European Journal of Obstetrics & Gynecology and Reproductive Biology 243 (2019) 67–71

Contents lists available at ScienceDirect

European Journal of Obstetrics & Gynecology and Reproductive Biology

journal homepage: www.elsevier.com/locate/ejogrb

Full length article

Inter- and intra-observer variability in fetal ductus venosus blood flow measurements in high-risk fetuses at 26–32 weeks

Clara M. Bruin[a,*], Wessel Ganzevoort[a], Ewoud Schuit[a,c], Nico A. Mensing van Charante[a,b], Hans Wolf[a]

[a] Department of Obstetrics and Gynaecology, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
[b] Department of Gynaecology, Dijklander Ziekenhuis, Hoorn, the Netherlands
[c] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

ARTICLE INFO

Article history:
Received 28 May 2019
Received in revised form 12 September 2019
Accepted 17 October 2019

Keywords:
Fetal growth restriction
Ductus venosus
Doppler ultrasonography
Reproducibility of ductus venosus pulsatility index
Correlation
Variability

ABSTRACT

Objectives: Early preterm fetal growth restriction is a significant contributor to perinatal morbidity and mortality. The ductus venosus pulsatility index for veins (DV PIV) is proposed as a monitoring tool because it appears to improve perinatal outcomes. The test characteristics and robustness of DV PIV have been inadequately described. The aim of this study was to investigate inter- and intra-observer variability of DV PIV.

Study design: Nineteen women with a gestational age between 26 and 32 completed weeks were included in this study. Doppler sonographic fetal assessment was performed by two independent maternal-fetal medicine specialists. Each sonographer alternately performed three flow tracings for each participant, in the absence of the other sonographer (six tracings in total per patient). DV PIV was calculated automatically from stored tracings by a third researcher. Inter- and intra-observer variability of DV PIV and limits of agreement were assessed using the Bland–Altman method. Comparison of the distribution was performed with Kendall's related samples test, and the intraclass correlation coefficient (ICC) was calculated.

Results: In total, 114 DV measurements were taken from 19 participants with a median age of 31 years [interquartile range (IQR) 26–34 years] at a median gestational age of 28 weeks (IQR 27–29 weeks). The proportional limits of agreement for intra-observer variation were −0.48 to 0.48 and −0.39 to 0.62 for the two observers. ICCs were 0.66 [95% confidence interval (CI) 0.42–0.84] and 0.68 (95% CI 0.45–0.85). The proportional limits of agreement for inter-observer variation were −0.29 to 0.19 with an ICC of 0.89 (95% CI 0.73–0.96).

Conclusion: Inter-observer variation was far less than intra-observer variation, probably due to mitigation of biological variation by averaging three measurements. DV PIV has acceptable test characteristics for use in a clinical setting when the average of at least three consecutive measurements is used.

Introduction

Fetal growth restriction (FGR) is a significant contributor to perinatal morbidity and mortality. The incidence of adverse outcomes is high, particularly in preterm gestational ages. Timing of delivery is pivotal in early preterm FGR fetuses, and depends on anticipated risks of antenatal death and neonatal complications. When estimated risks of antenatal death exceed anticipated neonatal risks, delivery is indicated. Unfortunately, monitoring the fetal condition in diagnosed early preterm FGR is a challenge, thereby complicating the timing of delivery [1].

At present, cardiotocography (CTG) and Doppler ultrasound of the umbilical artery are at the cornerstone of clinical evaluation for timing of delivery [2,3]. Additional use of the ductus venosus pulsatility index for veins (DV PIV) to indicate delivery appears to improve perinatal outcomes. In a multicentre, prospective, observational study, adverse perinatal outcome could be predicted at 0–1 days before delivery by DV PIV of 3 standard deviations (SD) below the mean [odds ratio (OR) 11.3, 95% confidence interval (CI) 2.3–57], and at 2–7 days before delivery by DV PIV of 2 SD below the mean (OR 3.0, 95% CI 0.8–12) [4]. In a pragmatic randomized controlled trial with strict monitoring schedules, waiting for either abnormal CTG or DV changes (whichever came first) was

* Corresponding author at: Amsterdam UMC, Department of Obstetrics and Gynaecology, Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands.
E-mail address: c.m.bruin@amsterdamumc.nl (C.M. Bruin).

https://doi.org/10.1016/j.ejogrb.2019.10.028
0301-2115/

associated with a more favourable 2-year outcome in early-onset FGR compared with monitoring with computerized CTG alone [5].

