The impact of diagnostic criteria on the reproducibility of the hysteroscopic diagnosis of the septate uterus: a randomized controlled trial

Janine G. Smit1,*, Sanne Overdijkink1, Ben W. Mol2, Jenneke C. Kasius1, Helen L. Torrance1, Marinus J.C. Eijkemans3, Marlies Bongers4, Mark H. Emanuel5, Michael Vleugels6, and Frank J.M. Broekmans1

1Department of Reproductive Medicine and Gynecology, University Medical Center Utrecht, PO Box 85500, Utrecht, GA 3508, The Netherlands
2The Robinson Institute, School of Paediatrics and Reproductive Health, University of Adelaide, Adelaide, SA 5000, Australia
3University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, Utrecht, CG 3584, The Netherlands
4Department of Obstetrics and Gynecology, Maxima Medical Centre, Veldhoven, DB 5504, The Netherlands
5Department of Obstetrics and Gynecology, Spaarne Hospital, Hoofddorp, TM 2134, The Netherlands
6Department of Obstetrics and Gynecology, Rivierenland Hospital, Tiel, WP 4002, The Netherlands

*Correspondence address. E-mail: jsmit6@umcutrecht.nl

Submitted on December 15, 2014; resubmitted on March 4, 2015; accepted on March 25, 2015

STUDY QUESTION: Does the use of diagnostic criteria in the hysteroscopic diagnosis of a septate uterus improve inter-observer agreement?

SUMMARY ANSWER: Pre-set diagnostic criteria slightly improve the inter-observer reproducibility of hysteroscopy in diagnosing a uterine septum, although agreement remains moderate.

WHAT IS KNOWN ALREADY: The inter-observer agreement on the hysteroscopic diagnosis of the septate uterus has been reported to be poor.

STUDY DESIGN, SIZE, DURATION: From April 2013 until May 2014, a randomized controlled comparative inter-observer study was performed. A total of 191 gynecologists from 43 countries took part.

PARTICIPANTS/MATERIALS, SETTINGS, METHODS: Each gynecologist was asked to assess 10 video recordings of hysteroscopy procedures with a specific focus on the internal uterine shape. The hysteroscopies had been performed in subfertile women and women with recurrent miscarriage. The recordings contained images of uterine cavities primarily diagnosed as septate, arcuate or normal. Participating gynecologists were randomized into two groups: one group received diagnostic criteria for a septate uterus before assessment of the videos (DC group), whereas the other group assessed the recordings without instruction (no DC group). The inter-observer agreement, expressed as the intra-class correlation coefficient (ICC), was compared between groups. Main outcomes were the inter-observer agreement on the uterine shape and the necessity of surgical correction.

MAIN RESULTS AND THE ROLE OF CHANCE: Eighty-six observers were randomized to the DC group and 105 to the no DC group. The ICCs in the diagnosis of a septum were 0.59 versus 0.52, in the DC group and the non-DC group, respectively (P-value: 0.002). The overall agreement on the need for surgical correction was found to be moderate (DC ICC 0.43 versus no DC 0.39, P-value: 0.70). Most importantly, once a septate uterus had been diagnosed, the agreement on the need for surgery was poor in both groups (DC ICC 0.05 versus no DC ICC 0.02, P-value: 0.78).

LIMITATIONS, REASONS FOR CAUTION: We used video recordings rather than studying real-time hysteroscopic procedures, which may have influenced the accuracy of the assessments.

WIDER IMPLICATIONS OF THE FINDINGS: The reproducibility of hysteroscopy for the diagnosis of a septate uterus is moderate, even with the use of standardized criteria. The fact that the agreement among physicians on both the diagnosis of a uterine septum, as well as the decision to resect such septum after hysteroscopy is moderate, may imply that hysteroscopy is insufficient as single tool to diagnose and decide on treatment of a septate uterus.
Office hysteroscopy is increasingly used as a screening tool in subfertile women. The reported prevalence of intrauterine abnormalities in this population shows considerable variation (11–45%) (Oliveira et al., 2003; Bosteels et al., 2010; Fatemi et al., 2010). This variation is not only seen for small, unsuspected abnormalities, but also for profound malformations such as the septate uterus (Grimbizis et al., 2001). Although the American Fertility Society (AFS) classification is used worldwide to categorize uterine anomalies, an exact definition of class V (septate uterus) and VI (arcuate uterus) abnormalities is not provided (AFS, 1998). In contrast, the new European Society of Human Reproduction and Embryology (ESHRE) and European Society for Gynaecological Endoscopy (ESGE) classification defines a septate uterus as a uterus with a normal outline and an internal indentation at the fundal midline exceeding 50% of the uterine wall thickness (Grimbizis et al., 2013). This classification is of limited use for hysteroscopy, as the thickness of the uterine wall cannot be determined during this procedure.

