3D US, 3D hysterosonography and MR are useful in the diagnosis of septate and bicornuate uterus. Although there is grade B evidence that septate uterus increases the risk of miscarriage, preterm birth, fetal malposition, fetal growth restriction, abruption and perinatal mortality (Practice Committee of the American Society for Reproductive Medicine, 2016), there is not enough evidence to assure that septate uterus is a cause of infertility (grade C evidence), and therefore septostomy has not been demonstrated to improve reproductive or obstetric prognosis. Moreover, septostomy may have secondary effects like adherences which, although uncommon, are difficult to prevent. For this reason, the ASRM is currently working on improving the guidelines for management of the septate uterus.

To conclude, the coronal view of the uterus acquired by 3D US allows the measurement of the uterine wall thickness at the level of the fundus, in addition to the indentation of the cavity and the indentation of the fundus, showing a very good inter-observer repeatability. Additionally, it allows the classification of the UM by the ESHRE/ESGE criteria showing a good inter-observer reproducibility in most cases. However, further studies in routine populations should be performed to analyse the performance of this technique in a low-risk population.

ACKNOWLEDGEMENTS
We would like to express our gratitude to Professor Grigoris F. Grimbizis, First Department of Obstetrics & Gynecology, Aristotle University of Thessaloniki, Greece, for comments that greatly improved the manuscript.
Hospital Universitario de Torrejón ensured the research was conducted according to the Declaration of Helsinki for Medical Research involving Human Subjects.

CONFLICT OF INTEREST
The authors declare no competing interests.

REFERENCES

1. Acien P, Acien M. Diagnostic imaging and cataloguing of female genital malformations. Insights Imaging 2016;7:713–726.
2. Bermejo C, Martinez Ten P, Cantarero R, et al. Three-dimensional ultrasound in the diagnosis of Mullerian duct anomalies and concordance with magnetic resonance imaging. Ultrasound Obstet Gynecol 2010;35:593-601.
3. Bermejo C, Martínez-Ten P, Recio M, et al. Three-dimensional ultrasound and magnetic resonance imaging assessment of cervix and vagina in women with uterine malformations. Ultrasound Obstet Gynecol 2014;43:336–345.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; i:307-310.
5. Di Spiezio Sardo A, Campo R, Gordts S, et al. The comprehensiveness of the ESHRE/ESGE classification of female genital tract congenital anomalies: a systematic review of cases not classified by the AFS system. Human Reprod 2015;30:1046-1058.
6. Graupera B, Pascual MA, Hereter L, et al. Accuracy of three-dimensional ultrasound compared with magnetic resonance imaging in diagnosis of Müllerian duct anomalies using ESHRE-ESGE consensus on the classification of congenital anomalies of the female genital tract. Ultrasound Obstet Gynecol 2015; 46: 616-622.
7. Grimbizis G, Gordts S, Di SpiezoSardo A, et al. The ESHRE/ESGE consensus on the classification of female genital tract congenital anomalies. Hum Reprod 2013;28:2032-2044.
8. Grimbizis GF, Di Spiezio Sardo A, Saravelos SH, et al. The Thessaloniki ESHRE/ESGE consensus on diagnosis of female genital anomalies. Gynecol Surg 2016; 13: 1-16. Human Reprod 2016;31:2-7.

9. Lantz CA, Nebenzahl E. Behavior and interpretation of the k statistic: Resolution of the two paradoxes. J Clin Epidemiol 1996;49:431-434.

10. Ludwin A, Ludwin I. Comparison of the ESHRE-ESGE and ASRM classifications of Müllerian duct anomalies in everyday practice. Hum Reprod 2016;30:569-580.

11. Practice Committee of the American Society for Reproductive Medicine; American Society for Reproductive Medicine, Birmingham, Alabama. Uterine septum: a guideline. Fertil Steril 2016;3:530-540.

12. The American Fertility Society. The American Fertility Society classifications of adnexal adhesions, distal tubal obstruction, tubal occlusion secondary to tubal ligation, tubal pregnancies, Mullerian anomalies and intrauterine adhesions. Fertil Steril 1988;49:944-955.

13. Venetis, C, Papadopoulos S, Campo R, Gordts S, Tarlatzis B, Grimbizis G. Clinical implications of congenital uterine anomalies: a meta-analysis of comparative studies. BioMedicine Online 2014;29:665–683.