While good repeatability of DV PIV is of importance to substantiate use in a clinical setting, its test characteristics and robustness have been inadequately described. Therefore, the aim of this study was to investigate inter- and intra-observer variability of DV PIV.

Materials and methods

This study was conducted at the Department of Obstetrics of the Academic Medical Center, Amsterdam, the Netherlands between May 2010 and August 2011. All women with a gestational age between 26 and 32 completed weeks were eligible for inclusion in this study. Gestational age was determined using a first-trimester ultrasound dating scan. Women were included in this study when they were hospitalized for fetal monitoring with FGR [6], pregnancy-induced hypertension, pre-eclampsia or HELLP syndrome. This is not a consecutive cohort because women had to give consent for an elaborate procedure and all examiners had to be available. Fetal heart rate tracings near the time of measurement had to be normal at study inclusion. Medical ethics approval was not required under Dutch law at that time.

Study procedures

Doppler sonographic fetal assessment was performed with a Philips IU-22 (Philips, Amsterdam, The Netherlands). Doppler flow velocity waveforms were recorded according to international standards [7,8]. The DV was visualized either in a mid-sagittal longitudinal plane of the fetal trunk or in the oblique transverse plane through the upper abdomen, depending on the fetal position. The sample volume was positioned at its origin from the umbilical vein, where colour Doppler indicated the highest velocities [9]. Ultrasonographic assessment was performed independently by HW (Observer A) and WG (Observer B). Both are maternal-fetal medicine specialists with ample experience in Doppler sonographic fetal assessment, including measuring DV PIV. When both investigators were present, eligible women were asked to participate in the study. Both examiners alternately performed three measurements for each participant in the absence of the other investigator. Each measurement consisted of three to four waveforms and was stored on the hard disk of the ultrasound machine. The screen was cleared before the other investigator returned for the next measurement. Automated calculation of DV PIV was done off-line by NM from the recorded waveforms after completion of the measurement session. This ensured that the observers were not aware of the results of their own and of each other's measurements during the DV measurement procedure.

The 95th centile for gestational age at measurement was used as a cut-off between normal and abnormal [9].

Fetal biometry was measured and the Hadlock IV formula was used to estimate fetal weight [10]. Maternal parameters, including age, body mass index (BMI), hypertensive disorders, parity, admission diagnosis and gestational age, were also recorded.

Intra-observer repeatability

Three DV PIV values were recorded by Observer A and three by Observer B in each measurement session. For each observer, the mean of their three measurements was calculated. The absolute difference of each measurement from this mean was calculated, and the proportional difference was calculated by dividing the absolute difference by the mean of the three measurements. A Bland–Altman plot was created with the proportional difference on the Y-axis and the mean value on the X-axis [11]. From the distribution of the proportional differences, the fifth centile and the 95th centile were calculated to determine the proportional limits of agreement. The distribution of measurements was compared by Kendall's related samples test for each observer. The intra-class correlation coefficient (ICC) was calculated using the measurements of each observer.

Inter-observer agreement

For inter-observer agreement, the means of the three measurements for each observer for each measurement session were used. Proportional differences between the means were calculated by dividing the difference by the average of the mean values of each observer. A proportional Bland–Altman plot was created, plotting the proportional differences between the means against the average of the mean values. From the distribution of the proportional differences, the fifth centile and the 95th centile were calculated to determine the proportional limits of agreement. The distribution of the two measurement series was compared by Kendall's related samples test. ICC was calculated between Observer A and Observer B. Agreement of the observers for discrimination between normal and abnormal DV PIV (cut-off at the 95th percentile of DV PIV) was calculated.

