Intraobserver and interobserver reproducibility of third trimester uterine artery pulsatility index

Laura Marchi1 | Claudia Gaini1 | Chiara Franchi1 | Federico Mecacci1 | Caterina Bilardo2 | Lucia Pasquini1

1 Unit of Fetal Medicine and Diagnosis, Department for Women and Child Health, Careggi University Hospital, University of Florence, Florence, Italy
2 Department of Obstetrics and Gynaecology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands

Correspondence
Laura Marchi, Unit of Fetal Medicine and Diagnosis, Department for Women and Child Health, Careggi University Hospital, University of Florence, Largo Brambilla 3, Florence 50141, Italy.
Email: laura.marchi84@gmail.com

Abstract

Objectives: The primary aim of the study was to investigate intraobserver and interobserver reproducibility of uterine artery (UtA) pulsatility index (PI) in the third trimester of pregnancy. The secondary aim of the study was to examine whether high maternal body mass index (BMI) or gestational age (GA) influence the reliability of this measurement.

Methods: Singleton pregnancies in women with known BMI were recruited between 30+0 and 40+0 weeks. UtA PI Doppler measurements were performed prospectively, independently, and blindly by 2 Fetal Medicine Foundation–accredited operators. Intraobserver and interobserver reproducibility was assessed by concordance correlation coefficient (CCC) and intraclass correlation coefficient (ICC); Bland–Altman plots were built and limits of agreement (LoA) were calculated. The analysis was performed for both raw numbers and percentiles. To estimate the intraoperator and interoperator agreement in defining normal and pathological measurements, the assessments were divided in 2 categories—UtA PI <95th percentile/ ≥95th percentile—and Cohen’s kappa coefficients were calculated. Results were interpreted according to the cutoffs reported by the True Reproducibility of Ultrasound Techniques review. Correlation between maternal BMI and GA and accuracy of UtA measurements was studied with Spearman’s correlation coefficient.

Results: Measurements were available in 101 women. For intraobserver reproducibility, ICCs and CCCs were calculated for raw values and percentiles and were 0.912 and 0.835, and 0.837 and 0.716, respectively. For interobserver reproducibility, ICCs and CCCs were 0.809 and 0.732, and 0.677 and 0.576, respectively. This indicates a poor–moderate reproducibility of third trimester UtA PI. LoA were also wide (from a minimum of −0.30–0.35 to a maximum of −0.53–0.62). Cohen’s kappa coefficients were 0.478 and 0.418, showing a moderate intraoperator and interoperator agreement in distinguishing between normal and pathological values. No correlation was found between maternal BMI and GA and reproducibility of the measurements.

Conclusions: Intraobserver and interobserver reproducibility of third trimester UtA PI as assessed by ICC, CCC, and LoA is only moderate-poor. The agreement between operators in defining pathological and normal measurements is moderate.

1 | INTRODUCTION

The clinical utility of uterine artery (UtA) Doppler indices assessed during the third trimester of pregnancy in identifying fetuses at risk of adverse outcome is increasingly reported in literature. Several authors have indeed demonstrated that small fetuses with abnormal (≥95th centile) mean UtA pulsatility index (PI) during the third trimester are at increased risk of adverse perinatal outcome and show placental signs of underperfusion. Although evidences on the usefulness of this measurement are growing, the accuracy of UtA PI evaluation at third trimester has never been investigated. This issue should not be overlooked since unjustified confidence in the reliability of a diagnostic test may bring to misleading conclusions, thus threatening clinical practice. A recently published article on Nature pointed at a

FMF World Congress 2016 ISUOG World Congress 2016.
“Reproducibility Issue” in biological sciences. Obstetric measurements, and Doppler in particular, are not immune from these concerns, and a debate is currently ongoing on which cutoffs should be used to interpret reproducibility studies.