A septate uterus results from abnormal resorption of the Mullerian ducts during fetal life. High-quality studies on the clinical impact of the septate uterus in fertile women are lacking but the condition seems to be associated with adverse pregnancy outcomes (increased miscarriage (Raga et al., 1997; Chan et al., 2011), preterm delivery and breech presentation rates (Homer et al., 2000; Fox et al., 2014)). The effect of a septum on fertility is less clear. The prevalence of the septate uterus in the infertile population is similar compared with a population of patients who underwent investigation of the uterus for other indications (e.g. sterilization, pelvic pain and abnormal bleeding) (Chan et al., 2011). This suggests that a uterine septum does not play a role in the process of conception as such. Not only the septate uterus, but also the arcuate uterus is characterized by a certain degree of indentation of the fundus into the uterine cavity. The arcuate uterus is a variation of normal uterine anatomy and is not associated with adverse pregnancy outcomes or infertility (Lin et al., 2002; Mucowski et al., 2010; Gergolet et al., 2012; Grimbizis et al., 2013). Due to the differences in pregnancy outcome, correct diagnosis of both conditions is essential. In a recent study, however, the agreement on the diagnosis of the septate uterus based on hysteroscopy was demonstrated to be poor with differentiation between a septate and arcuate uterus appearing to be especially difficult (Smit et al., 2013).

The data mentioned above suggest that the reproductive outcome of women diagnosed with a uterine septum is compromised. These suggestions, together with the fact that a uterine septum can easily be corrected, have led to the performance of septum resection in daily practice, thus assuming that this optimizes pregnancy chances. Surprisingly, and maybe unfortunately, randomized controlled trials evaluating the effectiveness of hysteroscopic septum resection have never been performed, although currently such a trial is under execution in The Netherlands (NTR number 1676). Although randomized clinical trials are lacking, septum resection can only be effective if the diagnosis of a uterine septum can be made accurately. In absence of a perfect diagnostic gold standard, reproducibility of the diagnosis is a prerequisite for an accurate diagnosis, which in itself is needed to allow subsequent treatment to be effective.

The development of standardized definitions for the hysteroscopic diagnosis of the septate uterus might increase consensus among gynecologists, and thereby the accuracy of office hysteroscopy as a screening tool for uterine abnormalities. Therefore, the aim of this study is to assess the reproducibility of the hysteroscopic diagnosis of a uterine septum, both without and with the instruction for criteria to diagnose a septate uterus.

**Materials and Methods**

We performed a randomized controlled reproducibility study involving 191 gynecologists from 43 countries. An email was sent to all ESGE members with an e-mail invitation to participate in the DUTCHIES (Diagnostic Criteria for the Hysteroscopic diagnosis of the Septate uterus) study. We asked them to assess 10 hysteroscopy recordings with specific focus on the internal uterine shape and to fill out a short questionnaire for each recording. Participants could access the questionnaire and the hysteroscopy recordings by using the link in the e-mail. The questionnaire was built using an open-source survey application ([https://www.limesurvey.org](https://www.limesurvey.org)). In addition, gynecologists visiting a national Dutch conference (19th Dutch/Flemish Doelen Conference on Infertility, Gynecology and Obstetrics, [http://www.igoandpractice.nl](http://www.igoandpractice.nl)) were asked to participate in the study and assess the recordings on computers at the study stand. Characteristics of the included observers were recorded, including date of birth, level of medical specialization, years of experience performing hysteroscopy, the estimated amount of hysteroscopies performed, country of OB-GYN residency and country of current affiliation.