Data analysis

Descriptive statistics were used to describe population characteristics. Categorical outcomes were presented as absolute number and percentage of the total study population, and continuous variables were presented as median and interquartile range (IQR).

Statistical analysis was performed using SPSS Version 25 (IBM Corp., Armonk, NY, USA). A p-value <0.05 was considered to indicate statistical significance.

Power calculation

No formal power calculation was made before the study commenced. Based on the sample sizes of previous similar studies in the literature, it was estimated that this study needed approximately 20 participants with three DV PIV measurements per sonographer.

Literature review

A literature search was undertaken to compare the present findings with those from other studies. The following terms were used: interobserver[Title/Abstract]) OR inter-observer [Title/Abstract] AND agreement[Title/Abstract] OR repeatability[Title/Abstract] OR reliability[Title/Abstract] OR reproducibility[Title/Abstract] AND pulsatility index[Title/Abstract] OR ultrasound sonography[Title/Abstract] AND pregnancy[MeSH Terms] Filters: Humans

Results

Nineteen women participated in this study, resulting in 114 measurements. Table 1 shows demographic and obstetric details of these women. The majority of participants were admitted with FGR based on severe placental insufficiency (90%), which was reflected in absent or reversed end-diastolic umbilical flow in 27% of women. The measurements were performed at a median gestational age of 28 weeks (IQR 27–29 weeks). Both examiners were able to assess DV PIV successfully in all participants.

Intra-observer variation for both observers is shown by Bland–Altman plots in Fig. 1a and b for both observers. The distribution of

Table 1
Demographic and clinical characteristics.

Characteristics	
n	19
Age, years	31 (26–34)
Nulliparity	10 (53%)
Body mass index, kg/m^2	25.4 (21.3–29.1)
Admission diagnosis	
Pre-eclampsia	8 (42%)
Imminent preterm delivery	1 (5%)
Fetal growth restriction [19]	17 (90%)
Measurement session	
Gestational age, weeks	28 (27–29)
Estimated fetal weight, g	860 (790–976)
Umbilical artery pulsatility index	1.57 (1.16–2.00)
Absent or reversed end-diastolic flow	5 (27%)
Middle cerebral artery pulsatility index	1.40 (1.21–1.72)
Umbilical/cerebral ratio	0.99 (0.72–1.57)

Data expressed as *n* (%) or median (interquartile range).

the three measurements for each observer were the same (related samples Kendall's coefficient of concordance $p = 0.81$ for Observer A and $p = 0.34$ for Observer B). The proportional limits of agreement were −0.48 to 0.48 for Observer A and −0.39 to 0.62 for Observer B. ICC was 0.66 (95% CI 0.42–0.84) for Observer A and 0.68 (95% CI 0.45–0.85) for Observer B.

Inter-observer variation is shown by Bland–Altman plot in Fig. 2. The distribution of the two measurement series was the same (related samples Kendall's coefficient of concordance $p = 0.82$). The proportional limits of agreement were −0.29 to 0.19 and ICC was 0.89 (95% CI 0.73–0.96). Observer A measured DV PIV ≥95th percentile in seven women, compared with six women for Observer B. Agreement between the observers was 84%: Observer A measured two abnormal values that were normal by Observer B, and Observer B measured an abnormal value in one woman while Observer A did not. The observers agreed on five abnormal measurements and 11 normal measurements.

In a separate analysis, maternal BMI, gestational age at inclusion or inclusion diagnosis had no influence on inter-observer measurement variation (data not shown).