The aim of our study was to assess the reliability of UtA Doppler during the late (30⁺⁰ to 40⁺⁰ w’ gestation) third trimester of pregnancy investigating its intraobserver and interobserver reproducibility. The analysis was performed on raw numbers and percentiles and qualitatively dividing the measurements in normal/pathological values. Moreover, we examined if an association between reproducibility and maternal body mass index (BMI) or gestational age (GA) exists, attempting to investigate whether some particular condition worsen the reliability of UtA PI measurement.

2 | METHODS

This prospective, observational, cross-sectional study was performed at the Unit of Fetal Medicine and Diagnosis of the Careggi University Hospital, Florence. The study was approved by the local hospital ethics committee (CEAVC protocol number 10388).

Singleton pregnancies at a GA between 30⁺⁰ and 40⁺⁰ weeks determined by first trimester ultrasound scan, without major fetal malformation nor chromosomal anomalies, in women with recorded BMI, undergoing a routine, planned scan at the Unit, presumably in absence of contractions, were enrolled in the study.

UtA PI measurements were performed by 2 experienced operators (C.G., operator A, and L.M., operator B) who both hold the Fetal Medicine Foundation certificate in Doppler ultrasound, using a Voluson E8 ultrasound machine (GE Healthcare Austria GmbH & Co OG Zipf) equipped with a 2-5 MHz convex transabdominal probe (4C-D H4001BC).

Doppler assessments were performed in absence of fetal movements and voluntary suspended maternal breathing. To obtain the measurements, the recommendations of the International Society of Ultrasound in Obstetrics and Gynecology were followed: the probe was placed longitudinally in the lower lateral quadrant of the abdomen, angled medially; color flow mapping was therefore used to identify UtAs as it is seen crossing the external iliac artery. Measurements were taken 1 cm downstream from this crossover point, using the built-in program, from at least 3 consecutive uniform waveforms, with a sample gate set at 2 mm and an insonation angle as close to 0° as possible and always below 30°. Every operator performed 2 scans alternated by an insonation angle as close to 0° as possible and always below 30°. Every operator performed 2 scans alternated by the other observer. The acquisitions were blind and completely independent. The possible confounding effect of a significant change in maternal heart rate (MHR) on the UtA PI and therefore on the evaluation of the reproducibility was assessed through a pairwise comparison with repeated measures.

Measurements were recorded as raw numbers; thereafter, percentiles were calculated according to published nomograms. The analysis was performed for both raw values and centiles.

The estimated sample size calculated to achieve an intraclass correlation coefficient (ICC) ≥ 0.7 with sufficient precision (95% confidence interval [CI] width = 0.2) was 101 measurements. The measurement of each UtA was considered separately for the analysis. Intraobserver reproducibility was calculated comparing the first and second measurement of operator A, for interobserver reproducibility the average of the first and second measurement of operators A and B were used.

Intraobserver and interobserver reproducibility were assessed by both concordance correlation coefficient (CCC) and ICC, absolute agreement, 2-way mixed, average measurements, with 95% CI. Bland-Altman plots were built and limits of agreement (LoA) were also reported. To assess the agreement between the observers in defining pathological or normal values, we divided the measurements in the 2 categories (PI ≥ 95° centile and <95° centile) and calculated the Cohen’s kappa coefficient.

For interpretation of ICC, CCC, and Cohen’s kappa, we used recently published cutoff values: ICC and CCC: ≤ 0.70 very poor, 0.70 to 0.90 poor, 0.90 to 0.95 moderate, 0.95 to 0.99 good, ≥ 0.99 very good; Cohen’s kappa: ≤ 0.20 very poor, 0.21-0.40 poor, 0.41-0.60 moderate, 0.61-0.80 good, ≥ 0.81 very good. The relationship between BMI and GA and reproducibility was investigated with Spearman’s correlation coefficient. For this purpose, reproducibility was defined as the difference between the first and the second measurement of operator A. P < .05 was considered to be statistically significant.

Statistical analysis was performed using SPSS version 22.0 (SPSS Inc, Chicago, Illinois).

3 | RESULTS

A total of 101 consecutive pregnancies were enrolled. The acquisition of UtA PI was successful in all patients. Since we considered each UtA separately, the sample was actually the double of the number of patients enrolled—202 measurements.