**Hysteroscopy recordings**

The video recordings of hysteroscopy procedures had been performed in subfertile women or women with recurrent miscarriage and were placed online after they had been made anonymous. The recordings were performed in the context of the InSIGHT (Significance of routine Hysteroscopy prior to a first IVF Treatment cycle, register number: NCT01242852) and TRUST trial (The Randomized Uterine Septum Transection Trial, NTR number 1676). The studies were approved by the institutional review board and informed consent was obtained.

The selected recordings contained images of uterine cavities previously diagnosed as septate, arcuate or normal based on hysteroscopy and laparoscopy or MRI. Each recording was diagnosed and selected by one of the four leading experts in the field (F.B., M.H.E., M.B. and M.V.). The procedures had been performed in the early follicular phase of the cycle (cycle Day 3–12) in three hospitals. In one hospital, hysteroscopies were performed at a random moment in the menstrual cycle. The instruments and distention medium used varied between the hospitals. Karl Storz continuous flow hystoscopes 5-mm and 5.5-mm or 8-mm Olympus hysteroscopes were used. Each procedure was recorded. The recordings were edited so that every recording started at the passage of the internal os and ended before leaving the external os of the cervix.

**Key words:** hysteroscopy / septate uterus / diagnostic criteria / inter-observer agreement / reproducibility
**Definition of the septate uterus**

In our previous study, we asked gynecologists for their opinion on the value of several hysteroscopy characteristics of a septate uterus (Smit et al., 2013). Based on this study and after an extensive literature search, several diagnostic criteria were defined that may distinguish between a normal, arcuate and septate uterus. The systematic search was conducted in PubMed, Embase and the Cochrane library using the following keywords and their synonyms: (‘uterine malformation’ or ‘septate uterus’ or ‘arcuate uterus’) and (‘classification system’ or ‘diagnostic criteria’ or ‘definition’). An additional search, using web of Science and Scopus, was conducted in order to reveal any missed articles. The search resulted in nine relevant articles (Buttram and Gibbons, 1979; Buttram, 1983; Saleem, 2003; Pui, 2004; Troiano and McCarthy, 2004; Oppelt et al., 2005; Mueller et al., 2007; Gubbini et al., 2009; Faivre et al., 2012). None of the articles described criteria specific for the hysteroscopy diagnosis of a septate uterus but characteristics for the septate uterus in general or criteria used for MRI and three-dimensional sonography. After the literature search for diagnostic criteria was completed, a proposal of possible diagnostic criteria was compiled by J.S. and S.O. Any disagreement between the two authors was resolved through discussion with J.K. and F.B. Subsequently, members of the general board and the executive board of the ESGE and the members of the ESHRE special interest group of reproductive surgery were asked to judge the proposed diagnostic criteria and add suggestions that could possibly improve those criteria. Ten of the 25 members that were contacted responded to our request.

The process resulted in three diagnostic criteria (Fig. 1). The first diagnostic criterion deals with the extent of the intracavitary development. Fundal bulging protruding <30% into the uterine cavity suggests an arcuate uterus, whereas 30% or more fundal bulging is suspect for a septate uterus. The second criterion is the angle at the central point of the indentation of the fundus. An obtuse angle (>90°) indicates an arcuate uterus, whereas a sharp angle of 90° or less indicates a septate uterus. The last criterion considered the width of the bulging structure compared with the distance between both tubal orifices. An intercornual distance of 30% or more suggests an arcuate uterus, whereas an intercornual distance of <30% suggests a septate uterus.

**Evaluation**

Standardized assessment forms were used for the evaluation of the hysteroscopy recordings (Supplementary, Fig. S1). The assessment forms contained three questions per recording about the quality of the recording (good, intermediate and poor), internal shape of the uterine cavity (normal, arcuate and septate) and the observers’ opinion or advice on the need for surgical correction (yes and no). All observers were instructed to assess the recordings presuming that a normal outer contour of the uterus had been diagnosed by laparoscopy or MRI. Observers in the DC group were offered the pre-set diagnostic criteria before assessment of the video recordings. The observers were instructed that they had to use the diagnostic criteria as formulated by the study to diagnose the uterine shape and that the diagnostic criteria were available after every video recording. The diagnostic criteria were explained by illustrations of the diagnostic criteria (Fig. 1). The observers in the DC group were informed that the proposed diagnostic criteria could possibly improve those criteria. Ten of the 25 members that were contacted responded to our request. The group without DC assessed the recordings without pre-set definitions of the