The literature search found 80 papers, 10 of which were aimed at inter-observer variability of fetal arterial or venous pulsatility index measurements (Table 2). ICCs for the three studies that assessed first-trimester DV PIV differed largely between studies, with a range from 0.2 to 0.9. Both studies on umbilical artery pulsatility index measurements showed rather disappointing ICCs of 0.4 and 0.6, similar to the single study aimed at middle cerebral artery pulsatility index measurements with ICC of 0.3. Four studies assessed uterine artery pulsatility index measurements and reported ICC of 0.6–0.9, with little difference between vaginal and abdominal measurement techniques. The authors did not find any studies on DV PIV in the second or third trimester. Half of the

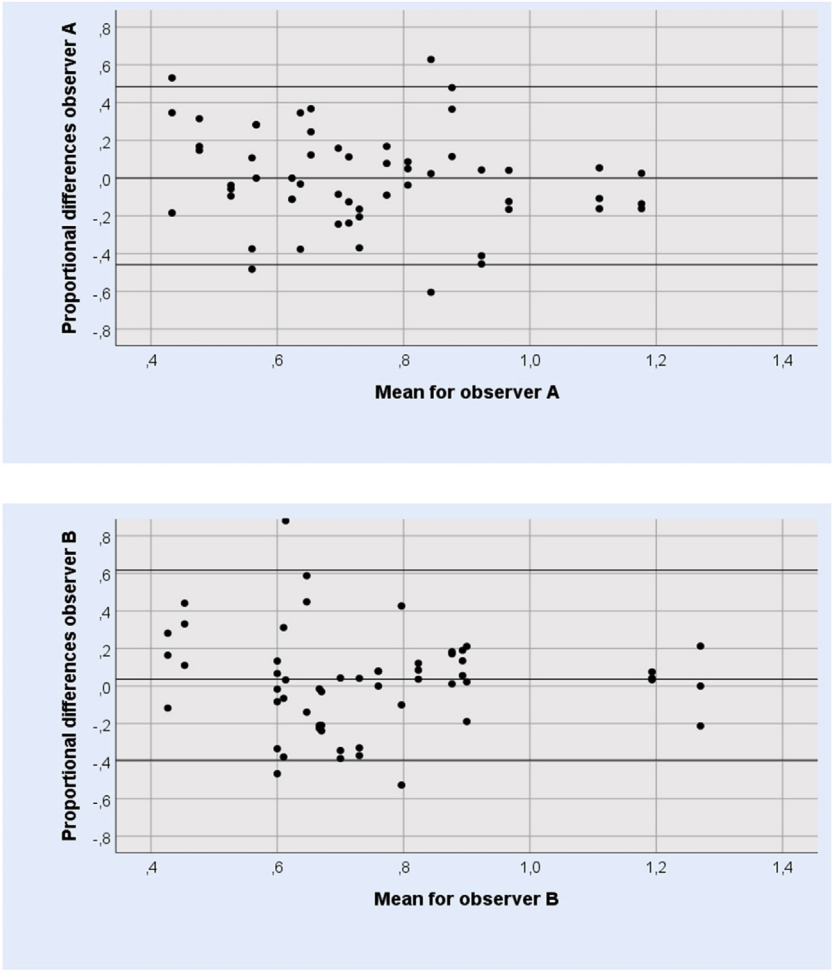

Fig. 1. Proportional difference from the mean for venous pulsatility index measurements by (a) Observer A and (b) Observer B. The upper and lower bold line represent the fifth and 95th percentiles (limits of agreement), and the middle bold line is the median.