The characteristics of the population are reported in Table 1. Median GA at enrolment was 34⁺¹³ weeks. Our sample included, as
expected considering the incidence in the general population, 7 cases of growth restriction. MHR did not change significantly between the 2 examination of operator A \((P = .13)\) and between the first examination of operator A and that of operator B \((P = .85)\), therefore excluding any possible influence of this factor on reproducibility.

Table 2 shows the ICCs, CCCs, and LoAs for intraobserver and interobserver reproducibility calculated for raw numbers. Figures 1 and 2 show the Bland-Altman plots pertaining to these values. Table 3 shows the ICCs, CCCs, and LoAs for intraobserver and interobserver reproducibility calculated for percentiles. Figures 3 and 4 show the concerning Bland-Altman plots.

According to the cutoffs set by the TRUST, only the intraobserver reproducibility assessed by ICC calculated using raw numbers has to be considered moderate, all the rest of the values show a poor or very poor reproducibility; the LoAs are wide as well, ranging from -0.30 to 0.62.

The Bland-Altman plots built using percentiles show that the intraoperator and interoperator agreement is worst around the 50th centile but improves for the extreme values, therefore suggesting a better concordance for values in the pathological range (namely, UtA PI ≥95th percentile). To verify this finding, data were divided in 2 categories (normal-pathological). A total of 20 and 28 of the UtAPI evaluations were assessed as pathological by operators A and B, respectively. Cohen’s kappa coefficients were calculated to give a quantitative measure of the magnitude of agreement. The Cohen’s kappas were 0.478 and 0.418, respectively, for intraoperator and interoperator agreement; therefore, the concordance should be interpreted as moderate according to the selected criteria.

Spearman’s correlation coefficient did not show any significant association between elevated BMI and reproducibility \((P = .677)\) nor between increasing GA and reproducibility \((P = .563)\).

## 4 | DISCUSSION

Our results suggest that the reproducibility of UtA PI measurement in the third trimester assessed by ICC, CCC, and LoA is generally poor. The reproducibility does not seem to be influenced by maternal BMI or by GA at measurement. Cohen’s kappa coefficient showed a moderate agreement in distinguishing the pathological from the normal values.

### TABLE 1  Population characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Median (Range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal age, y</td>
<td>34 (18-43)</td>
</tr>
<tr>
<td>Maternal BMI, kg/m²</td>
<td>26 (21-46)</td>
</tr>
<tr>
<td>Gestational age at enrolment, wk</td>
<td>34.1 (30.0-40.0)</td>
</tr>
<tr>
<td>Small for gestational age fetuses</td>
<td>7 (6.9%)</td>
</tr>
</tbody>
</table>

### TABLE 2  ICCs (95% CI), CCCs (95% CI), and LoAs calculated with raw numbers

<table>
<thead>
<tr>
<th></th>
<th>ICC</th>
<th>CCC</th>
<th>LoA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intraobserver reproducibility</td>
<td>0.912 (0.882-0.934)</td>
<td>0.837 (0.789-0.875)</td>
<td>-0.30-0.35</td>
</tr>
<tr>
<td>Interobserver reproducibility</td>
<td>0.809 (0.744-0.858)</td>
<td>0.677 (0.592-0.747)</td>
<td>-0.34-0.41</td>
</tr>
</tbody>
</table>

### TABLE 3  ICCs (95% CI), CCC (95% CI), and LoAs calculated with percentiles

<table>
<thead>
<tr>
<th></th>
<th>ICC</th>
<th>CCC</th>
<th>LoA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intraobserver reproducibility</td>
<td>0.835 (0.780-0.876)</td>
<td>0.716 (0.640-0.778)</td>
<td>-0.44-0.52</td>
</tr>
<tr>
<td>Interobserver reproducibility</td>
<td>0.732 (0.641-0.799)</td>
<td>0.576 (0.472-0.663)</td>
<td>-0.53-0.62</td>
</tr>
</tbody>
</table>