![Figure 1](https://academic.oup.com/humrep/article-abstract/30/6/1323/617164/30612361716 by Biomedical Library User on 11 January 2019)

**Figure 1** Diagnostic criteria for the septate uterus. (A) Fundal bulging protruding <30% in the uterine cavity suggests an arcuate uterus. Thirty percent or more fundal bulging suggests a septate uterus. (B) An obtuse angle (>90°) indicates an arcuate uterus. An acute angle (<90°) indicates a septate uterus. (C) The width of the bulging structure compared with the distance between the tubal orifices suggests an arcuate uterus when ≥30% and a septate uterus when <30%.
septate and arcuate uterus. Two authors (J.S. and S.O.) screened all assessment forms. Duplicates were removed and observers were excluded from analysis if ≤ 4 video recordings were assessed. Only questionnaires filled out before the second randomization were used when an observer was randomized more than once.

Outcomes

The main outcome of this study was the inter-observer agreement on the hysteroscopic diagnosis of the septate and arcuate uterus in the group with and without pre-set DC. Moreover, the inter-observer agreement on the necessity for surgical correction and the opinion on the quality of the recordings were assessed.

Sample size calculation

We calculated the required sample size by computer simulation. For various scenarios of values of sample size and the relevant difference in ICC, random datasets were simulated and the power was assessed using the fraction of simulated datasets in which the mixed model analysis revealed a statistical significance difference. The inter-observer agreement on the hysteroscopic diagnosis of the septate and arcuate uterus appeared to be fair in a previous study (ICC: 0.27) (Smit et al., 2013). The level of agreement would have to improve to an ICC of at least 0.35 in the group with DC to be considered clinically relevant. With 10 hysteroscopy recordings being used, the number of observers needed to have 90% power to detect such a difference, with a Bonferroni correction because we will look at four different determinants (Shape, Correction, Correction if septate and Quality) ($\alpha = 0.05/4 = 0.0125$), is 90 per study arm.

Statistical analysis

Descriptive statistics were used to analyse baseline characteristics. We used Mann–Whitney U-tests and $\chi^2$ test to compare continuous and dichotomous data, respectively. The inter-observer agreement on the uterine shape and necessity for correction was calculated for both groups. Because the aim of the study was to investigate the impact of diagnostic criteria on the level of inter-observer agreement, the agreement within both groups was compared and not the agreement between both groups. The inter-observer agreement was expressed as the intraclass correlation coefficient (ICC). The ICC is an approximation of the overall weighted k and was calculated to mean k-values (Fleiss and Cohen, 1973). A k-value (and therefore ICC) of <0.20 represents poor agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, while a value of 0.81–1.00 indicates almost perfect agreement (Landis and Koch, 1977). The ICC was calculated from a linear mixed model as ICC = variance (recording)/[Variance (recording) + residual variance]. A linear mixed model was used to calculate the variance using the video recording as a random effect, thereby adjusting for within person repetitions. The difference in reliability between the group with and without DC was computed by a linear mixed model using the group as fixed effect, including different residual variances per group. A formal likelihood ratio test was used to determine the impact of DC on the ICC by comparing the model with and without DC. Statistical analysis was performed using SPSS version 20.0 (SPSS Inc.) and R version 3.1.0 (http://www.r-project.org/).

Subgroup analysis

To assess the impact of the differentiation between the normal and arcuate uterus on the inter-observer agreement, subgroup analysis was performed in which the normal and arcuate uterus were categorized as one diagnosis (normal or arcuate versus septate). To investigate the general opinion on the necessity for surgical correction of a septate uterus among participants in this study, an additional analysis was performed. In this subgroup analysis, recordings were only included if a septate uterus was diagnosed.

Results

Randomization for this study was open from April 2013 until May 2014. A total of 389 observers were randomized, of which 30 observers assessed the recordings during the Dutch Doelen Conference. The randomization link for online assessment of the video recordings was used 359 times. After online randomization, 54 assessments were excluded because of double randomization (i.e. when the randomization link was used more than once by the same observer). 125 participants were excluded because they did not assess any of the video recordings, 19 observers were excluded because the questionnaire was incomplete (assessment of ≤ 4 video recordings). Finally, 191 observers from 43 different countries were included in the analysis, of which 86 participants were included in the DC group and 105 in the no DC group (Fig. 2).