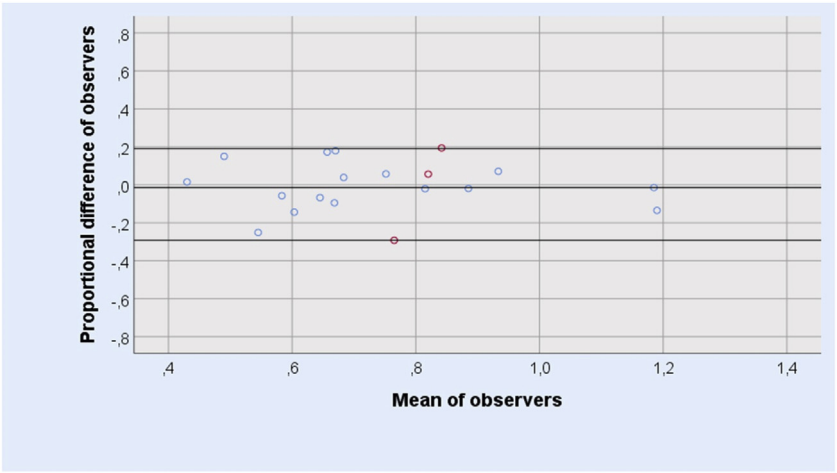

Fig. 2. Bland–Altman plot of the proportional difference between the two observers. The upper and lower bold line represent the fifth and 95th percentiles, and the middle bold line is the mean. The vertical line represents the 95th percentile of ductus venosus pulsatility index for veins (DV PIV) at approximately 28 weeks. The red dots are discrepant for the 95th percentile cut-off of DV PIV.

Table 2
Summary of studies that assessed inter-observer variability for fetal Doppler measurements.

Study (year of publication)	Blood vessel	Number of patients	Gestational age (weeks)	Population	No. of samples for comparison	Blinded	Inter-observer ICC (95% CI)
Current study (2019)	DV PIV	19	26–32	FGR	3	Yes	0.9 (95% CI 0.7–1.0)
Borrell (2007) [15]	DV PIV	35	11–14	High risk for chromosomal defects	3	Unclear[c]	0.7 (95% CI 0.5–0.9)
Mavrides (2001) [16]	DV PIV	24	10–14	Termination of pregnancy	3	Not mentioned	0.2 (95% CI not described)
Prefumo (2001) [14]	DV PIV	54	11–14	High risk for chromosomal defects	2	Yes	0.9 (95% CI not described)
Figueras (2006) [18]	UmbA PI	100	24–40	Low-risk singleton	1	Not mentioned	0.6 (95% CI 0.4–0.7)[d]
Scherjon (1993) [17]	UmbA PI	14	29–42	FGR and threatened preterm labour	2	Unclear[c]	0.4 (95% CI not described)
Ferreira (2015) [20]	UtA PI	96	14–27	Singleton pregnancy	1	Yes	0.8 (0.7–0.9)[a]
Hollis (2001) [21]	UtA PI	47	10–14	Termination of pregnancy	1	Yes	0.6 (95% CI not described)[b]
Marchi (2016) [22]	UtA PI	101	11–14	Low-risk singleton	1	Yes	0.9 (0.7–1.0)[a]
Marchi (2017) [23]	UtA PI	101	30–40	Singleton pregnancy	2	Yes	0.7 (95% CI not described)
Figueras (2006) [24]	MCA PI	100	24–40	Low-risk singleton	1	Not mentioned	0.3 (0.1–0.6)

DV PIV, ductus venosus pulsatility index for veins; UmbA PI, umbilical artery pulsatility index; UtA PI, uterine artery pulsatility index; MCA PSV, middle cerebral artery peak systolic velocity; FGR, fetal growth restriction; ICC, interclass correlation coefficient; CI, confidence interval.
[a] Small differences in ICC between transvaginal and transabdominal technique.
[b] Minimal differences between left and right UtA.
[c] Observers were blinded to each other's results, but unclear if they were blinded to their own results.
[d] This measurement was taken at a free loop; small differences in location of measurement in the umbilical cord are described.

studies only used single measurements for comparison between two observers.

Discussion

This study shows that while individual measurements show relatively large variation, the mean value of a series of three measurements has far less inter-observer variability. Maternal BMI, gestational age at inclusion or inclusion diagnosis did not influence this finding. It is assumed that variation of the individual measurements is based on the normal biological variation of cardiovascular control, probably because there was approximately 5 min between measurements due to the alternating procedure, and each measurement contained only 3–4 waveforms. Averaging a number of consecutive measurements adjusts for this biological variation. This is probably relevant for most measurements of biological parameters that are subject to measurement variation (e.g. this is common practice for blood pressure measurement).

