model. Furthermore, the model was not tested in individual participating centres, instead the whole cohort was used. As our model has been designed for universal implementation into clinical practice and not to be restricted to use in Centres for Assisted Reproduction (CARE) clinics alone, adding centre into the model as a predictor would have limited the application of the model. In general terms, we would agree that the inclusion of centre variability may effect model performance (Bouwmeester et al., 2013). However, in the future we plan to externally validate our model on IVF data from other populations in order to assess its transportability to non-private clinics and other geographical regions.

References


R.K. Dhillon1,*, D.J. McLemon2, P.P. Smith1, K. Dowell3, S. Fisher3, J. Deeks4, S. Bhattacharya1 and A. Coomarasamy1
1School of Clinical and Experimental Medicine, Academic Department, University of Birmingham, Birmingham Women’s Hospital, Birmingham B15 2TG, UK
2Institute of Applied Health Sciences, School of Medicine and Dentistry, Foresterhill, Aberdeen AB25 2ZD, UK
3CARE (Centres for Assisted Reproduction) John Webster House, 6 Lawrence Drive, Nottingham Business Park, Nottingham NG8 6PZ, UK
4School of Health and Population Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

*Correspondence address. E-mail: rima.dhillon@doctors.org.uk

doi:10.1093/humrep/dew091

Advanced Access publication on April 15, 2016

Reliability of hysteroscopy-based diagnosis of septate, arcuate and normal uterus: estimate or guestimate?

Sir,

Recently, Smit et al. (2015) questioned the reliability of hysteroscopy as a method for distinguishing between septate, arcuate and the normal uterine cavity. In our opinion, the independent indicators (Mokkink et al., 2010; Kottner et al., 2011) and comparison of the results of Smit et al. with results of a preceding study from the same author group (Smit et al., 2013) indicate a high risk of bias mainly related to: (i) the inappropriate choice of subjects (selection bias) and (ii) the small sample size of subjects, which may not have allowed for correct and precise estimation of reliability.

Selection bias and its confirmation

The sample, including cases and raters should be representative of the target population (Kottner et al., 2011). However, in both Smit et al. (2013) and Smit et al. (2015) hysteroscopy videos seem to be arbitrarily selected. The prevalence and clinical representation of the features of interest have a major influence on the reliability and agreement. For instance a high proportion of borderline cases of arcuate/septate uteriases lowers the reliability.

The prevalence of all malformations does not exceed 6.1–8.0% (95% CI: 3.9–12%) in an infertile population (Chan et al., 2011). In the selected sample of subjects in Smit et al. (2013) and Smit et al. (2015), the malformations rate is higher than the upper 95% CI limit for estimating prevalence of defects in infertile or even high-risk populations (12 and 49%, respectively; Chan et al., 2011), as illustrated in Fig. 1. If consensus between experts/raters is used to determine the true diagnosis, as in a Delphi procedure (agreement >67%), it can be concluded that in Smit et al. (2015) and Smit et al. (2013), three of ten (30%) and one of nine (11%) uteri, respectively, had hysteroscopic features indicative of

Figure 1 The selection bias and effect of sample size in the international reliability trials of hysteroscopy. The small sample of arbitrarily selected hysteroscopy videos is significantly different in regards to the prevalence of specific uterine morphologies compared with the potential target population of women for screening and diagnostic use of hysteroscopy; i.e. (i) infertile, (ii) miscarriage, and (iii) mixed infertility and recurrent miscarriages (high-risk population). Therefore, the results of randomized controlled trial (Smit et al., 2015) cannot be generalized.

(a) The prevalence of borderline cases in real populations was estimated based on the assumption that the frequency of borderline cases should not be higher than 50% of estimated prevalence of septate and arcuate uterus using optimal diagnostic tests in real populations (Chan et al., 2011). (b) The number of borderline cases and other morphology (septate, arcuate and normal uterus) was estimated by consensus between raters (agreement <67% and >67%, respectively), who participated in the trials.
septate uterus (Fig. 1). In both studies, no case met the broad criteria for normal or arcuate uterus (agreement, < 67%). Thus, the studied population seems inappropriate (70 and 89% borderline cases in Smit et al. (2015) and Smit et al. (2013), respectively, with a risk of 100% malformations without normal uterus by participated independent raters’ consensus). To avoid selection bias, subjects or cases should be selected randomly from the population of interest.

**Effect of sample size of subjects**

Further, the sample size of subjects (N = 10 in Smit et al., 2015 and N = 9 in Smit et al., 2013) has a significant impact on the precision of reliability estimation. For smaller sample sizes, the reliability value range is very large with correspondingly wide CIs. Therefore, the small sample size may have been a limitation. The authors did not calculate 95% CIs, although they are most important for interpretation. The CIs should be disclosed. Additionally, only an estimate of agreement (e.g., proportions of agreement) is able to determine whether, and to what degree, the raters actually agree (Kottner et al., 2011, Kraemer, 2014), and should also be disclosed.

According to Smit et al., the differences in the reliability (intraclass correlation coefficient (ICC)) of hysteroscopy between the first and second study (Smit et al., 2013: more borderline cases, ICC = 0.27, CIs unknown; Smit et al., 2015: less borderline cases, group without criteria, ICC = 0.52, CIs unknown) are related to the recent ESHRE-ESGE classification system. However, this reasoning might be problematic, due to known ESHRE-ESGE paradox (Ludwin and Ludwin, 2015), and the highlighted limitations in both study design.

**Conclusion**

The methods employed by Smit et al. (2013, 2015) might be inappropriate to determine the reliability of hysteroscopy in the target (infertile) or high-risk population. We agree with the authors that the reliability of subjective hysteroscopic assessment without ‘true’ measurements using calibrated instruments for identification of borderline forms of arcuate/septate uteri is nearly impossible. However, as Kraemer (2014) stated, if the study design has limitations and the reliability is not estimated precisely, its findings cannot be trusted. Thus, the reliability of hysteroscopy still remains to be confirmed.

**References**


**Reply: Reliability of hysteroscopy-based diagnosis of septate, arcuate and normal uterus: estimate or guestimate?**

Sir,

We want to thank Ludwin & Ludwin for the interest in our papers ‘The impact of diagnostic criteria on the reproducibility of the hysteroscopy diagnosis of the septate uterus: a randomized controlled trial’ (Smit et al., 2015) and ‘The international agreement study on the diagnosis of the septate uterus at office hysteroscopy in infertile patients’ (Smit et al., 2013). We want to respond to the concern of Ludwin & Ludwin about the risk of bias in our papers.

**Selection bias**

First of all, concern is voiced that the video recordings are not representative of the target population. These video recordings were selected from a population of subfertile women and women with a history of recurrent miscarriage who all underwent hysteroscopy. Recordings judged by experts in the field to be examples of a distinct septum were included in the study, as was a sample of video recordings judged to show a normal or arcuate uterus. We do not agree that these are by definition borderline cases. The reasoning applied by Ludwin & Ludwin is flawed in this respect: concluding from disagreement between raters that these are borderline cases is incorrect: consensus between raters can only be used to determine the true status if the subject matter lends itself to consensus. If raters are not able to judge consistently, there will be a lack of consensus, which our studies illustrate. In an ideal situation, all recordings would have been included in the randomized trial, however, this would have significantly influenced feasibility of the trial. Furthermore, observers were unaware of the relatively high prevalence of abnormalities at hysteroscopy. We agree that in general the value for intraclass correlation coefficient (ICC) depends on the population used. In our case, we would like to point out that by using a set of samples with a higher...