Observers

The group of observers consisted of 174 gynecologists and 17 residents. Their median [interquartile range] age was 46 years [39–55], the median number of years of hysteroscopy experience was 10 [5–20] and the median number of hysteroscopies performed was 600 [150–2000]. No significant differences in baseline characteristics were present between the DC and no DC group (Table I).

Inter-observer agreement

The quality of each recording (n = 10) was judged by every observer (n = 191). The quality of the recording was scored as good or intermediate in 77% of the cases (1468/1910). In both groups, agreement for recording 2 and 3 on the diagnosis of a septate uterus was near complete, 183/191 and 185/191, respectively (Fig. 3). These recordings showed a distinct clear septum (recording 2, Supplementary, Fig. S2). Agreement on the need for surgical correction of the uterine abnormality was moderate (ICC: 0.43 versus 0.39, P-value: 0.7) (Table II).

Subgroup analysis

The inter-observer agreement for the diagnosis of the uterine shape, categorized as normal or arcuate versus septate, was substantial for the DC group and moderate for the no DC group (ICC: 0.64 versus 0.58, P-value: 0.08). Most importantly, the agreement on the need for surgery when a septate uterus had been diagnosed was poor (DC ICC 0.05 versus no DC ICC: 0.02, P-value: 0.78).
Discussion

To our knowledge, this is the first study that investigates the impact of the application of pre-set criteria on the reproducibility of the hysteroscopic diagnosis of the septate uterus. This study shows that using diagnostic criteria improves the inter-observer agreement for the hysteroscopic diagnosis of the septate uterus, but that the level of agreement remains moderate. The clinical relevance of the improved diagnostic accuracy may therefore be questioned. The level of agreement became substantial when the arcuate uterus was incorporated in the definition of a normal uterus in the subgroup analysis, although this did not reach statistical significance. The inter-observer agreement was highest for video recordings showing uterine cavities with a distinct septum (recording 2 and 3, Fig. 3; Supplementary, Fig. S2). This suggests that differentiation of the normal or classical septate uterus from less explicit abnormalities, such as the arcuate uterus, is especially difficult. The agreement probably is better when diagnosing the extremes (e.g. total septum and normal uterus). Also, the consensus on the need for surgical correction was only moderate (ICC: 0.41), and extremely poor once a septum had been diagnosed (ICC: 0.05).

Strengths of this study are its randomized design and the international setting providing a proper representation of the current global agreement on the septate uterus. The gynecologists that participated were representative for the overall group of physicians specialized in reproductive surgery. As shown in Table 1, the participants originated from 43 countries from different continents and were experienced with performing hysteroscopy. Additionally, the diagnostic criteria were formulated based upon expert opinion and the current literature. Furthermore, a power calculation was performed. An alternative study design would have been to investigate the agreement of observers before and after training for diagnostic criteria. However, this study design would lead to methodological problems such as carry-over effects. Some limitations of this study also need to be addressed. We excluded participants after randomization if videos were not assessed and consequently, predominantly well-motivated gynecologists may have been included in the study. On the other hand, even motivated physicians may have been disturbed while preparing for the questionnaire and later on forgot to continue. This latter may well have been mostly according to chance effects and is not likely to have influenced results.