Agreement between the observers concerning classification of normal vs abnormal DV PIV at the 95th percentile was 84%, and the three discrepancies between the observers were close to the cut-off limit. Values around the cut-off for abnormality probably need repeated measurement at a later time for confirmation. The International Society of Ultrasound in Obstetrics and Gynecology practice guideline for Doppler ultrasonography in obstetrics [12] advises repeated measurements if there is an obvious discrepancy. This study provides evidence for that advice for DV measurements, and the authors recommend that at least three consecutive measurements should be made as discrepancies between measurements are frequent and reproducibility of the average of repeated measurements is much better.

This study used proportional differences between measurements rather than absolute differences, as interpretation is easier to generalize [13].

This study has several strengths. First, to the authors' knowledge, this is the first study to assess intra- and inter-observer variability of DV measurement in the late second and early third trimester. Second, the study had a prospective design and does not suffer from problems generally associated with retrospective studies [13]. Third, examiners were blinded to their own DV PIV measurements and the measurements from the other examiner, leading to unbiased assessment of DV PIV. Fourth, the range of DV PIV measurements amongst participants was considerable. This broad range of measurements strengthens the generalizability of the findings. Finally, all DV PIV measurements were taken from each patient in a single visit without moving the patient between measurements, supposedly resulting in low variation in fetal haemodynamics.

A limitation could be that both examiners had some clinical information on admitted patients, such as knowledge about underlying maternal or fetal disease, which may potentially influence DV PIV measurement.

The present results were compared with other research that assessed inter-observer variability of fetal Doppler measurements; the results of the relevant studies are summarized in Table 2. The variability of the present results was similar or better than in other studies. Two larger studies that assessed ICC of DV PIV in first-trimester ultrasounds found somewhat lower but similar results [14,15], and one study found a much lower ICC in first-trimester DV PIV [16]. These differences in ICC cannot be fully explained by the study technique, as all of these studies used two or more measurements. One reason for the results of the present study may be the use of multiple measurements. Also, improvements in ultrasound imaging may have a positive effect on inter-observer agreement, as the lowest ICCs are reported in the oldest studies.

ICC of the umbilical artery pulsatility index, reported by two studies, was far lower than optimal for repeatability studies [17,18]. Notwithstanding the fact that these results indicate rather poor reliability of the technique, umbilical artery Doppler measurements have become standard in antenatal care.

Conclusion

Inter-observer variation was far less than intra-observer variation, probably due to mitigation of biological variation by averaging three measurements. DV PIV has acceptable test characteristics for use in a clinical setting when the average of at least three consecutive measurements is used.

Funding

None.

Declaration of Competing Interest

None declared.