Abbreviations: CCC, concordance correlation coefficient; ICC, intraclass correlation coefficient; LoA, limits of agreement.
The use of UtA Doppler measurement as a tool for the diagnosis of placental intrauterine growth restriction is a relatively new acquisition in maternal-fetal medicine. Abnormal UtA Doppler was recently found to be associated with poorer perinatal outcome in small for gestational age (SGA) fetuses with normal umbilical artery Doppler.\textsuperscript{11-13} Cruz-Martinez et al\textsuperscript{2} found that at the time of SGA diagnosis, abnormal UtA Doppler identifies a subgroup of fetuses at risk of progression to abnormal brain Doppler findings. Moreover, a model that includes UtA measurement has been proposed to identify third trimester SGA fetuses with abnormal outcome and in need of intensive monitoring, even in the presence of normal umbilical artery Doppler and cerebroplacental ratio.\textsuperscript{14,15}

Although the clinician may be prone to take for granted that Doppler measurements are reliable, unfortunately, they are subject to a great variation. Doppler acquisitions have been shown to be poorly reproducible, and this is true not only for fetal Doppler but also for UtA measurement during the first and second trimester.\textsuperscript{16,17}

No study currently exists on the reliability of third trimester UtA Doppler. The acquisition of third trimester UtA PI may sometimes be challenging due to the vessel dislocation in the maternal abdomen with advancing gestation. In fact, the anatomical landmarks used in the second trimester (crossing of the external iliac artery) are more difficult to be used (although in our sample, the acquisition was feasible in all patients). Moreover, higher maternal BMI and patient's discomfort may hamper the accuracy of the Doppler sampling.

Our study is the first to investigate the reproducibility of UtA PI measurement in the third trimester and shows that reliability evaluated through ICC, CCC, and LoA is not satisfactory, at least according to the criteria of the TRUST review. We decided to adopt these criteria since these cutoffs were endorsed by the authors of the TRUST from the GRAAS study\textsuperscript{18} that up to now is the only paper in which guidelines for performing and reporting reliability and agreement studies were proposed. It was as well suggested in the TRUST that the analyses should be performed using percentiles instead of raw values. This method is indeed probably more rigorous; the results we obtained using the 2 different methods were anyway comparable.

Calculating the Cohen’s kappa coefficient, we aimed to assess what we think is the crucial point of studying the reliability of Doppler acquisition, which is the intraoperator and interoperator agreement in distinguishing the normal values from the pathological ones. In the everyday practice, indeed, what the clinician is interested in is not to miss a case of placental insufficiency or to erroneously assign a patient to the high-risk group. The Bland-Altman plots suggest a better intraoperator and interoperator agreement for extremes values; indeed, the agreement, according to the TRUST cutoffs, is moderate. Therefore, the clinical impact of the variability in UtA PI assessment is probably limited and would probably be even less relevant when evaluated in the context of the Uterine Artery Score, a combination of UtA PI and notch assessment that several authors have demonstrated to identify fetuses at higher risk of adverse perinatal outcome.\textsuperscript{11,12,19}

We expected that the reproducibility would be worse for higher BMI and more advanced GA; Spearman’s coefficient, however, did not show any significant association between BMI and GA and agreement, contradicting our hypothesis.

The main strengths of our work are that the examinations were performed blindly and completely independently. Moreover, for the first time in reliability studies in Obstetrics and Gynaecology, the statistical analysis was performed not only with raw values but also with percentiles. The possible limitation of our study is that the cutoffs we have chosen may be considered arbitrary since there is not yet a consensus on these, although they are widely used.\textsuperscript{6,7}

5 | CONCLUSIONS

According to the criteria set by the TRUST study, both intraobserver and interobserver reproducibility of UtA PI measured in the third trimester are poor; the clinical impact may however be limited since the Cohen’s kappa coefficient show a moderate agreement. Clinician should be aware of this, and repetition of measurements to confirm the results before clinical decisions are taken is mandatory.
REFERENCES