Randomization was performed to reduce the risk of confounding by
creating equality in the observer groups. By the exclusion of participants, the distribution between groups could have been disrupted if differential dropout was present. However, this seems to be unlikely, because no significant differences were present in baseline characteristics, while the dropout rate (52% (n = 94) in the DC group and 50% (n = 104) in the no DC group (Figure 2)), and reasons for exclusion were equally distributed between both groups. Another possible limitation is that the sample of observers might have been a sample of convenience rather than a random sample. Unfortunately, this is an inevitable risk of an online questionnaire as participation will depend on the willingness of the gynecologists to participate. However, it is barely feasible to obtain a truly random sample from a human population, as people are allowed to choose whether they participate or not. Furthermore, the use of video recordings to assess the internal uterine shape might be considered as a limitation. The reproducibility of hysteroscopy may have been affected because the participant was dependent on the skills of the hysteroscopy performer for the assessment of the uterine shape. However, the hysteroscopies were performed by four experienced gynecologists and the quality of the recordings was judged to be good or intermediate in 77% of the cases. Moreover, the observers assessed the hysteroscopy recordings under the baseline information that the outer contour of the uterus was normal. In practice, hysteroscopy is often performed prior to laparoscopy and this information is not available. The final study limitation is that the proposed diagnostic criteria cannot be objectified by measurement during hysteroscopy and therefore might be difficult to use. Although the diagnostic criteria can support the right diagnosis, assessment of the internal uterine shape remains subjective. This could have influenced the extent of improvement in agreement when diagnostic criteria are used. The impact of diagnostic criteria may therefore be higher for diagnostic modalities, such as three-dimensional sonography and MRI, where different features of the septate uterus can be measured.

Table 1  Characteristics of participants Categorical data are given as number (%), continuous data are given as median [interquartile range].

<table>
<thead>
<tr>
<th></th>
<th>Diagnostic criteria</th>
<th>No diagnostic criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of participants</td>
<td>86</td>
<td>105</td>
</tr>
<tr>
<td>Specialist OBGYN</td>
<td>80 (93)</td>
<td>94 (90)</td>
</tr>
<tr>
<td>Resident OBGYN</td>
<td>6 (7)</td>
<td>11 (11)</td>
</tr>
<tr>
<td>Age in years</td>
<td>45 [37–55]</td>
<td>47 [39–54]</td>
</tr>
<tr>
<td>Years of experience</td>
<td>10 [5–19]</td>
<td>10 [5–20]</td>
</tr>
<tr>
<td>performed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residence (no. of participants per continent)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>63 (73)</td>
<td>86 (81)</td>
</tr>
<tr>
<td>Asia</td>
<td>13 (15)</td>
<td>8 (8)</td>
</tr>
<tr>
<td>South America</td>
<td>4 (5)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Middle East</td>
<td>2 (2)</td>
<td>9 (9)</td>
</tr>
<tr>
<td>Africa</td>
<td>1 (1)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Unknown</td>
<td>3 (3)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

Figure 3  Assessments on the internal uterine shape. This figure shows the assessments by the groups of observers with or without the pre-set diagnostic criteria, of the internal uterine shapes shown in the video recordings. Blue: normal uterine cavity, green: arcuate uterus, brown: septate uterus.
The level of agreement in this study was higher for both the DC and no DC group (ICC: 0.59 and ICC: 0.52) compared with our previous study investigating the inter-observer agreement for the diagnosis of the septate uterus at hysteroscopy (ICC: 0.27) (Smit et al., 2013). This difference might be explained by the recently published classification system for uterine malformations by the ESGE and ESHRE (Grimbizis et al., 2013). This classification system gives a precise description of the septate uterus and might have led to more consensus, on the exact definition of the septate uterus among gynecologists worldwide. At the moment of our first inter-observer study, only the AFS-classification system was available for the classification of uterine malformations. This classification system lacks a precise description of the septate and arcuate uterus making it hard to distinguish between both conditions. Differences in the quality of the recordings or hysteroscopy procedures might also explain the different level of agreement. However, the opinion of the observers on the quality of the recordings in both studies was comparable (this study: 77% good/moderate, Smit et al., 2013: 81% good/moderate). In current practice, hysteroscopy combined with laparoscopy is still considered to be the gold standard for diagnosing the septate uterus. However, based upon this study the reproducibility in diagnosing the septate uterus should be questioned. Other diagnostic modalities such as three-dimensional sonography and MRI have proved to be highly accurate in diagnosing uterine malformations, including the septate uterus (Ghi et al., 2009; Bermejo et al., 2010). Also, the reproducibility of both techniques appears to be almost perfect. For three-dimensional sonography a $\kappa$-value of 0.97 (95% CI: 0.94–1.00) for diagnosing uterine malformations has been described and for MRI a high level of agreement for diagnosing uterine cavity pathology was found (mean $\kappa$: 0.97, 95% CI: 0.91–1.00) (Dueholm et al., 2002; Salim et al., 2003). Also for exact measurements, such as septum size, MRI and three-dimensional ultrasound reveal comparable results (Ata et al., 2014). Compared with hysteroscopy, three-dimensional sonography and MRI seem to be good or even better alternatives to diagnose the septate uterus. The gold standard for diagnosing the septate uterus should therefore be reconsidered and specified by the ESGE/ESHRE classification system. This might be necessary to achieve uniformity in the diagnosis and treatment of uterine malformations. The new classification system for uterine malformations published by the ESGE and ESHRE may have led to more consensus on the exact definition of the septate uterus in general. However, it cannot be used for hysteroscopy because the uterine wall thickness is used as reference point. The diagnostic features as formulated in the new classification system can be measured by both three-dimensional sonography and MRI.