References

[1] Pollack RN, Divon MY. Intrauterine growth retardation: definition, classification, and etiology. Clin Obstet Gynecol 1992;35:99–107.
[2] Hecher K, Hackeloer BJ. Cardiotocogram compared to Doppler investigation of the fetal circulation in the premature growth-retarded fetus: longitudinal observations. Ultrasound Obstet Gynecol 1997;9:152–61.
[3] Karsdorp VH, van Vugt JM, van Geijn HP, et al. Clinical significance of absent or reversed end diastolic velocity waveforms in umbilical artery. Lancet 1994;344:1664–8.
[4] Bilardo CM, Wolf H, Stigter RH, et al. Relationship between monitoring parameters and perinatal outcome in severe, early intrauterine growth restriction. Ultrasound Obstet Gynecol 2004;23:119–25.
[5] Lees CC, Marlow N, van Wassenaer-Leemhuis A, et al. 2 year neurodevelopmental and intermediate perinatal outcomes in infants with very preterm fetal growth restriction (TRUFFLE): a randomised trial. Lancet 2015;385:2162–72.
[6] Gaillard R, de Ridder MA, Verburg BO, et al. Individually customised fetal weight charts derived from ultrasound measurements: the Generation R Study. Eur J Epidemiol 2011;26:919–26.
[7] Pearce JM, Campbell S, Cohen-Overbeek T, Hackett G, Hernandez J, et al. Reference ranges and sources of variation for indices of pulsed Doppler flow velocity waveforms from the uteroplacental and fetal circulation. Br J Obstet Gynaecol 1988;95:248–56.
[8] Hsieh YY, Chang CC, Tsai HD, Tsai CH. Longitudinal survey of blood flow at three different locations in the middle cerebral artery in normal fetuses. Ultrasound Obstet Gynecol 2001;17:125–8.
[9] Hecher K, Campbell S, Snijders R, Nicolaides K. Reference ranges for fetal venous and atrioventricular blood flow parameters. Ultrasound Obstet Gynecol 1994;4:381–90.
[10] Hadlock FP, Harrist RB, Sharman RS, Deter RL, Park SK. Estimation of fetal weight with the use of head, body, and femur measurements – a prospective study. Am J Obstet Gynecol 1985;151:333–7.
[11] Altman DG, Bland JM. Measurement in medicine – the analysis of method comparison studies. J R Stat Soc 1983;32:307–17.
[12] Bhide A, Acharya G, Bilardo CM, et al. ISUOG practice guidelines: use of Doppler ultrasonography in obstetrics. Ultrasound Obstet Gynecol 2013;41:233–9.
[13] Martins WP, Nastri CO. Interpreting reproducibility results for ultrasound measurements. Ultrasound Obstet Gynecol 2014;43:479–80.
[14] Prefumo F, De Biasio P, Venturini PL. Reproducibility of ductus venosus Doppler flow measurements at 11–14 weeks of gestation. Ultrasound Obstet Gynecol 2001;17:301–5.
[15] Borrell A, Perez M, Figueras F, Meler E, Gonce A, Gratacos E. Reliability analysis on ductus venosus assessment at 11–14 weeks' gestation in a high-risk population. Prenat Diagn 2007;27:442–6.
[16] Mavrides E, Holden D, Bland JM, Tekay A, Thilaganathan B. Intraobserver and interobserver variability of transabdominal Doppler velocimetry measurements of the fetal ductus venosus between 10 and 14 weeks of gestation. Ultrasound Obstet Gynecol 2001;17:306–10.
[17] Scherjon SA, Kok JH, Oosting H, Zondervan HA. Intra-observer and inter-observer reliability of the pulsatility index calculated from pulsed Doppler flow velocity waveforms in three fetal vessels. Br J Obstet Gynaecol 1993;100:134–8.
[18] Figueras F, Fernandez S, Eixarch E, et al. Umbilical artery pulsatility index: reliability at different sampling sites. J Perinat Med 2006;34:409–13.
[19] Gordijn SJ, Beune IM, Thilaganathan B, et al. Consensus definition of fetal growth restriction: a Delphi procedure. Ultrasound Obstet Gynecol 2016;48:333–9.
[20] Ferreira AE, Mauad Filho F, Abreu PS, Mauad FM, et al. Reproducibility of first- and second-trimester uterine artery pulsatility index measured by transvaginal and transabdominal ultrasound. Ultrasound Obstet Gynecol 2015;46:546–52.
[21] Hollis B, Mavrides E, Campbell S, Tekay A, Thilaganathan B. Reproducibility and repeatability of transabdominal uterine artery Doppler velocimetry between 10 and 14 weeks of gestation. Ultrasound Obstet Gynecol 2001;18:593–7.
[22] Marchi L, Zwertbroek E, Snelder J, Kloosterman M, Bilardo CM. Intra- and inter-observer reproducibility and generalizability of first trimester uterine artery pulsatility index by transabdominal and transvaginal ultrasound. Prenat Diagn 2016;36:1261–9.
[23] Marchi L, Gaini C, Franchi C, Mecacci F, Bilardo C, Pasquini L. Intraobserver and interobserver reproducibility of third trimester uterine artery pulsatility index. Prenat Diagn 2017;37:1198–202.
[24] Figueras F, Fernandez S, Eixarch E, et al. Middle cerebral artery pulsatility index: reliability at different sampling sites. Ultrasound Obstet Gynecol 2006;28:809–13.