Correct diagnosis of the septate uterus is necessary because of the differences in pregnancy outcome. Although, treatment of the septate uterus is still subject of debate, in some clinics septum resection is performed on regular basis. The accuracy of hysteroscopy in diagnosing the septate uterus is less than expected. Therefore, hysteroscopy as a single tool to diagnose a septate uterus is considered to be insufficient. Also, treatment of the septum may better be based on further diagnostic steps such as MRI or three-dimensional ultrasound. Misclassifying a septate uterus might lead to overtreatment and unnecessary harm to the patient. The results of our study jeopardize the value of studies on the effectiveness of septum resection for women with subfertility or recurrent miscarriage. When the same woman has a septum and needs surgery according to one gynecologist, while another gynecologist does not diagnose that septum, overall treatment effect will be limited as one woman gets offered resection with one doctor, but not with another. This implicates that the potential treatment effect can at maximum be 50% of what would be achieved when all women were correctly diagnosed. The total absence of reproducibility for the indication for resection once a septate uterus has been diagnosed, indicates that the test-treatment indication is completely ineffective. There are no RCTs comparing septum resection to no treatment, with one trial ongoing (TRUST trial, NTR number 1676). In our opinion, completion of this trial should be top priority.

In conclusion, diagnostic criteria slightly improve the reproducibility of hysteroscopy in diagnosing the septate uterus but agreement is still only moderate. The fact that the agreement among physicians on both the diagnosis of a uterine septum as well as the advice to resect such septum was moderate, may imply that hysteroscopy is insufficient as a single tool to diagnose a septate uterus. The diagnosis should be confirmed by additional diagnostic techniques such as three-dimensional ultrasound or MRI. In fact, under current conditions, if two doctors have a different opinion on the diagnosis of a septum and subsequent resection, at least one of the patients seen by these two doctors receives inappropriate treatment.

**Supplementary data**

Supplementary data are available at http://humrep.oxfordjournals.org/.

**Acknowledgements**

We would like to thank the participating gynecologists for their contribution to this study.

**Authors’ roles**

J.S. selected and edited the video recordings, was responsible for the execution of the study, interpreted analysis and took the lead in writing the manuscript; J.S., J.K., S.O. and F.B. formulated diagnostic

---

**Table II** Inter-observer agreement.

<table>
<thead>
<tr>
<th>Intra-class correlation coefficient</th>
<th>Overall</th>
<th>Diagnostic criteria</th>
<th>No diagnostic criteria</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape of uterine cavity (normal, arcuate or septate)</td>
<td>0.55</td>
<td>0.59</td>
<td>0.52</td>
<td>0.002</td>
</tr>
<tr>
<td>Need for surgical correction of uterine cavity</td>
<td>0.41</td>
<td>0.43</td>
<td>0.39</td>
<td>0.7</td>
</tr>
</tbody>
</table>

*DC versus no DC. Corresponding with a relative difference in between-observer variation of 0.90 for shape of uterine cavity (i.e. variation in the DC group is 0.9 of the variation in the no DC group) and 0.98 for need for surgical correction.
criteria, S.O. searched the electronic databases for diagnostic criteria; M.J.E. provided statistical support and performed analysis; J.K. selected the video recordings and revised several drafts of the manuscript; H.T. and B.W.M. revised several drafts of the manuscript; F.B., M.H.E., M.B. and M.V. performed the hysteroscopies and revised the manuscripts. All authors confirmed with the final version of the manuscript.

Funding
No funding was received for this study.

Conflict of interest
No conflicts of interest are present for any of the authors.

References